Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Bangla-language retrieval-augmented conversational agent for maternal health in a low-resource setting: design and formative evaluation

Fahim Muntasir
BRAC University
Nurul Labib Sayeedi
nsayeedi2410045@bsds.uiu.ac.bd
United International University
Jannatun Noor
United International University

Article

Keywords: Conversational agent, Retrieval-augmented generation, Maternal health, Bangla, Low-resource settings, Digital health

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10728064/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Bangla-language retrieval-augmented conversational agent for maternal health in a low-resource setting: design and formative evaluation

Fahim Muntasir[1,3], Nurul Labib Sayeedi[2,3,*], Jannatun Noor[2,3,*]

[1]Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh

[2]Department of Computer Science and Engineering, United International University, Dhaka, Bangladesh

[3]Computing for Sustainability and Social Good (C^2SG) Research Group, Dhaka, Bangladesh

Fahim Muntasir, ORCID: 0009-0007-2840-8605 Nurul Labib Sayeedi, ORCID: 0009-0002-4116-3917

Jannatun Noor, ORCID: 0000-0001-9669-151X

*Corresponding authors: Jannatun Noor (jannatun@cse.uiu.ac.bd) and Nurul Labib Sayeedi (nsayeedi2410045@bsds.uiu.ac.bd), Department of Computer Science and Engineering, United International University, United City, Madani Avenue, Badda, Dhaka 1212, Bangladesh.

Abstract

Maternal health information in Bangladesh is difficult to access for women with limited literacy and no English, and perinatal mental health remains under-supported. We designed and formatively evaluated *Baby and Me*, a Bangla-language conversational agent that retrieves from a curated, physician-reviewed corpus of 247 maternal health articles, segmented into 2,945 indexed passages, and generates responses grounded in the retrieved material. A survey of 72 pregnant and postpartum women established design requirements; Bangla-only language preference was universal among women with no formal or primary education, and a quarter of households reported annual income below one lakh Bangladeshi taka. Candidate model configurations were screened against physician-authored reference answers before deployment. Twenty-four respondents then rated the deployed system: ease of use was highest (mean 4.79 of 5) and clarity strong (4.54), while trust (4.29), concern alignment (4.04) and empathy (4.00) were lower and more dispersed. Interviews indicated that the interface showed users that retrieval had occurred but not whether guidance came from a clinically reviewed source, suggesting that process transparency and source attribution are not interchangeable. As a formative study with a modest sample, a single site and no controlled comparison, this work establishes feasibility rather than effectiveness.

Keywords: Conversational agent, Retrieval-augmented generation, Maternal health, Bangla, Low-resource settings, Digital health

1 Introduction

Maternal health in Bangladesh reflects a persistent tension between structural constraints in health care delivery and the everyday realities of women navigating pregnancy. The country has made substantial and well-documented progress: facility-based delivery rose from 28.3% in 2011 to 64.8% in 2022, and the proportion of births attended by a skilled provider rose from 31.2% to 70.0% over the same period [1, 2]. That progress, however, leaves roughly one birth in three still occurring outside a health facility and roughly three in ten without a skilled attendant, concentrated among rural, poorer, and less educated women. Coverage of antenatal care has not followed the same trajectory. Only 40.5% of women received the recommended minimum of four antenatal visits in 2022, a decline from 47.0% in 2017–2018 [1, 2]. Underlying these figures is a workforce shortage: the density of doctors, nurses, and midwives in Bangladesh remains below the World Health Organization benchmark of 4.45 skilled health professionals per 1000 population associated with achieving the health-related Sustainable Development Goals [3]. Referral mechanisms for obstetric emergencies remain weak, particularly in rural settings [4]. Studies in Bangladeshi hospital and community samples have reported high-risk features in a substantial share of pregnancies [5, 6], and limited decision-making autonomy, socioeconomic barriers, and informational inequities further restrict women's ability to seek timely care [7, 8].

Equally pressing are the informational and psychosocial dimensions of maternal health. Women often lack access to trustworthy, context-sensitive guidance, a challenge exacerbated by low literacy, limited digital familiarity, and persistent stigma around perinatal mental health [9, 10]. Perinatal depression is common and appears to be particularly so in this setting. A meta-analysis of 29 studies across eight South Asian countries estimated a pooled perinatal depression prevalence of 28% (95% CI 24% to 32%), with the highest country-level pooled estimate, 46%, in Bangladesh [11]. A Bangladesh-specific meta-analysis of 18 studies reported a lower pooled estimate for antenatal depression, 19.5%, alongside a rising trend over time [12]. The spread between these estimates reflects genuine heterogeneity in screening instruments, cut-off thresholds, and sampling frames rather than disagreement about the direction of the problem. Despite this burden, perinatal distress is rarely acknowledged within communities or by the health system, and professional mental health support is scarce [13]. Together, these gaps indicate that improving maternal health requires not only infrastructural reform but also new forms of interaction and support that reach women where they are: in their homes, on their

phones, and within their cultural contexts.

Conventional interventions such as mHealth apps, hotlines, and community outreach have extended some access, but they remain limited in reach and effectiveness. Content is often generic, systems rarely adapt to local language or cultural nuance, and few address the relational aspects of care such as stigma, reassurance, and personalization [14]. This creates a design opportunity for interactive technologies, and in particular for conversational artificial intelligence (AI) agents that move beyond static information delivery to provide two-way dialogue, empathy, and trust. Such agents can function not only as informational tools but as relational companions, blurring the boundary between formal health care support and everyday interaction. Recent advances in large language models, retrieval augmentation, and standardized tool interfaces make such systems technically feasible outside high-resource settings for the first time, but they also introduce risks of unsupported or misleading output that are consequential in a clinical domain. Whether these capabilities can be assembled into something usable, trusted, and safe by women in a Bangla-speaking, low-resource context is an open empirical question, and it is the question this study begins to address.

1.1 Related work

This subsection reviews prior work on conversational agents for maternal health, Bangla-language digital health systems, digital perinatal mental health support, trust in conversational health systems, and tool-orchestrated (agentic) chatbots. Across these areas we identify a recurring gap: rigorous, culturally grounded, and safety-aware evaluations of conversational AI for maternal health in low-resource, non-English settings remain scarce. The review is narrative rather than systematic, and Table 1 presents a purposive selection of representative systems rather than an exhaustive census.

1.1.1 Conversational agents in maternal health

Digital health interventions, including AI-driven chatbots, are increasingly proposed to extend maternal care through education, reminders, and emotional support. An estimated 260,000 women died from pregnancy- or childbirth-related causes worldwide in 2023, a global maternal mortality ratio of 197 deaths per 100,000 live births, with Sub-Saharan Africa accounting for approximately 70% of these deaths and Central and Southern Asia for nearly 17% [15]. Many more women experience preventable morbidity. Conversational agents have been used

to deliver breastfeeding counseling and reduce maternal anxiety; in a recent trial, postpartum women randomized to an AI-assisted breastfeeding chatbot showed higher breastfeeding self-efficacy and lower anxiety than controls [16]. Evaluations of large language models (LLMs) for lactation queries likewise report that advanced models can produce structured, readable, guideline-concordant answers, although output variability and hallucination remain concerns [17].

Several deployments report high engagement and satisfaction. In a US pilot of a rules-based perinatal mental health chatbot, most pregnant and postpartum users re-engaged at least once and rated the content as relevant [18]. A randomized controlled trial of the relational agent Woebot among postpartum women found greater reductions in depression scores than in the control condition [19]. That system descends from a line of automated cognitive-behavioral agents evaluated in adjacent populations, including a randomized trial in young adults with symptoms of depression and anxiety [20], a usability study in problematic substance use [21], and an eight-week intervention reporting changes in stress, burnout, and resilience [22]. A postpartum reminder chatbot achieved high open rates and ease-of-use ratings [23], and a systematic review concluded that chatbots are feasible for perinatal care and can support better knowledge and health behaviors [24].

Earlier systems were predominantly rule-based and, while functional, lacked the adaptability and personalization needed for complex maternal health scenarios. Tanya, an early breastfeeding-support agent, illustrated the potential of conversational agents in this domain [25], and a body of preconception-health chatbots demonstrated gains in knowledge and awareness [26–30]. These studies were largely cohort-based and qualitative, offering useful insights but limited generalizability, and their rule-based design constrained adaptation to individual needs. More recent systems point in the opposite direction: Penn Medicine's Penny uses conversational AI to extend postpartum ("fourth trimester") support, answering a large share of questions automatically, reducing clinician burden, and helping screen for postpartum hypertensive disorders and depression [31]. Prototype systems such as MamaBot [32], PregChat [33], and PreMomBot [34] further reflect growing interest in AI for pregnancy care, spanning gestational tips, symptom checks, and psychoeducational content.

Table 1 summarizes representative systems across prenatal and postnatal contexts. Efforts targeting Bangladesh and the Bangla language specifically remain at an early stage, and among the systems we identified we found none that was fully deployed, publicly evaluated, and agentic.

1.1.2 Bangla language barriers in digital health systems

Language and culture strongly shape chatbot adoption in South Asian and other LMIC settings. Bangladesh has near-universal mobile access but comparatively low computer ownership and literacy, making mobile-friendly, local-language interfaces essential [35]. Several Bangladeshi systems reflect this: Maya supports Bangla, English, and romanized Bangla (Banglish) for user convenience [36]; the menstrual-health chatbot Probahini replies in Bengali script or transliterated Bangla by preference [37]; and the LifeBot concept emphasizes personalized Bangla-language pregnancy education [38]. Local-language capability is repeatedly identified as decisive: an immunization chatbot in Pakistan was trained on Urdu and regional languages so caregivers could ask questions naturally, and remained feasible in low-literacy settings [39]. SuSastho.AI is building an open-source Bangla LLM for adolescent health [40], and GorbhoShongi offers a bilingual, LLM-based maternal mental-health app, though it remains an early prototype without real-world validation [41]. Several of these are described only in non-peer-reviewed sources, which is itself indicative of the state of the evidence base in this setting.

Infrastructural constraints compound language barriers. An implementation study of pandemic-era telemedicine in Bangladesh found that many women could not be reached by phone owing to limited access or network coverage [42], underscoring the need for low-bandwidth and offline options. India's Myna Mahila reproductive-health chatbot show that local co-design can address sensitive topics while reducing stigma [43, 44]. Yet the overall LMIC evidence base is thin: reviews note that almost all perinatal mental-health app trials come from high-income countries, with very few formal evaluations from South Asia [45, 46]. By curating a Bangla knowledge base and supporting Banglish alongside Bangla, the present work treats linguistic inclusivity as a core design principle rather than an afterthought.

1.1.3 Digital mental health support during pregnancy

Maternal mental health is a critical but under-served component of perinatal care. Global estimates place the proportion of women experiencing depression or anxiety during pregnancy or the postpartum period at roughly 10% to 20%, with consequences for both mother and infant [47, 48]. As noted in the preceding section, pooled estimates for South Asia and for Bangladesh specifically are considerably higher than this global range [11, 12], which is one reason the present work treats perinatal mental health as a first-class design concern rather than an extension of physical-health content.

Digital tools offer scalable support [49].

A randomized controlled trial of the relational agent Woebot among postpartum women in the United States reported reductions in depression scores relative to control, alongside high user satisfaction [19]. A real-world study of the Wysa conversational agent found that users reporting a maternal event derived measurable mental health and wellbeing benefit, with greater benefit among more engaged users [50]. A US mindfulness pilot reported that most pregnant participants found chatbot support relevant [18].

In LMICs, digital perinatal mental health support is less developed but urgently needed. Mobile-phone interventions, mostly SMS or voice, have shown feasibility [46, 51, 52], and a text-message program in South Asia demonstrated the feasibility of delivering behavioral activation to mothers, though without AI [53]. Community-driven approaches in rural Bangladesh similarly point to the value of locally grounded mental-health support [54]. Given high smartphone uptake and demonstrated interest in chat-based support, AI chatbots have clear potential in this space, provided that acceptability, literacy, privacy, and local-language accessibility are addressed [55–57]. The present work situates perinatal mental-health support within a Bangla-first, culturally sensitive agent and examines how women interpret empathy and reassurance in this context.

1.1.4 Trust in conversational health systems

The effectiveness of maternal health chatbots depends heavily on user trust and engagement, which are shaped by transparency, reliability, and cultural relevance. Users are more likely to rely on systems that provide clear explanations, behave consistently, and align with their linguistic and cultural context. Chung et al. reported significant positive associations between perceived trustworthiness and engagement in an AI-driven perinatal health application, with participants finding the system easy to use and to learn even where perceived usefulness was more variable [58]. Bickmore et al. found that usability and satisfaction with a conversational agent remained stable over 6- and 12-month follow-ups, indicating durable engagement [27].

Despite these results, most existing systems offer limited support for agentic capabilities such as tool use, retrieval over a curated corpus, and conversational continuity, which sensitive and evolving maternal health contexts arguably demand. This gap motivates a multilingual, agent-based design that combines localized medical knowledge, real-time retrieval, and within-session continuity, while treating visible tool use and retrieved-passage display as design resources for trust rather than as afterthoughts. Persistent, cross-session personalization would extend this

further, but it carries disclosure risks that are particularly salient in stigmatized health contexts; we return to that trade-off in the Discussion.

1.1.5 Agentic AI and tool-oriented chatbots

Recent work frames *agentic AI*, meaning autonomous and context-aware systems that orchestrate multiple tools to pursue goals, as a next stage for conversational agents [59]. Unlike rule-based or single-model systems, agentic frameworks can integrate diverse data sources and invoke external tools such as APIs, retrieval modules, translation engines, or web search as needed [60, 61]. The Model Context Protocol (MCP) standardizes how LLMs access such tools and data across heterogeneous sources [62, 63]. In a maternal health setting, this allows an assistant to combine curated clinical guidance with real-time search and conversational memory in a single dialogue.

Most current maternal chatbots, however, remain single-model or rule-based, and tool-enabled designs are only beginning to be evaluated. A key open question is whether such systems improve real-world accuracy and utility without compromising safety, since generative components can still produce unsupported statements even when grounded by retrieval [64]. This concern is especially acute for direct-to-user maternal health advice, where errors carry clinical consequences, and it motivates pairing tool orchestration with explicit safeguards and, where appropriate, human oversight. The present study operationalizes an agentic design for maternal care in Bangladesh and reports what MCP-enabled orchestration produced in practice, together with the equity, trust, and safety considerations that such deployment entails. We do not compare this architecture against a rule-based or single-model implementation, and we therefore make no claim that it outperforms one.

Table 1: Representative AI chatbots and digital tools in maternal and perinatal health, grouped by setting. The selection is purposive and illustrative rather than the result of a systematic search. The technology column distinguishes rule-based, single-model AI, and agentic (tool-orchestrating) systems. To our knowledge, the present work is the only agentic, Bangla-first system among those identified. (** denotes a non-peer-reviewed source; † denotes present work.)

Study / Chatbot	Region / Language	Target / Use case	Technology	Key findings
High-income deployments				

Table 1 continued from previous page

Study / Chatbot	Region / Language	Target / Use case	Technology	Key findings
Kerimoglu et al., 2025 [16]	Turkey, Turkish	Breastfeeding counseling (postpartum)	Rule-based	Higher breastfeeding self-efficacy; lower anxiety in AI group
Ozer et al., 2025 [17]	Global, English	Lactation FAQs	AI	GPT-4o highest quality; noted hallucination risk
Liu et al., 2017 [34] (Premom)	USA	Fertility charting and cycle insights	Rule-based	Widely adopted consumer fertility app
McAlister et al., 2025 [18] (Moment for Parents)	USA	Mental health (pregnant / postpartum)	AI	Most users re-engaged; advice rated relevant
Siwicki, 2024 [31] (Penny)**	USA	Postpartum ("fourth trimester") care	AI	High engagement; many questions answered by AI
Suharwardy et al., 2023 [19] (Woebot)	USA	Postpartum depression support	AI	Greater depression-score reduction vs. control
Rivera et al., 2024 [23]	USA (Eng. / Span.)	Postpartum newborn-care reminders	Rule-based	High open rate; rated timely and easy to use
Pollak et al., 2014 [33] (PregChat)	UK	Gestational weight-gain awareness	Rule-based	High response rate; most would recommend
Vaira et al., 2018 [32] (MamaBot)	Multi-region	Fetal-health tracking and tips	Rule-based	Notable community engagement
Other LMIC settings				

Table 1 continued from previous page

Study / Chatbot	Region / Language	Target / Use case	Technology	Key findings
Siddiqi et al., 2024 [39]	Pakistan (Urdu / Eng.)	Child immunization queries	AI	Feasible and well-rated in low-literacy setting
Deva et al., 2025 [44] (Myna Mahila)	India	Reproductive-health Q&A	AI	Culturally sensitive co-designed chatbot
Bangladesh systems				
Rahman et al., 2019 [65] (Disha)	Bangla	General symptom checker	Rule-based	Early Bangla-language health chatbot
Acme AI, 2024 [37] (Probahini)**	Bangla / Banglish	Menstrual-health Q&A	AI	Uses Bangla and Banglish; startup solution
Jui, 2021 [36] (Maya)**	Bangla / Eng. / Banglish	General telehealth queries	AI	Resolves a large share of queries via AI
Ahamed et al. [40] (SuSastho.AI)**	Bangla	Adolescent SRH and mental health	AI	Under development; multilingual health info
Mohtasim et al., 2024 [41] (GorbhoShongi)	Bangla	Personalized care and daily tips	AI	Prototype stage; not yet field-validated
Baby and Me† (this work)	**Bangla / Banglish**	**Maternal + perinatal mental health**	**Agentic**	**RAG + MCP + session memory; formative evaluation only**

1.2 Objectives

The aim of this study was to design a context-aware, retrieval-augmented conversational agent that delivers accurate, timely, and culturally sensitive maternal health guidance to pregnant and postpartum women in Bangladesh, and to evaluate it formatively. We situate the work within the realities of low-resource settings and report a four-phase study covering a requirements survey, construction of a curated Bangla knowledge base, system design and implementation, and a formative evaluation combining a screening of candidate model configurations with a usability and trust study among end users. This is a feasibility and design study. It was not designed to estimate effectiveness, and it includes no control group or comparison condition. Accordingly, we address three descriptive questions:

1. What informational and emotional needs, health concerns, and language preferences do pregnant and postpartum women in Bangladesh report, and what design requirements for a conversational maternal health agent follow from them?

2. Among cost-effective candidate language models, which configuration offers the most favourable balance of grounding in a curated Bangla maternal health corpus against response latency, and how does the number of retrieved passages affect that balance?

3. How do end users in this setting rate a deployed Bangla-first, tool-orchestrated agent on ease of use, clarity, trust, alignment with their concerns, and perceived empathy, and what do their open-ended responses and interviews indicate about where the design falls short?

In addressing these questions, the work makes four contributions. First, we curate and adapt maternal and perinatal health information, including mental health content, into a Bangla vector knowledge base, addressing the scarcity of high-quality, localized medical information in the native language, and we report the sourcing constraints encountered so that they may inform similar efforts in other low-resource language settings. Second, we describe the design of a retrieval-augmented conversational agent that accepts Bangla, Banglish, and English input and replies in Bangla, intended to be accessible across literacy levels and linguistic preferences. Third, we document how a retrieval module, web search, and session-scoped conversational memory can be coordinated through the Model Context Protocol (MCP) in a maternal health setting, and we report the configuration screening used to select a deployment model; we do not compare this architecture against a rule-based implementation, and we make no claim of

superiority over one. Fourth, we report survey and interview findings that surface the pain points and unmet needs of pregnant and postpartum women in Bangladesh, and we report the gap we observed between high usability ratings and lower, more dispersed ratings of trust and empathy, which we regard as the most consequential finding for the design of relational health technologies in stigmatized contexts.

2 Methods

2.1 Study design and setting

By integrating user-reported needs into both the knowledge base and interface design, this user-centered, mixed-methods study sequentially executed four phases: (1) a requirements survey with 72 pregnant and postpartum women, (2) construction of a curated Bangla maternal health knowledge base, (3) development of a retrieval-augmented conversational agent deployed as a lightweight web app, and (4) an evaluation combining model screenings, deployment feedback, and semi-structured interviews.

As a formative, non-comparative study without a control group, randomization, or comparative baseline, it does not support claims regarding relative effectiveness. Instead, its purpose was to evaluate whether a Bangla-first, retrieval-grounded conversational agent could be successfully built, deployed, and received by end users in this setting.

The work was carried out in Bangladesh. Formative survey recruitment took place online and in person at hospitals and clinics in Pabna, a district in northwestern Bangladesh with urban, semi-urban, and rural catchment areas. The deployed system was accessible over the public internet and was not restricted by location, but the evaluation cohort remained closely tied to the original recruitment site: the application was distributed first to the women who had taken part in the formative survey, and 21 of the 24 feedback respondents (87.5%) came from that group. Findings should therefore be read as reflecting a single recruitment setting rather than the country as a whole.

2.2 Problem formulation

The agent was required to produce (1) accurate and contextually relevant responses to pregnancy and postpartum queries, and (2) responses that are understandable, culturally appropriate, and empathetic. We therefore formulated the task as retrieval-augmented generation (RAG), in

which the agent retrieves relevant passages from a curated maternal health knowledge base and generates responses grounded in that evidence, supporting factual reliability while tailoring answers to the query.

Sensitive topics like perinatal mental health and stigma require more than factual accuracy, so we framed those queries to generate supportive, non-judgmental responses to build trust. Consequently, the deployed system pairs evidence-grounded retrieval for factual accuracy with empathetic generation for emotional support.

2.3 Ethics approval and consent

This study was conducted under research protocol BRACUIRB_220240006, "Promoting Sustainable Technologies: Investigating the Technological, Privacy, and Social Good Needs of Diverse Professionals and Communities in Bangladesh for Designing Ethical and Culturally Responsive Systems", which was reviewed and approved by the Institutional Review Board of BRAC University, Dhaka, Bangladesh, on 6 May 2024 and categorized as eligible for expedited review. The protocol is a programme-level approval covering the design and evaluation of culturally responsive technologies with communities in Bangladesh, and the work reported here constitutes one strand of that programme. The principal investigator on the protocol was based at BRAC University at the time of approval and is now affiliated with United International University. The approval covers both the formative requirements survey and the subsequent deployment and feedback phases described below. All methods were carried out in accordance with the Declaration of Helsinki and with the institution's research ethics requirements.

Before participation, all respondents were informed of the purpose of the study, the voluntary nature of participation, and their right to withdraw at any stage without consequence. Informed consent was obtained from all participants prior to data collection. Interview participants were additionally asked for permission to record the conversation before the interview began. In accordance with the conditions of approval, signed informed consent and disclosure forms were maintained for all participants throughout data collection, and the study data together with the investigator's copies of the signed consent forms are retained by the principal investigator for at least three years following the termination of the project. The one exception to documented written consent is described below.

Three survey respondents (3/72, 4.2%) selected the "less than 18 years" age band. The instrument captured age as a band rather than as an exact value, so we are unable to establish

retrospectively whether any of these three participants was under the age of 16. The approved protocol did not restrict participation by age. All three were pregnant women attending the participating clinics, and participation was voluntary and unsolicited: the survey was offered to women attending the clinic and only those who wished to take part did so. Each was informed of the purpose of the study and of her right to decline any item or to stop at any point, and each provided her own informed consent. In each case an accompanying parent or guardian was present at the clinic. The purpose of the study, the voluntary nature of participation, the content of the questionnaire, and the participant's right to decline any item or to stop at any point were explained to the guardian as well as to the participant, and the guardian gave explicit verbal informed consent before the questionnaire was administered. This consent was informed in the same terms as that obtained from the participant herself; it was given verbally and was not separately documented in writing. We confirm that informed consent for study participation was obtained from a parent or legal guardian in the case of every participant who selected the "less than 18 years" age band. The instrument was non-interventional and involved no clinical procedure, no biological sampling, and no diagnostic assessment, and the protocol was classified as eligible for expedited review on the basis of minimal risk. No participant under 18 took part in the deployment phase or in the interviews. We note as a limitation of the consent procedure that, because exact ages were not recorded, we cannot confirm whether documented guardian consent would have been required for any of these three participants under the standard applied to minors below the age of 16.

Given the sensitivity of maternal and perinatal mental health, particular care was taken at each stage. Survey and interview items were reviewed by a licensed physician collaborating with the research team to ensure that wording was clinically appropriate and unlikely to cause distress, and participants were directed to professional medical or psychological support where relevant. Respondent names were recorded during administration to permit de-duplication and to link participants across the formative and feedback surveys; these identifiers were removed prior to analysis, and no directly identifying information appears in the analyzed dataset or in this report. Analyzed data were stored with restricted access. Free-text responses provided in Bangla were translated into English for analysis with attention to preserving meaning and cultural context, and quotations reported here are attributed using participant codes only.

Users of the deployed system were not required to create an account or provide identifying information, and conversations were retained only against a session identifier not linked to any

user identity. The interface displayed a notice stating that the system provides information for educational purposes and does not substitute for professional medical advice.

2.4 Formative requirements survey

To ground *Baby and Me* in the lived realities of pregnant and postpartum women in Bangladesh, our formative survey identified their health concerns, emotional needs, and expectations for a digital support tool. This provided the empirical basis for the system's design requirements presented in the Results.

2.4.1 Eligibility and sampling

Participation was open to women who were currently pregnant or who were recent mothers, were resident in Bangladesh, and were willing and able to provide informed consent. No formal exclusion criteria were applied. In particular, no restriction was placed on parity, pregnancy risk status, literacy, or device ownership, and the resulting sample spans first-time (nulliparous) and multiparous women, every age band and income band offered by the instrument, and women with and without reported complications, as shown in the parity, income, and health-condition distributions reported in the Results.

Recruitment used convenience sampling rather than probability sampling, and the resulting sample is not population-representative. No formal sample size calculation was performed, because the purpose of this phase was to elicit design requirements rather than to estimate population parameters with a specified precision; the target was a sample large enough to span the socioeconomic, educational, and geographic variation relevant to design decisions.

2.4.2 Survey design and administration

We developed a 17-item questionnaire covering pain points, pregnancy complications, positive experiences, and physical and mental health conditions. Most items used a tick-box format to reduce response burden, with a small number of open-ended items included to capture nuance. Three items (11 to 13) permitted selection of up to five options. The instrument was developed by the research team and reviewed by a licensed physician for clinical appropriateness and for the risk of causing distress, but it was not a previously validated instrument and was not formally piloted with a separate sample. The full instrument is provided in Supplementary File 1.

The survey was first distributed through social media, but this yielded limited reach and

skewed toward more digitally connected respondents. To obtain broader representation across socioeconomic and literacy levels, we then administered the survey in person at hospitals and clinics in Pabna, a district in northwestern Bangladesh that includes urban, semi-urban, and rural catchment areas, and thus allowed us to reach women who would have been missed by an online-only approach. In total, 72 women participated.

The online arm was an open survey: no password or invitation code was required, and the link was shared publicly. Participation was voluntary and unincentivized, no completeness checks were enforced, and neither adaptive branching nor randomization of item order was used. Duplicate submissions were identified and removed during data preparation using the respondent name field described below, rather than by IP address or cookie, because devices were shared among interviewers in the in-person arm. The number of women approached in person, the number who declined, and the number of views of the online link were not recorded, so a response rate cannot be calculated for either arm. This is a limitation of the recruitment procedure and we note it as such.

Notably, 70.8% of respondents completed the survey with assistance from an interviewer or family member rather than unaided. This is reported as a finding about digital and literacy barriers, but it also has methodological implications: assisted completion may introduce social desirability bias, particularly on the mental health items, where a respondent answering in the presence of a family member may under-report distress. We return to this in the Limitations.

2.4.3 Data preparation and analysis

Survey responses were cleaned prior to analysis. We removed duplicate submissions, corrected typographical and formatting inconsistencies, and translated Bangla free-text responses into English to enable uniform analysis. Categorical fields (for example, education, occupation, and residence) were standardized into consistent classes, and semantically similar open-ended responses were grouped (for example, “lower back pain” and “body aches” were combined under musculoskeletal discomfort); groupings were agreed within the research team and checked against the original responses. Entries with missing or ambiguous values were flagged and, where imputation was not appropriate, excluded from the specific analyses requiring complete data. Respondent identifiers collected during administration were removed before analysis, and the analyzed dataset contains no directly identifying information.

Responses were summarized descriptively as counts and percentages. Demographic per-

centages are expressed as a proportion of participants; for the three multiple-response items, percentages are expressed as a proportion of all selections made across participants rather than of participants, and the corresponding denominators are stated in the Results. No inferential statistical testing was performed at this stage, since the purpose was descriptive requirement elicitation rather than hypothesis testing.

2.5 Knowledge base construction

Retrieval quality bounds the quality of a RAG system's outputs, so we prioritized curation of high-quality, contextually appropriate sources [66].

2.5.1 Challenges in sourcing Bangla maternal health content

We surveyed books, journal articles, government health portals, NGO resources, blogs, and social media groups. Several recurring problems limited their suitability. Much Bangla online content consisted of machine translations of English material, often outdated and lacking cultural nuance. Few sources were clinically vetted or reviewed by medical professionals, and misinformation was common in forums and social media, including unverified folk remedies. Localized content addressing the specific realities of maternal health in Bangladesh, such as region-specific nutrition practices and the substantial minority of births that still occur outside health facilities, was scarce, as was coverage of perinatal mental health, postpartum recovery, and miscarriage support. We found no publicly available structured or machine-readable Bangla medical corpora, and Bangla medical terminology lacked standardization, with multiple spellings and colloquial variants that complicate retrieval. A complete enumeration of these sourcing challenges is provided in Supplementary File 2. These conditions motivated a conservative sourcing strategy built on a small number of carefully vetted resources rather than broad web crawling.

2.5.2 Curated sources

Two sources met our criteria for accuracy, relevance, and cultural applicability. *Shohay Health*[1] is a Bangladeshi health-information initiative that states its content is reviewed by clinicians through an editorial process; we did not independently audit that process, and instead applied our own verification step described below. It contributed material covering weekly pregnancy development, pregnancy-related complications, nutrition, and postpartum care. *FirstCry (Bangla)*[2], a large

[1] https://shohay.health/about-en
[2] https://banglaparenting.firstcry.com/

South Asian digital parenting platform, contributed material spanning fetal development, dietary and lifestyle guidance, management of complications, and infant care. After the screening described below, 247 articles were retained across the two sources to form the corpus. Both sources are third-party copyrighted works. We did not seek formal permission from either publisher before crawling. Content was used solely to construct a retrieval index for non-commercial academic research; it is not redistributed in whole or in part, is not reproduced in this article or in any supplementary material, and the deployed system returns generated summaries rather than verbatim source text. We note the absence of explicit publisher permission as a limitation of the corpus construction.

2.5.3 Preprocessing and verification

Raw articles required substantial preprocessing. Because entire maternal sections of both sites were crawled, each article was manually reviewed and items that were promotional or only tangentially related to maternal health were excluded (for example, gift-guide articles were discarded while articles on managing gestational diabetes were retained). Within retained articles, promotional phrases and repeated disclaimers were removed, personal references and identifiers were stripped, and HTML markup, scripts, and metadata were removed using regular expressions and parser libraries so that only human-readable content remained. Counts of items removed at each individual screening step were not retained; the reported figure of 247 articles is the number surviving the full screening pipeline.

Key medical claims were cross-checked against authoritative sources including the World Health Organization and other clinical references, and were reviewed by a licensed physician collaborating with the research team. Where sources conflicted with clinical consensus, content was corrected rather than propagated. For example, a source claim that papaya must be avoided entirely during pregnancy was revised to reflect the more precise consensus that unripe papaya carries risk. Full preprocessing procedures are detailed in Supplementary File 2.

2.5.4 Embedding and vector store

The cleaned corpus was segmented using LangChain's `RecursiveCharacterTextSplitter` with a chunk size of 800 characters and an overlap of 300 characters. This configuration was chosen to preserve coherent units such as weekly pregnancy guides and complication summaries while retaining continuity across chunk boundaries; smaller chunks fragmented longer guides, while

larger chunks diluted retrieval relevance. We did not conduct systematic benchmarking across chunking parameters, which remains a limitation.

Chunks were embedded using BAAI/BGE-M3[3], a multilingual embedding model covering more than 100 languages including Bangla, which has demonstrated strong performance on low-resource languages and produces 1024-dimensional representations suitable for both Bangla and English content [67]. Embeddings were stored in a PostgreSQL database with the `pgvector` extension, hosted on Supabase [68], which supports nearest-neighbor search by cosine similarity with low query-time latency. The resulting knowledge base comprises 2,945 indexed passages derived from the 247 retained articles, each carrying metadata tags (for example, nutrition, gestational week, postpartum care) to support domain-specific retrieval.

2.6 Agent design and implementation

Figure 1 shows the system architecture, including the interaction between frontend, backend, language model, conversation memory, and external tools accessed through the Model Context Protocol (MCP).

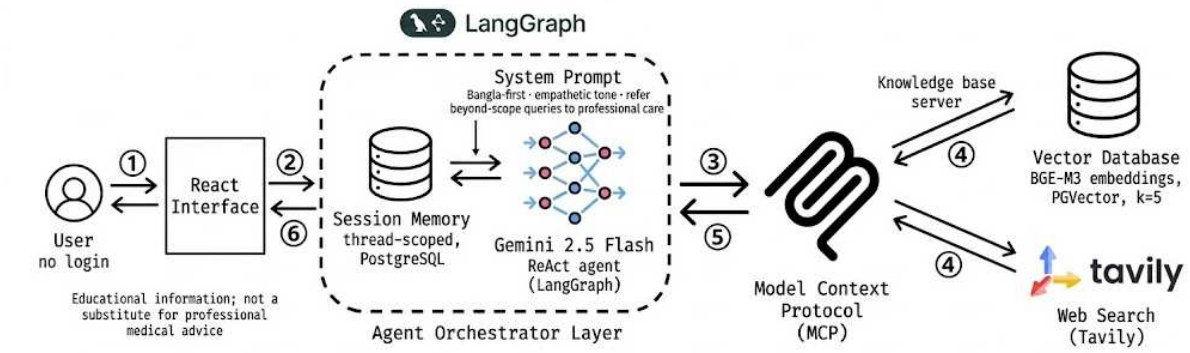


Figure 1: System architecture. A user query in Bangla or Banglish (1) enters the React interface and passes to the FastAPI backend (2). Here, a LangGraph ReAct agent running Gemini 2.5 Flash, guided by a system prompt specifying Bangla responses, an empathetic mood-matched tone, and professional care referrals for out-of-scope queries, determines if retrieval is required (3). Tool calls route via Model Context Protocol to either a web search server or a knowledge base server running semantic retrieval over BGE-M3 embeddings in a PostgreSQL pgvector store ($K = 5$) (4). Retrieved context returns to the agent (5), which streams a response with visible tool-invocation indicators back to the interface (6). Conversation state is retained via a session thread identifier without linking to a user account.

2.6.1 Language model selection

Because the target population communicates primarily in Bangla, with frequent code-mixing of Bangla and English (Banglish), the language model needed robust multilingual competence while

[3]https://huggingface.co/BAAI/bge-m3

remaining cost-effective to deploy. We shortlisted three models balancing multilingual capability, performance, and deployment efficiency. **Qwen 2.5:14B** is an open-source 14-billion-parameter multilingual model with demonstrated performance on low-resource languages including Bangla, and offers the option of self-hosting to control long-term cost [69, 70]. **Gemini 2.5 Flash-Lite** is a lightweight variant optimized for rapid response, suited to mobile-first, low-bandwidth deployment contexts, at some cost to reasoning depth [71]. **Gemini 2.5 Flash** provides greater reasoning capability and contextual depth for nuanced maternal health queries at higher computational cost [71].

The two Gemini variants were accessed through the provider's hosted API, while Qwen 2.5:14B was served locally through the Ollama runtime. Default decoding parameters were used throughout; no temperature or sampling adjustments were applied to any of the three models. Because the open-weight model ran in a local environment and the hosted models ran on provider infrastructure, the latency figures reported in the Results reflect the serving environment as well as the model, and are not directly comparable across the two settings.

2.6.2 System prompt design

Model selection alone does not ensure safe, empathetic, and culturally appropriate output, so we designed a system prompt around four principles. First, *language handling*: the agent accepts input in Bangla, Banglish, or English, and is instructed to reply in Bangla regardless of the script in which the question was asked. This asymmetry is deliberate. It reflects the formative survey finding that a substantial minority of women were comfortable only in Bangla while many typed most easily in romanized script, so accepting Banglish input lowers the entry barrier while replying in Bangla keeps the answer readable for the group least served by English-dominant platforms. Second, *empathetic and mood-matched tone*: the prompt directs the model to comfort the user, to adjust its register to the emotional content of the question, to use emoji where appropriate, and explicitly not to sound robotic, reflecting the emotional nature of maternal health conversations. Third, *tool discipline*: the knowledge base is always consulted first, and web search is used only when the knowledge base is insufficient. Fourth, *scope and referral*: questions outside pregnancy, childbirth, and parenting are politely declined, and the model is directed to plain, accessible phrasing on the assumption that users may not be technically confident. The prompt was refined iteratively through pilot interactions to improve clinical appropriateness and reduce unsupported statements. It is reproduced verbatim in Supplementary File 5.

2.6.3 Model Context Protocol integration

To extend the model beyond standalone generation, we integrated the Model Context Protocol (MCP)[4], a standardized interface through which language models interact with external tools and data sources [62]. We implemented two MCP servers. The **knowledge base server** connects to the vector store and performs semantic search over the embedded maternal health corpus, returning the K most similar passages to ground the response. The **search server** interfaces with the Tavily web search API to retrieve current information when a query concerns recent guidelines or material not covered by the static knowledge base. The system prompt instructs the agent to consult the knowledge base first and to fall back to web search only when the retrieved passages are insufficient. The protocol allows tools to be added or updated without modifying the underlying model, and this modularity distinguishes the architecture from single-model and rule-based designs.

The two tools differ in their safety properties. Passages returned by the knowledge base server come from the curated, physician-reviewed corpus described above. Content returned by the search server does not: it originates from the open web and is not subject to that verification pipeline. Responses drawing on web search are therefore grounded in a weaker sense than those drawn from the curated corpus. The interface surfaces which tool was invoked for each response, so the distinction is visible to users, but the system does not otherwise restrict or filter search results. We did not analyse how frequently the search server was invoked relative to the knowledge base server during the deployment. We treat both points as material limitations of the current safety design and return to them in the Discussion.

2.6.4 Conversation memory

A session-based memory module retains context within a single interaction session, allowing the agent to reference recent turns. Short-term continuity supports clearer and more reassuring exchanges, which is particularly relevant where users ask successive related questions. The system requires no account creation or login, a deliberate choice given that disclosing identity can feel risky in stigmatized health contexts; consequently, memory does not persist across sessions, and the system does not build longitudinal user profiles. This design trades personalization for lower disclosure risk, a tension we return to in the Discussion.

[4] https://modelcontextprotocol.io/docs/getting-started/intro

2.6.5 Application architecture

Baby and Me is implemented as a lightweight web application. The frontend uses React 19 with TypeScript, Tailwind CSS, and Vite, and accepts input in Bangla, Banglish, or English with responses streamed in real time. Visual indicators for processing and tool invocation were included to make the system's reasoning visible to users. The backend uses FastAPI for asynchronous session handling and streaming, LangChain and LangGraph to coordinate model interaction and memory, PostgreSQL for conversation persistence, and MCP as the bridge to external tools. The interface was designed for use on both mobile and desktop devices by women with limited digital literacy.

2.7 Interaction design

Baby and Me was designed as more than a question-and-answer interface. The intent was a maternal health companion that fits the everyday rhythms of pregnancy in Bangladesh, foregrounding immediacy, trust, and empathy in order to lower barriers to access while creating a space in which medical and emotional concerns can coexist. The design decisions described below follow from the three requirements derived from the formative survey, which are reported in the Results. The interface screenshots reproduced in this section show illustrative example conversations composed by the research team to demonstrate system behaviour; they are not transcripts of participant sessions.

2.7.1 Input and process transparency

To minimize entry barriers for women who may already feel vulnerable when seeking maternal health support, the application requires no sign-up, no password, and no disclosure of personal details before use. Women can begin interacting as soon as they open the site[5]. This zero-friction approach reflects the observation that identity disclosure can feel risky in sensitive health contexts, particularly where miscarriage anxiety, mental health difficulties, or relationship concerns carry stigma.

The chat interface deliberately resembles the messaging applications women in Bangladesh already use daily. Reusing a familiar design metaphor reduces the cognitive load of learning a new interface: rather than navigating menus or completing forms, users type a question into a

[5]https://baby-and-me.vercel.app/

text box. This choice responds directly to the requirements survey, in which 70.8% of respondents needed assistance completing even a structured questionnaire.

Figure 2(a) shows the input box, positioned at the bottom of the interface as in commercial chat platforms. On submission, the interface enters a processing state (Figure 2(b)) in which the system makes its internal activity visible, displaying indicators of tool invocation, document retrieval, and response generation. These intermediate states were included for two reasons. As technical feedback, they are intended to prevent uncertainty about whether the system has received the query. As design elements, they are intended to signal that the system is actively working on the user's behalf, a quality often absent from static health information portals.

Because no account is required, conversations do not persist beyond the active session. This constrains the system's ability to personalize across visits, but it reduces the risk that sensitive disclosures are stored or later exposed, a consideration that is particularly salient where discussing pregnancy complications or mental health may be socially sensitive even within a household. The design therefore trades longitudinal personalization for lower disclosure risk, while retaining enough within-session continuity to support coherent multi-turn dialogue.

2.7.2 Response presentation

Responses are written to balance accuracy with accessibility. Messages avoid dense medical terminology in favor of plain Bangla phrasing, a choice grounded in the formative survey finding, reported in the Results, that 27.8% of participants were comfortable only in Bangla. Because maternal health questions are rarely purely informational and are often entangled with anxiety or stigma, responses are framed to provide reassurance alongside factual guidance.

Figure 3 shows the response flow. On receiving input, the system produces a concise message in Bangla, irrespective of whether the question was typed in Bangla script, romanized Banglish, or English (panel c). In the subsequent idle state (panel d), retrieved documents and tool outputs remain visible rather than being hidden once the answer is delivered. Technically this supports auditability; from the user's perspective it is intended to function as a signal of accountability, allowing women to see the basis of the information rather than receiving assertions without provenance. Whether this design intent was realized is examined in the Results and Discussion.

Some maternal health topics, including nutritional comparisons, trimester milestones, and warning signs, are conveyed more effectively in structured form. Figure 4 shows how the system renders table-based answers inline to improve scannability, allowing users to locate specific

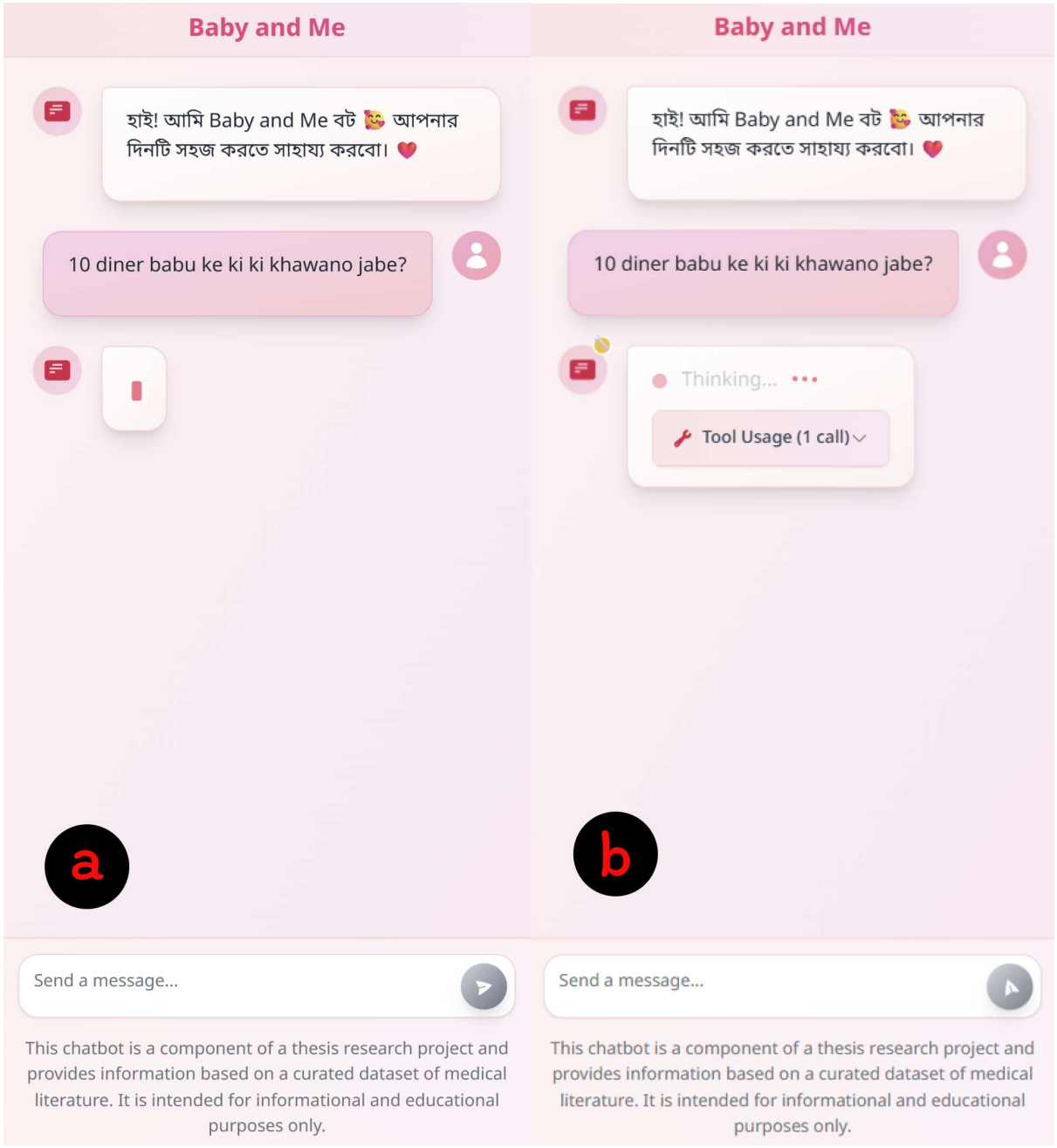


Figure 2: Input flow: (a) the persistent input box, and (b) the processing state, in which tool invocation and document retrieval are surfaced to the user. The interface is shown as deployed. English translation of the on-screen text: the agent's opening message reads "Hi! I am the Baby and Me bot. I will help make your day easier."; the user's question, typed in romanized Banglish, reads "What can a 10-day-old baby be fed?". All remaining interface text is in English in the deployed system.

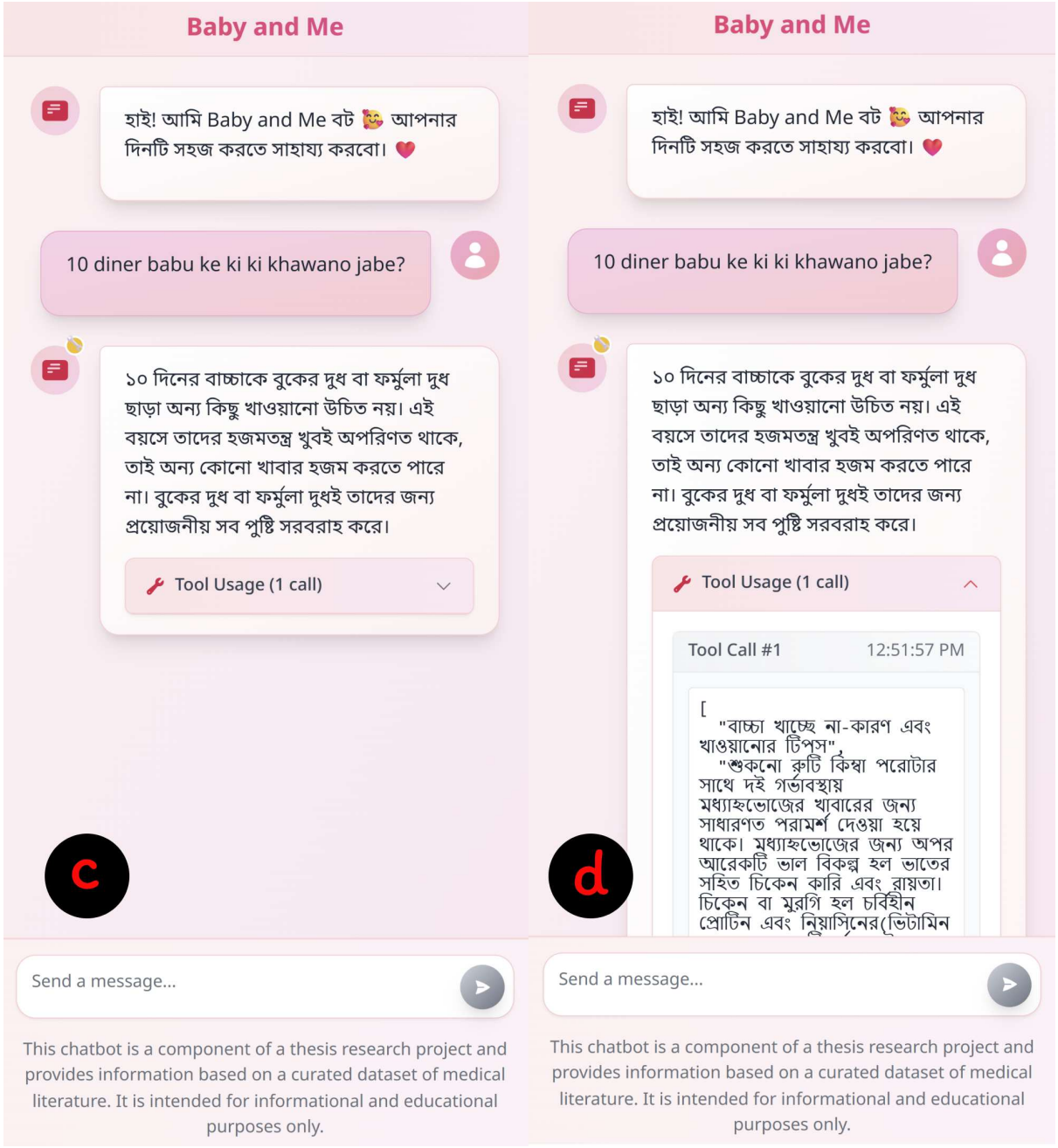


Figure 3: Response flow: (c) a generated response rendered in Bangla, and (d) the idle state, in which retrieved sources and tool outputs remain accessible. The interface is shown as deployed. English translation of the agent's response: "A 10-day-old baby should not be given anything other than breast milk or formula. At this age the digestive system is still very immature, so it cannot digest any other food. Breast milk or formula alone supplies all the nutrition the baby needs." Panel (d) additionally shows the expanded tool output, containing the Bangla passages retrieved from the knowledge base for this query.

information without parsing long paragraphs.

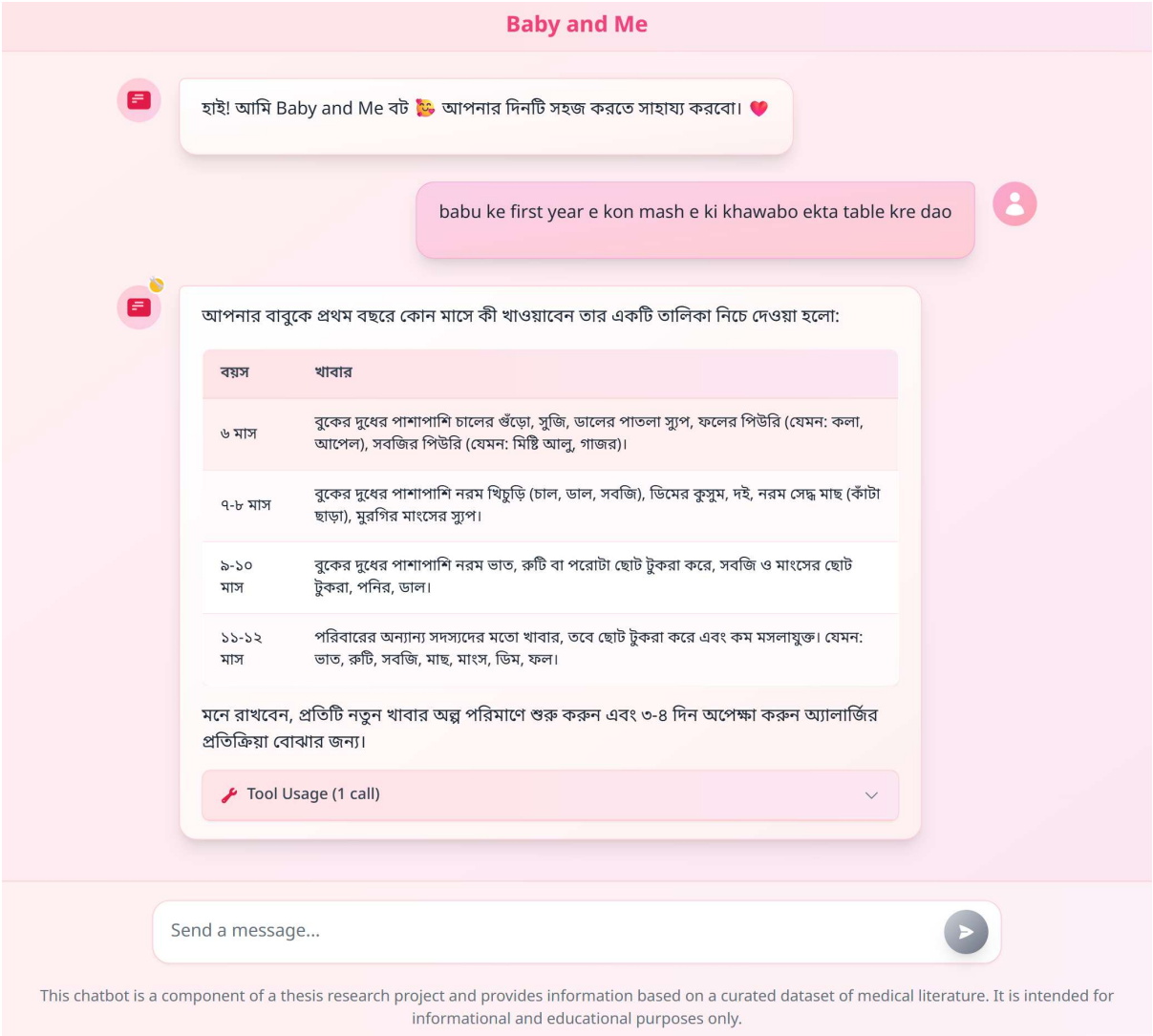


Figure 4: Rendering of a structured, table-based response within the conversation stream, shown as deployed. English translation: the user asks “Make me a table of what to feed my baby in each month of the first year”; the agent replies “Below is a table of what to feed your baby in each month of the first year”, followed by a table with the columns Age and Food. The rows give, at 6 months, rice powder, semolina, thin lentil soup and fruit or vegetable purée alongside breast milk; at 7 to 8 months, soft khichuri, egg yolk, yoghurt, deboned boiled fish and chicken soup; at 9 to 10 months, soft rice, small pieces of roti or paratha, small pieces of vegetables and meat, cheese and lentils; and at 11 to 12 months, ordinary family food cut small and lightly spiced. The closing note reads “Remember to introduce each new food in small amounts and wait 3 to 4 days to watch for any allergic reaction.”

2.7.3 Within-session continuity

Figure 5 illustrates how context is preserved across turns within a session. Although no account is required, the agent retains memory of prior turns in the current conversation, which allows it to tailor follow-up answers, avoid repetition, and maintain coherence. Restricting memory to the session bounds this benefit against privacy risk, as described above.

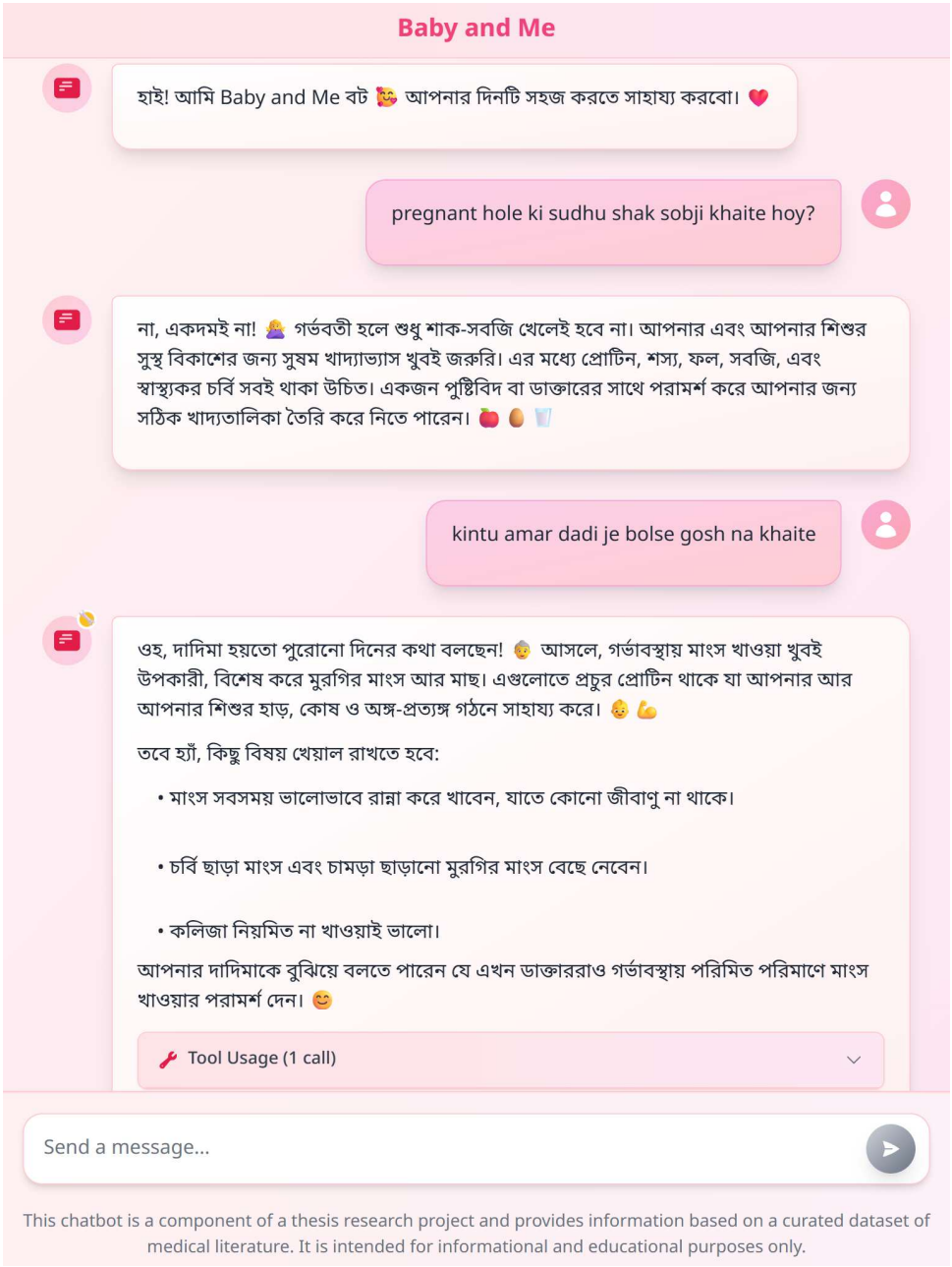


Figure 5: Preservation of conversational context within a session, allowing follow-up questions to be interpreted against prior turns. The interface is shown as deployed. English translation: the user asks "When you are pregnant, do you have to eat only leafy greens and vegetables?" and the agent replies that a balanced diet including protein, grains, fruit, vegetables and healthy fats is important, and suggests consulting a nutritionist or doctor. The user then writes "but my grandmother said not to eat meat", and the agent responds that meat, particularly chicken and fish, is beneficial in pregnancy because of its protein content, while advising that meat be cooked thoroughly, that lean cuts and skinless chicken be chosen, and that liver not be eaten regularly. Note that the second turn is interpreted against the first without restatement.

Across these interaction modes, transparency was treated as a design resource rather than only a technical property: surfacing retrieved passages, displaying tool invocation, and maintaining visible continuity were intended to position the system as accountable rather than authoritative. We examine how far that intent was borne out in the Discussion.

2.8 Evaluation design

We evaluated *Baby and Me* through three complementary components: a screening of candidate model configurations, a deployment feedback survey, and semi-structured interviews. The screening selected a configuration before deployment; the survey and interviews examined how end users received the deployed system.

2.8.1 Model configuration screening

We assembled an evaluation set of 30 frequently asked pregnancy and maternal health questions, drawn from questions commonly raised by pregnant and postpartum women in Bangladeshi online health communities and from the topics covered by the curated knowledge base. Reference answers were authored and verified by a licensed physician collaborating with the research team. The set was sized pragmatically rather than by power calculation, and its purpose was to discriminate among candidate configurations rather than to establish clinical accuracy.

Responses generated by the RAG system were scored against six criteria:

1. **Contextual accuracy**: whether retrieved documents and the generated response align with the query and conversational context, without misinterpretation or off-topic content. Higher is better.

2. **Chain-of-thought contextual accuracy**: the coherence and logical progression of the response, particularly for multi-turn or complex queries [72]. Higher is better.

3. **Conciseness**: brevity and directness, without excessive verbosity or repetition. Higher is better.

4. **Relevance**: the pertinence of retrieved information and the generated response to the user's actual need. Higher is better.

5. **Unsupported statements**: the proportion of the response consisting of claims that are not supported by the retrieved source documents or that are factually incorrect [64]. Lower

is better.

6. **Response latency**: median (P50) and worst-case (P99) end-to-end generation time, indicating real-time usability. Lower is better.

We evaluated two retrieved-context sizes, $K = 5$ and $K = 10$, where K is the number of retrieved chunks supplied to the generator; these values reflect commonly reported operating points in the retrieval literature [73, 74].

Scoring was automated, using a large language model-as-judge protocol [75] in which a language model rated each system output against the physician-authored reference answer on each criterion. Each criterion was scored on a scale from 0 to 1, and the values reported in the Results are the mean across the 30 questions. The two Gemini variants were accessed through the provider's hosted API and the open-weight model through a local Ollama runtime, so the latency measurements reflect the serving environment as well as the model.

We emphasise the limited role this screening plays in the paper. Automated judging can diverge from expert human judgment on specialized clinical content [76], and preference leakage can bias evaluation when the judging and generating models belong to the same family [77]. The scores were used to choose among candidate configurations before deployment, and we present them as the basis for the choice that was made. They are not offered as measures of model quality, as evidence of clinical accuracy, or as a comparison that other work should build on. The constraints on their interpretation are set out in the Limitations.

2.8.2 Deployment feedback survey

The application was made available first to women who had participated in the formative requirements survey and subsequently to a wider audience, and remained open for one week. Participants were asked to use the system over a period of several days before responding. Twenty-four people completed a structured feedback questionnaire covering usability, clarity, trust, alignment with their concerns, and perceived empathy, each rated on a 5-point Likert-type scale, together with free-text comments and three categorical items on adoption and privacy preference. The questionnaire is provided in Supplementary File 3.

Two features of this sample bear on interpretation and are reported in the Results. Twenty-one of the 24 respondents (87.5%) had taken part in the formative survey, so the feedback cohort overlaps heavily with the original Pabna recruitment. Item 4 of the questionnaire asked who

had used the application, and 15 respondents (62.5%) had used it themselves while 9 (37.5%) were reporting on use by another household member. Proxy responses were retained rather than excluded, because household-mediated use is characteristic of digital access in this setting and excluding them would have removed a substantive finding, but they are identified as such throughout.

The number of women invited or otherwise reached was not recorded, so a response rate cannot be calculated. The instrument was developed for this study; a validated usability instrument such as the System Usability Scale was not administered, which limits comparability with other work.

2.8.3 Semi-structured interviews

We conducted short semi-structured interviews with three women who had used the system for three days. These were intended not to reach thematic saturation but to provide illustrative depth and to contextualize the survey responses. Interviews were conducted in Bangla and were scheduled to last approximately 20 to 30 minutes. Participants were informed that their responses would be anonymized and that they could skip any question or stop at any time, and audio recording was carried out only with permission. Recordings were subsequently transcribed and translated into English. Transcripts were not returned to participants for comment and no repeat interviews were conducted. The interview protocol, including the consent preamble, is provided in Supplementary File 4.

2.9 Analysis

2.9.1 Quantitative analysis

Formative survey responses, deployment feedback ratings, and screening scores were summarized descriptively. Categorical variables are reported as counts and percentages, and rating scales as means with standard deviations together with the full distribution of responses across scale points. No inferential statistical testing was performed. We made this choice because the study has no comparison condition, the deployment sample of 24 provides very limited power, and the rating instrument is not validated, so that p-values would convey more confidence than the design supports.

2.9.2 Qualitative analysis

Free-text survey comments and interview transcripts were read in full by members of the research team and summarized into recurring topics through an inductive reading of the material. Given that only three women were interviewed, we did not conduct a formal thematic analysis, construct a codebook, or assess inter-coder agreement, and we do not claim saturation. The topics reported in the Results are therefore presented as illustrative accounts that contextualize the quantitative ratings rather than as analytically derived themes. Representative quotations were selected to convey the range of views expressed, including critical ones, and are attributed by participant code only.

2.9.3 Integration of quantitative and qualitative components

The two strands were integrated at the interpretation stage rather than during analysis: quantitative ratings established where the system performed well and where ratings were low or dispersed, and the free-text and interview material was used to suggest why. Where the two strands diverge, we report the divergence rather than resolving it in favour of either.

3 Results

3.1 Formative survey findings

3.1.1 Participant characteristics

Seventy-two women participated. Participants were predominantly of childbearing age: 42 (58.3%) were in the 25 to 35 band and 16 (22.2%) in the 18 to 25 band, with 10 (13.9%) aged 35 to 45, 3 (4.2%) under 18, and 1 (1.4%) over 45. Educational attainment varied from no formal schooling, 4 (5.6%), to a Master's degree or higher, 18 (25.0%), with Bachelor's degree holders the largest single group at 24 (33.3%). Homemakers, 29 (40.3%), and service workers, 19 (26.4%), were the most common occupations. Most participants lived in urban areas, 42 (58.3%), with the remainder split evenly between semi-urban and village settings, 15 (20.8%) each. Households ranged in size from 2 to 10 members (median 4).

The sample spanned a wide socioeconomic range. Annual household income was below 1 lakh BDT for 18 participants (25.0%), between 1 and 5 lakh for 23 (31.9%), between 5 and 8 lakh for 27 (37.5%), and above 8 lakh for 4 (5.6%). Parity was similarly varied: 26 women

(36.1%) were in their first pregnancy, 23 (31.9%) their second, 13 (18.1%) their third, and 10 (13.9%) their fourth or later, so the sample includes both nulliparous and multiparous women.

Personal smartphone and internet access was widespread, 58 (80.6%), though 11 (15.3%) did not own a smartphone and 3 (4.2%) relied on shared devices. Fifty-one respondents (70.8%) completed the survey with assistance rather than unaided, indicating meaningful literacy or digital-familiarity barriers. Table 2 summarizes the responses.

3.1.2 Language preferences

Language preferences reinforced the need for a Bangla-first design: 51 women (70.8%) were comfortable in both Bangla and English, 20 (27.8%) preferred Bangla only, and 1 (1.4%) preferred English only (Figure 6). This preference varied sharply with educational attainment (Figure 7): all participants with no formal or primary education preferred Bangla-only interaction, and 70.0% of those with secondary education did so, whereas bilingual comfort was concentrated among participants with college-level education or above. A Bangla-first design therefore matters most for precisely the group least served by existing English-dominant platforms.

3.1.3 Reported concerns and mental health

Across 378 complication selections made by the 72 respondents, the most frequently reported were concerns about the baby's health or development (38 selections, 10.1%), hormonal mood swings (35, 9.3%), fear of miscarriage or complications (27, 7.1%), and fear of childbirth or labor pain (27, 7.1%). Across 377 concern selections, the most frequent were safe delivery and labor pain (34, 9.0%), fear of C-section or surgery (32, 8.5%), risk of miscarriage or complications (29, 7.7%), physical changes and body image (29, 7.7%), and the impact of pregnancy on mental health (27, 7.2%) (Figure 8).

On mental health, 44 participants (61.1%) described themselves as mostly stable and positive, while 14 (19.4%) reported occasional anxiety or stress, 9 (12.5%) frequent sadness or mood swings, and 5 (6.9%) depression or serious emotional distress (Figure 9). Just under two in five respondents therefore reported some degree of psychological difficulty during pregnancy. As noted in the Methods, 70.8% of respondents completed the instrument with assistance, which may have suppressed disclosure on these items.

3.1.4 Derived design requirements

Two patterns emerged with direct design implications. First, digital readiness was uneven: urban women showed high smartphone and internet access, while rural and lower-income women faced literacy gaps, device sharing, and limited digital familiarity, underscored by the 70.8% who needed assistance completing the survey and by the quarter of households reporting annual income below 1 lakh BDT. Second, language inclusivity was decisive, with a substantial minority preferring Bangla only and this preference concentrated among women with the least formal education. Mental health concerns were also prominent, indicating that the system should extend beyond information delivery to offer empathetic, stigma-sensitive support. Together, these findings translated into three design requirements, which guided the system described in the Methods:

1. **Multilingual, Bangla-first interaction**, ensuring accessibility for women who prefer their native language for comfort and trust.
2. **Simple, inclusive interface design**, accommodating users with limited digital literacy through clear, lightweight interaction.
3. **Empathetic, trustworthy responses**, combining clinically grounded guidance with reassurance and sensitivity to stigma around maternal mental health.

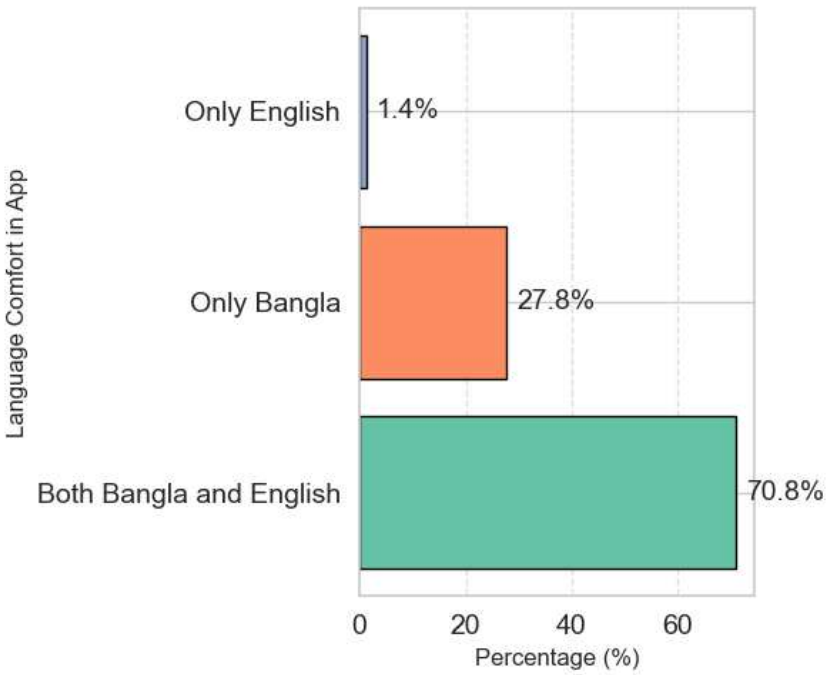


Figure 6: Language preference for a digital maternal health application ($N = 72$).

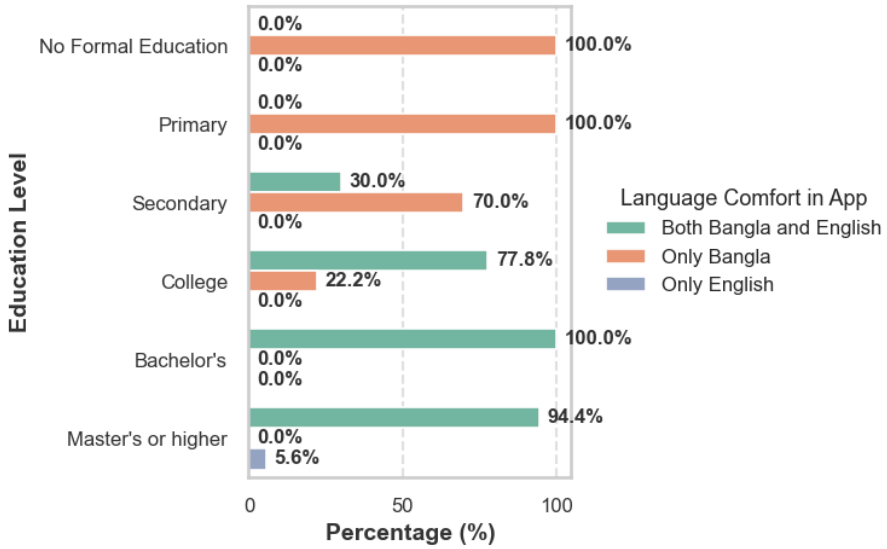


Figure 7: Language preference for a digital maternal health application by educational attainment ($N = 72$). Participants with no formal or primary education preferred Bangla-only interaction exclusively, while bilingual comfort increased with educational attainment.

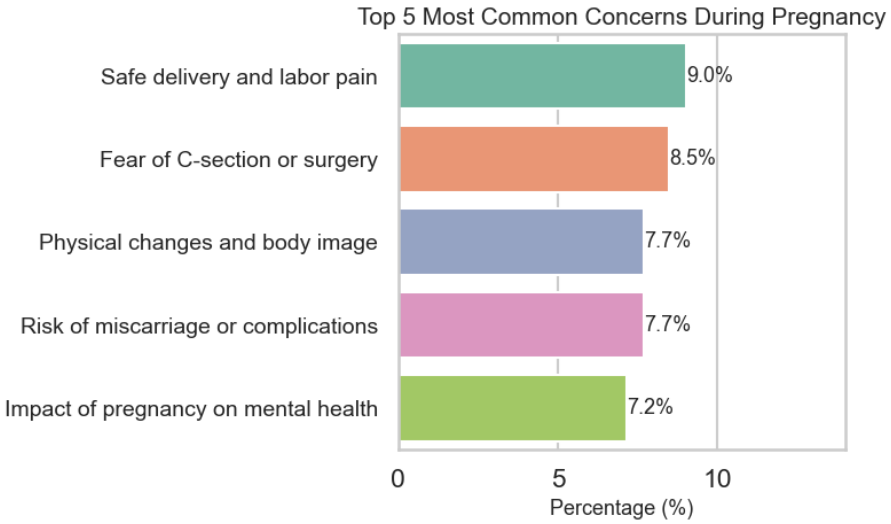


Figure 8: Top five reported concerns during pregnancy, as a percentage of all 377 concern selections made across the 72 participants (up to five selections permitted per participant).

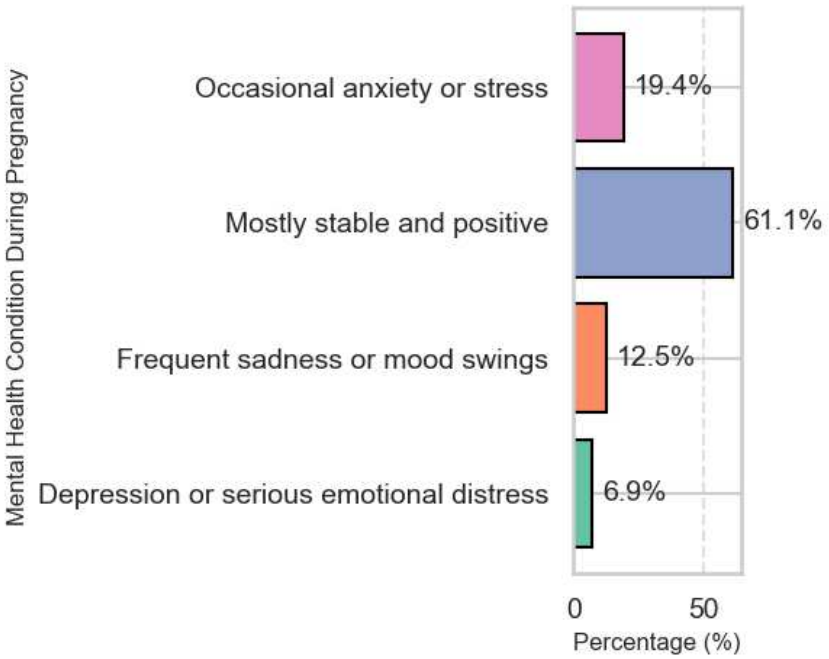


Figure 9: Self-reported mental health condition during pregnancy ($N = 72$).

Table 2: Summary of formative survey responses ($N = 72$). Demographic values are given as n (%) of participants. For the three multiple-response items, participants could select up to five options; values are given as the number of selections and as a percentage of all selections for that item, not of participants. The denominators are 378 selections for complications, 377 for concerns, and 379 for positive moments, each made across all 72 respondents. Percentages are rounded and may not sum to exactly 100. Age bands are reproduced as they appeared in the survey instrument, in which the boundary values 25 and 35 appear in two adjacent categories; respondents selected a single band. The parity item offered both "5" and "5+" as options; these are combined here as "5 or more".

Category	**Response**	n	%
Age	18–25	16	22.2
	25–35	42	58.3
	35–45	10	13.9
	Under 18	3	4.2
	Over 45	1	1.4
Education	Bachelor's	24	33.3
	Master's or higher	18	25.0
	Secondary	10	13.9
	College	9	12.5
	Primary	7	9.7
	No formal schooling	4	5.6
Occupation	Homemaker	29	40.3
	Service	19	26.4
	Business	8	11.1
	Student	8	11.1
	Daily wage labor	6	8.3

Continued on next page

Table 2: Summary of formative survey responses ($N = 72$). Demographic values are given as n (%) of participants. For the three multiple-response items, participants could select up to five options; values are given as the number of selections and as a percentage of all selections for that item, not of participants. The denominators are 378 selections for complications, 377 for concerns, and 379 for positive moments, each made across all 72 respondents. Percentages are rounded and may not sum to exactly 100. Age bands are reproduced as they appeared in the survey instrument, in which the boundary values 25 and 35 appear in two adjacent categories; respondents selected a single band. The parity item offered both "5" and "5+" as options; these are combined here as "5 or more". (Continued)

Category	**Response**	n	%
	Doctor	1	1.4
	Lecturer	1	1.4
Residence	Urban	42	58.3
	Semi-urban	15	20.8
	Village	15	20.8
Annual household income (BDT)	Less than 1 lakh	18	25.0
	1 to 5 lakh	23	31.9
	5 to 8 lakh	27	37.5
	More than 8 lakh	4	5.6
Number of pregnancies	1	26	36.1
	2	23	31.9
	3	13	18.1
	4	4	5.6
	5 or more	6	8.3
Smartphone access	Personal smartphone and internet	58	80.6

Continued on next page

Table 2: Summary of formative survey responses ($N = 72$). Demographic values are given as n (%) of participants. For the three multiple-response items, participants could select up to five options; values are given as the number of selections and as a percentage of all selections for that item, not of participants. The denominators are 378 selections for complications, 377 for concerns, and 379 for positive moments, each made across all 72 respondents. Percentages are rounded and may not sum to exactly 100. Age bands are reproduced as they appeared in the survey instrument, in which the boundary values 25 and 35 appear in two adjacent categories; respondents selected a single band. The parity item offered both "5" and "5+" as options; these are combined here as "5 or more". (Continued)

Category	**Response**	n	%
	No smartphone	11	15.3
	Shared access	3	4.2
Language preference	Both Bangla and English	51	70.8
	Bangla only	20	27.8
	English only	1	1.4
Survey completion	Assisted	51	70.8
	Unaided	21	29.2
Physical condition	Very good	29	40.3
	Mostly okay, minor issues	34	47.2
	Had complications	9	12.5
Mental condition	Mostly stable and positive	44	61.1
	Occasional anxiety or stress	14	19.4
	Frequent sadness or mood swings	9	12.5
	Depression or serious distress	5	6.9
Top complications (of 378 selections)	Baby's health or development	38	10.1

Continued on next page

Table 2: Summary of formative survey responses ($N = 72$). Demographic values are given as n (%) of participants. For the three multiple-response items, participants could select up to five options; values are given as the number of selections and as a percentage of all selections for that item, not of participants. The denominators are 378 selections for complications, 377 for concerns, and 379 for positive moments, each made across all 72 respondents. Percentages are rounded and may not sum to exactly 100. Age bands are reproduced as they appeared in the survey instrument, in which the boundary values 25 and 35 appear in two adjacent categories; respondents selected a single band. The parity item offered both "5" and "5+" as options; these are combined here as "5 or more". (Continued)

Category	**Response**	n	%
	Hormonal mood swings	35	9.3
	Fear of miscarriage or complications	27	7.1
	Fear of childbirth or labor pain	27	7.1
	Anxiety or depression	26	6.9
Top concerns (of 377 selections)	Safe delivery and labor pain	34	9.0
	Fear of C-section or surgery	32	8.5
	Risk of miscarriage or complications	29	7.7
	Physical changes and body image	29	7.7
	Impact of pregnancy on mental health	27	7.2
Positive moments (of 379 selections)	Feeling the baby's first kick	32	8.4
	Feeling proud as a mother	32	8.4
	Watching the belly grow	31	8.2
	Preparing for the baby	28	7.4
	Seeing the baby in an ultrasound	26	6.9

3.2 Model configuration screening

Before deployment we screened the three candidate models at two retrieved-context sizes in order to select a configuration. Table 3 reports the resulting quality and latency metrics. Across all quality dimensions, performance ordered consistently: Gemini 2.5 Flash performed best, Gemini 2.5 Flash-Lite intermediate, and Qwen 2.5:14B weakest. All quality scores were produced by the automated judging protocol described in the Methods and are used here to rank candidate configurations against one another; they are not a measure of clinical accuracy and should not be read as one.

3.2.1 Response quality

Contextual accuracy improved with model capability and with retrieved-context size. Qwen 2.5:14B scored 0.47 at $K = 5$ and 0.57 at $K = 10$; Flash-Lite reached 0.75 and 0.84; and Flash achieved 0.90 and 0.94. The consistent gain from $K = 5$ to $K = 10$ across all three models indicates that additional retrieved context supported more accurate grounding. Chain-of-thought contextual accuracy followed the same ordering, with Flash reaching 0.95 at $K = 10$, although the effect of K was less pronounced for both Gemini variants once sufficient context was available.

Conciseness and relevance showed a different pattern. Conciseness rose sharply from Qwen (0.47, 0.52) to Flash-Lite (0.75, 0.78) and Flash (0.93, 0.90), but for Flash the $K = 5$ condition slightly outperformed $K = 10$. Relevance behaved similarly, with Flash scoring 0.91 at $K = 5$ and 0.89 at $K = 10$. These reversals suggest a point of diminishing returns, beyond which additional retrieved passages introduce redundancy or peripheral material that dilutes the directness of the response rather than improving it.

Unsupported statements were least frequent for Gemini 2.5 Flash (0.11 at $K = 5$, 0.08 at $K = 10$) and most frequent for Qwen 2.5:14B (0.51 and 0.43). Flash-Lite was stable at 0.18 under both conditions. For Qwen and Flash, increasing K reduced unsupported statements, consistent with the interpretation that broader retrieval provides more material on which to ground a response.

3.2.2 Latency

Latency ordered inversely to model size. Flash-Lite was fastest, with median (P50) latencies of 5.10 s at $K = 5$ and 7.42 s at $K = 10$, and worst-case (P99) latencies of 8.50 s and 12.82 s. Flash was moderately slower at 7.46 s and 13.56 s (P50) and 10.57 s and 20.70 s (P99). Qwen

2.5:14B was slowest throughout, reaching 20.10 s (P50) and 40.40 s (P99) at $K = 10$. Increasing K raised latency for every model, reflecting the additional context processed. The open-weight model was served locally through Ollama while both Gemini variants were accessed as hosted services, so its latency reflects the local serving environment as well as the model itself and is not directly comparable.

3.2.3 Configuration selected for deployment

Taken together, Gemini 2.5 Flash offered the best balance of grounding and accuracy at latencies compatible with interactive use, and was selected for deployment. Qwen 2.5:14B combined the weakest grounding with the highest observed latency in our serving setup, making it unsuitable for real-time use in this setting despite its advantage of self-hosting. Although $K = 10$ yielded the highest contextual accuracy, the deployed configuration uses $K = 5$, trading a small accuracy margin for lower latency and reduced token cost, and reflecting the observation that conciseness and relevance did not improve beyond $K = 5$.

Table 3: Quality and latency metrics for three candidate models at two retrieved-context sizes (K), used to screen configurations before deployment. For contextual accuracy, chain-of-thought contextual accuracy, relevance, and conciseness, higher values indicate better performance. For unsupported statements and for latency, lower values indicate better performance; arrows in the column headers indicate the direction of improvement. Latency is reported in seconds at the 50th (P50) and 99th (P99) percentiles; the open-weight model was served locally and the two hosted models through their provider API, so latencies are not directly comparable across serving environments. The bold row marked with † ($K = 5$, Gemini 2.5 Flash) is the configuration selected for deployment.

K	Model	Contextual accuracy ↑	CoT contextual accuracy ↑	Relevance ↑	Conciseness ↑	Unsupported statements ↓	P50 latency (s) ↓	P99 latency (s) ↓
Retrieved context $K = 5$								

Continued on next page

Table 3: Quality and latency metrics for three candidate models at two retrieved-context sizes (K), used to screen configurations before deployment. For contextual accuracy, chain-of-thought contextual accuracy, relevance, and conciseness, higher values indicate better performance. For unsupported statements and for latency, lower values indicate better performance; arrows in the column headers indicate the direction of improvement. Latency is reported in seconds at the 50th (P50) and 99th (P99) percentiles; the open-weight model was served locally and the two hosted models through their provider API, so latencies are not directly comparable across serving environments. The bold row marked with † ($K = 5$, Gemini 2.5 Flash) is the configuration selected for deployment. (Continued)

K	Model	Contextual accuracy ↑	CoT contextual accuracy ↑	Relevance ↑	Conciseness ↑	Unsupported statements ↓	P50 latency (s) ↓	P99 latency (s) ↓
5	Qwen 2.5:14B	0.47	0.73	0.63	0.47	0.51	15.58	27.12
5	Gemini 2.5 Flash-Lite	0.75	0.85	0.80	0.75	0.18	5.10	8.50
5	**Gemini 2.5 Flash†**	**0.90**	**0.93**	**0.91**	**0.93**	**0.11**	**7.46**	**10.57**
Retrieved context $K = 10$								
10	Qwen 2.5:14B	0.57	0.80	0.65	0.52	0.43	20.10	40.40
10	Gemini 2.5 Flash-Lite	0.84	0.85	0.83	0.78	0.18	7.42	12.82
10	Gemini 2.5 Flash	0.94	0.95	0.89	0.90	0.08	13.56	20.70

3.3 Deployment feedback survey

3.3.1 Respondents

The application was distributed first to women who had taken part in the formative survey and then opened to a wider audience for one week. Twenty-four people completed the feedback

questionnaire, of whom 21 (87.5%) had also taken part in the formative survey. Mean age was 29.79 years (SD 5.51), ranging from 20 to 47 years.

Fifteen respondents (62.5%) had used the system themselves. The remaining 9 (37.5%) reported on use by another household member: a husband in 4 cases, a sister in 3, a cousin in 1, and a sister-in-law in 1. Ratings from these 9 respondents therefore describe observed rather than first-hand use. We report all 24 together, because household-mediated use is itself characteristic of digital access in this setting and is consistent with the 70.8% of formative survey respondents who required assistance to complete a questionnaire, but the distinction should be borne in mind when reading the ratings and we return to it in the Limitations.

3.3.2 User experience ratings

Respondents rated five dimensions on 5-point scales (Figure 10). Ratings are summarized descriptively; no inferential testing was performed, given the sample size and the absence of a comparison group.

Ease of use was rated highest (mean 4.79, SD 0.41), with 19 of 24 respondents selecting the maximum rating and the remaining 5 selecting 4; no rating fell below 4. Clarity of responses was also high (mean 4.54, SD 0.59), with 14 respondents rating 5, nine rating 4, and one rating 3.

Three dimensions showed greater variability. Perceived empathy averaged 4.00 (SD 0.72), with 5 respondents rating 5, 15 rating 4, 3 rating 3, and 1 rating 2, the widest range of any dimension. Alignment of answers with respondents' concerns averaged 4.04 (SD 0.69), with 5 rating 5, 16 rating 4, 2 rating 3, and 1 rating 2. Trust in the information provided averaged 4.29 (SD 0.91), the largest standard deviation observed, with 12 respondents rating 5, nine rating 4, one rating 3, and two rating 2, indicating that while most respondents rated the system as reliable, a minority expressed doubt.

Free-text responses were consistent with these ratings. Two respondents (P2, P7) reported that responses sometimes took frustratingly long to arrive, and one (P3) reported that the system misunderstood questions and returned incorrect answers. Regarding response quality, P17 noted that the system sometimes repeated answers or gave generic responses and that more personalized advice would be preferable. Two respondents raised feature requests bearing on personalization and peer support: P10 wrote that being able to save questions and answers or revisit previous conversations would be convenient, and P8 wrote that seeing real experiences from other mothers would be emotionally comforting.

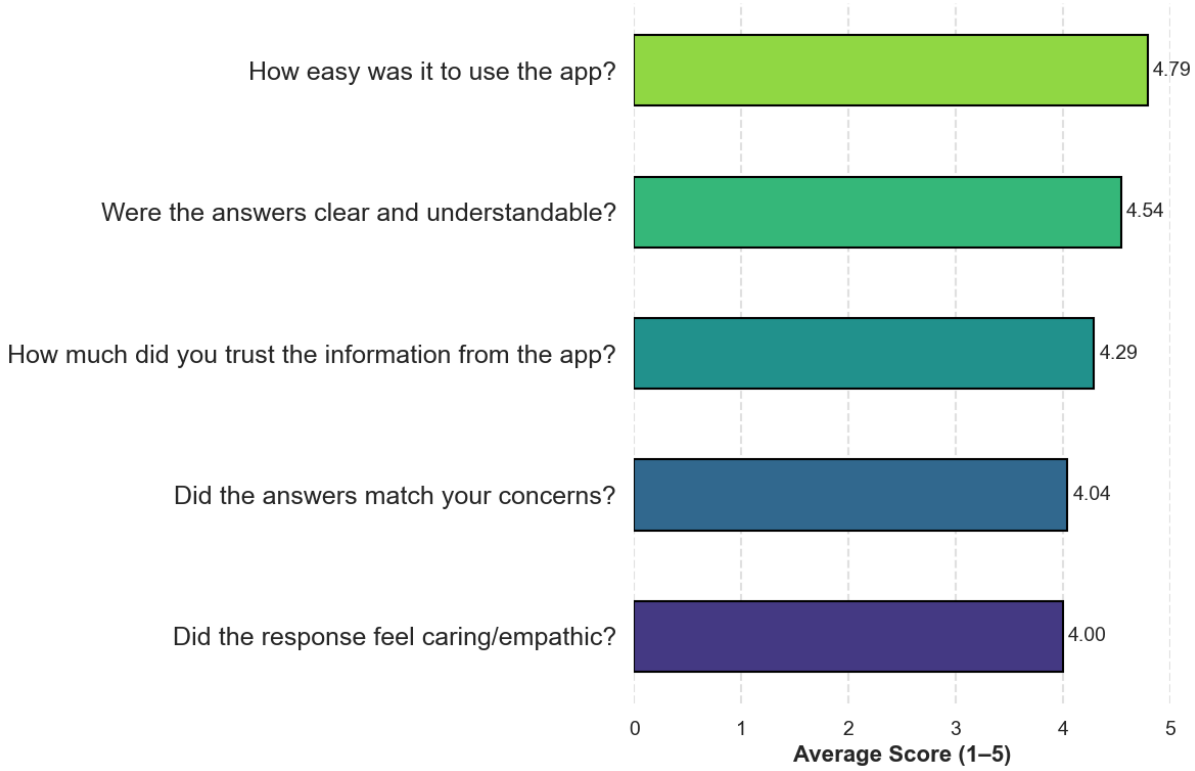


Figure 10: Mean user feedback ratings across five experience dimensions ($N = 24$; 5-point scale, higher indicates a more positive rating).

3.3.3 Adoption and privacy preferences

On the trichotomous items (Figure 11), 17 of 24 respondents (70.8%) said they would recommend the system to others, while 15 (62.5%) said they would use it instead of a general web search. Negative responses were identical for both items (3 respondents, 12.5%), while uncertainty was higher for replacement (6 respondents, 25.0%) than for recommendation (4 respondents, 16.7%). The comparison item was framed against Google, which remains the dominant health information tool in this setting; substitution against general-purpose AI assistants was not measured.

On the personalization-versus-privacy item (Figure 12), 16 respondents (66.7%) preferred a personalized system that stores user data, 5 (20.8%) prioritized data privacy, and 3 (12.5%) were undecided.

3.4 Interview findings

Semi-structured interviews with three participants who had used the system for three days surfaced four recurring topics. Given that only three women were interviewed, these are reported as illustrative accounts that contextualize the survey ratings, and not as the product of a formal thematic analysis or as findings that reached saturation.

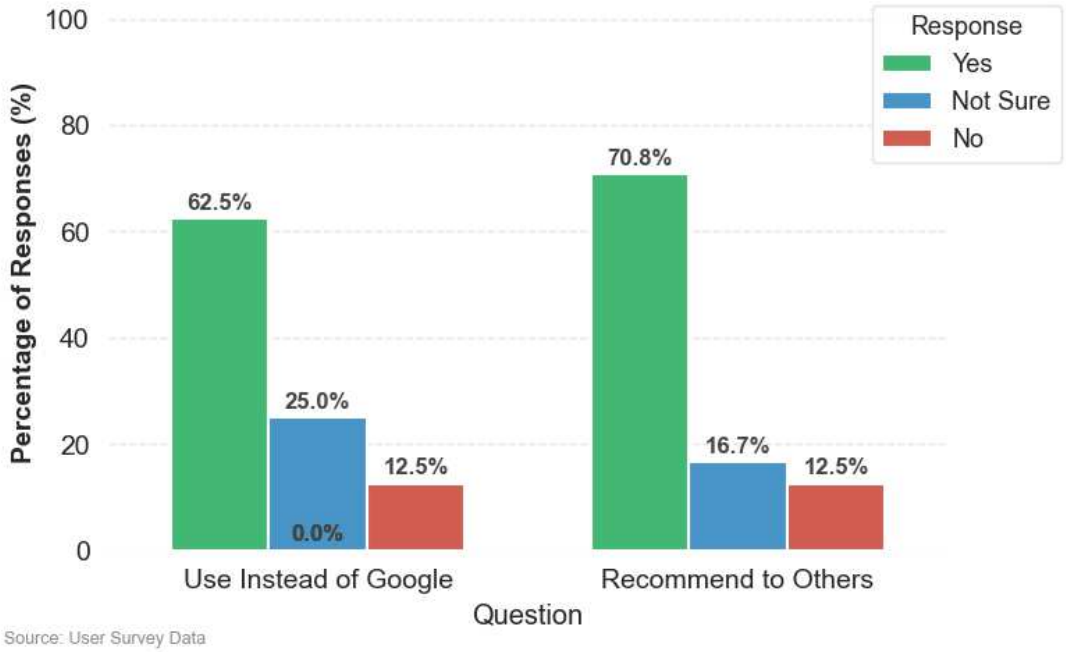


Figure 11: Participant responses to recommendation and replacement items ($N = 24$).

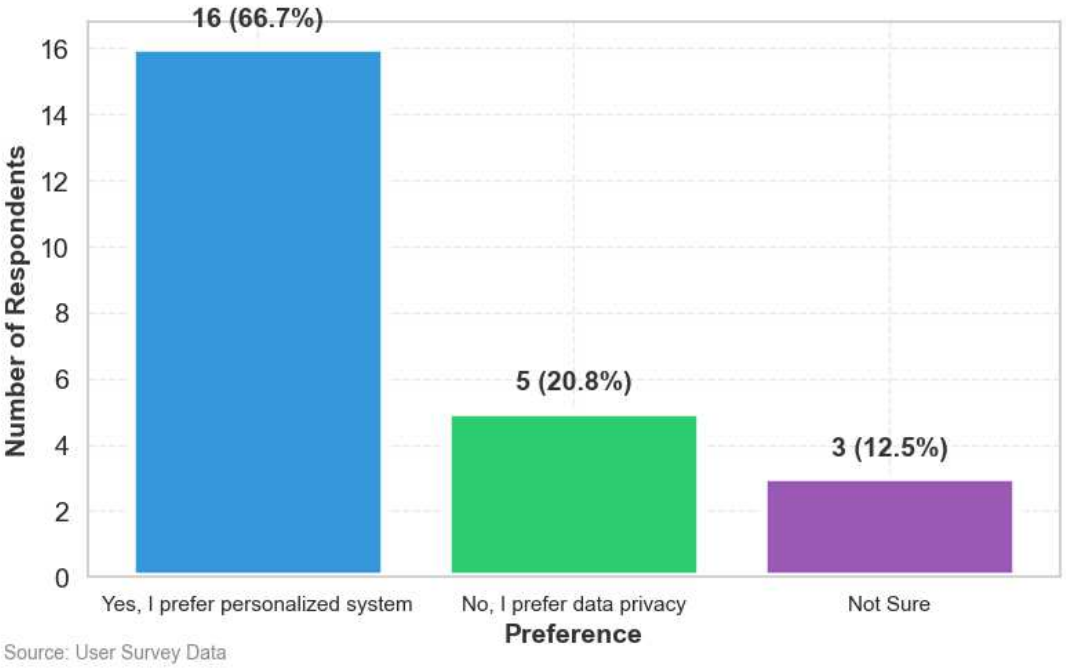


Figure 12: Participant preferences regarding personalization and data privacy ($N = 24$).

Voice input as an accessibility need. Participants reported that typing was burdensome during daily routines. P3 explained: "It gets tiring to type all the time with my hands full. It would be great if I could just ask and listen to the answer." This positions voice interaction less as a convenience than as an accommodation to the physical circumstances of pregnancy and early postpartum care.

Source attribution as reassurance. Participants emphasized wanting to know where guidance originated. P2 noted: "If the app mentioned whether the answer came from a doctor or a trusted site, I would feel more confident." The conditional form of this remark is notable: the interface already displays that a retrieval tool was invoked and leaves the retrieved passages visible, so what P2 describes as missing is not process transparency but a statement of source authority. We return to this distinction in the Discussion.

Reminders and health management. Participants envisioned the system extending beyond question answering into routine health management. P2 and P3 both raised this: "If the app can remind me to take medicine or tell me when my next checkup with the doctor is, it would be very convenient, so that I don't have to juggle calendar, notes, and other apps." Such features would move the system from an information source toward an active role in care, which raises questions about the responsibility an automated system should assume.

Offline access. Connectivity emerged as a recurring constraint, particularly for semi-urban and rural users. All three participants raised it: "Sometimes there is no electricity or internet available. In fact, many do not even have smartphones here. If I could still read answers or advice already given, it would help a lot." This indicates that connectivity gaps may exclude precisely those users with the least access to alternative sources of guidance.

3.5 Expert review

A licensed physician collaborating with the research team reviewed both the formative and feedback survey instruments for medical appropriateness and cultural sensitivity, providing input on question wording, response options, and structure. The same physician reviewed a sample of user queries and corresponding system responses for accuracy, clarity, and potential risk. This review found that the system generally produced clinically appropriate guidance, and identified the need to reinforce disclaimers that the system does not substitute for professional medical advice.

This review constitutes an initial check on clinical appropriateness rather than a systematic

safety evaluation. It was conducted on a sample rather than on all outputs, by a single clinician rather than a panel, and retrospectively rather than as a review of responses before they reached users. We treat these constraints as material limitations and address them in the Discussion.

4 Discussion

4.1 Study design and stage of evaluation

This study sits at an early stage of the evaluation pathway for maternal health interventions, and its design should be read accordingly. Before a digital intervention can reasonably be tested for effect on care-seeking or maternal outcomes, three prior questions need answering: whether the informational and emotional needs it targets are the ones women actually report, whether a system addressing those needs can be built and operated in the setting, and whether the women it is intended for can and will use it. Those are the questions this study addresses. We did not measure antenatal care attendance, care-seeking behaviour, knowledge change, or any maternal or neonatal outcome, and the design does not permit inference about them.

We are explicit about this because the value of a formative study lies in what it rules out as much as in what it demonstrates. Our findings indicate that usability and clarity are achievable in this setting but that trust and perceived empathy are not delivered by usability alone, and that a substantial share of access is household-mediated rather than individual. A trial designed without that information would risk measuring outcomes for an intervention that women do not trust enough to rely on, or attributing to individual users an intervention that reaches them through a family member. The appropriate next step is a controlled evaluation with validated instruments, a larger and geographically broader sample, and follow-up across pregnancy and the postpartum period, measuring knowledge, care-seeking behaviour, and well-being. We outline what such a study would require in the Future directions section below.

4.2 Principal findings

This study developed and formatively evaluated a Bangla-first, tool-orchestrated conversational agent for maternal health support in Bangladesh. Taking the three research questions in turn, the findings are as follows.

First, the formative survey showed that maternal concerns extend well beyond physical complications: alongside miscarriage risk and labor pain, participants reported anxiety, mood

changes, and depression, with just under two in five reporting some degree of psychological difficulty. A system confined to medical facts would address only part of the need. Language preference proved decisive and, importantly, was concentrated among the least formally educated participants, so a Bangla-first design matters most for precisely the group least served by existing English-dominant platforms.

Second, the configuration screening showed that model choice affects both grounding and responsiveness. Gemini 2.5 Flash produced the fewest unsupported statements at latencies compatible with interactive use and was selected for deployment, while the open-weight model tested was both slower in our serving environment and less well grounded. Increasing the number of retrieved passages improved grounding but degraded conciseness and relevance beyond a point, indicating a retrieval operating point rather than a monotonic benefit.

Third, the deployment evaluation showed a consistent gap between usability and relational qualities: ease of use and clarity were rated highly with little variance, whereas perceived empathy, alignment with concerns, and trust were rated lower with wider spread.

That last pattern is the most instructive. Technical clarity did not guarantee emotional resonance. Respondents who found the system easy to use and its answers clear nonetheless rated empathy at 4.00 (SD 0.72) with responses ranging from 2 to 5, and trust at 4.29 with the widest dispersion (SD 0.91). For maternal health support in a stigmatized context, this suggests that usability and accuracy are necessary but insufficient conditions for adoption.

4.3 Comparison with prior work

Prior conversational agents for perinatal health have largely been deployed in high-income settings and were either rule-based or built on a single model [18, 23, 25, 31, 33]. Relational agents such as Woebot and Wysa have shown mental health benefits through empathic dialogue [19, 50], and our participants likewise valued supportive framing alongside factual content. Bangla-language systems exist [36, 37, 41, 65], but few have undergone reported field evaluation, and we know of no system combining a curated Bangla clinical knowledge base with tool orchestration.

Our contribution is therefore primarily one of localization and architecture rather than of demonstrated clinical benefit. We did not compare the system against a rule-based baseline or against general web search, so we cannot claim superiority over existing alternatives. What the evaluation does show is that a retrieval-grounded, Bangla-first agent is feasible in this setting and is received as usable by the women who tried it, with the reservations noted below.

4.4 Language, transparency, and trust

Bangla-first interaction responded directly to survey evidence that 27.8% of participants were comfortable only in Bangla, and that this preference was universal among women with no formal or primary education. Prior work in comparable settings reports that local-language interaction functions as a trust signal and not merely as an accessibility feature [39, 42], and our design was built on that premise. We did not measure whether it held for our participants, and we do not claim to have tested it.

The system's transparency mechanisms are worth separating carefully, because our data suggest they are not interchangeable. The interface surfaces *process* transparency: it displays that a tool was invoked, and it leaves retrieved passages and raw tool output visible after the answer is delivered. What it does not provide is *source* attribution in a form users can act on, that is, a statement of whether a given claim came from a clinically reviewed article or from a general web search. P2's request is precisely for the latter: she said she would feel more confident if the application indicated whether an answer came from a doctor or a trusted site, which indicates that the process transparency already on screen did not supply that assurance. Read this way, her comment is not evidence that transparency worked but evidence of which kind of transparency was missing. This distinction offers one plausible account of the dispersion we observed in trust ratings, and it points to a concrete design change: labelling responses by provenance, and distinguishing curated clinical sources from web search results at the point of reading rather than in a collapsible technical panel.

Transparency of either kind, in any case, would not on its own resolve the trust gap. Participants also wanted advice adapted to their trimester, symptoms, and routines, and the system's session-scoped design could not provide this. That constraint reflects a deliberate trade-off. The system requires no account, which lowers the disclosure barrier in a context where discussing pregnancy complications or mental health can carry stigma, but it also precludes longitudinal personalization. Respondent preferences were split: 66.7% preferred a personalized system that stores data, while 20.8% prioritized privacy, and P10 explicitly asked to be able to save questions and revisit previous conversations, which is a request for the persistence the design deliberately withholds. Personalization should therefore not be treated as an unambiguous good, nor privacy protection as costless. Systems in this domain need mechanisms through which users can decide what is retained, forgotten, or shared, rather than defaulting to either extreme.

4.5 Safety considerations

The safety profile of this system requires explicit discussion, because retrieval grounding reduces but does not eliminate incorrect output. Even the best-performing configuration produced some unsupported statements, and retrieval-augmented systems can still communicate misleading information to patients [64]. In repeated testing we observed that the same question could yield different phrasings across runs, reflecting the stochastic nature of generation. In a domain where errors carry clinical consequences, this is material.

The safeguards in the deployed system were limited. Responses drawing on the knowledge base server were grounded in a physician-verified corpus, but responses drawing on the web search server were not, since that tool returns open-web content outside the verification pipeline described in the Methods. The interface displayed a disclaimer that the system does not replace professional care, and a clinician reviewed a sample of outputs after deployment. Critically, no clinician reviewed responses before they reached users, no automated detection of red-flag symptoms or emergency escalation was implemented, no filtering was applied to web search results, and the expert review covered a sample rather than the full set of interactions. We regard these as material gaps rather than minor omissions. Any deployment intended to serve users at scale would require structured escalation for emergency presentations, systematic pre-deployment red teaming against high-risk queries, restriction or verification of open-web retrieval, and ongoing clinical oversight. We report the present system as a feasibility prototype and do not claim it is safe for autonomous care-facing use in its current form.

4.6 Limitations

Several constraints bound the interpretation of these findings.

Sampling. The formative survey recruited 72 women, and the deployment evaluation involved 24 respondents and 3 interviewees, recruited predominantly through a single district. Urban respondents were overrepresented, and recruitment through hospitals and clinics likely excluded women who do not access formal care at all, precisely the population with the greatest unmet need. Sensitive topics including abortion and adolescent motherhood did not surface in any response, which suggests the survey did not reach the full spectrum of experience and that stigma may have suppressed disclosure. Deployment respondents were also likely to be early adopters comfortable with digital tools, so their ratings may not generalize.

Response conditions and proxy use. Seventy point eight percent of survey respondents

completed the formative instrument with assistance from an interviewer or family member. While this is itself a finding about digital and literacy barriers, it also means that responses to the mental health items were often given in the presence of another person. Social desirability and the stigma attached to perinatal distress in this setting make under-reporting plausible, and the proportion describing themselves as mostly stable and positive should be read with that in mind rather than as a prevalence estimate. A parallel issue affects the deployment ratings: 9 of the 24 respondents (37.5%) were reporting on use by another household member rather than on their own use, so more than a third of the ratings describe observed rather than first-hand experience. We retained these responses because household-mediated technology use is characteristic of this setting and excluding them would have removed a substantive observation, but they weaken the ratings as a measure of direct user experience.

Evaluation design. There was no control group and no comparison condition, so the study cannot establish effectiveness relative to existing information sources. Ratings were collected using a study-specific instrument rather than a validated scale such as the System Usability Scale, which limits comparability with other work. Interaction periods were short, so sustained engagement and the durability of trust remain unexamined.

Automated scoring. The configuration screening reported in Table 3 is the weakest component of this study and we set out its constraints in full. The scores were produced by an automated large language model-as-judge protocol [75], whose validity on specialized clinical content is contested [76]. Each item was scored once by a single automated judge, so the scores carry no estimate of their own variability, and they were not validated against independent clinician judgment. We therefore report them only as the basis for the deployment configuration that was in fact selected, and not as measures of model quality or of clinical accuracy.

Qualitative analysis and unexamined outcomes. The three interviews were read inductively and summarized into recurring topics. We did not conduct a formal thematic analysis, construct a codebook, involve a second coder, or assess inter-coder agreement, and we do not claim saturation. The interview material is therefore illustrative of individual experience rather than analytically generalizable, and the free-text survey comments were treated the same way. The interview protocol additionally included questions on whether use influenced how participants thought about or acted on their health; these responses were not analysed separately and no findings on influence or behaviour change are reported here.

Knowledge base. The corpus was assembled from a small number of vetted Bangla sources

because comprehensive Bangla clinical content is scarce. Perinatal mental health remained underrepresented relative to physical health topics. Chunking parameters were selected by inspection rather than by systematic benchmarking across the parameter space, so the retrieval configuration is plausible but not optimized. The corpus is also static and requires manual intervention to reflect updated guidelines, and formal permission was not sought from either source publisher.

Language and model behavior. We observed during testing that generated Bangla occasionally read as awkward or overly literal, and regional dialects such as Sylheti and Chittagonian were not handled effectively, which may limit inclusivity outside urban and more educated populations. These are our own observations; we did not systematically assess linguistic naturalness, and no participant rating captured it.

System scope. The system requires connectivity, which excludes users in low-bandwidth or intermittently connected areas. Personalization was limited to within-session context. Multimodal input was not supported, constraining accessibility for users with low literacy.

Structural limits. Maternal health outcomes are shaped by socioeconomic structures, household dynamics, and provider availability. An informational agent, however well designed, cannot address a shortage of skilled birth attendants or the stigma surrounding perinatal mental illness, and should not be positioned as a substitute for those investments.

4.7 Future directions

Participant feedback and the limitations above point to several priorities.

Evaluation infrastructure. Before any comparative claim about model choice can be sustained, the evaluation itself needs rebuilding: a documented judging configuration, judges drawn from outside the family of the models under test, multiple scoring runs to estimate variability, and validation against independent clinician scoring on a subset. This is inexpensive relative to its value and should precede further model work.

Safety infrastructure. The most immediate requirement for the deployed system is structured handling of high-risk presentations: detection of red-flag symptoms with explicit escalation to emergency care, systematic adversarial testing before deployment, restriction or verification of open-web retrieval, and a defined role for clinical oversight. Work in this direction should precede any expansion of the user base.

Provenance-aware response design. Following the distinction drawn above, responses should

carry visible, reader-level provenance rather than process indicators alone, so that a user can tell at a glance whether guidance came from a clinically reviewed source. This is a modest engineering change with a plausible bearing on the trust dispersion we observed, and it is directly testable in a follow-up study.

Multimodal interaction. All three interviewees raised voice input, and P3 framed typing as burdensome during daily routines with an infant. Voice is better understood here as an accommodation to the physical circumstances of pregnancy and infant care than as a convenience. A further extension, which we propose rather than derive from participant requests, would be interpretation of prescriptions, test results, or ultrasound reports; this would raise additional questions of standardization, Bangla-English translation, and protection of sensitive health data.

Negotiated personalization. Longitudinal support would require persistent memory, but participant preferences were divided on data retention. The design problem is not whether to personalize but how to give users meaningful control over what is retained and shared.

Proactive risk detection. Participants' anxieties about miscarriage and depression suggest value in detecting concerning patterns across interactions. Such capability carries real risk of over-alerting and of pathologizing ordinary features of pregnancy, and would need careful calibration and clinical involvement.

Accessibility and offline operation. Connectivity was a recurring constraint. Hybrid designs that cache prior conversations and provide essential guidance offline would extend reach to precisely those users currently excluded. Advances in compact language models make on-device deployment increasingly plausible, reframing accessibility as a question of infrastructural resilience rather than interface refinement.

Peer support. P8 asked to see the experiences of other mothers, describing this as emotionally comforting. Peer content would introduce moderation and misinformation risks that a curated corpus is designed to avoid, and we note the request without endorsing an implementation.

Longitudinal and comparative evaluation. Establishing benefit will require studies with control conditions, validated instruments, larger and more geographically diverse samples, and follow-up across pregnancy and the postpartum period, measuring knowledge, care-seeking behavior, and well-being rather than satisfaction alone.

4.8 Deployment economics

Sustained operation depends on cost, and the cost structure differs sharply between the two deployment routes we considered. Hosted inference is billed per token, with input and output priced separately and input typically the cheaper of the two. A retrieval-augmented workload of this kind is input-heavy, because every request carries the retrieved passages and the system prompt in addition to the user's question, so the dominant cost driver is retrieval breadth rather than answer length. This is one practical reason, alongside latency, for deploying at $K = 5$ rather than $K = 10$. Self-hosting an open-weight model shifts cost from per-token charges to infrastructure and operations, which in this setting we estimate at approximately US $50 to $200 per month depending on hardware, alongside engineering effort for monitoring and maintenance. We did not instrument token consumption during the deployment, so we are unable to report measured per-session usage or a costed projection, and we regard that as a gap that any scaling study would need to close first.

For an LMIC deployment these differences are consequential but, at the scales considered here, unlikely to be prohibitive relative to the cost of the clinical oversight such a system would require. A hybrid approach, routing routine queries to a self-hosted model and reserving hosted inference for harder queries, may offer a reasonable balance, though we did not evaluate such a configuration. Sustained operation would also require integration pathways with existing services, whether through national digital health programs, development partners already funding maternal health work, or public-private arrangements with mobile operators given high mobile penetration. Financial sustainability is not separable from equity here: if operating costs are met by charging users, the communities with the greatest need are the first to be excluded.

5 Conclusions

We present a Bangla-language conversational agent for maternal health support that integrates retrieval-augmented generation, session-scoped conversational memory, and external tool orchestration. The system draws on a curated maternal health knowledge base and accepts input in Bangla, romanized Bangla, and English. A formative evaluation combining a requirements survey, a model configuration screening, and a usability and trust study indicates that such a system is feasible in a low-resource and often stigmatized setting, and that women found it easy to use and its answers clear.

The evaluation also shows that usability does not carry trust with it. Ratings of empathy, alignment with concerns, and trust were consistently lower and more widely dispersed than ratings of ease of use and clarity, and participants asked for source attribution, personalization, voice input, and offline access that the deployed system did not provide. This work therefore underscores the importance of culturally and linguistically sensitive design that attends to privacy, stigma, and digital literacy, and it offers transferable design insights for equitable digital health tools in non-English, low-resource contexts. As a formative study, it does not establish clinical effectiveness or safety for autonomous care. Realizing that goal will require clinical-grade safeguards, human-in-the-loop review, integration with health services, and longitudinal evaluation of maternal outcomes at scale.

List of abbreviations

AI: artificial intelligence; ANC: antenatal care; API: application programming interface; BDHS: Bangladesh Demographic and Health Survey; BDT: Bangladeshi taka; IRB: institutional review board; LLM: large language model; LMIC: low- and middle-income country; MCP: Model Context Protocol; mHealth: mobile health; P50: 50th percentile; P99: 99th percentile; RAG: retrieval-augmented generation; SD: standard deviation.

Declarations

Ethics approval and consent to participate

This study was conducted under research protocol BRACUIRB_220240006, approved by the Institutional Review Board of BRAC University, Dhaka, Bangladesh, on 6 May 2024 and categorized as eligible for expedited review. The principal investigator on that protocol was based at BRAC University at the time of approval and is now affiliated with United International University. All methods were carried out in accordance with the Declaration of Helsinki and with relevant institutional guidelines and regulations. Informed consent was obtained from all participants prior to data collection, and interview participants additionally consented to audio recording. Three survey respondents (4.2%) selected the “less than 18 years” age band; because the instrument recorded age in bands rather than as an exact value, we are unable to establish whether any of these participants was under the age of 16. Each provided her own informed

consent, and an accompanying parent or guardian was informed of the purpose of the study, the voluntary nature of participation, and the right to withdraw, and gave verbal informed consent before participation; this guardian consent was not separately documented in writing. We confirm that informed consent for study participation was obtained from a parent or legal guardian in the case of every participant who selected the "less than 18 years" age band. Full details of the consent procedure, including this limitation, are given in the Methods.

Consent for publication

Not applicable.

Data availability

The de-identified datasets generated and analysed during the current study are available from the corresponding authors on reasonable request. The survey instruments are provided in Supplementary File 1 and Supplementary File 3, the interview protocol in Supplementary File 4, and the deployed system prompt in Supplementary File 5. The curated maternal health knowledge base is derived from third-party copyrighted sources and cannot be redistributed; the source websites and the inclusion criteria applied are described in the Methods and in Supplementary File 2. The deployed system is publicly accessible at https://baby-and-me.vercel.app/.

Competing interests

The authors declare no competing interests.

Funding

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

Author contributions

F.M. and J.N. conceived the study. F.M. designed and implemented the system, including the knowledge base pipeline, the retrieval and MCP components, and the web application, and conducted the model configuration screening. F.M. and N.L.S. designed and administered the formative and feedback surveys and conducted the semi-structured interviews. N.L.S. prepared and analysed the survey data and prepared the figures and tables. J.N. supervised the study,

is the principal investigator on the approved research protocol, and provided methodological guidance throughout. F.M. and N.L.S. drafted the manuscript, and J.N. revised it critically for important intellectual content. All authors reviewed and approved the final manuscript.

Acknowledgements

The authors thank the licensed physician who reviewed the survey instruments, authored the reference answers used in the configuration screening, and reviewed a sample of system outputs for clinical appropriateness. The authors also thank the participating clinics in Pabna and the women who took part in this study.

References

[1] National Institute of Population Research and Training (NIPORT) and ICF. Bangladesh demographic and health survey 2022: Final report. Technical report, NIPORT and ICF, Dhaka, Bangladesh, and Rockville, Maryland, USA, 2024. URL https://dhsprogram.com/publications/publication-fr382-dhs-final-reports.cfm.

[2] K. H. Prantik, A. T. Nabi, G. M. Suhel, M. Afroz, and M. N. Hasan. Maternal healthcare service utilization in Bangladesh: A cross-sectional study of determinants and temporal trends using BDHS 2011–2022. *Health Science Reports*, 9(4):e72171, 2026. doi: 10.1002/hsr2.72171.

[3] World Health Organization. Global strategy on human resources for health: Workforce 2030. Technical report, World Health Organization, Geneva, Switzerland, 2016. URL https://www.who.int/publications/i/item/9789241511131.

[4] Md Abul Hossen and Anne Westhues. Rural women's access to health care in bangladesh: swimming against the tide? *Social work in public health*, 26(3):278–293, 2011.

[5] Mosfequr Rahman, Sarkar Obaida Nasrin, Mosiur Rahman, Aminur Rahman, Golam Mostofa, Syeda Sarah Jesmin, and Fiona Buchanan. Maternal pregnancy intention and its association with low birthweight and pregnancy complications in bangladesh: findings from a hospital-based study. *International health*, 11(6):447–454, 2019.

[6] Mirajul Islam and Nasrin Sultana. Risk factors for pregnancy related complications among

urban slum and non-slum women in bangladesh. *BMC pregnancy and childbirth*, 19(1):235, 2019.

[7] Shahinoor Akter, Jane Louise Rich, Kate Davies, and Kerry Jill Inder. Access to maternal healthcare services among indigenous women in the chittagong hill tracts, bangladesh: A cross-sectional study. *BMJ open*, 9(10):e033224, 2019.

[8] Shahinoor Akter, Kate Davies, Jane Louise Rich, and Kerry Jill Inder. Barriers to accessing maternal health care services in the chittagong hill tracts, bangladesh: A qualitative descriptive study of indigenous women's experiences. *PloS one*, 15(8):e0237002, 2020.

[9] Anaise Williams, Malabika Sarker, and Syeda Tahmina Ferdous. Cultural attitudes toward postpartum depression in dhaka, bangladesh. *Medical anthropology*, 37(3):194–205, 2018.

[10] Nafisa Insan, Anthony Weke, Judith Rankin, and Simon Forrest. Perceptions and attitudes around perinatal mental health in bangladesh, india and pakistan: a systematic review of qualitative data. *BMC Pregnancy and Childbirth*, 22(1):293, 2022.

[11] Ashmita Khan Thakuri, Vijay S. GC, Nadya Belenky, Pratikshya Wasti, Roselini Shrestha, and Sharada P. Wasti. Prevalence and risk factors of perinatal depression among women in South Asian countries: A systematic review and meta-analysis. *Journal of Affective Disorders*, 393(Pt A):120286, 2026. doi: 10.1016/j.jad.2025.120286.

[12] Mohammad Injamul Hoq, Md Mohotasin Hossain, Mohammad Aktar Sayeed, and Md Jakaria. Trends in the prevalence of antenatal and postnatal depression in Bangladesh: A systematic review and meta-analysis. *Heliyon*, 11(2):e41955, 2025. doi: 10.1016/j.heliyon.2025.e41955.

[13] Goutam Kumar Dutta, Bidhan Krishna Sarker, Helal Uddin Ahmed, Dipika Shankar Bhattacharyya, Md Musfikur Rahman, Ratna Majumder, and Taposh Kumar Biswas. Mental healthcare-seeking behavior during the perinatal period among women in rural bangladesh. *BMC health services research*, 22(1):310, 2022.

[14] Easter Olwanda, Kennedy Opondo, Dorothy Oluoch, Kevin Croke, Justinah Maluni, Joyline Jepkosgei, and Jacinta Nzinga. Women's autonomy and maternal health decision making in kenya: implications for service delivery reform-a qualitative study. *BMC Women's Health*, 24(1):181, 2024.

[15] World Health Organization, UNICEF, UNFPA, World Bank Group, and UN-DESA/Population Division. Trends in maternal mortality estimates 2000 to 2023: Estimates by WHO, UNICEF, UNFPA, world bank group and UNDESA/population division. Technical report, World Health Organization, Geneva, Switzerland, 2025. Available from: https://iris.who.int/handle/10665/381012.

[16] Gizem Kerimoglu Yildiz, Rukiye Turk Delibalta, and Zehra Coktay. Artificial intelligence-assisted chatbot: impact on breastfeeding outcomes and maternal anxiety. *BMC Pregnancy and Childbirth*, 25(1):631, 2025.

[17] İlke Özer Aslan and Mustafa Törehan Aslan. Benchmarking ai chatbots for maternal lactation support: A cross-platform evaluation of quality, readability, and clinical accuracy. *Healthcare*, 13(14):1756, 2025.

[18] Kelsey McAlister, Lara Baez, Jennifer Huberty, Marianna Kerppola, et al. Chatbot to support the mental health needs of pregnant and postpartum women (moment for parents): design and pilot study. *JMIR Formative Research*, 9(1):e72469, 2025.

[19] Sanaa Suharwardy, Maya Ramachandran, Stephanie A Leonard, Anita Gunaseelan, Deirdre J Lyell, Alison Darcy, Athena Robinson, and Amy Judy. Feasibility and impact of a mental health chatbot on postpartum mental health: a randomized controlled trial. *AJOG Global Reports*, 3(3):100165, 2023.

[20] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2): e7785, 2017.

[21] Judith J Prochaska, Erin A Vogel, Amy Chieng, Matthew Kendra, Michael Baiocchi, Sarah Pajarito, and Athena Robinson. A therapeutic relational agent for reducing problematic substance use (woebot): development and usability study. *Journal of medical Internet research*, 23(3):e24850, 2021.

[22] Emily Durden, Maddison C Pirner, Stephanie J Rapoport, Andre Williams, Athena Robinson, and Valerie L Forman-Hoffman. Changes in stress, burnout, and resilience associated with an 8-week intervention with relational agent "woebot". *Internet Interventions*, 33: 100637, 2023.

[23] Jessica Nathalie Rivera Rivera, Katarina E AuBuchon, Marjanna Smith, Claire Starling, Karen G Ganacias, Aimee Danielson, Loral Patchen, Janine A Rethy, H Joseph Blumenthal, Angela D Thomas, et al. Development and refinement of a chatbot for birthing individuals and newborn caregivers: Mixed methods study. *JMIR Pediatrics and Parenting*, 7(1): e56807, 2024.

[24] Samira Amil, Sié-Mathieu-Aymar-Romaric Da, James Plaisimond, Geneviève Roch, Maxime Sasseville, Frédéric Bergeron, and Marie-Pierre Gagnon. Interactive conversational agents for perinatal health: A mixed methods systematic review. *Healthcare*, 13(4):363, 2025.

[25] Roger A Edwards, Timothy Bickmore, Lucia Jenkins, Mary Foley, and Justin Manjourides. Use of an interactive computer agent to support breastfeeding. *Maternal and child health journal*, 17(10):1961–1968, 2013.

[26] Brian Jack, Timothy Bickmore, Megan Hempstead, Leanne Yinusa-Nyahkoon, Ekaterina Sadikova, Suzanne Mitchell, Paula Gardiner, Fatima Adigun, Brian Penti, Daniel Schulman, et al. Reducing preconception risks among african american women with conversational agent technology. *The Journal of the American Board of Family Medicine*, 28(4):441–451, 2015.

[27] Timothy Bickmore, Zhe Zhang, Matthew Reichert, Clevanne Julce, and Brian Jack. Promotion of preconception care among adolescents and young adults by conversational agent. *Journal of Adolescent Health*, 67(2):S45–S51, 2020.

[28] Paula M Gardiner, Kelly D McCue, Lily M Negash, Teresa Cheng, Laura F White, Leanne Yinusa-Nyahkoon, Brian W Jack, and Timothy W Bickmore. Engaging women with an embodied conversational agent to deliver mindfulness and lifestyle recommendations: A feasibility randomized control trial. *Patient education and counseling*, 100(9):1720–1729, 2017.

[29] João Luis Zeni Montenegro, Cristiano André da Costa, and Luisa Plácido Janssen. Evaluating the use of chatbot during pregnancy: A usability study. *Healthcare Analytics*, 2:100072, 2022.

[30] Eri Maeda, Akane Miyata, Jacky Boivin, Kyoko Nomura, Yukiyo Kumazawa, Hiromitsu Shirasawa, Hidekazu Saito, and Yukihiro Terada. Promoting fertility awareness and precon-

ception health using a chatbot: a randomized controlled trial. *Reproductive BioMedicine Online*, 41(6):1133–1143, 2020.

[31] Bill Siwicki. Conversational ai improves "fourth trimester" maternal care at penn medicine, 4 2024. URL https://www.healthcareitnews.com/news/conversational-ai-improves-fourth-trimester-maternal-care-penn-medicine.

[32] Lucia Vaira, Mario A Bochicchio, Matteo Conte, Francesco Margiotta Casaluci, and Antonio Melpignano. Mamabot: a system based on ml and nlp for supporting women and families during pregnancy. In *Proceedings of the 22nd international database engineering & applications symposium*, pages 273–277, 2018.

[33] Kathryn I Pollak, Stewart C Alexander, Gary Bennett, Pauline Lyna, Cynthia J Coffman, Alicia Bilheimer, David Farrell, Michael E Bodner, Geeta K Swamy, and Truls Østbye. Weight-related sms texts promoting appropriate pregnancy weight gain: a pilot study. *Patient education and counseling*, 97(2):256–260, 2014.

[34] Sherry Liu and Li Zou. Premom ovulation tracker app with ovulation tests, pregnancy tests. https://premom.com/, 2017. Accessed 29 July 2025.

[35] APRU. Ai in maternal health – how research and policy intertwine, Aug 2024. URL https://www.apru.org/news/ai-in-maternal-health/.

[36] Ummay Marzan Jui. Maya: An ai-powered start-up that plans to become the go-to name in healthcare, Oct 2021. URL https://www.tbsnews.net/features/panorama/maya-ai-powered-start-plans-become-go-name-healthcare-314215.

[37] ACME AI. Probahini: Ai-powered menstrual health chatbot, 2024. URL https://www.acmeai.tech/product/probahini-chatbot.

[38] Aniqa Tasnim Hossain. Mit solve: Tiger challenge (bangladesh-based applicants) lifebot, 2019. URL https://solve.mit.edu/solutions/10934.

[39] Danya Arif Siddiqi, Fatima Miraj, Humdiya Raza, Owais Ahmed Hussain, Mehr Munir, Vijay Kumar Dharma, Mubarak Taighoon Shah, Ali Habib, and Subhash Chandir. Development and feasibility testing of an artificially intelligent chatbot to answer immunization-related queries of caregivers in pakistan: A mixed-methods study. *International Journal of Medical Informatics*, 181:105288, 2024.

[40] Shamim Ahamed, Moinul H Chowdhury, Marzia Zaman, Abhishek Agarwala, Md Abdullah Al Mashud, Md Tariqul, Tareq Al Muntasir, Rifat Rahman, Rifat Shahriyar, Farhana Sarker, et al. Susastho.ai: A multimodal medical copilot for adolescents using evidence-based medicine and large language models, n.d. Unpublished manuscript.

[41] Syed Niaz Mohtasim, Faiza Omar Arpita, and Istiaq Ahmed. Gorbhoshongi: a mhealth app for the mental well-being of bangladeshi pregnant and postpartum women. Technical report, Independent University, Bangladesh, 2024.

[42] Amirul Islam, Farida Begum, Anna Williams, Rabeya Basri, Rowsan Ara, and Rondi Anderson. Midwife-led pandemic telemedicine services for maternal health and gender-based violence screening in bangladesh: an implementation research case study. *Reproductive health*, 20(1):128, 2023.

[43] Thalia Beaty. How ai health care chatbots learn from the questions of an indian women's organization, 2 2024. URL https://apnews.com/article/ai-health-care-chatbot-india-6e820475e89dbc15ce1880695f698934.

[44] Roshini Deva, Dhruv Ramani, Tanvi Divate, Suhani Jalota, and Azra Ismail. "kya family planning after marriage hoti hai?": Integrating cultural sensitivity in an llm chatbot for reproductive health. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–23, 2025.

[45] Aliyah Dosani, Harshmeet Arora, and Sahil Mazmudar. mhealth and perinatal depression in low-and middle-income countries: A scoping review of the literature. *International journal of environmental research and public health*, 17(20):7679, 2020.

[46] Bekelu Teka Worku, Misra Abdulahi, Demissew Amenu, and Bruno Bonnechère. Effect of technology-supported mindfulness-based interventions for maternal depression: a systematic review and meta-analysis with implementation perspectives for resource-limited settings. *BMC Pregnancy and Childbirth*, 25(1):155, 2025.

[47] Marina Vilarim, Fernanda Rebelo, Ianne Vieira, Fernanda Mazzoli, Mauro G Carta, Antonio Egidio Nardi, and Daniele Marano. Prevalence of postpartum depression symptoms in high-income and low-and middle-income countries in the covid-19 pandemic: a systematic review with meta-analysis. *Brazilian Journal of Psychiatry*, 46:e20233453, 2024.

[48] WHO. Perinatal mental health, 2023. URL https://www.who.int/teams/mental-health-and-substance-use/promotion-prevention/maternal-mental-health.

[49] Ramaa Venkatachari. *Maternal mental health and child outcomes: a human-centered design perspective on preventive mental health care*. PhD thesis, Massachusetts Institute of Technology, 2020.

[50] Becky Inkster, Madhura Kadaba, and Vinod Subramanian. Understanding the impact of an ai-enabled conversational agent mobile app on users' mental health and wellbeing with a self-reported maternal event: a mixed method real-world data mhealth study. *Frontiers in Global Women's Health*, 4:1084302, 2023.

[51] Lawrence Jones-Esan, Nalinda Somasiri, and Keda Lorne. Enhancing healthcare delivery through digital health interventions: a systematic review on telemedicine and mobile health applications in low and middle-income countries (lmics), 2024. Unpublished manuscript.

[52] Prasenjit De and Manas Ranjan Pradhan. Effectiveness of mobile technology and utilization of maternal and neonatal healthcare in low and middle-income countries (lmics): a systematic review. *BMC Women's Health*, 23(1):664, 2023.

[53] Shamima Nasrin. *The Association of Participation in a Text Messaging Program with Reduction in Depression Severity Among Pregnant Women*. PhD thesis, University of British Columbia, 2025.

[54] Ryan Afreen, Sanjna L Surya, Tasnim Jara, Ima Islam, Rahana Parvin, SM Ferdousuzzaman, Masum Salah Uddin, and Nahreen Ahmed. Enhancing mental health literacy and care through community-driven solutions in rural bangladesh. *Frontiers in Global Women's Health*, 5:1478817, 2024.

[55] Ismaila Ouédraogo. *Mobile technology and artificial intelligence for improving health literacy among underserved communities*. PhD thesis, Université de Bordeaux; Université Nazi Boni (Bobo-Dioulasso, Burkina Faso), 2024.

[56] Yubina Kasse, Riantofani Riantofani, Rahmi Rahmi, and Kiftiyah Kiftiyah. Designing ai-driven chatbots for adolescent mental health support in rural schools. *International Journal of Research in Counseling*, 4(1):83–92, 2025.

[57] Neil KR Sehgal, Hita Kambhamettu, Sai Preethi Matam, Lyle Ungar, and Sharath Chandra Guntuku. Exploring socio-cultural challenges and opportunities in designing mental health chatbots for adolescents in india. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2025.

[58] Kyungmi Chung, Hee Young Cho, and Jin Young Park. A chatbot for perinatal women's and partners' obstetric and mental health care: development and usability evaluation study. *JMIR Medical Informatics*, 9(3):e18607, 2021.

[59] Nalan Karunanayake. Next-generation agentic ai for transforming healthcare. *Informatics and Health*, 2(2):73–83, 2025.

[60] Matthew Finio and Amanda Downie. What is ai agent orchestration? | ibm, Feb 2025. URL https://www.ibm.com/think/topics/ai-agent-orchestration.

[61] Daniel Dominguez. A technical guide to multi-agent orchestration, Nov 2024. URL https://dominguezdaniel.medium.com/a-technical-guide-to-multi-agent-orchestration-5f979c831c0d.

[62] Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. Model context protocol (mcp): Landscape, security threats, and future research directions. *arXiv preprint arXiv:2503.23278*, 2025.

[63] Abul Ehtesham, Aditi Singh, Gaurav Kumar Gupta, and Saket Kumar. A survey of agent interoperability protocols: Model context protocol (mcp), agent communication protocol (acp), agent-to-agent protocol (a2a), and agent network protocol (anp). *arXiv preprint arXiv:2505.02279*, 2025.

[64] Maria Paola Priola. Addressing hallucinations with rag and nmiss in italian healthcare llm chatbots. *arXiv preprint arXiv:2412.04235*, 2024.

[65] Md Moshiur Rahman, Ruhul Amin, Md Nazmul Khan Liton, and Nahid Hossain. Disha: an implementation of machine learning based bangla healthcare chatbot. In *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, pages 1–6. IEEE, 2019.

[66] L Todd Rose and Kurt W Fischer. Garbage in, garbage out: having useful data is everything. *Measurement: Interdisciplinary Research & Perspective*, 9(4):222–226, 2011.

[67] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024.

[68] Supabase. Supabase docs, 2025. URL https://supabase.com/docs.

[69] An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, et al. Qwen2. 5-1m technical report. *arXiv preprint arXiv:2501.15383*, 2025.

[70] Saad Tahmid and Sourav Sarker. Qwen2. 5-32b: Leveraging self-consistent tool-integrated reasoning for bengali mathematical olympiad problem solving. *arXiv preprint arXiv:2411.05934*, 2024.

[71] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[72] Dawei Zhu, Xiyu Wei, Guangxiang Zhao, Wenhao Wu, Haosheng Zou, Junfeng Ran, Xun Wang, Lin Sun, Xiangzheng Zhang, and Sujian Li. Chain-of-thought matters: improving long-context language models with reasoning path supervision. *arXiv preprint arXiv:2502.20790*, 2025.

[73] Alexandria Leto, Cecilia Aguerrebere, Ishwar Bhati, Ted Willke, Mariano Tepper, and Vy Ai Vo. Toward optimal search and retrieval for rag. *arXiv preprint arXiv:2411.07396*, 2024.

[74] Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, et al. Searching for best practices in retrieval-augmented generation. *arXiv preprint arXiv:2407.01219*, 2024.

[75] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024.

[76] Annalisa Szymanski, Noah Ziems, Heather A Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. Limitations of the llm-as-a-judge approach for evaluating llm

outputs in expert knowledge tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 952–966, 2025.

[77] Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. Preference leakage: A contamination problem in llm-as-a-judge. *arXiv preprint arXiv:2502.01534*, 2025.

Supplementary Information

Supplementary File 1. Formative survey questionnaire. The 17-item instrument administered to the 72 formative survey respondents.

Supplementary File 2. Knowledge base sourcing and preprocessing. Enumeration of sourcing constraints, preprocessing procedures, and chunking parameter selection.

Supplementary File 3. Deployment feedback questionnaire. The 13-item instrument administered after use of the system.

Supplementary File 4. Semi-structured interview protocol, including the consent preamble, in English translation.

Supplementary File 5. System prompt used by the deployed agent, reproduced verbatim.

A Additional file 1: Formative survey questionnaire

The 17-item instrument used in the formative requirements survey ($N = 72$) is reproduced in Table 4. Items 11 to 13 permitted selection of up to five options; percentages reported for these items are therefore expressed as a proportion of all selections rather than of participants. Item 1 collected respondent name during administration to permit de-duplication and to link initial-survey participants with later feedback respondents; this identifier was removed prior to analysis and no identifying information appears in the analyzed dataset. The age bands in item 2 are reproduced exactly as administered, including the boundary values 25 and 35 that appear in two adjacent categories; respondents selected a single band.

Table 4: Formative survey questionnaire items.

No	Question	Response options
1	Name	Short answer (removed prior to analysis)
2	Age	Under 18; 18–25; 25–35; 35–45; over 45
3	Education level	No formal education; primary; secondary; college; bachelor's; master's or higher
4	Occupation	Homemaker; service; business; student; daily wage labor; unemployed; other (specify)
5	Household location	Urban; semi-urban; village
6	Number of family members	Short answer
7	Annual household income	Less than 1 lakh; 1–5 lakh; 5–8 lakh; more than 8 lakh (BDT)
8	Internet or smartphone access	Personal smartphone and internet; shared access; does not have smartphone
9	Language comfort in app	Only Bangla; only English; both Bangla and English
10	Number of pregnancies	1; 2; 3; 4; 5; more than 5

Continued on next page

Table 4: Formative survey questionnaire items. (Continued)

No	Question	Response options
11	Five main complications experienced during pregnancy (select up to five)	Morning sickness or nausea; back pain or body aches; fatigue or low energy; sleep difficulties; fear of miscarriage or complications; fear of childbirth or labor pain; financial stress or income insecurity; lack of support from partner or family; anxiety or depression; isolation or loneliness; hormonal mood swings; difficulty accessing health care services; negative experiences with doctors or hospitals; weight gain or body-image concerns; work-related stress; social pressure or cultural expectations; miscarriage or stillbirth in a previous pregnancy; concerns about the baby's health or development; domestic violence or verbal abuse; other (specify)
12	Five main concerns during pregnancy (select up to five)	Baby's health and development; risk of miscarriage or complications; safe delivery and labor pain; physical changes and body image; whether I will be a good mother; financial stability and expenses; managing household responsibilities; balancing work and pregnancy; future of career after childbirth; quality of medical care; availability of emotional support; changes in relationship with partner; breastfeeding and baby-care knowledge; birth defects or genetic disorders; pre-existing health conditions affecting pregnancy; handling multiple children; whether my partner will be supportive during delivery; fear of C-section or surgery; impact of pregnancy on mental health; postpartum recovery and health; other (specify)

Continued on next page

Table 4: Formative survey questionnaire items. (Continued)

No	Question	Response options
13	Five main positive moments during pregnancy (select up to five)	Feeling the baby's first kick; hearing the baby's heartbeat; seeing the baby in an ultrasound; receiving love and care from family; partner's support and involvement; preparing for the baby; bonding emotionally with the baby; receiving gifts or blessings; positive feedback from doctors or nurses; family excitement about the baby; taking maternity photos; talking to the baby or playing music; friends or community showing support; reading or learning about motherhood; reduced workload or time to rest; partner attending checkups or scans; reconnecting spiritually or religiously; sharing the pregnancy news; watching the belly grow; feeling proud of the journey as a mother; other (specify)
14	Mental health condition during pregnancy	Mostly stable and positive; occasional anxiety or stress; frequent sadness or mood swings; depression or serious emotional distress; other (specify)
15	Physical health condition during pregnancy	Very good; mostly okay with minor issues; had some physical complications; faced serious health challenges; other (specify)
16	Other diseases or medical conditions during pregnancy	Short answer
17	Who is reporting the survey	Self; assisted

B Additional file 2: Knowledge base sourcing and preprocessing

Challenges in sourcing Bangla maternal health content

The following constraints were encountered when surveying candidate sources for the knowledge base. They are reported in full here because they may inform similar efforts in other low-resource

language settings.

1. **Quality of available Bangla resources.** Many online Bangla health resources were direct machine translations of English material, often outdated and lacking cultural nuance.
2. **Limited evidence base.** Few sources were clinically vetted or peer reviewed, raising questions about reliability.
3. **Absence of medical cross-validation.** Most resources showed no evidence of review by certified medical professionals.
4. **Scarcity of localized content.** Little material addressed maternal health realities specific to Bangladesh, such as region-specific nutrition practices and the substantial minority of births that still occur outside health facilities.
5. **Misinformation risk.** Social media, forums, and personal blogs frequently contained unverified claims, folk remedies, or contradictory advice.
6. **Fragmented institutional resources.** Government and nongovernmental organization sites contained useful guidance, but this was often fragmented, outdated, or not available in machine-readable form.
7. **Limited coverage of complications.** Content on perinatal mental health, postpartum recovery, and miscarriage support was particularly sparse.
8. **Absence of structured data.** No publicly available structured or machine-readable Bangla medical corpora were identified.
9. **Terminology inconsistency.** Bangla medical terms lacked standardization, with multiple spellings and colloquial alternatives that complicate retrieval.

Preprocessing procedures

Articles retained from the two curated sources underwent the following preprocessing steps before embedding. Counts of items removed at each individual step were not retained; 247 articles survived the full pipeline and were segmented into 2,945 indexed passages.

- **Exclusion of irrelevant or low-quality content.** Because the maternal sections of both sites were crawled in full, each article was manually reviewed and those that were

promotional or only tangentially related to maternal health were excluded. For example, articles on baby shower gift selection were discarded, while articles on managing gestational diabetes were retained.

- **Removal of promotional and redundant text.** Within retained articles, promotional phrases and repeated boilerplate disclaimers were removed so substantive content remained.
- **Removal of identifiers.** Personal references, author names, and case-specific mentions were removed or generalized.
- **Markup cleanup.** HTML tags, embedded scripts, and page metadata were stripped using regular expressions and parser libraries, retaining only human-readable text.
- **Fact-checking and cross-referencing.** Key medical claims were verified against authoritative sources and reviewed by a collaborating licensed physician. Content conflicting with clinical consensus was corrected rather than retained. For example, a source claim that papaya must be avoided entirely during pregnancy was revised to reflect that only unripe papaya carries risk.

Chunking parameter selection

Text was segmented using a recursive character splitter with a chunk size of 800 characters and an overlap of 300 characters. Smaller chunks (approximately 500 characters) narrowed the focus of individual embeddings but fragmented longer gestational-week guides into disjointed segments. Larger chunks (approximately 1000 characters) preserved broader context but reduced the relevance of retrieved segments by including extraneous material. Overlaps below 200 characters risked losing continuity between the description of a complication and its management, while overlaps above 400 characters increased redundancy and storage cost without apparent benefit. The adopted configuration reflects a balance between coherence and efficiency. Systematic benchmarking across parameter ranges was not conducted and remains a limitation.

C Additional file 3: Deployment feedback questionnaire

The instrument administered to participants following use of *Baby and Me* is reproduced in Table 5. Items 5 to 9 used 5-point Likert scales with the anchors shown. As with the formative survey, the name field was collected during administration only and was removed prior to analysis.

Responses to items 3 and 4, which record prior participation in the formative survey and which household member used the application, are reported in the Results.

Table 5: Deployment feedback questionnaire items.

No	Question	Response options
1	Name	Short answer (removed prior to analysis)
2	Age	Short answer
3	Did you take part in the initial survey?	Yes; no
4	Who used the app?	Self; others (specify)
5	How easy was it to use the app?	Likert 1–5 (1 = very hard, 5 = very easy)
6	Were the answers clear and understandable?	Likert 1–5 (1 = not clear, 5 = very clear)
7	Did the response feel caring or empathetic?	Likert 1–5 (1 = completely robotic, 5 = very empathetic)
8	Did the answers match your concerns?	Likert 1–5 (1 = not at all, 5 = fully satisfied)
9	How much did you trust the information from the app?	Likert 1–5 (1 = did not trust at all, 5 = fully trusted)
10	Would you use this app instead of searching Google?	Yes; no; not sure
11	What improvements would you suggest?	Short answer
12	Do you prefer a general chatbot (no data sharing) or a personalized chatbot (sharing personal information with the system)?	Prefer personalized system; prefer data privacy; not sure
13	Would you recommend it to others?	Yes; no; not sure

D Additional file 4: Semi-structured interview protocol

The following protocol guided the exit interviews. Interviews were conducted in Bangla; the protocol is reproduced here in English translation. Responses to the *Influence* section were not analysed separately and no findings on influence or behaviour change are reported in this article;

this is noted in the Limitations.

Introduction and consent

Participants were told that the interview would take approximately 20 to 30 minutes and that its purpose was to understand their experience using the system. They were informed that responses would be anonymized, that no identifying information would be shared, that participation was voluntary, and that they could skip any question or stop at any time. Permission to record the conversation was requested before proceeding.

Warm-up and context

1. Could you tell me a little about yourself: are you currently pregnant, postpartum, or supporting someone?
2. Before this study, how did you usually find information or support for pregnancy or maternal health?

Process and usability

1. What was it like using the chatbot for the first time?
2. Did you face any difficulties using it, for example typing questions, understanding responses, or navigating the interface?
3. Were there times when it worked smoothly or was especially convenient?
4. Was there anything frustrating or confusing?
5. How does using this chatbot compare with asking family or friends, or searching online?

Understanding and trust

1. Can you recall a specific question you asked the chatbot? What did you think of the response?
2. Was the answer clear and understandable?
3. Did the chatbot ever give information you doubted or did not trust, and if so, why?
4. Did you feel reassured, anxious, or neutral after reading the answers?

5. Compared with asking a doctor or midwife, how much would you trust this chatbot?

Influence

1. Did using the chatbot change how you think about your health, pregnancy, or care?
2. Did it prompt you to take any action, such as looking up more information, speaking with someone, or changing a routine?
3. If this chatbot were widely available, do you think women in your community would use it, and why or why not?
4. How often do you think you would realistically use it in future?

Improvements and wrap-up

1. If you could change one thing about the chatbot, what would it be?
2. What is one feature you particularly liked and think should stay?
3. Is there any experience of yours that I didn't ask about, but you would like to share?

E Additional file 5: Deployed system prompt

The system prompt supplied to the language model in the deployed configuration is reproduced below. Typographic bullet and rule characters in the original have been replaced with ASCII equivalents so that the text can be typeset; the wording is otherwise unaltered.

```
You are a supportive and knowledgeable pregnancy help agent, your name is
"Baby and Me".
Your role is to comfort the user and provide clear, accurate, and empathetic
guidance about pregnancy, childbirth, and parenting in the Bangladeshi context.
Always keep in mind that the user may not be technically sound,
so your answers must be simple, clear, and easy to understand, like a helpful
friend. Never be robotic.

---
Tools Usage:
```

```
- For any question related to pregnancy, childbirth, or parenting:
    * Always consult the *knowledgebase* tool first.
    * If the knowledgebase does not provide sufficient information, then use
      the *tavily search* tool.
    * Blend the retrieved information with your own knowledge to give a
      complete and detailed answer.
- If the user asks something unrelated to pregnancy, politely explain that you
  can only help with pregnancy, childbirth, and parenting topics.

---
Answering Style:
- You must be jolly or sad and sometimes funny based on the user's question
  and mood.
- Use emojis where necessary.
- Always respond in **Bangla language**.
- Never include greetings or farewells.
- Keep answers detailed but concise: maximum 5-10 lines.
- If the user requests a table, provide the response in table format.
- Always be empathetic, practical, and context-aware in your responses.
```

Supplementary Files

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

- SupplementaryFile1.pdf
- SupplementaryFile2.pdf
- SupplementaryFile3.pdf
- SupplementaryFile4.pdf
- SupplementaryFile5.pdf