Designing a Bangla Conversational AI Agent for Maternal Health Using Model Context Protocol

by

Fahim Muntasir
23266020

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
M.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
September 2025

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Fahim Muntasir

23266020

Approval

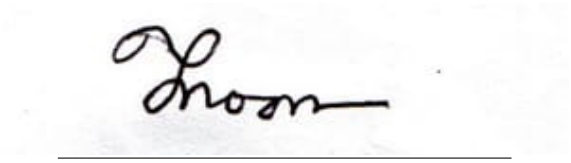
The thesis/project titled "Designing a Bangla Conversational AI Agent for Maternal Health Using Model Context Protocol" submitted by

1. Fahim Muntasir (23266020)

Of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on September 22, 2025.

Examining Committee:

Supervisor:
(Member)



Dr. Jannatun Noor

Associate Professor
Department of Computer Science and Engineering
United International University

External:
(Member)

Dr. A. B. M. Alim Al Islam

Professor
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology

Internal:
(Member)

Moin Mostakim

Senior Lecturer
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Dr. Md Sadek Ferdous

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Associate Professor
Department of Computer Science and Engineering
Brac University

Ethics Statement

This thesis was conducted in strict adherence to the ethical codes and research practices established by BRAC University and in alignment with broader international research ethics guidelines. Given the sensitivity of maternal health and perinatal mental health, particular care was taken to ensure that all stages of the study respected the dignity, privacy, and well-being of participants.

The research protocol was reviewed and approved by the BRAC University Institutional Review Board (IRB) under reference number (220240006), dated (May 6, 2024). Before participation, all respondents were informed about the purpose of the study, the voluntary nature of their involvement, and their right to withdraw at any stage without consequence. Informed consent was obtained verbally and in writing where applicable.

To protect participant confidentiality, all personal identifiers were removed, survey responses were anonymized, and data were securely stored with restricted access. Translated responses were handled with care to ensure cultural sensitivity and accuracy. Given that topics of pregnancy and mental health can be emotionally challenging, questions were framed to avoid harm, and participants were provided with resources for professional medical or psychological support when needed.

Throughout this work, I have ensured that all sources of secondary material are properly acknowledged and cited. As the sole author, I take full responsibility for maintaining research integrity and for addressing any breaches of ethical conduct that may arise. This study aims not only to contribute academically but also to uphold the trust placed in research by the women who shared their experiences.

Abstract

Maternal health in Bangladesh faces persistent challenges, including limited access to skilled providers, informational gaps, and stigma around perinatal mental health. Prior studies show that digital health tools remain constrained by generic content, English-dominated design and neglect of maternal mental health, limiting trust and engagement. We present *Baby and Me*; the first Model Context Protocol-enabled agentic AI system designed for maternal health in Bangladeshi contexts. The system delivers personalized, empathetic guidance in Bangla and Banglish by combining retrieval-augmented generation with a clinically curated knowledge base, web search, and conversational memory. A survey with 72 women revealed frequent worries about miscarriage, anxiety, and mood changes, highlighting the need for empathetic, accessible support. Evaluation of our prototype showed high contextual accuracy, low hallucination, and strong user satisfaction, with participants valuing empathy and trust while requesting greater personalization. Our findings extend human-AI interaction research by demonstrating how culturally grounded, agentic AI can serve not only as an informational tool but also as a relational companion, offering design insights for equitable health technologies. It paves the way for scalable interventions in low-resource settings, with future directions to enhance maternal outcomes. The chatbot can be accessed in this ***link***.

Keywords: Agentic AI; Model Context Protocol; Maternal Health; Retrieval-Augmented Generation; Perinatal Mental Health; SDG 3 (Good Health and Well-being); SDG 10 (Reduced Inequalities); SDG 9 (Industry, Innovation and Infrastructure)

Acknowledgement

I would like to begin by expressing my deepest gratitude to my supervisor, Dr. Jannatun Noor, whose guidance, encouragement, and insightful feedback have been invaluable throughout the development of this thesis. Her patience and mentorship not only shaped the direction of my research but also inspired me to pursue it with dedication.

My heartfelt thanks go to my wife, whose unwavering support, love, and constant encouragement sustained me through the challenges of this journey. Her belief in me has been a source of strength at every step.

I am also thankful to my teachers, colleagues, and friends who have contributed with their advice, feedback, and encouragement. Finally, I owe my sincerest appreciation to my parents and family for their prayers, support, and sacrifices, without which this accomplishment would not have been possible.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter provides the foundation for the thesis by situating the research within its broader context. It begins by outlining the background and motivation behind the study, highlighting the challenges in maternal health and the opportunities for AI-driven solutions. The objectives are then presented, followed by a description of the research design and the key challenges encountered. The chapter also summarizes the main contributions of the work and concludes with an outline of the thesis structure.

1.1 Background

Bangladesh's health care system faces substantial structural and resource limitations, which are particularly evident in the domain of maternal health. Pregnancy-related care remains highly neglected, with insufficient availability of skilled providers, inadequate infrastructure, and poor referral mechanisms for obstetric emergencies [2]. Access to accurate and timely health information is also severely constrained, especially for women in rural and marginalized communities, where reliable sources of reproductive and maternal health guidance are scarce [14], [20]. These systemic weaknesses collectively contribute to delayed care-seeking, inadequate pregnancy monitoring, and heightened risks of adverse maternal and neonatal outcomes [11].

Here, home deliveries remain predominant, with approximately 71% of births occurring outside health facilities. However, only 4% of these are attended by trained birth professionals. Nationally, the density of health workers is 2424 constitutes merely 17% of the global benchmark, underscoring a severe shortage of skilled birth attendants, particularly in rural settings [54]. Approximately 41.5% of women in Bangladesh experience high-risk pregnancies characterized by multiple complications [16], [18]. Contributing factors include limited decision-making autonomy, early or late maternal age at childbirth, high parity, and short interpregnancy intervals, all of which heighten the likelihood of adverse outcomes. Concurrently, only about half of women receive adequate antenatal care during pregnancy [1], [19]. Women with lower levels of education, limited income, and residence in rural areas often experience restricted autonomy in reproductive health decision-making, which in turn increases their vulnerability to adverse pregnancy outcomes.

Another important aspect, mental health in the perinatal period, remains heav-

ily stigmatized [13]. Although postpartum depression is relatively common [7], it is rarely acknowledged within communities or by the health care system. Social stigma, coupled with fear of judgment and internalized shame, discourages many women from seeking professional support [28], [29]. These barriers are further compounded by low levels of awareness and the scarcity of trained mental health providers, resulting in the underdiagnosis and undertreatment of maternal mental health conditions.

Altogether, these findings illustrate that maternal health in Bangladesh is constrained not only by systemic weaknesses in health infrastructure but also by socio-economic and informational inequities. The convergence of high reliance on home births, inadequate antenatal care, and limited reproductive autonomy highlights an urgent need for innovative, scalable interventions that can bypass structural barriers. In this context, technological solutions—particularly artificial intelligence and agentic systems—offer significant potential to expand access to reliable, context-sensitive maternal health information. My study builds on this premise by designing a retrieval-augmented agent that can deliver timely, trustworthy guidance to pregnant women, thereby complementing existing health services and mitigating risks associated with informational and structural gaps.

1.2 Problem Statement and Motivation

Despite significant policy attention to maternal health in Bangladesh, critical challenges persist that prevent women from receiving adequate care during pregnancy and postpartum [30]. A severe shortage of skilled health professionals, particularly in rural and marginalized regions, means that most women rely on home births without trained assistance, and many receive inadequate or no antenatal care [78]. These systemic weaknesses contribute to preventable complications, delayed care-seeking, and poor maternal and neonatal outcomes. Structural solutions alone, such as policy reforms or increased funding, have yet to overcome the entrenched gaps in service delivery.

Equally pressing is the informational gap: women often lack access to timely, trustworthy, and context-specific maternal health guidance [50]. Existing interventions, including mobile health applications, hotlines, and community outreach programs, have delivered limited impact due to infrastructural bottlenecks, generic content delivery, and the absence of personalization. These limitations are particularly acute in rural and marginalized communities, where women face barriers of low literacy, restricted autonomy in health-related decision-making, and socio-economic constraints that further restrict their access to professional care.

Maternal mental health remains an even more neglected dimension. Conditions such as perinatal depression are common but underrecognized due to cultural stigma, fear of judgment, and the scarcity of trained mental health providers. Consequently, women experiencing psychological distress are rarely identified or supported within existing health frameworks, leaving a critical gap in comprehensive maternal health care. Addressing maternal health in Bangladesh thus requires not only physical and clinical interventions but also accessible, stigma-free emotional support.

These realities underscore the urgent need for innovative approaches that can extend the reach of health services while empowering women with reliable, personalized guidance. Technological interventions, particularly those that leverage artificial intelligence, offer a unique opportunity to address these challenges. Unlike static mobile health tools or rule-based chatbots, retrieval-augmented and agent-based conversational systems can deliver evidence-based knowledge in multilingual and culturally sensitive forms, while also personalizing interactions through conversational memory and real-time decision-making.

The motivation for this study, therefore, lies in exploring how an empathetic, agentic AI system can bridge existing structural and informational gaps, enhance the accessibility and trustworthiness of maternal health guidance, and ultimately contribute to reducing preventable maternal and neonatal risks in resource-constrained contexts such as Bangladesh. To explore this potential in the Bangladeshi context, this study investigates the following research questions:

1. How does the integration of multilingual support (Bengali and Banglish) and culturally contextualized knowledgebases affect the accessibility, user engagement, and perceived trustworthiness of AI-driven maternal health guidance systems among pregnant and new mothers in Bangladesh?

2. To what extent does an agent-based conversational system with a tool orchestration, conversational memory, and real-time decision-making capabilities improve the quality and personalization of maternal health support compared to traditional rule-based chatbot systems?

1.3 Objectives

The overarching aim of this study is to design and develop a context-aware, retrieval-augmented AI agent capable of providing accurate, timely, and culturally sensitive maternal health guidance to pregnant and postpartum women in Bangladesh. By leveraging technological innovation, the study seeks to address the dual challenges of structural limitations in health care delivery and informational inequities that hinder women's access to reliable maternal health support.
To achieve this primary aim, the study pursues the following specific objectives:

1. **Knowledgebase Integration:** To construct a comprehensive, evidence-based knowledgebase encompassing maternal health and perinatal mental health information tailored to the socio-cultural context of Bangladeshi women.

2. **Agent Design and Accessibility:** To develop a user-friendly, interactive AI agent that delivers personalized guidance, ensuring accessibility for women with low literacy levels, limited digital familiarity, or socio-cultural barriers to seeking care.

3. **Evaluation of Effectiveness:** To assess the agent's impact on women's knowledge, decision-making autonomy, and early care-seeking behavior during pregnancy and the postpartum period.

Through these objectives, the study aims to provide a scalable technological intervention that complements existing health services, mitigates informational and structural gaps, and contributes to improved maternal and neonatal health outcomes in resource-limited settings.

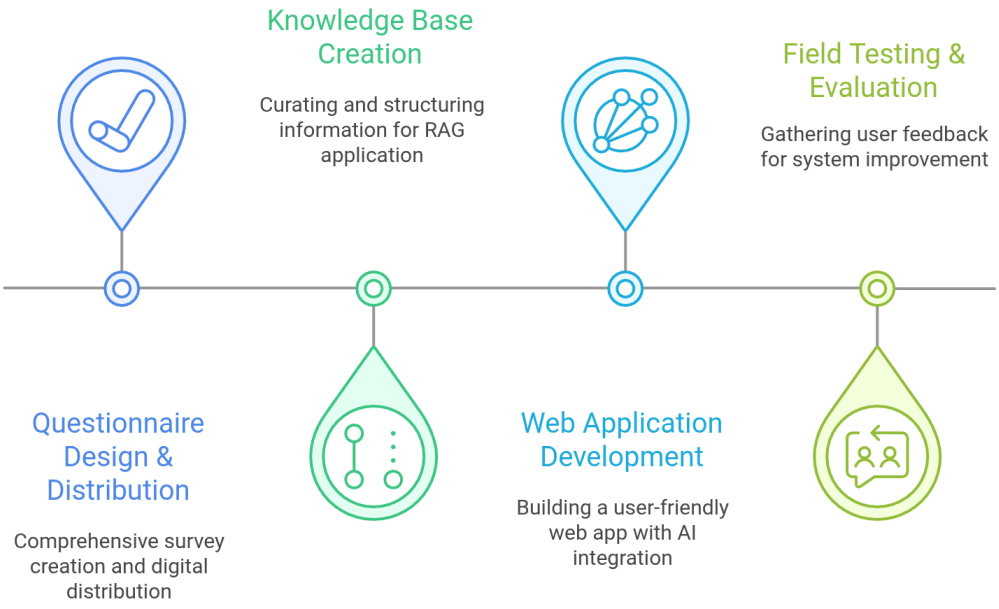


Figure 1.1: Research design summary

1.4 Research Design

To achieve the goals of our research, we have adopted a user-centered, exploratory design that combines both qualitative and quantitative methods to understand the lived experiences of pregnant women and to develop a responsive AI-powered support system. Fig. 1.1 shows the summary of the research design.

1. **Questionnaire Design & Distribution**: The study began with the design of a comprehensive questionnaire to capture the emotional, informational, and physical needs of pregnant women. It explored concerns like health complications, mental well-being, support systems, and technology accessibility, and was digitally distributed to a diverse participant pool.

2. **Knowledgebase Creation**: Based on survey insights, relevant information was sourced from public and semi-structured online resources (medical blogs, forums, informational websites). This content was carefully curated, cleaned,

and structured to form a domain-specific knowledgebase for a Retrieval-Augmented Generation (RAG) application.

3. **Web Application Development**: A no-login, lightweight web application was developed, featuring an intuitive front-end for accessibility. The backend integrates a natural language AI model with the knowledgebase using an RAG architecture, enabling the system to retrieve context-aware information from natural language queries.

4. **Field Testing & Evaluation**: The final phase involves field testing the application by introducing it to real users (pregnant women and new mothers). Their feedback on usability, relevance, emotional tone, and perceived helpfulness is collected to evaluate the system's effectiveness and guide future iterations.

5. **Holistic Development**: This participatory and iterative approach ensures the AI tool is not only technically functional but also contextually appropriate and empathetically aligned with users' needs, making it a human-centric and impactful AI solution.

1.5 Research Challenges

We faced numerous challenges throughout different stages of the research process:

1. **Collecting survey data across socio-economic groups:** Reaching participants from diverse socio-economic backgrounds proved challenging due to disparities in literacy, digital access, and willingness to engage with research activities. Many women from rural and marginalized communities were hesitant to participate, either because of limited awareness of the study's purpose or cultural barriers that restrict open discussion about pregnancy-related issues. Ensuring representation across different income, education, and geographic segments required extensive outreach and careful coordination.

2. **Identifying the pain points of pregnant mothers and their concerns:** Understanding the lived experiences of pregnant women went beyond simple data collection. Many participants expressed concerns indirectly or hesitated to speak openly due to stigma surrounding maternal mental health and reproductive health discussions. Extracting genuine pain points required building trust, ensuring confidentiality, and designing survey questions that were sensitive yet probing enough to capture nuanced challenges.

3. **Obtaining reliable medical data in the Bangla language:** High-quality, structured medical information in Bangla was scarce. Most available resources were either in English or fragmented across different platforms, making it difficult to build a curated knowledgebase tailored for Bangladeshi women. Translating and contextualizing medical guidelines into accurate, culturally sensitive Bangla terminology posed an additional challenge, as direct translations often failed to capture nuances understood by local communities.

4. **Selecting a suitable large language model for Bangla conversations:** Many state-of-the-art language models are optimized for English and perform poorly in low-resource languages like Bangla. Identifying a model that could understand and generate natural, context-aware Bangla responses required significant experimentation. Balancing linguistic proficiency with computational efficiency was an added challenge, as models with better Bangla performance often demanded high resource overheads.

5. **Enabling effective tool integration for the agent:** Designing the agent to use multiple tools—such as retrieval modules, real time search, and conversational memory—posed technical difficulties. Ensuring that the agent could orchestrate these tools efficiently, without producing irrelevant or redundant outputs, required iterative fine-tuning. Additionally, aligning tool usage with the conversational flow in Bangla, while maintaining accuracy and trustworthiness, proved to be a complex engineering task.

1.6 Contributions

To our knowledge, this is the first research to incorporate Agentic AI with Bangla knowledgebase for maternal care. This thesis makes several contributions at the intersection of maternal health and artificial intelligence, with a particular focus on the Bangladeshi context:

1. **Contextual Knowledgebase Development:** A curated 'vector knowledgebase' was constructed using clinically vetted Bangla maternal and perinatal health information, including mental health content. This addresses the scarcity of high-quality, localized medical resources in the native language.

2. **Design of a Multilingual, Culturally Sensitive Agent:** A retrieval-augmented conversational agent was developed with support for Bangla and Banglish, ensuring accessibility for women across literacy levels and linguistic preferences. The system was designed to deliver guidance that aligns with cultural sensitivities and social realities.

3. **Advancement in Tool-Orchestrated Conversational Systems:** The agent integrates multiple tools—including retrieval modules, real time search, and conversational memory—demonstrating how tool orchestration and real-time decision-making can improve the quality and personalization of maternal health support compared to traditional rule-based systems.

4. **User-Centered Insights into Maternal Health Needs:** Through surveys and user feedback, the study identifies pain points and unmet needs of pregnant and postpartum women in Bangladesh, contributing new empirical knowledge about maternal health challenges from a user perspective.

Collectively, these contributions illustrate how AI-driven, agentic systems can complement existing health services, bridge informational inequities, and provide scalable support for maternal and perinatal health in resource-limited settings.

1.7 Thesis Outline

This thesis is organized into ten chapters, each addressing a key stage of the research. This chapter introduces the problem, provides background context, and highlights the motivation behind the study. It also states the research goals, outlines the limitations, and presents the contributions along with the analytical framework. The remaining chapters are structured as follows:

Chapter 2 reviews related work and prior studies on maternal health technologies. It surveys global interventions, with a particular focus on conversational agents and chatbots used in similar domains.

Chapter 3 details the system requirements understanding. It discusses the initial survey, highlights user needs, and identifies the design requirements that informed the development of the system.

Chapter 4 describes the research methodology, including survey instruments, analytical methods, and the design of both the frontend and backend frameworks of the system.

Chapter 5 introduces the interaction with the *Baby and Me* bot, illustrating its functionalities and user interfaces.

Chapter 6 presents the technical evaluation of the system, measuring the chatbot's performance across different contexts and models.

Chapter 7 focuses on end-user evaluation, analyzing responses from participants and assessing usability, trust, and perceived empathy.

Chapter 8 discusses the overall findings of the research, integrating insights from both technical evaluation and user studies.

Chapter 9 outlines the limitations of the study and identifies directions for future research.

Chapter 10 concludes the thesis by summarizing the contributions, reflecting on the implications, and suggesting potential improvements derived from user feedback.

Chapter 2

Related Work

This chapter provides a comprehensive overview of the existing literature and related work in the field of maternal health support systems, particularly focusing on conversational agents and their applications in pregnancy care. We will explore the current state of research, identify gaps in existing solutions, and highlight the need for culturally and linguistically inclusive systems that can effectively address the unique challenges faced by pregnant and new mothers in Bangladesh.

2.1 The Role of Conversational Technology in Maternal Health

Digital health interventions, including AI-driven chatbots, are increasingly proposed to improve maternal care by extending access, education, and emotional support. Globally, about 295,000 women die annually from pregnancy-related complications (with 95% of deaths in low- and middle-income countries), and many more experience preventable morbidity [48]. AI – especially conversational agents – could help fill gaps in knowledge and services, providing timely information on pregnancy and postpartum care and screening for mental health issues. For example, AI chatbots have been developed to deliver breastfeeding counseling and address maternal anxiety: in a recent Turkish trial, postpartum mothers randomized to an AI-assisted breastfeeding chatbot had significantly higher breastfeeding self-efficacy and success rates and lower anxiety than controls [68]. Similarly, an evaluation of large language models (ChatGPT, Gemini, Copilot) for lactation queries found that advanced models (GPT-4o) produced structured, readable, guideline-concordant answers, although variability and "hallucination" risk remain [72]. These results illustrate both the promise and current limitations of generative AI for maternal health education.

Many chatbot studies report high user engagement and satisfaction. In the US pilot of "Moment for Parents", a rules-based perinatal mental health chatbot, 63.9% of pregnant/postpartum users reengaged at least once, and 93% rated the content relevant [70]. A larger U.S. RCT (using Woebot, a conversational CBT agent) found that postpartum women had greater reductions in PHQ-9 depression scores over six weeks (-1.32 vs. -0.13) and high satisfaction (91%) [8], [27], [33]. Another chatbot-based outreach study sent postpartum Recovery reminders at discharge: about 63% of recipients opened the chatbot messages and 80% found them easy to use [53].

These Findings confirm that perinatal chatbots can engage users and positively influence behaviors (e.g. increased information, decline in anxiety), consistent with a systematic review, which found chatbots feasible for perinatal care advice (breastfeeding, nutrition, stress management, etc.) and associated with better knowledge and behaviors [60].

Earlier studies mostly focused on rule-based chatbots, which, while functional, lacked the adaptability and personalization needed for complex maternal health scenarios. 2013 Chatbot "Tanya" [4] was one of the early examples that demonstrated the potential of conversational agents in this domain, providing support for breastfeeding and infant care. A lot of studies were mostly focused on the use of chatbots in informing about preconception health complexities. Jack et al., Bickmore et al., Gardiner et al., Montenegro et al., and Maeda et al. explored the use of chatbots for preconception health education, highlighting their potential to improve knowledge and awareness among users [6], [9], [21], [23], [31]. These studies were cohort-based and primarily qualitative in nature, providing valuable insights but lacking generalizability. As these were rules-based chatbots, they were limited in their ability to adapt to individual user needs and preferences.

Penn Medicine has implemented a conversational AI chatbot, "Penny," to enhance postpartum maternal care, which is often called the fourth trimester [56]. This AI-powered tool provides automated, personalized responses to new parents' questions via text message, reducing the burden on clinicians and allowing them to focus on more complex cases. The program has shown high patient engagement and satisfaction, with over 70% of questions answered correctly by the AI, and has helped identify new onset postpartum hypertensive disorders and screen for depression, particularly benefiting underserved populations.

Recent years have witnessed a surge in conversational agents aimed at supporting maternal and reproductive health. Systems such as MamaBot [12], PregChat [5], and PreMomBot [10] demonstrate growing interest in leveraging AI for pregnancy care. MamaBot provides weekly gestational tips, fetal milestone tracking, and nutrition guidance, offering basic yet useful functionality for expecting mothers. PregChat, still under beta testing, expands on this by including symptom checks, emergency protocols, medication guidance, and antenatal exercise suggestions features rooted in clinical knowledge. PreMomBot places a notable emphasis on maternal mental health, integrating both prenatal and postnatal care pathways alongside psychoeducational content.

Table 2.1 presents a comparison of several studies designed to address prenatal and postnatal contexts. However, research efforts specifically targeting Bangladesh and utilizing the Bangla language remains in their nascent stages, with no fully deployed or publicly accessible systems currently available.

Table 2.1: Illustrating AI chatbots and digital tools in maternal health contexts. (** denotes non-peer-reviewed source.)

Study/ Chatbot	**Region/ Language**	**Target/ Use-Case**	**Tech (AI vs Rule)**	**Key Findings**
Kerimoglu et al.,2025 [68]	Turkey, Turkish	Breastfeeding counseling (postpartum)	Rule-based AI chatbot on smartphone	Increased exclusive BF rates and self-efficacy; lower anxiety in AI group
Ozer et al.,2025 [72]	(Global test, English)	Breastfeeding FAQs (lactation)	AI chatbots (LLMs: GPT-4o, Copilot, etc)	GPT-4o had highest quality/accuracy (95% factual); noted risk of hallucinations
Liu et al.,2017 [10] (Premom App)	USA	digital ovulation test reader; intelligent fertility charting; customized cycle insights	Rule Based	More than 8 million women have used the Premom app
Mcalister et al., 2025 [70] (Moment for parents)	USA	Mental health support (pregnant/ postpartum)	AI Chatbot	63.9% users re-engaged; 93.3% found advice relevant, high satisfaction
Siwicki et al., 2024 [56] (Penny)	USA	Enhance postpartum maternal care	AI Chatbot	Seen high user engagement; 70% correctness achieved by AI model.
Suharwardy et al., 2023 [36] (Woebot)	USA	Postpartum depression support	AI chatbot	Greater PHQ-9 score reduction in chatbot vs. control; 91% user satisfaction

Continued on next page

Table 2.1: Illustrating AI chatbots and digital tools in maternal health contexts. (** denotes non-peer-reviewed source.) (Continued)

Rivera et al., 2024 [53]	USA (English/ Spanish)	Postpartum newborn care reminders	Rule-based SMS chatbot	63% open-rate; 80% found chatbot messaging timely and easy to use
Siddiqi et al., 2024 [55]	Pakistan (Urdu/ English)	Child immunization info (LMIC example)	AI/NLP chatbot (WhatsApp)	30.7% caregivers used bot; 90% satisfaction; feasible in low-literacy setting
Vaira et al., 2018 [12] (Mama Bot)	USA	Tracks fetal health and share tips	Rule Based WhatsApp chatbot	Saw exciting interactions in African communities
Pollak et al., 2014 [5] (Pregchat)	UK	SMS-texting for Gestational Weight Awareness	Rule based text Message	86% responded to texts; 80% said they would recommend this program to a friend
Rahman et al., 2019[17] (Disha)	Bangladesh (Bengali)	General health symptom checker	Rule-based closed-domain chatbot	Early Bangla-language health chatbot; diagnoses diseases, tracks health
Acme AI, 2024 [39] Probahini**	Bangladesh (Bangla/ Banglish)	Menstrual health QA	AI chatbot (multilingual)	Uses Bangla and romanized Bangla ("Banglish"); local startup solution; presumably improved info access

Continued on next page

Table 2.1: Illustrating AI chatbots and digital tools in maternal health contexts. (** denotes non-peer-reviewed source.) (Continued)

Jui et al., 2021[26] Maya**	Bangladesh (Bangla, English, Banglish)	General telehealth queries	AI-powered platform chatbot	Resolves 70% of queries via AI (95% accuracy)
Ahmed et al., 2023 [80] SuSastho.AI**	Bangladesh (Bangla)	Adolescent SRH and mental health	Planned LLM-based chatbot	Under development: multilingual AI for sexual/ reproductive/ mental health info for teens
Mohtasim et al., 2024 [49] (Gorbho Shongi)	Bangladesh (Bangla)	Personalized care and daily tips	AI Chatbot	prototype stage

2.2 Bangla Language Barriers in Digital Health Systems

In South Asia and other LMIC settings, language and culture greatly affect chatbot adoption. Bangladesh, for instance, has near-universal mobile phone access (100% penetration) but low computer ownership and literacy barriers [40]. AI tools must therefore use mobile-friendly interfaces and local languages. Several initiatives in Bangladesh demonstrate this: Maya is a Bangladeshi "digital doctor" platform whose chat interface supports Bangla, English, and even Romanized Bangla ("Banglish") for user convenience [26]. Similarly, the menstrual-health chatbot Probahini deliberately replies in Bengali script or in transliterated Bangla to suit user preference [39]. A recent "LifeBot" concept proposed in Bangladesh emphasizes a Bangla-language pregnancy chatbot to educate mothers with personalized messages, noting that a Bengali-language chatbot is a "smart solution" to overcome information gaps [15].

Local-language capability is crucial: Hussain et al. (2023) found that Bablibot was trained on Urdu and regional languages so caregivers in Pakistan could ask questions naturally [55]. They conclude that chatbots can be tailored to LMICs by training on local language datasets, making them feasible even in low literacy settings. In Bangladesh, the SuSastho.AI project is explicitly building an open-source LLM that operates in Bangla (and other local languages) to serve adolescents' sexual and mental health needs [80]. GorbhoShongi is a culturally tailored mHealth app designed to support maternal mental well-being in Bangladesh, featuring a bilingual interface and an LLM-based chatbot [49]. It offers personalized care and daily tips for pregnant and postpartum women. However, the app is still in early prototype stages and lacks real-world user validation. These examples underscore that the use of Bangla/Banglish and other regional languages can greatly broaden access – an English-only bot would exclude many. Likewise, a Lancet case study of a pandemic-

era telemedicine program in Bangladesh found that limited phone/internet access remained a major barrier (many women "were not reachable by phone due to lack of access or network coverage") [35], reinforcing the need for offline and low-bandwidth solutions in LMICs.

Beyond Bangladesh, other South Asian efforts show promise. In India, NGOs are piloting AI chatbots for women's reproductive health: for example, Mumbai's Myna Mahila Foundation trained local women to help develop a chatbot answering questions on sexual/reproductive health [41], [62]. Adoption studies suggest that culturally tailored chatbots can address sensitive topics while reducing stigma. However, overall, the evidence in LMICs is sparse. A recent review of perinatal depression apps noted almost all existing trials were in high-income countries; only one was in a lower-middle-income country (India), and none in low-income settings [22], [76]. This gap extends to chatbots: most published chatbot studies are from North America or Europe, and very few are formal evaluations from South Asia. The nascent projects above show momentum, but rigorous evaluations in LMIC contexts remain an urgent research need.

2.3 Digital Mental Health Support During Pregnancy

Maternal mental health is a critical component of perinatal care. Globally, about 10–20% of women experience depression or anxiety during pregnancy or postpartum [34], [58]. Untreated perinatal mental health issues harm mothers and infants (e.g., affecting infant development). Digital tools offer scalable support, and AI chatbots have begun to address this [24]. For example, Woebot for Pregnancy (a spin-off of the Woebot chatbot) was tested in a U.S. RCT: while the primary outcome showed a moderate decrease in PHQ-9 scores for the chatbot group [33], the difference with usual care was small, although nearly all users reported satisfaction. The Wysa study of 51 users (in the UK) found that highly engaged pregnant/postpartum women had significantly greater reductions in self-reported depression over 6 weeks than less-engaged users [44], suggesting benefit from regular interaction. Another pilot in the U.S. (Moment for Parents) showed pregnant individuals were willing to use a chatbot for anxiety/mindfulness: 93% found it relevant [70].

In LMICs, digital perinatal mental health is less developed but urgently needed. Mobile-phone–based interventions (mostly SMS or phone calls) have been tried [32], [45], [76]. One text-message study in India showed the feasibility of delivering behavioral activation to depressed mothers, but did not use AI [71]. Some non-AI chat interventions exist in South Asia (e.g., motivational messages in India), but published data are scarce. Given smartphone uptake (85% worldwide) and demonstrated interest in chat-based support, there is a strong rationale to extend AI chatbots for mental health to pregnant women in LMICs.

Chatbots could also assist non-psychological needs affecting mental health. For For example, some maternal chatbots deliver medical and pregnancy information – empowering women with knowledge that may indirectly alleviate anxiety. Moreover,

some chatbots include empathic features or link users to counselors. Crucially, deployment in LMICs must consider acceptability and privacy: users in Bangladesh have different literacy and privacy concerns than Western populations. Early evidence suggests they will use AI support if it is local-language and free [51], but formal studies of mental-health outcomes are lacking [67], [73].

2.4 Conversational Interfaces and Trust in AI Systems

The effectiveness of conversational agents in maternal health is significantly influenced by user trust and engagement. Trust in AI systems is crucial for ensuring that users feel comfortable relying on these tools for sensitive health information. Factors such as transparency, reliability, and cultural relevance play a vital role in building trust. Studies have shown that users are more likely to engage with AI systems that provide clear explanations, demonstrate reliability, and align with their cultural and linguistic contexts.

Chung et al. [25] found significant positive correlations between perceived trustworthiness and user engagement in AI-driven health applications, emphasizing the importance of trust in enhancing user experience and satisfaction. Their study investigating their chatbot's perceptions revealed that while participants generally found the chatbot easy to use ($M = 5.34$, $SD = 0.73$) and easy to learn ($M = 6.35$, $SD = 0.71$), about half reported it as useful ($M = 4.87$, $SD = 1.11$) or expressed overall satisfaction ($M = 4.90$, $SD = 1.26$). Bickmore et al. [21] conducted longitudinal assessments of chatbot performance indicating that usability remained relatively stable, with 62.9% at 6 months and 67.9% at 12 months (non-statistically significant). Similarly, user satisfaction with the chatbot demonstrated consistently high rates, at 80.0% and 85.7% at 6 and 12 months, respectively, with no statistically significant difference observed over time.

Despite their potential, these systems fall short in supporting the full spectrum of agentic features—such as personalized reasoning, memory, and multimodal tool integration—that are crucial for handling sensitive and evolving maternal health contexts. The current landscape thus reveals critical gaps in culturally adaptive, linguistically inclusive, and technically advanced maternal health chatbots. These limitations underscore the need for a robust, multilingual, agent-based system that combines real-time support, conversational continuity, and localized medical knowledge to serve pregnant and new mothers more effectively in under-resourced settings.

2.5 Agentic AI and Tool-Oriented Chatbots

The latest AI advances point toward "agentic AI" – systems that autonomously orchestrate multiple tools – as the future of chatbots. Traditional chatbots are either rule-based or single-model LLMs; agentic AI can do more. A recent perspective defines agentic AI as autonomous, context-aware systems that can integrate diverse data sources and actively pursue goals [66]. For example, an agentic healthcare assistant might recognize a patient's symptoms, then call an API to a medical database,

schedule an appointment, and send reminders – all without human prompting. IBM explains that agent orchestration involves specialized AI agents each designed for tasks (e.g. one for NLP, one for scheduling, one for retrieving guidelines) [64]. Such agents "call" external tools (APIs, calculators, medical records, web searches) as needed to fill gaps in knowledge [43]. In practice, this means a perinatal health chatbot could potentially tap into up-to-date clinical protocols, language translation APIs, symptom-checker engines, or even health sensors. New agent protocol called Model context protocol (MCP) has revolutionized the tool access by LLMs, making them adapt with information from any source [63], [65].

In summary, the technical frontier is moving from rule-based to standalone chatbots to autonomous agentic AI systems with multiple modules. While most current maternal chatbots use a single AI or rule engine, developers are exploring multi-agent designs. For example, some research prototypes embed chatbots within broader "health assistants" that can hand off to doctors or schedule visits if needed. The literature is just emerging: one review of AI chatbots for healthcare notes that agent frameworks (combining LLMs with plugins) could enhance accuracy and utility. As these tool-enabled chatbots evolve, future studies will need to evaluate whether they truly improve outcomes and how to ensure safety and equity in automated maternal health support.

AI chatbots in maternal healthcare are gaining attention due to their potential to provide timely information and support. However, further research is needed to ensure their ethical integration and local language adaptations. While some evidence shows acceptability in high-income settings, more rigorous trials are needed in high-risk LMIC populations. Future chatbots may incorporate local clinical data or decision rules.

Chapter 3

System Requirement Understanding

To ground the design of our maternal health conversational agent in the lived realities of pregnant women in Bangladesh, we began with a systematic exploration of their needs, challenges, and expectations. This phase was critical not only for surfacing the informational and emotional gaps faced by expectant mothers, but also for translating these insights into concrete design requirements for our system.

Figure 3.1: Inside of the hospital premises

Figure 3.2: Initial survey data collection

3.1 Initial Survey Design and Administration

An exploratory survey was designed to identify pain points, complications, precious positive moments, and both physical and mental health conditions that shape women's pregnancy experiences. The instrument contained 17 questions, primarily

in a tick-box format to reduce response burden, while including a few open-ended items to capture nuance. The form was first distributed through social media platforms, though the reach was limited. To ensure broader representation, we subsequently conducted in-person surveys at hospitals and clinics in Pabna, Bangladesh (Fig. 3.1 and 3.2).

In total, 72 women participated, spanning diverse socio-economic, educational, and age categories. Both currently pregnant women and recent mothers were included. Appendix A presents the questionnaire used for this survey. Our goal was to identify what sort of thinking the expectant mothers mostly had. The survey thus captured not only quantitative patterns but also the everyday concerns and joys of pregnancy as expressed by Bangladeshi women. It served as the empirical foundation for framing the problem space and guided the formulation of our system's requirements.

3.1.1 Survey Data Cleaning

Before any form of analysis, the raw responses underwent multiple layers of data cleaning to enhance their usability and eliminate inaccuracies.

- **Removal of Duplicate Entries:** Initially, the dataset was scanned for exact duplicate rows that may have resulted from accidental multiple submissions or form refresh errors. These duplicates were identified and removed to avoid bias and ensure each participant was only represented once.
- **Correction of Typographical and Input Errors:** Manual review was conducted on open-ended and categorical fields where participants had typed their responses. This included correcting spelling errors, inconsistent capitalization, and spacing issues to maintain uniformity across entries.
- **Translation of Bangla Responses to English:** Since the survey allowed participants to respond in Bangla, these responses were systematically translated to English to preserve semantic accuracy. This was necessary for the uniformity during analysis.
- **Standardization and Encoding of Categorical Variables:** Categorical columns such as education level, occupation, and household location were standardized into predefined classes. This step included mapping multiple variations of similar responses (e.g., "homemaker", "housewife") into a single, consistent category. Encoding was then applied to convert these standardized values into numerical or label-based formats suitable for analysis.
- **Grouping of Similar Items:** Responses from open-ended or multi-select questions (e.g., pregnancy complications, emotional concerns) often included a wide variety of phrasings. A semantic grouping process was applied to cluster similar responses under unified categories. For instance, "lower back pain" and "body aches" were grouped under "musculoskeletal discomfort" to reduce redundancy and increase interpretability.
- **Handling Missing or Ambiguous Data:** Responses with missing or ambiguous entries were flagged and reviewed. In cases where imputation was

not feasible, these entries were excluded from specific analyses that required complete data.

- **Anonymization and Privacy Preservation:** We didn't collect any personal information to ensure compliance with ethical research standards and participant confidentiality.
- **Data Type Correction and Consistency Checks:** Fields representing numerical values (e.g., age, number of pregnancies) were explicitly cast into appropriate data types. Range checks and logical constraints were also applied to validate consistency (e.g., age values were checked to fall within a humanly possible range).

These procedures ensured that the final dataset was not only clean and consistent but also more analytically meaningful and ethically sound.

Table 3.1: Summary of initial survey responses ($N = 72$). Percentages rounded.

Category	**Response**	**%**
Age distribution	18–25	22
	25–35	58
	35+	20
Education	No formal schooling	6
	Secondary	25
	Undergraduate	44
	Masters+	25
Occupation	Homemakers	40
	Service workers	26
	Others	34
Residence	Urban	58
	Semi-urban	21
	Rural	21
Smartphone access	Personal smartphone	80
	Shared/none	20
Language preference	Bangla only	28
	Bangla + English	71
Top complications reported	Fear of miscarriage	7.1
	Labor pain	7.1
	Mood swings	9.3
	Anxiety	6.9
	Depression	6.9

Continued on next page

Table 3.1: Summary of initial survey responses ($N = 72$). Percentages rounded. (Continued)

Top concerns	Miscarriage risk	9
	Physical changes	7.7
	Baby's health	7.7
Positive highlights	Feeling baby's first kick	8.4
	Sense of pride and joy	8.4
	Watching the belly grow	8.2

3.2 Demographics, Access, and Maternal Health Concerns

3.2.1 Univariate Analysis

Most participants were of childbearing age (58% aged 25–35), with varied education levels and occupations, predominantly homemakers (40%) and service workers (26%). The sample was largely urban (58%), with 80% owning smartphones, though language preferences emphasized the need for Bangla-first design (28% Bangla-only). Reported concerns centered on miscarriage risk, physical changes, and the baby's health, while moments of joy included the baby's first kick and feelings of maternal pride, underscoring pregnancy as a time of both vulnerability and empowerment. An overview of the responses is given in Table 3.1.

While fig. 3.12 and 3.13 indicate that a majority of respondents reported being in mostly good physical and mental health, this should not be misinterpreted as reducing the need for supportive systems. Pregnancy is a dynamic and vulnerable period, and even women who appear "mostly stable" may face episodes of anxiety, uncertainty, or lack of information. Moreover, the minority reporting complications or serious emotional distress represents a critically important group that requires additional support. Thus, the proposed system is not designed exclusively for "healthy" mothers but rather to provide inclusive assistance across the full spectrum of maternal experiences—from reassurance and preventive guidance for the majority, to contextual, empathetic support for those facing difficulties.

While fig. 3.15 and 3.16 align qualitatively with common themes in pregnancy literature (e.g., fears of birth and miscarriage), the reported percentages are lower than epidemiological data, potentially due to [e.g., sample demographics, question wording, or focus on primary vs. any concern]. For comparison, the WHO reports 10% of pregnant women experiencing mental disorders globally [37]. This suggests our findings may underestimate prevalence in diverse populations. These results highlight the emotional burden of pregnancy, even if percentages appear modest, and underscore the need for targeted support (e.g., mental health screening). Future research could standardize survey methods to better align with broader statistics.

1. **Age Distribution:** The majority of respondents (58%) were between 25 and 35 years of age. The next largest age groups were 18–25 years (22%) and 35–45 years (14%). A smaller proportion of participants fell into the under 18 (4%) and over 45 (1%) age brackets.

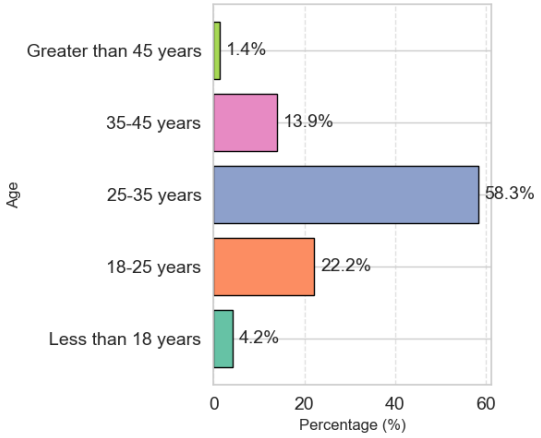


Figure 3.3: Age distribution

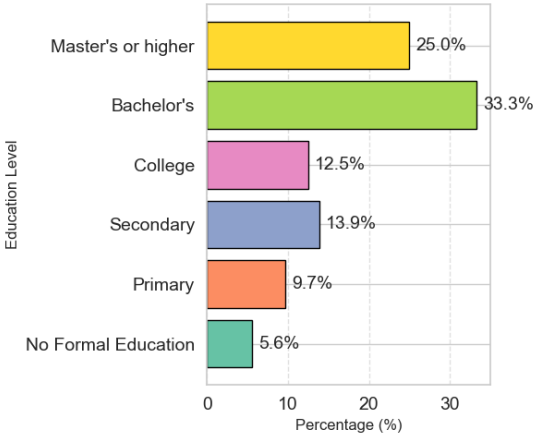


Figure 3.4: Education distribution

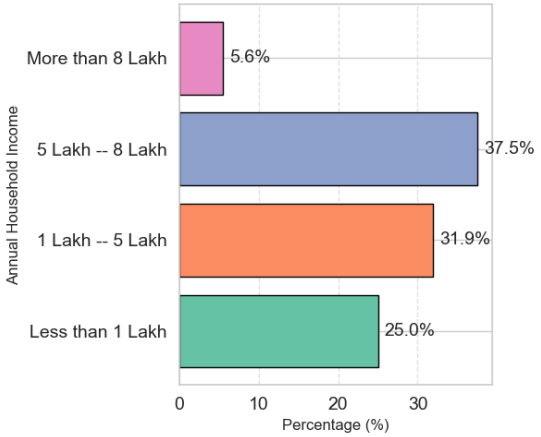


Figure 3.5: Annual household income

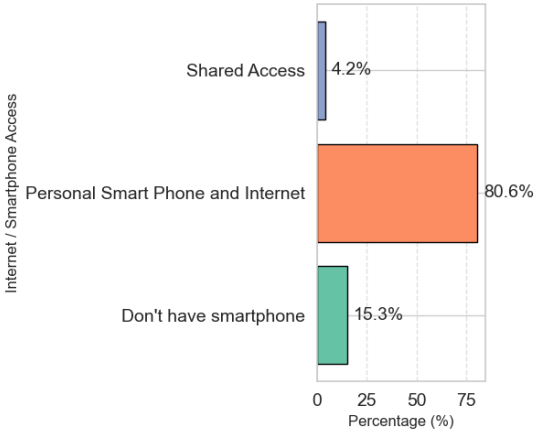


Figure 3.6: Internet access distribution

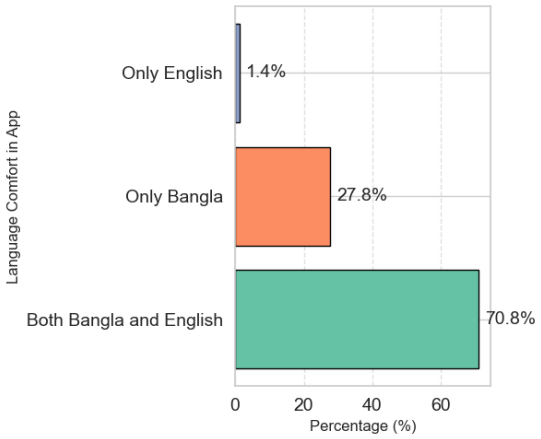


Figure 3.7: Language preference in application

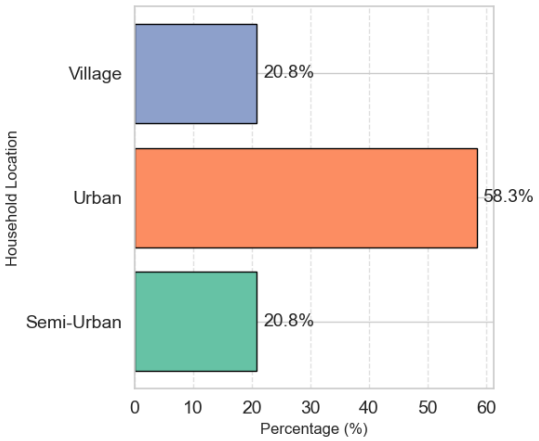


Figure 3.8: Location distribution

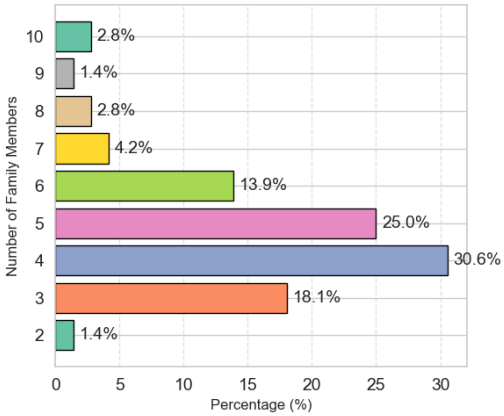


Figure 3.9: Distribution of number of family members

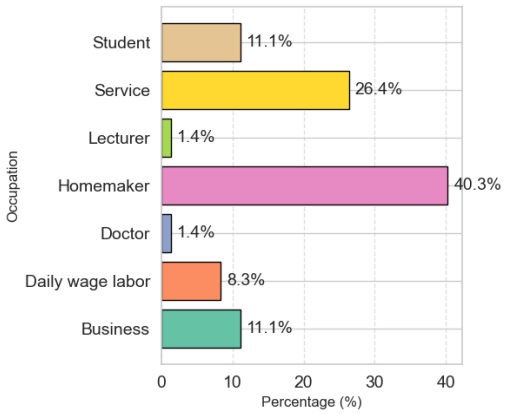


Figure 3.10: Occupation distribution

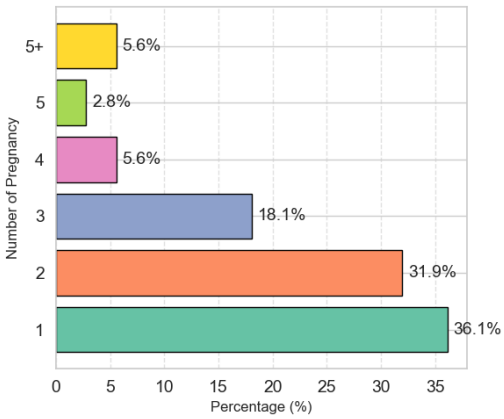


Figure 3.11: Number of pregnancy

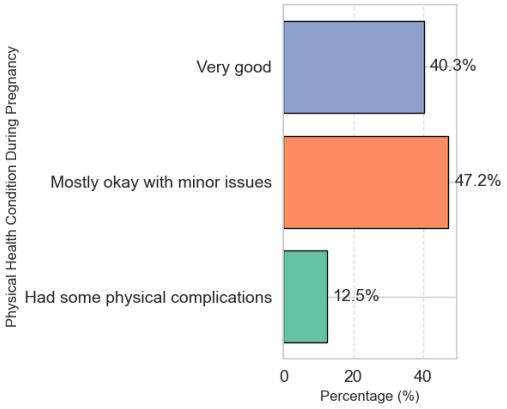


Figure 3.12: Physical condition during pregnancy

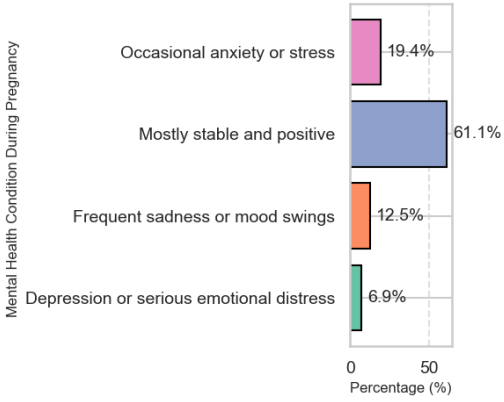


Figure 3.13: Mental health condition during pregnancy

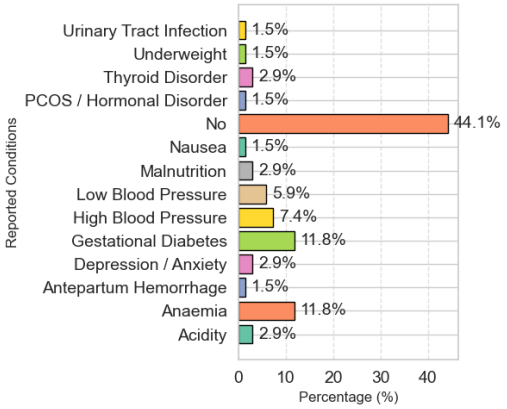


Figure 3.14: Reported other conditions during pregnancy

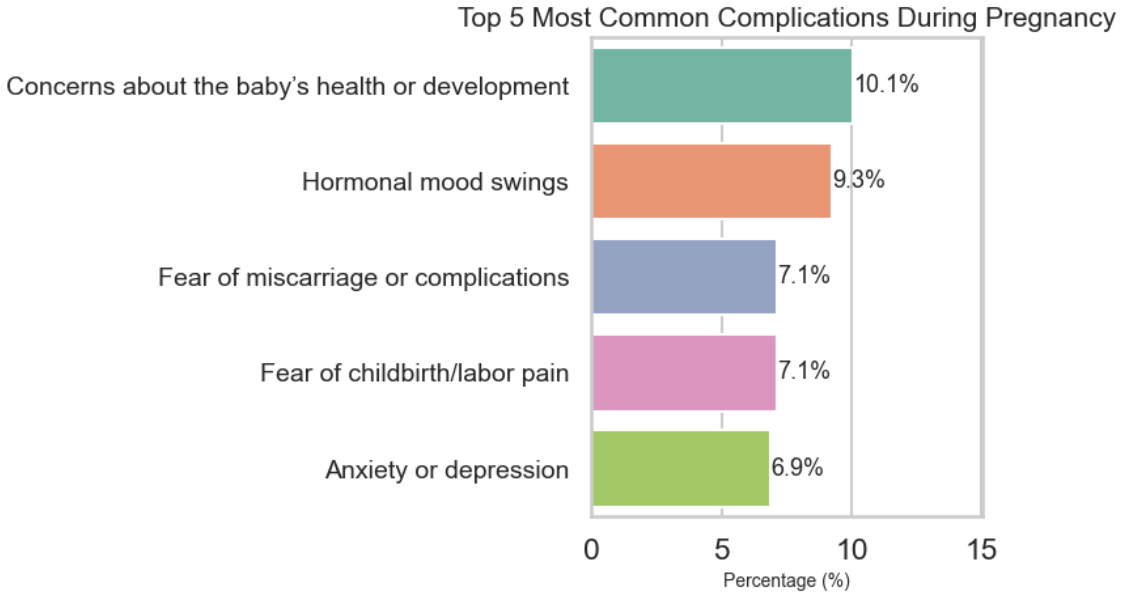


Figure 3.15: Top 5 reported complications by participants

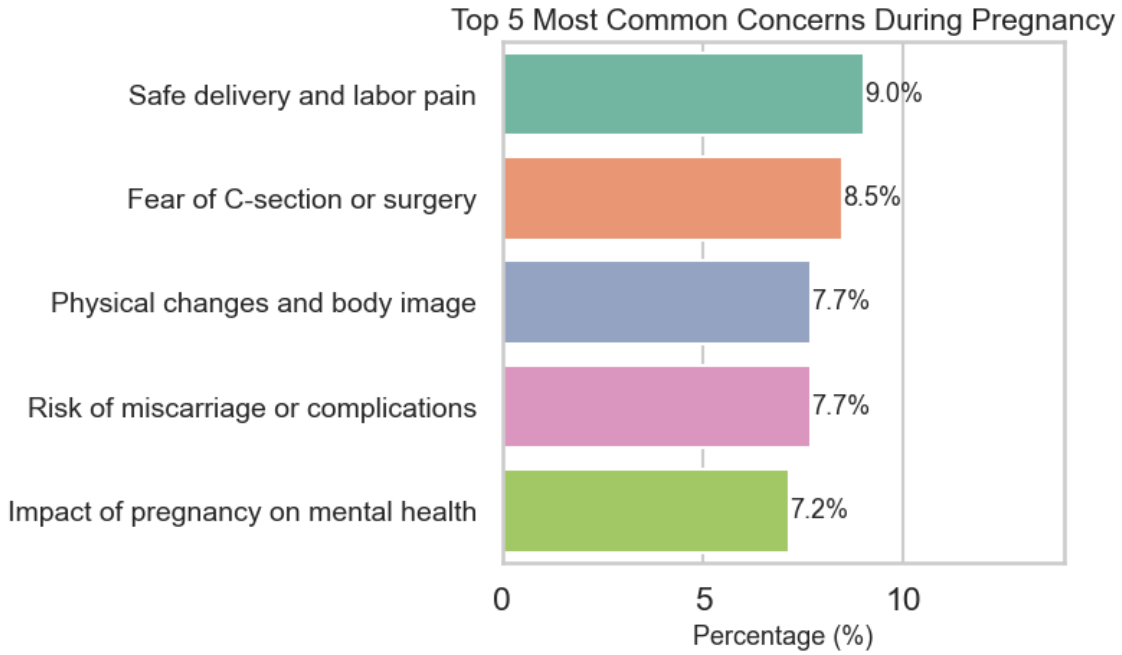


Figure 3.16: Top 5 reported concerns by participants

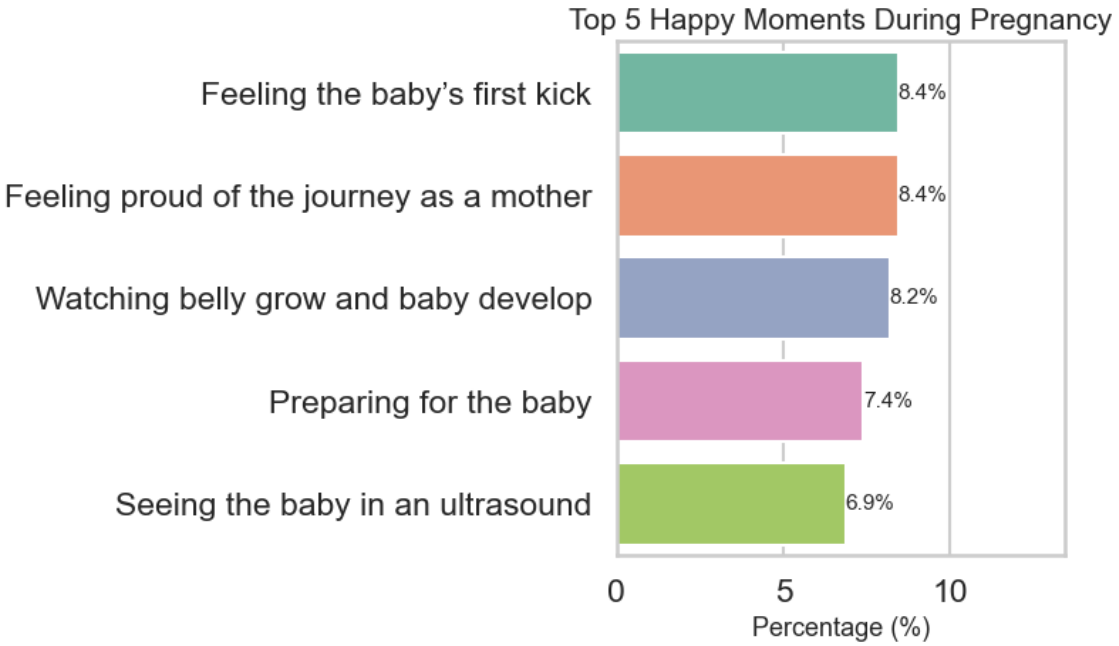


Figure 3.17: Top 5 happy aspects of pregnancy by participants

2. **Educational Attainment:** The participants exhibited a range of educational backgrounds. Thirty-three percent held a Bachelor's degree, and 25% possessed a Master's degree or higher. The remaining participants reported secondary (14%), college-level (12%), primary (10%), or no formal education (6%).

3. **Occupational Status:** Homemaker was the most prevalent occupation among the participants, accounting for 40% of the sample. Service roles were the second most common (26%), followed by business and student categories (11% each). A small number of participants were engaged in daily wage labor, lecturing, or medical professions.

4. **Geographic Residence:** The majority of participants resided in urban areas (58%). The remaining respondents were equally distributed between village (21%) and semi-urban (21%) locations.

5. **Household Size:** The typical household comprised 3 to 5 members, with a household size of 4 being the most frequently reported.

6. **Household Income:** Seventy percent of the participants reported an annual household income between 1 and 8 Lakh BDT. Twenty-five percent earned less than 1 Lakh BDT annually, and only 5% indicated an income exceeding 8 Lakh BDT.

7. **Digital Access:** A significant majority of participants (80%) reported having access to personal smartphones with internet connectivity. The remaining 15% did not own a smartphone, and 5% shared access to a device.

8. **Language Preference for Digital Platforms:** A substantial 71% of participants expressed a preference for using both Bangla and English within a

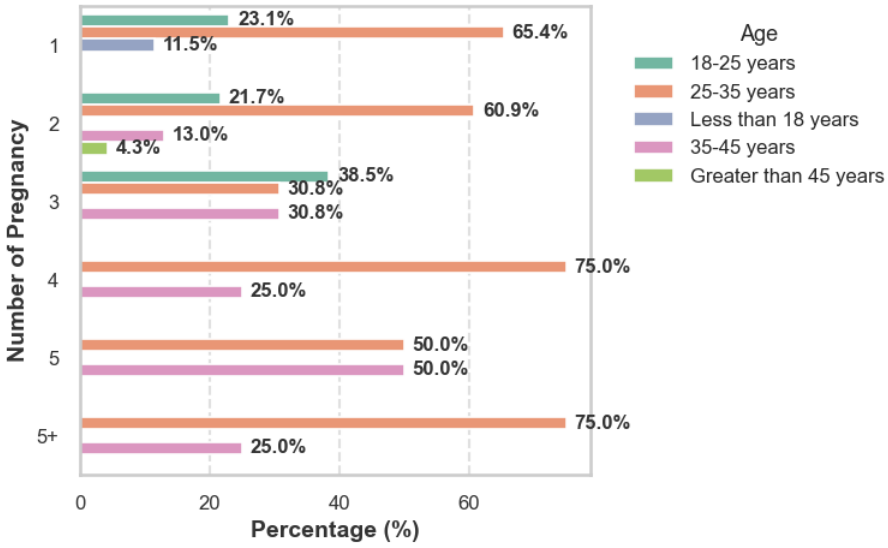


Figure 3.18: Age distribution by number of pregnancies

digital application. Twentyeight percent indicated a preference for Bangla only, while a single respondent preferred English exclusively.

9. **Survey Response Assistance:** The majority of survey responses (71%) were provided with assistance, suggesting a potential presence of literacy or digital accessibility barriers among the participant group.

10. **Top Complications:** The most frequently reported complications included fear of miscarriage (25%), labor pain (25%), and hormonal mood swings (22%). Other notable complications were anxiety (19%) and depression (17%).

11. **Top Concerns:** Participants expressed various concerns, with the most common being the risk of miscarriage (30%), physical changes (30%), and the health of the baby (25%). Other concerns included financial stability (20%) and support from family (15%).

12. **Top Happy Aspects:** The most cherished aspects of pregnancy included the baby's first kick (8.4%), maternal pride (8.4%), and watching belly grow (8.2%). Other positive experiences included preparing for the baby (7.4%) and seeing the baby in an ultrasound (6.9%).

3.2.2 Bivariate Demographic Analysis

Figure 3.18 shows the age distribution by number of pregnancies. Younger participants (18–25 years) predominantly reported 1–2 pregnancies, while women aged 25–35 displayed a broader distribution, including cases of 3–5+ pregnancies. Among 35–45 years, higher parity was more common, with up to 10% reporting 4 or more pregnancies. The categories under 18 and above 45 years both showed restricted patterns (single pregnancy and exclusively 2 pregnancies, respectively), which may reflect small sample sizes or demographic peculiarities.

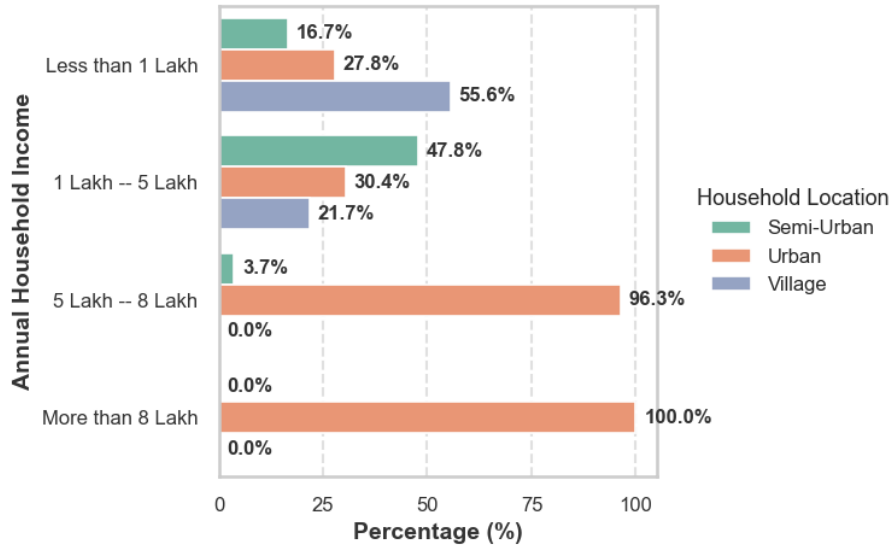


Figure 3.19: Income by household distribution of participants

Annual household income distributions across locations are presented in Figure 3.19. Semi-urban households were concentrated in the 1–5 Lakh range, with notable low-income representation. Urban households exhibited greater diversity, including higher-income brackets with 9.5% above 8 Lakh, absent in rural and semi-urban groups. Village households were predominantly low-income, with two-thirds earning less than 1 Lakh, highlighting rural economic vulnerability.

Family size by income distribution is illustrated in Figure 3.20. The lowest income group (below 1 Lakh) exhibited the largest families, with a median size of six members. As income increased, family sizes decreased, with the highest income bracket (above 8 Lakh) showing a tight distribution around four members. This inverse relationship between income and family size reflects well-established demographic trends.

The unexpectedly low proportion of personal smartphone and internet access among the greater than 8 lakh income group (Fig. 3.21 and 3.24) is best interpreted as an artifact of sample size rather than a true socioeconomic pattern. Only 4–5 respondents fell into this highest income category, compared with substantially larger numbers in other brackets. With such small representation, even one or two atypical responses (e.g., reporting reliance on computers or tablets rather than smartphones) disproportionately skew the percentages. This outlier effect underscores the importance of cautious interpretation and suggests that further studies with larger samples of affluent households are needed to validate these patterns.
Figure 3.21 demonstrates disparities in internet access across locations. Urban areas reported universal personal smartphone and internet access, while semi-urban households showed high but incomplete coverage, with some still relying on shared access or lacking smartphones. In contrast, villages revealed a stark digital divide, with the majority either without smartphones or dependent on shared access, and only a small minority with personal digital connectivity.

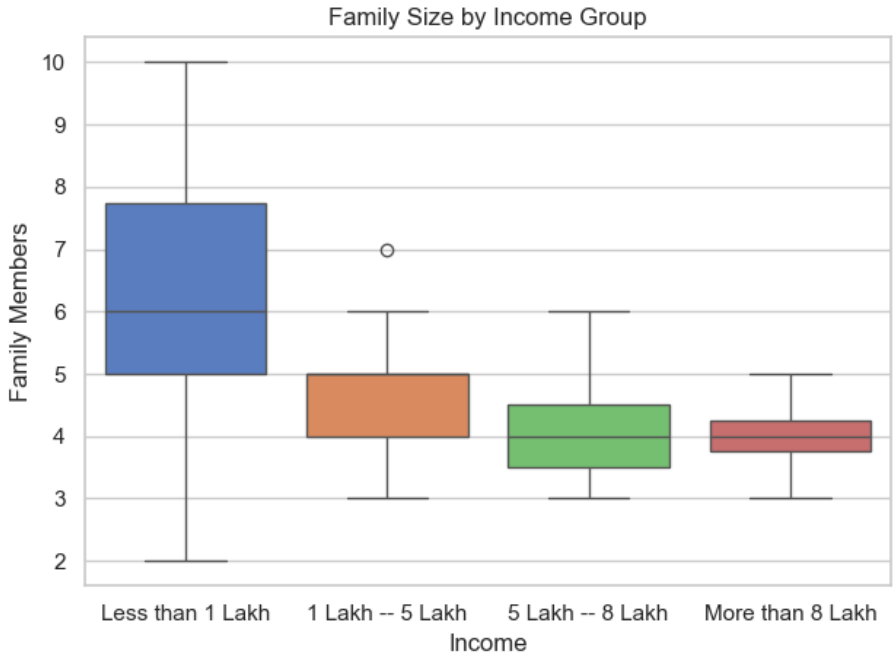


Figure 3.20: Income by family size distribution of participants

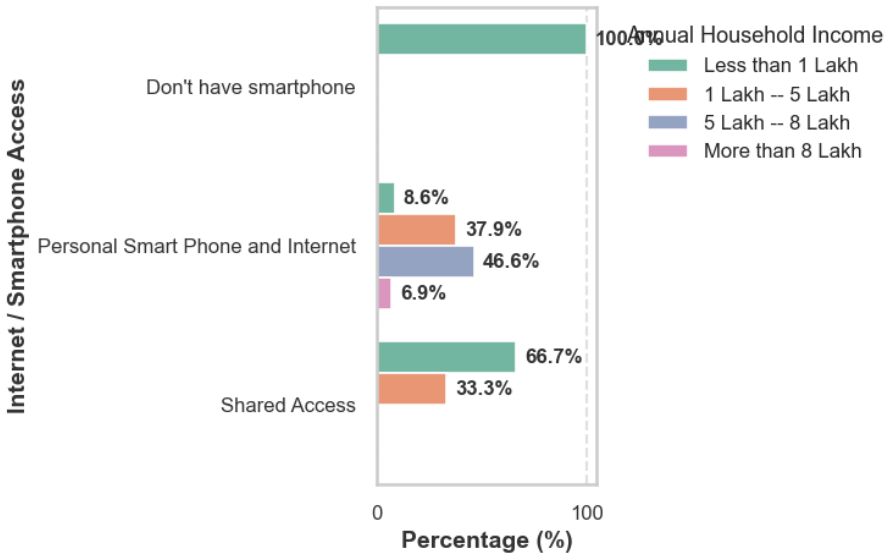


Figure 3.21: Internet access by household distribution of participants

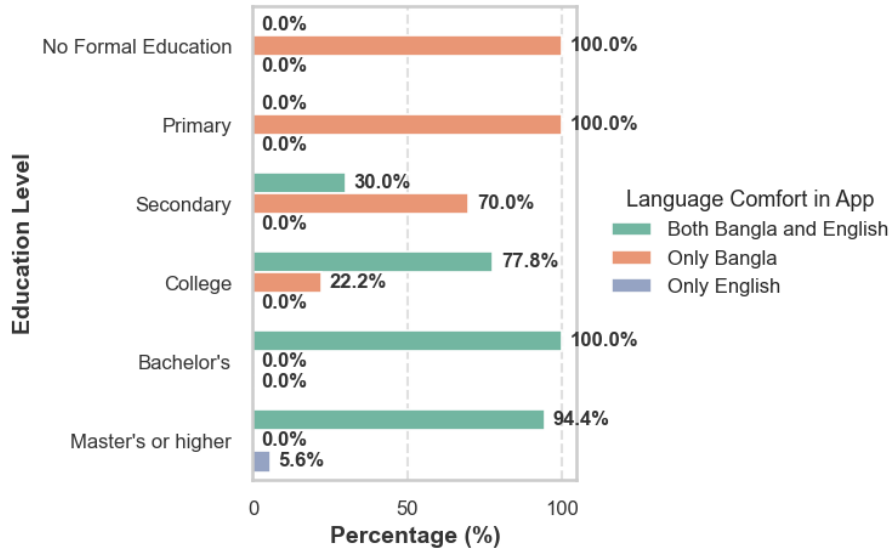


Figure 3.22: Language preference by education level among participants

Figure 3.26 explores mental health in relation to parity. Women with five or more pregnancies showed elevated levels of sadness, mood swings, and anxiety. Those with two or three pregnancies presented mixed outcomes, including both higher distress and stable conditions. Women with only one pregnancy showed limited variation but included reports of sadness and stability. These findings suggest complex associations between parity and mental health, shaped by broader socio-economic and support factors.

The presented figures offer insights into participant characteristics across education levels, internet access, and pregnancy distribution. Figure 3.22 highlights language preferences for app usage, indicating that participants with higher education levels (Bachelor's, Master's or higher) predominantly prefer both Bangla and English, at 100.0% and 94.4% respectively, while those with lower or no formal education (No Formal Education, Primary) exclusively prefer only Bangla (100.0%). Figure 3.23 details survey reporting methods, revealing a strong correlation between higher education levels and self-reporting, with 83.3% of Bachelor's degree holders and 66.7% of Master's or higher degree holders reporting themselves. Conversely, individuals with no formal education or primary education were entirely assisted (100.0%).

Fig. 3.24 examines internet and smartphone access based on the number of pregnancies, showing that a significant portion of participants across all pregnancy counts utilize personal smartphones and internet or shared access. However, those with fewer pregnancies (e.g., 1 or 2) also exhibit a notable percentage without a smartphone (9.1% and 8.7% respectively). Finally, Fig. 3.25 illustrates pregnancy distribution across household locations. Participants with fewer pregnancies (1 or 2) are more concentrated in urban and semi-urban areas, whereas those with higher parity are associated with village households. It is important to note that the percentages are calculated within each pregnancy-count group; therefore, all participants reporting exactly five pregnancies and all participants reporting more than five pregnancies were from villages, resulting in 100% values for the village category in both cases.

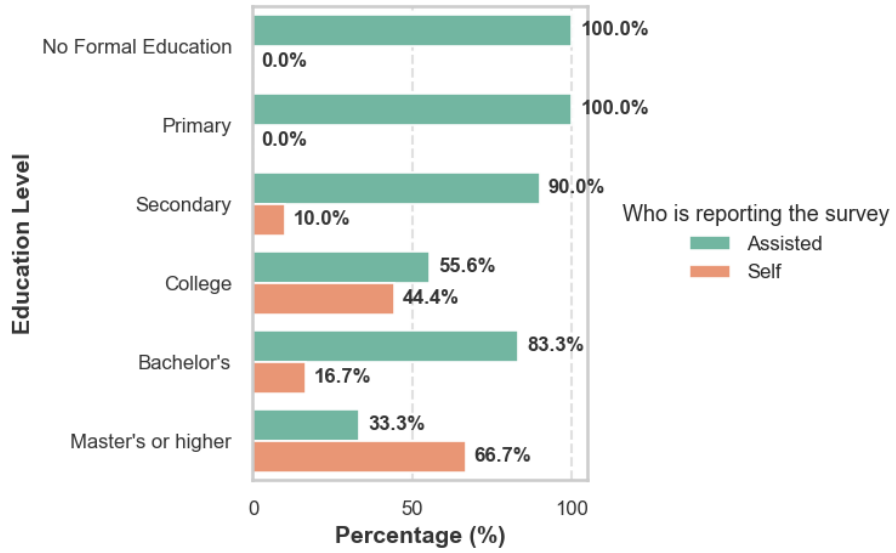


Figure 3.23: Who is reporting the survey across different Education levels

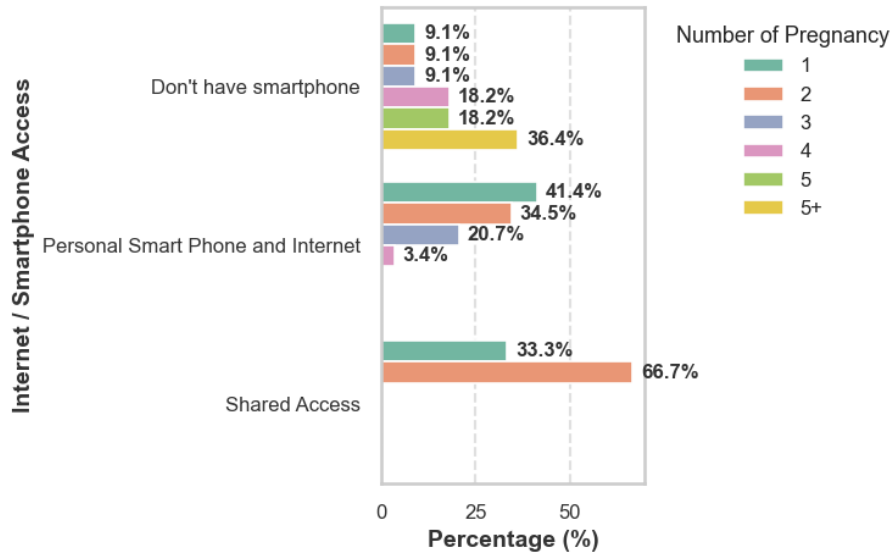


Figure 3.24: Internet access across different pregnancy distribution

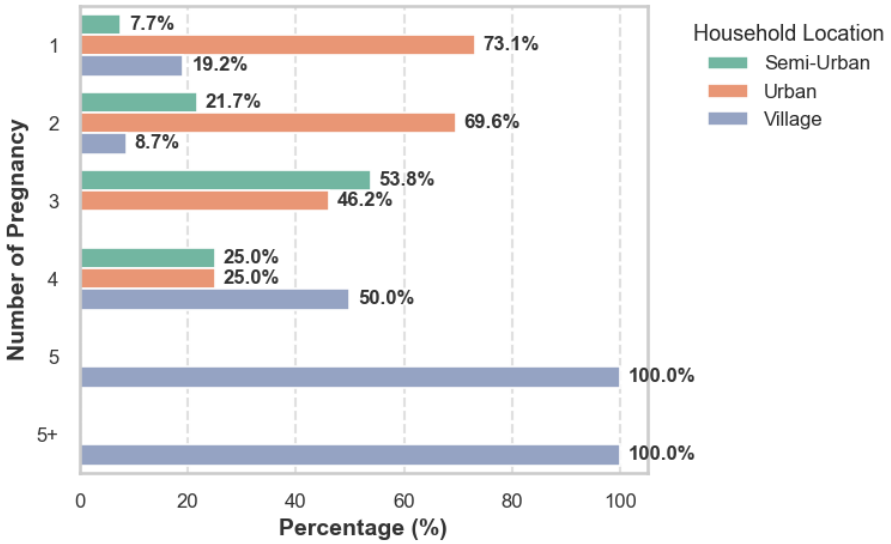


Figure 3.25: Pregnancy distribution across household location

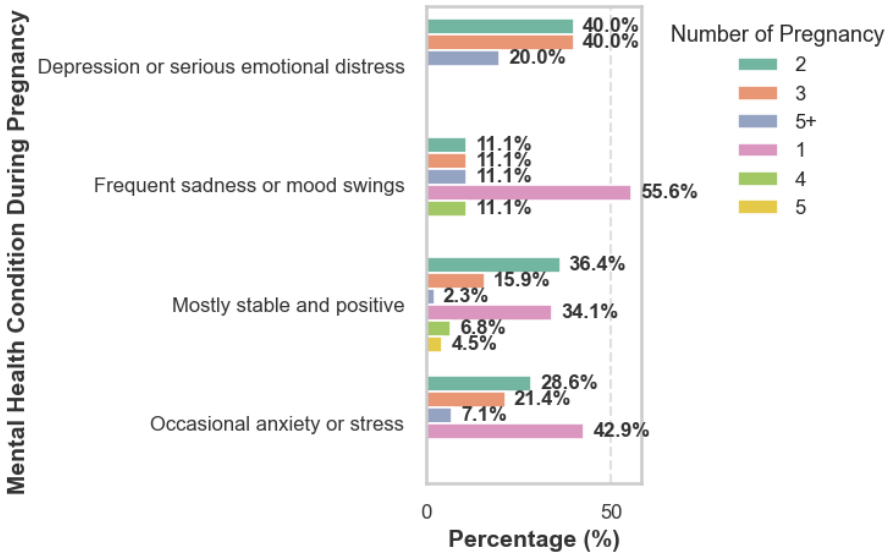


Figure 3.26: Mental health condition during different pregnancy distributions

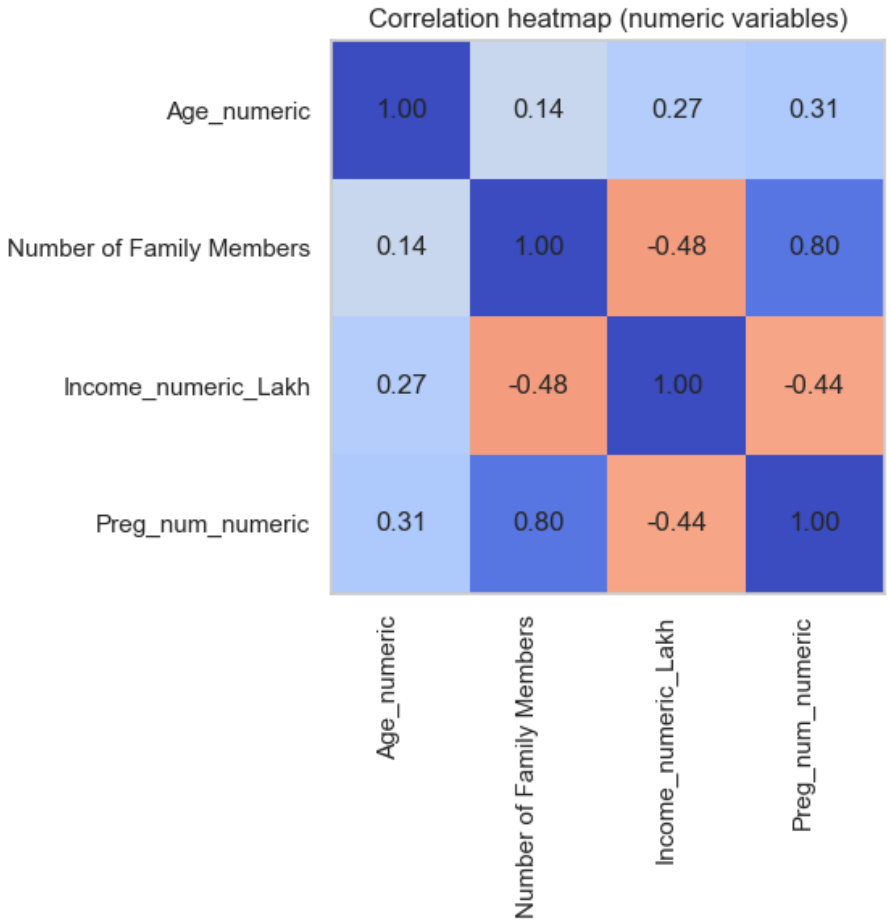


Figure 3.27: Correlation between different numerical columns

This bar chart in Fig. 3.26 visualizes the distribution of various mental health conditions during pregnancy, categorized by the number of pregnancies. "Depression or serious emotional distress" was equally reported by 40% of women with 2 and 3 pregnancies, while "Frequent sadness or mood swings" was most common among those with 1 pregnancy (55.6%). A "Mostly stable and positive" mental state was reported by 36.4% of women with 2 pregnancies and 34.1% with 1 pregnancy, whereas "Occasional anxiety or stress" was most prevalent in women with 1 pregnancy (42.9%).

3.2.3 Correlation Analysis

Figure 3.27 illustrates the correlation matrix among the key numerical variables of the survey dataset: age, number of family members, income, and number of pregnancies. The analysis reveals that the number of pregnancies is strongly and positively correlated with family size ($r = 0.80$), which aligns with the demographic reality that larger households in Bangladesh generally include multiple children. Age also shows moderate positive correlations with both income ($r = 0.27$) and number of pregnancies ($r = 0.31$). In contrast, household income exhibits moderate negative correlations with both family size ($r = -0.48$) and number of pregnan-

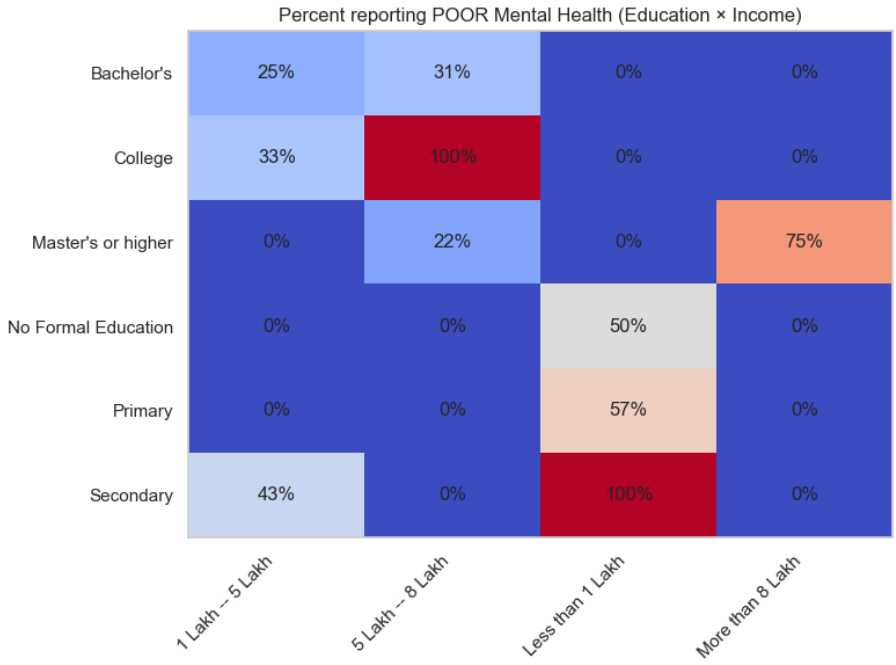


Figure 3.28: Connection of poor mental health with education and income Level

cies ($r = -0.44$). These findings indicate that lower-income families tend to have larger household sizes and more pregnancies, reflecting the structural inequities that exacerbate maternal health vulnerabilities.

Figure 3.28 explores the association between poor maternal mental health, education level, and household income. The results indicate a complex interaction between socio-economic factors. Women with college-level education but the lowest income (<1 Lakh BDT) reported the highest prevalence of poor mental health (100%). Similarly, women with only primary or no formal education also reported poorer mental health outcomes (57% and 50% respectively) when belonging to the lowest income bracket. Strikingly, even participants with Master's or higher education reported poor mental health at high rates (75%) when belonging to higher-income households, suggesting that societal and career-related pressures may also contribute to distress. Middle-income groups (1–5 Lakh BDT) generally demonstrated comparatively lower levels of reported poor mental health across most education levels.

Taken together, these two figures demonstrate that maternal mental health is embedded within broader socio-economic structures. Income, education, and family size interact in shaping maternal well-being, with both resource scarcity and psychosocial pressures contributing to distress. These insights underline the importance of designing empathetic and context-aware support systems, such as the proposed conversational AI agent, which can address not only informational gaps but also the social realities of expectant mothers in Bangladesh.

3.3 Design Implications

The findings reveal a dual tension. On the one hand, urban women demonstrate high digital readiness, with smartphones and internet access widespread. On the other, rural and low-income contexts remain constrained by literacy gaps, device-sharing,

and limited digital familiarity. Indeed, 71% of participants required assistance in completing the survey, suggesting that any technological intervention must assume minimal prior digital literacy.

Language inclusivity emerged as non-negotiable. A substantial minority of women preferred Bangla-only interactions, while many others were bilingual but still viewed Bangla as more natural and trustworthy. Language thus acts not just as a medium of communication but also as a signal of cultural authenticity and care.

Mental health concerns surfaced as particularly pressing, with a notable proportion of women reporting anxiety, mood swings, or depression. These insights suggest that a successful system must go beyond physical health information delivery also to provide empathetic and stigma-sensitive support for psychological well-being.

Taken together, these observations translate into three central design requirements for our system:

1. **Multi-language support with Bangla-first design** – ensuring accessibility for women who may not be fluent in English or who prefer their native language for trust and comfort.

2. **Simplicity and inclusivity in interface design** – accommodating users with limited digital literacy through clear, intuitive, and lightweight interaction modalities.

3. **Empathetic and trustworthy responses** – combining clinically grounded medical guidance with reassurance, acknowledgment of emotional states, and sensitivity to cultural stigma around mental health.

These requirements form the foundation for the design and implementation of our maternal health AI agent. They also underscore that the problem is not merely one of information delivery, but rather of building a relational technology that resonates with the lived realities of Bangladeshi women navigating pregnancy.

Chapter 4

Methodology

4.1 Research Design

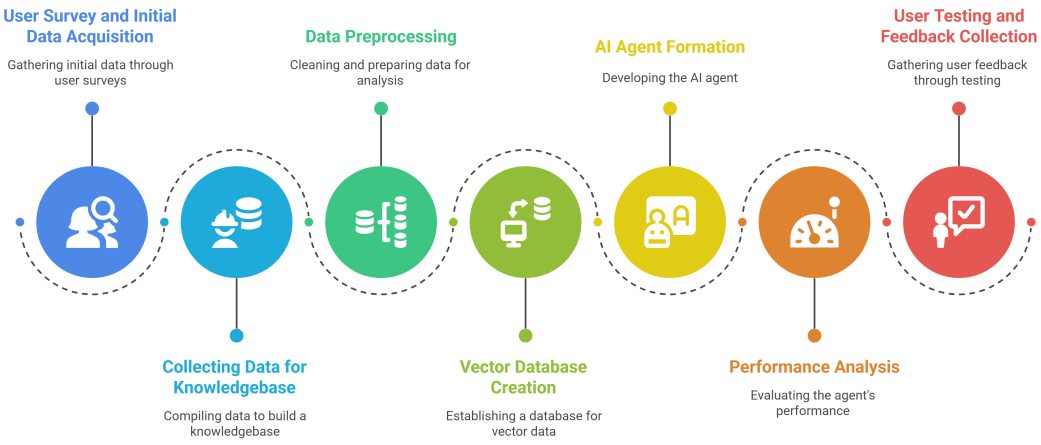


Figure 4.1: Research design steps

This research adopts a user-centered, exploratory design combining both qualitative and quantitative methods to understand the lived experiences of pregnant women and to develop a responsive AI-powered support system. The overall methodology is structured into several steps, as shown in Fig. 4.1

The study began with the design of a comprehensive questionnaire aimed at capturing the emotional, informational, and physical needs of pregnant women. The questions were crafted to explore a wide range of pregnancy-related concerns, including health complications, mental well-being, support systems, and technology accessibility. The questionnaire was distributed digitally to a diverse pool of participants to ensure variation in socioeconomic, educational, and geographical backgrounds.

Based on the insights gained from the survey responses, relevant information was sourced from a variety of public and semi-structured online resources, including medical blogs, forums, and informational websites. This content was carefully curated, cleaned, and structured to form the foundation of a domain-specific knowledgebase

for a Retrieval-Augmented Generation (RAG) application. This ensured that the AI system would be grounded in both real-world user concerns and reliable, vetted knowledge.

Following the data preparation phase, a no-login, lightweight web application was developed. The application features an intuitive front-end interface designed for accessibility and usability, particularly by users with limited digital literacy. The backend integrates a natural language AI model with the knowledgebase using an RAG architecture. This enables the system to retrieve relevant, context-aware information in response to users' natural language queries.

To complete the research cycle, the final phase involves field testing the application by introducing it to real users—pregnant women and new mothers. Their feedback on usability, relevance, emotional tone, and perceived helpfulness is being collected to evaluate the system's effectiveness and guide future iterations. This participatory and iterative approach ensures that the AI tool is not only technically functional but also contextually appropriate and empathetically aligned with users' needs.

This research design enables a holistic, real-world, grounded development process where user input informs both the system's knowledge foundation and its interface design, leading to a more human-centric and impactful AI solution.

4.2 Knowledgebase Creation

A core requirement for building a retrieval-augmented generation (RAG) system for maternal health is the availability of a reliable and context-sensitive knowledgebase. Since the agent's responses are "Garbage In, Garbage Out", we prioritized the curation of high-quality, relevant data sources [3]. Fig. 4.2 illustrates the multi-stage pipeline used in this process.

4.2.1 Challenges in Sourcing Bangla Maternal Health Data

We explored a wide variety of sources, including books, journal articles, government health portals, NGO resources, blogs, and social media groups. However, this search revealed a set of recurring problems that limited their utility for integration into a knowledgebase:

1. **Quality issues in Bangla resources:** Many online resources were direct Google Translate conversions of English blogs, often outdated and lacking cultural nuance.
2. **Lack of evidence-based information:** Few sources were clinically vetted or peer-reviewed, raising concerns about reliability.
3. **Absence of medical cross-validation:** Most resources had not been reviewed by certified medical professionals.
4. **Scarcity of localized content:** Very little information addressed the unique maternal health realities of Bangladesh, such as high prevalence of home births or nutrition practices specific to rural contexts.

5. **Misinformation risks:** Social media, forums, and blogs often contain unverified claims, folk remedies, or contradictory advice.

6. **Incomplete institutional resources:** Some government and NGO websites contained useful guidelines, but these were fragmented, outdated, or difficult to access in machine-readable form.

7. **Limited coverage of complications:** Content on nuanced topics such as perinatal mental health, postpartum recovery, or miscarriage support was particularly sparse.

8. **Lack of structured data:** We found no publicly available structured datasets or machine-readable medical corpora in Bangla.

9. **Terminology inconsistencies:** Medical terms in Bangla lacked standardization, with multiple spellings or colloquial alternatives, creating potential confusion during retrieval.

These challenges underscore the difficulty of curating high-quality Bangla medical knowledge and reinforced the importance of selecting carefully vetted sources for our system.

4.2.2 Curated Data Sources

After filtering through numerous candidates, we identified two high-quality resources that met our criteria for accuracy, relevance, and cultural applicability.

1. **Shohay Health**[1], founded by Dr. Tasnim Jara [38], is a health-tech initiative committed to making reliable medical information accessible to Bangla-speaking communities. Its content is clinically reviewed and developed through a rigorous editorial pipeline. We extracted approximately 70 articles from Shohay Health, which covered topics such as weekly pregnancy development, pregnancy-related complications, nutrition guidelines, and postpartum care.

2. **FirstCry (Bangla)**[2], is one of South Asia's largest digital parenting platforms, offering guidance throughout pregnancy and early child development. We collected over 200 articles from this source, spanning a broad range of subjects: fetal development across weeks and months, dietary and lifestyle recommendations, management of complications, and infant care from birth through toddler years.

Together, these two resources provided a diverse, clinically informed, and culturally grounded dataset suitable for powering a retrieval-based system.

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

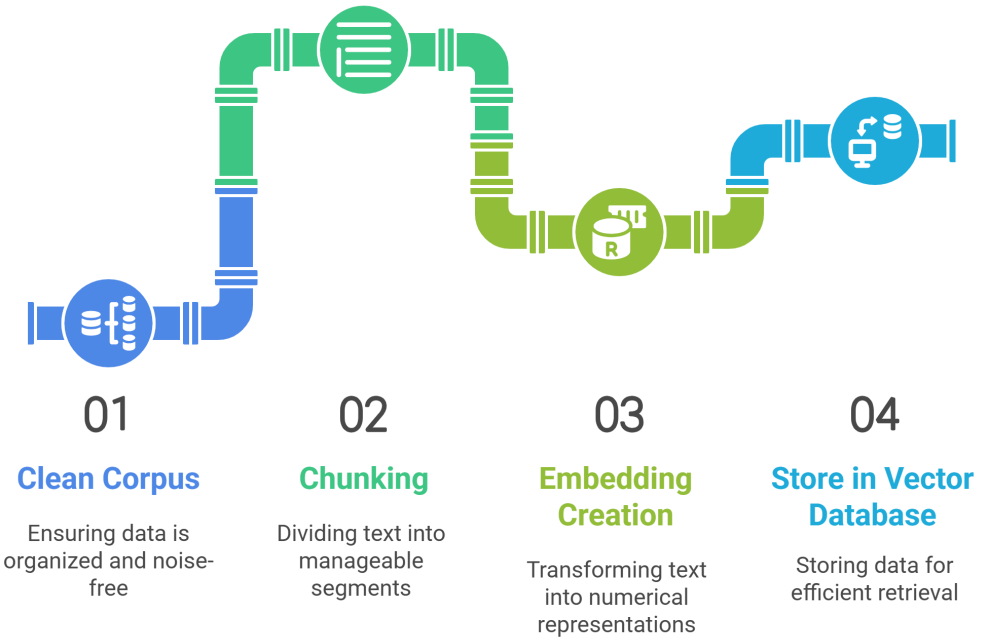


Figure 4.2: Knowledgebase Creation Pipeline

4.2.3 Preprocessing and Cleaning

The raw articles obtained from Shohay Health and FirstCry required extensive preprocessing before being integrated into the knowledgebase:

- **Exclusion of Irrelevant or Low-Quality Blogs:** Since the whole maternal section of the websites was crawled, we had to manually review each article to exclude any that were unnecessary, promotional, or tangentially related to maternal health. For example, articles such as *"Best Baby Shower Gift Ideas"* were discarded, while *"Managing Gestational Diabetes"* was retained due to clear clinical relevance.

- **Removal of Promotional or Redundant Phrases:** Inside the selected articles, there were promotional phrases (e.g., *"Buy now to get 20% off on supplements"*) and repeated disclaimers. These were systematically removed to focus on the substantive, medically relevant information.

- **Stripping Names and Identifiers:** Just as with the survey data, all personal references, usernames, and case-specific mentions were removed or generalized.

- **HTML and Script Cleanup:** Since all the sources were web-based, the raw data included HTML tags, JavaScript snippets, and metadata. A line such as `<div class="author">Written by... </div>` was stripped using regular expressions and parser libraries, retaining only the human-readable content.

- **Fact-Checking and Cross-Referencing:** Key medical claims and guidelines were fact-checked against authoritative sources (e.g., WHO, CDC, Mayo

Clinic) and a local medical expert to eliminate the risk of propagating misinformation. For example, an article statement like *"Papaya must be avoided completely in pregnancy"* was cross-verified, and adjusted to reflect the nuanced medical consensus that only unripe papaya poses risks.

Through this rigorous multi-stage cleaning process, the knowledgebase evolved into a high-quality, structured corpus tailored for safe, sensitive, and effective use in a medical-support AI application.

4.2.4 Embedding Creation

To enable fast and contextually relevant retrieval of maternal health information, the curated knowledgebase was transformed into a vectorized format using state-of-the-art embedding models. Embeddings are high-dimensional numerical representations of text that capture semantic meaning and relationships between concepts. By converting textual content into embeddings, the system can compare user queries against the knowledgebase not by surface-level keyword matching, but by underlying semantic similarity.

Model Selection

For generating embeddings, we employed the **BAAI/BGE-M3** model[1] , a multilingual embedding model trained across 100+ languages, including Bangla. Prior studies have demonstrated its strong performance in capturing nuanced semantics in low-resource languages like Bangla and Banglish [42]. Compared to older multilingual encoders such as LaBSE or mUSE, BGE-M3 offers higher retrieval accuracy and a denser vector space (1024 dimensions) suitable for both English and Bangla content. This ensured that our system could effectively handle bilingual queries.

Preprocessing and Chunking

Before embedding creation, the cleaned corpus underwent segmentation using LangChain's `RecursiveCharacterTextSplitter`. Articles were divided into overlapping chunks to balance context preservation with retrieval granularity. We set:

- `chunk_size` = 800 characters – large enough to capture coherent sections such as pregnancy week guides or complication summaries.
- `chunk_overlap` = 300 characters – ensuring that critical context (e.g., causes and solutions to a complication) was not lost at chunk boundaries.

This configuration produced semantically coherent text segments that improved the retrievability of detailed maternal health information.

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

Sensitivity of Chunking Parameters. Retrieval quality in RAG systems is often sensitive to the configuration of `chunk_size` and `chunk_overlap`. Smaller chunks (e.g., 500 characters) improved retrieval precision by narrowing the focus of semantic embeddings, but fragmented longer pregnancy week guides into disjointed pieces. Conversely, larger chunks (e.g., 1000 characters) preserved the broader context but occasionally reduced the relevance of retrieved segments by including extraneous details. The choice of overlap showed a similar trade-off: overlaps below 200 characters risked losing continuity around complication causes and remedies, while overlaps above 400 characters increased redundancy and storage costs without clear benefits. The adopted configuration (800/300), therefore, reflects a balance between coherence and efficiency for the maternal health corpus. However, systematic benchmarking across parameter ranges was beyond the scope of this thesis, and future work could evaluate retrieval accuracy, latency, and memory efficiency under different chunking strategies.

Embedding Pipeline

The embedding pipeline was implemented using the LangChain framework, which streamlined the integration of text loading, preprocessing, and embedding. Each text chunk was passed through the BGE-M3 encoder to generate a 1024-dimensional embeddings. These embeddings were stored in a **Supabase Vector Database**[74], chosen for its scalability, cloud-native deployment, and compatibility with similarity search operations. Supabase's Postgres-based vector extension allowed for efficient nearest-neighbor search using cosine similarity, ensuring low-latency retrieval at query time.

Resulting Knowledgebase

Through this process, we transformed approximately 270 maternal health articles (collected from Shohay Health and FirstCry) into a structured, vectorized knowledgebase. The resulting corpus comprised thousands of semantically indexed chunks, each linked to metadata tags (e.g., "nutrition," "week 20," "postpartum care") for domain-specific retrieval. This embedding-enabled knowledgebase serves as the backbone of our Retrieval-Augmented Generation (RAG) system, allowing user queries to be matched with contextually relevant, evidence-based maternal health guidance.

We are currently in the process of open sourcing the vector database as an *MCP server* as well as an *API*. Once ready, anyone can connect their application with the knowledgbase as an information source. The goal is to make bangla maternal health content as accessible as possible in the digital age.

4.3 Agent Design

This section details the design and implementation of the AI agent architecture for *Baby and Me*. Fig. 4.3 illustrates the overall system architecture, highlighting the interaction between the frontend, backend, large language model (LLM), conversation memory, and external tools via the Model Context Protocol (MCP).

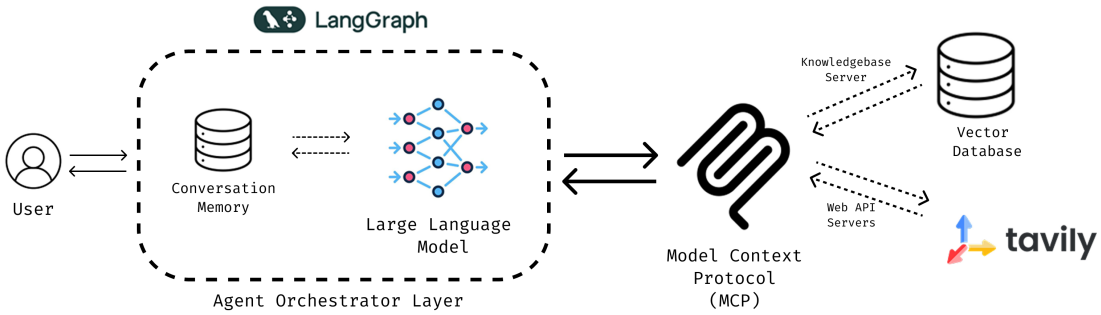


Figure 4.3: System architecture of *Baby and Me* bot, showing interaction between frontend, backend, LLM, conversation memory, and external tools via MCP.

While the current implementation demonstrates effective integration of the frontend, backend, LLM, conversation memory, and external tools via MCP, future work could include an ablation study to systematically evaluate the contribution of each component to the agent's overall performance. Such analysis would help identify the most critical modules and optimize resource allocation for real-world deployment.

4.3.1 Large Language Model Selection

Selecting the appropriate Large Language Model (LLM) was a critical decision in the design of *Baby and Me*. Since the target population primarily communicates in Bangla, with many also mixing Bangla and English (often referred to as Banglish), the chosen LLM needed to demonstrate robust multilingual competence while remaining cost-effective for deployment. After surveying a wide range of available LLMs, we shortlisted three models that balanced multilingual capability, performance, and deployment efficiency:

- **Qwen 2.5:14B** – An open-source 14 billion parameter multilingual model developed by Alibaba [57], [77]. Qwen 2.5 has demonstrated strong performance across low-resource languages, including Bangla, and offered the advantage of self-hosting to reduce long-term costs.
- **Gemini 2.5 Flash – Lite** – A lightweight variant of Google's Gemini 2.5 Flash family [61]. This model is optimized for efficiency and delivers rapid response times, making it suitable for low-resource deployment contexts (e.g., mobile-first environments in Bangladesh). Its smaller size comes at the cost of slightly less nuanced reasoning compared to larger counterparts, but its multilingual performance remains strong.
- **Gemini 2.5 Flash** – The larger, more capable sibling of Flash-Lite, this model provides enhanced reasoning, contextual depth, and richer generation capabilities. While computationally heavier, it was well suited for handling nuanced maternal health queries requiring complex reasoning or extended dialogue.

Together, these three models formed a balanced portfolio for experimentation: Qwen 2.5 offered openness and cost control, Gemini Flash-Lite provided speed and lightweight deployment, and Gemini Flash offered maximum performance for difficult queries.

4.3.2 System Prompt Design

Model selection alone was insufficient to guarantee that the chatbot would produce safe, empathetic, and culturally resonant outputs. We therefore invested in the design of a robust system prompt to guide LLM behavior. The system prompt emphasized three principles:

1. **Language Adaptability:** Respond in Bangla, English, or Banglish depending on the user's input language and preference, ensuring inclusivity.
2. **Empathetic Tone:** Provide not only factual responses but also reassurance and supportive language, reflecting the sensitive and emotional nature of maternal health conversations.
3. **Safety and Referral:** Recognize questions beyond the chatbot's scope (e.g., medical emergencies) and redirect users to professional medical care instead of attempting speculative answers.

The prompt was iteratively refined through pilot interactions to ensure that outputs were clinically appropriate, empathetic, and free from hallucinations. This process aligned with our design requirements of empathy, cultural sensitivity, and trustworthiness.

4.3.3 Model Context Protocol (MCP) Integration

To enhance the LLM's capabilities beyond standalone generation, we integrated the Model Context Protocol (MCP) [1] into our backend architecture. MCP provides a standardized interface for LLMs to interact with external tools and data sources, enabling more dynamic and context-aware responses [65]. For our case, we implemented two MCP servers: one for a search tool and another for a knowledgebase tool.

- **Search Tool:** This MCP server interfaces with a web search API to fetch real-time information. When the LLM encounters queries about recent medical guidelines, news, or statistics not covered in the static knowledgebase, it can invoke the search tool to retrieve up-to-date data.
- **Knowledgebase Tool:** This MCP server connects to our Supabase vector database containing the maternal health knowledgebase. The LLM can call this tool to perform semantic searches against the embedded corpus, retrieving relevant articles or sections to inform its responses.

The power of MCP lies in its ability to allow the LLM to reason about when and how to utilize these tools effectively, enhancing its overall performance and responsiveness. It also enables modularity, allowing us to add or update tools without retraining or modifying the LLM itself. MCP will allow users to integrate platforms they are already comfortable with, such as their calendars, notes, reminders, and health tracking apps.

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

4.3.4 Conversation Memory

To provide a coherent and contextually aware conversational experience, we implemented a session-based memory module. This module retains information within a single interaction session, allowing the AI agent to reference recent user inputs and responses. By maintaining short-term context, the chatbot can deliver more relevant replies and improve user engagement. This feature is particularly important in maternal health conversations, where continuity during a session supports clearer and more reassuring communication.

4.4 Application Design and Implementation

The *Baby and Me* system is built as a lightweight web application with two primary layers: a modern frontend that provides a user-friendly interface and a backend that orchestrates AI reasoning, data management, and tool integration.

4.4.1 Frontend Layer

The frontend of *Baby and Me* was designed to provide a simple, responsive, and culturally inclusive user experience. A modern web stack was adopted to achieve this: React 19 forms the foundation for modular components, TypeScript ensures type safety and maintainability, Tailwind CSS enables a consistent lovely design across devices, and Vite supports rapid development with fast hot reloads. The interface allows users to type in Bangla, Banglish, or English and immediately see responses streamed back in real time. Visual indicators, such as loading animations and tool-call states, were added to make the system's reasoning process transparent and trustworthy. The design goal was to ensure that even women with limited digital literacy could easily interact with the chatbot on both mobile and desktop devices.

4.4.2 Backend Layer

The backend serves as the orchestration hub of the system, handling user requests, AI reasoning, data persistence, and external tool integration. FastAPI provides the web framework, enabling asynchronous handling of multiple sessions and streaming responses. LangChain manages interaction with the large language model, and coordinates chains, agents, and memory. PostgreSQL is used to persist conversation history in structured form, ensuring continuity across sessions. The Model Context Protocol (MCP) acts as a bridge between the LLM and external tools, allowing requests such as web searches or database queries to be executed seamlessly. Together, these components ensure that every user query flows smoothly from input to reasoning to output, with responses that are not only factually grounded but also empathetic and contextually appropriate.
This layered and orchestrated design ensures that the chatbot is not only technically robust but also responsive.

4.5 Evaluation Criteria

We evaluated *Baby and Me* through three complementary lenses: (i) a technical evaluation of large language models (LLMs) across multiple performance criteria, (ii) an online survey study (N=20+) with pregnant and postpartum women who used the chatbot for several days, capturing broad feedback on usability, perceived utility, and overall experience, and, (iii) in-person semi-structured interviews (N=3) with selected participants from the survey to provide illustrative depth and contextual insights. Our evaluation aims to understand not only the system's technical performance, but also how women perceive the chatbot's usability, trustworthiness, and influence on health-related awareness and behaviors.

4.5.1 Technical Evaluation

Initially **LLM as Judge** method was used [47]. It involved using the LLM itself to evaluate the system's outputs against a set of predefined metrics. We Prepared a evaluation dataset of frequently asked pregnancy and maternal questions and answers from trusted websites like UCLA Health[1], NewYork-Presbyterian[2] and Shohay Health[3].
The evaluation criteria were designed to encompass multiple dimensions of the AI agent's functionality, including accuracy, relevance, and user experience with respect to latency. To assess these factors, we experimented with two different retrieved context window sizes (K), set to 5 and 10, which are widely regarded in the literature as optimal values [46], [59]. K referes to the number of matching chunks of data retrieved by the LLM. The evaluation process involved generating responses from the RAG system and then having the LLM assess these responses based on the established criteria. The criteria were as follows:

1. **Contextual Accuracy**: This metric evaluates the degree to which the retrieved documents and the generated response align with the user's query and the broader conversation context. It assesses whether the system correctly identifies and utilizes relevant information, avoiding misinterpretations or off-topic content.

2. **Chain of Thought Contextual Accuracy**: Building upon contextual accuracy, this criterion specifically examines the coherence and logical flow of information within the generated response, particularly in multi-turn conversations or complex queries. It evaluates if the RAG system maintains consistent understanding and progression of thought, ensuring that each part of the response is contextually sound and contributes to a comprehensive answer [79].

3. **Conciseness**: This criterion measures the brevity and directness of the generated responses. We assess whether the system provides necessary information without excessive verbosity, extraneous details, or repetitive phrasing, aiming for clarity and efficiency in communication.

[1]https://www.uclahealth.org/medical-services/birthplace/pregnancy-newborn-health/prenatal-education/your-pregnancy/pregnancy-frequently-asked-questions
[2]https://www.nyp.org/womens/pregnancy-and-birth/faq
[3]https://shohay.health/pregnancy/labor-delivery/delivery-faqs

4. **Relevance**: This metric quantifies the pertinence of the retrieved information and the generated response to the user's query. It evaluates whether the system delivers information that directly addresses the user's need, distinguishing between genuinely useful content and merely related but ultimately irrelevant passages.

5. **Hallucinations**: This critical criterion identifies instances where the RAG system generates information that is factually incorrect, unsubstantiated by the source documents, or fabricated [52]. It quantifies the occurrence of such erroneous outputs, which can severely undermine user trust and the system's utility.

6. **Response Latency (P50 and P99)**: These metrics measure the time taken for the RAG system to generate a response. P50 (50th percentile) represents the median response time, indicating that 50% of queries are answered within this duration. P99 (99th percentile) represents the worst-case response time, indicating that 99% of queries are answered within this duration. These metrics are crucial for evaluating the system's real-time usability and responsiveness under varying loads.

4.5.2 Feedback Survey Study

Baby and Me bot was deployed and sent among pregnant and postpartum women with a feedback survey to provide illustrative depth and contextual insights. After using the chatbot, 20+ participants gave their feedback. This feedback was collected through a structured questionnaire (table A.2) designed to capture their experiences and perceptions of the system. To validate the effectiveness and usability of the chatbot, we formulated three primary questions:

1. Process: Can women use the chatbot easily? Any friction?

2. Results: How do they interpret and trust the answers?

3. Influence: Does it change how they think/act about health?

4.5.3 Semi-structured Interviews

A short, semi-structured interviews were conducted with three women who had used *Baby and Me* for three days. These were designed not for saturation but to provide illustrative depth and to contextualize survey feedback. The protocol was originally in Bangla and has been translated into English. The protocol is given in Appendix B.

Chapter 5

Interaction with *Baby and Me* Bot

We designed the *Baby and Me* bot as more than a question–answer engine: it was intended to act as a maternal health companion that fits into the everyday rhythms of pregnancy in Bangladesh. The design foregrounded immediacy, trust, and empathy, aiming to lower barriers to access while creating a conversational space where medical and emotional concerns could co-exist. Figure 5.1 shows an end user (new mother) engaging with the system.

Figure 5.1: A mother trying the *Baby and Me* bot

5.1 Seamless Input and Transparent Interaction

A core design principle of the *Baby and Me* system was to minimize entry barriers for women who may already feel vulnerable or hesitant when seeking maternal health support. To this end, the application deliberately requires no sign-up, no password creation, and no disclosure of personal details before use. Women can begin interacting with the chatbot the moment they arrive on the platform[1]. This "zero-friction" approach reflects a recognition that identity disclosure can feel risky in sensitive health contexts, particularly where issues such as miscarriage anxiety, mental health struggles, or relationship concerns may carry significant stigma. By eliminating these barriers, the system signals that it is not a gatekeeper but a companion—ready to listen without conditions.

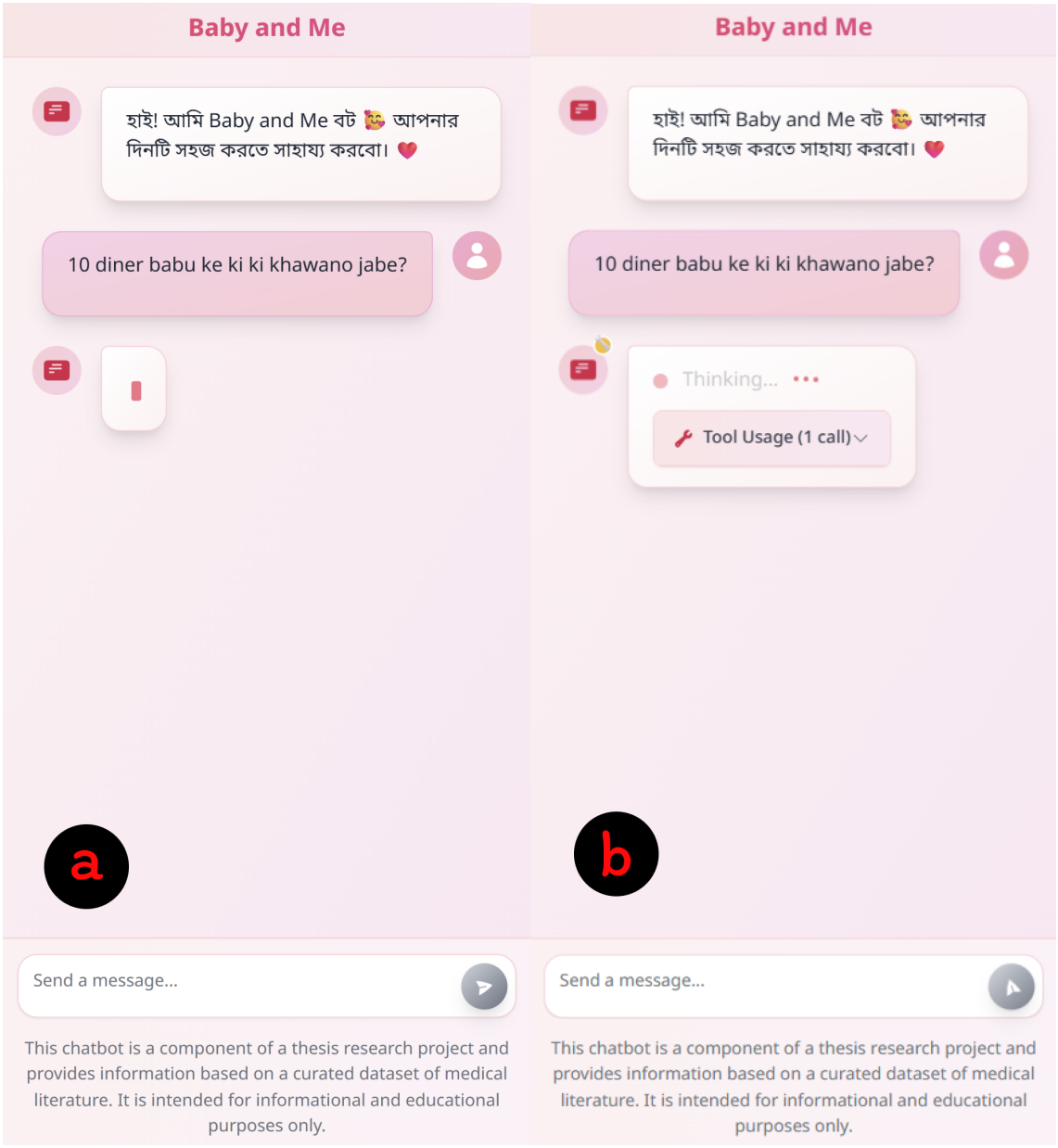


Figure 5.2: Input flow in *Baby and Me* Bot interface

[1] *Baby and Me* Web Application

The chat interface itself is intentionally familiar. It mimics the look and feel of everyday messaging applications that women in Bangladesh already use in their daily lives, such as WhatsApp, Messenger, or Imo. From an HCI perspective, leveraging a well-known design metaphor reduces the cognitive load required to understand how to operate the system. Instead of learning a new navigation style or filling in forms, users simply type a question into a text box. This design choice is not trivial: it directly addresses digital literacy barriers, which were strongly present in our survey results, where 71% of respondents reported needing assistance with even basic digital tasks.

Figure 5.2(a) illustrates the simple input box, which is always visible and positioned at the bottom of the interface to mirror commercial chat platforms. Once the user submits a query, the interface transitions into what we describe as the "thinking" state (Figure 5.2(b)). Here, the system makes its internal activity visible. Rather than silently processing in the background, the application shows animated signals of tool invocation, document retrieval, and content generation. These intermediate states operate on two levels. First, they are technical feedback mechanisms, preventing the user from wondering whether the system has frozen or ignored their query. Second, they function as affective design elements: small moments of reassurance that the agent is actively working on the user's behalf. In our user feedback, several participants described this visible responsiveness as giving them confidence that the system was "listening," a quality often missing in static health information portals. Finally, input design carries implications for privacy and safety. Because no login is required, conversations are ephemeral beyond the active session. This temporary memory lowers risks of unwanted surveillance or data breaches while still enabling the agent to sustain coherent multi-turn interactions. The balance is delicate: enough continuity to foster a sense of dialogue, but not so much persistence that women fear being permanently recorded. This balance is particularly salient in contexts where discussing pregnancy complications or mental health may be perceived as socially sensitive, even within the household.
Taken together, the input design of *Baby and Me* represents more than a technical gateway into the system. It is an intentional negotiation of accessibility, familiarity, transparency, and care. In making the act of asking a question feel safe, simple, and respected, the design transforms a utilitarian text box into a conduit for trust and empathy.

5.2 Output Interface

The output interface of *Baby and Me* was designed not only as a display mechanism for generated text, but as an affective and relational layer of the interaction. Whereas input establishes the ease of access and transparency of process, output must strike a careful balance between factual accuracy, emotional support, and cultural resonance. Responses are intentionally crafted to avoid dense medical jargon, favoring accessible, Bangla-first phrasing that allows women with different educational backgrounds to feel included. This linguistic choice was critical, as our survey revealed that nearly one-third of participants were comfortable only in Bangla, and many associated English medical terminology with exclusion or intimidation.
At the same time, the system recognizes that maternal health questions are rarely

purely informational. They are often entangled with anxiety, stigma, or the emotional weight of uncertainty. For this reason, each message is framed to provide reassurance alongside factual guidance. This dual design goal transforms output from mere information delivery into an empathetic exchange.

5.2.1 Generated Responses as Relational Messages

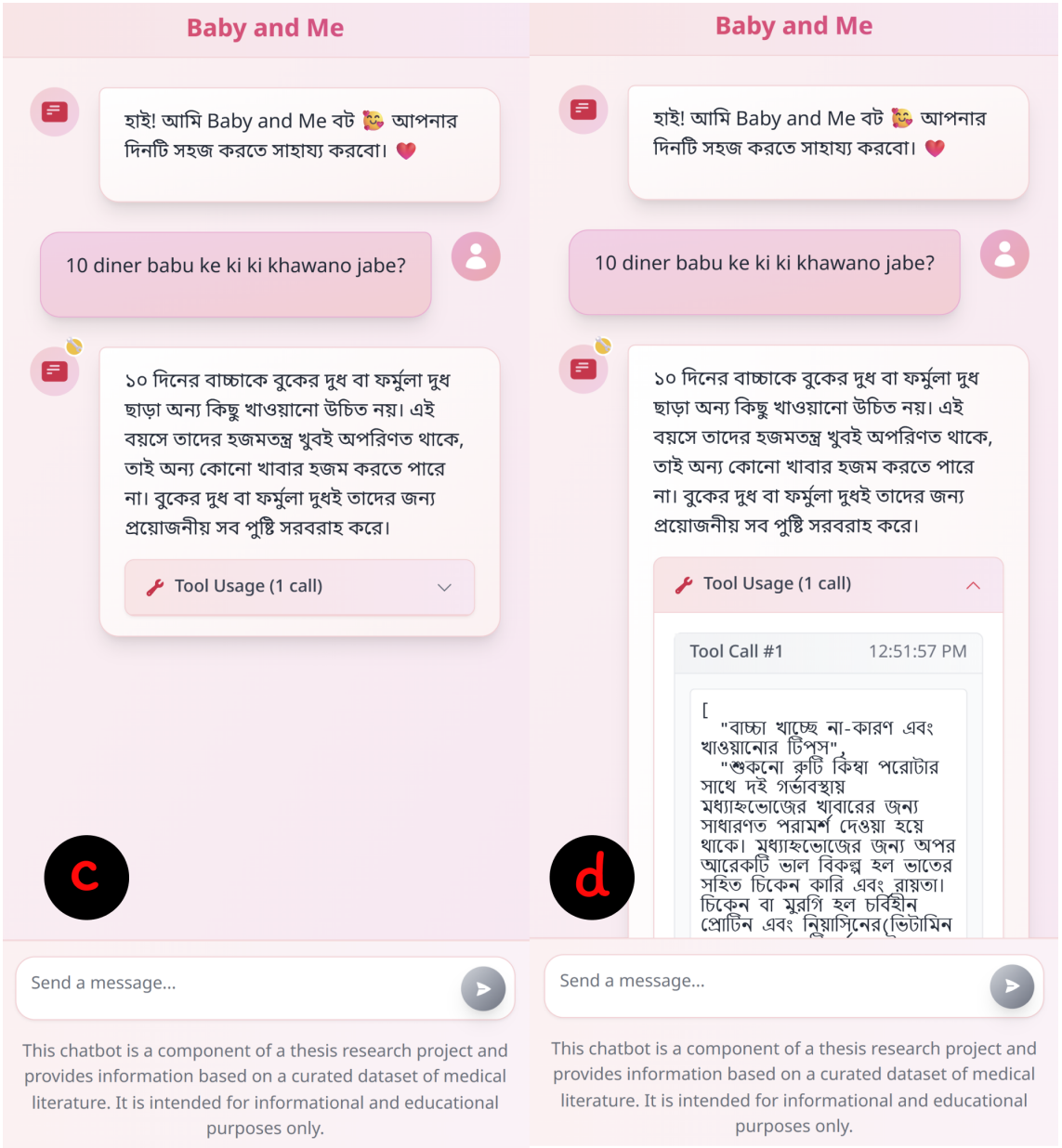


Figure 5.3: Output flow in *Baby and Me* Bot interface

Figure 5.3 shows the flow of generated responses within the interface. Upon receiving an input, the system generates a concise yet empathetic message (panel c), framed in simple Bangla or Banglish depending on user preference. By foregrounding accessibility, the bot avoids alienating women who may otherwise disengage if they encounter unfamiliar terminology. The design of this textual layer echoes prior HCI

work on conversational agents as *relational technologies*, where tone, phrasing, and reassurance contribute as much to trust as factual accuracy.
The idle state (panel d) illustrates another deliberate choice: the interface retains visibility of retrieved documents and tool outputs even after a response has been delivered. Rather than "hiding the machinery," the system surfaces its sources in an unobtrusive but accessible manner. From a technical standpoint, this aids auditability. From a social standpoint, however, it functions as an interaction ritual of accountability. Women reported that seeing the origin of the information made them feel respected, as though the system was willing to "show its work" rather than asserting authority unilaterally.

5.2.2 Tabular Responses for Structured Guidance

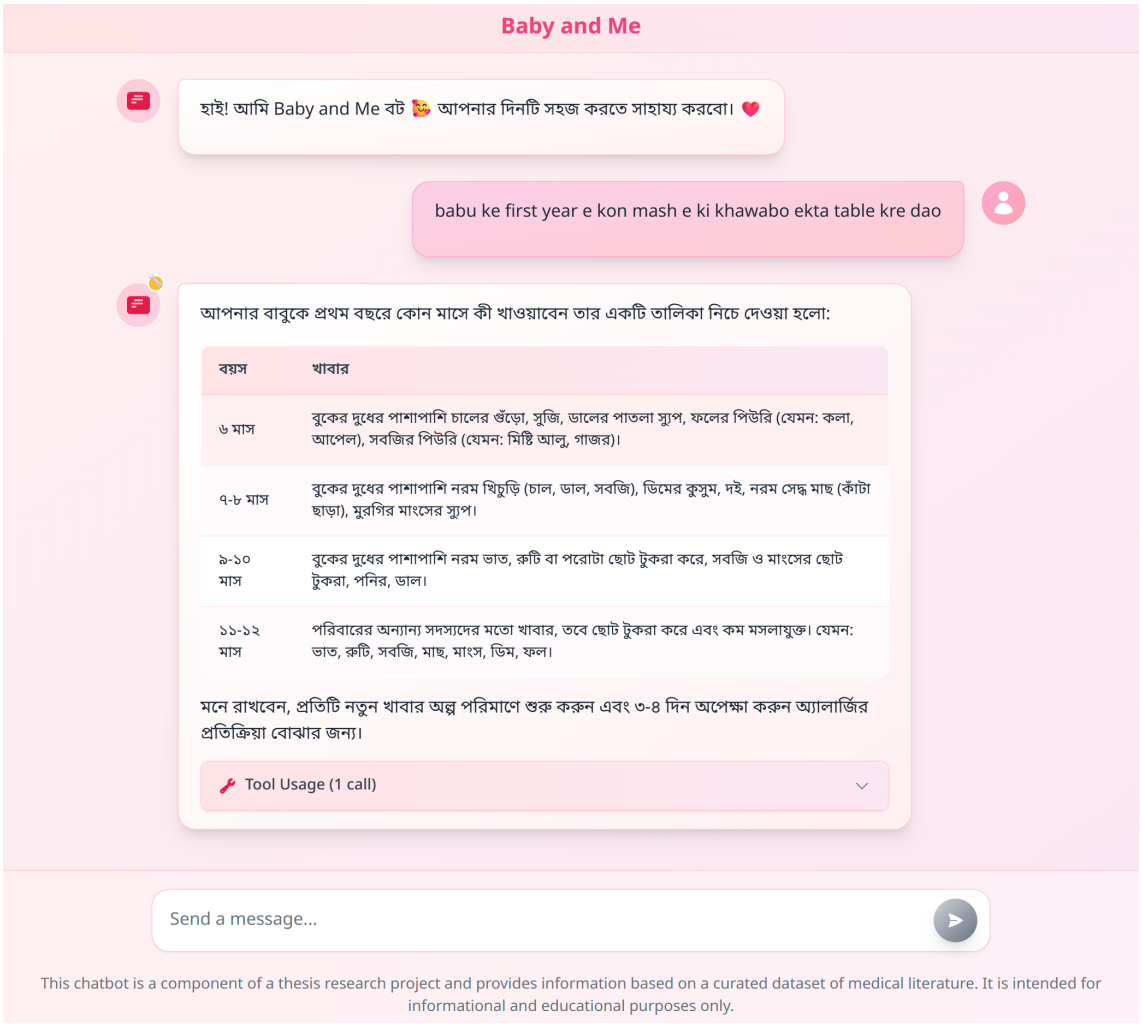


Figure 5.4: How a table-based answer is shown in the UI

Certain maternal health topics, such as nutritional comparisons, trimester milestones, or warning signs, are better conveyed in structured rather than narrative form. Figure 5.4 demonstrates how the system renders table-based answers directly in the chat stream. This approach serves two design purposes.
First, it enhances clarity: women can scan the rows and columns for quick reference rather than parsing long paragraphs. Second, it reinforces credibility: tables evoke

associations with clinical or educational material, lending weight to the information without overwhelming the user. Feedback from participants highlighted that this feature was particularly appreciated when comparing safe vs. unsafe foods, or identifying symptom categories, as it simplified decision-making for everyday concerns.

5.2.3 Continuity Through Session Memory

Figure 5.5 illustrates how the system preserves context throughout a session. Although no login or persistent user account is required, the agent maintains memory of prior turns within the current conversation. This modest form of continuity proved powerful in shaping women's perceptions of the system. Several participants described the bot as "remembering them," a subtle cue that transformed interaction from a transactional query-response exchange into an ongoing dialogue.
Session memory also supports more accurate responses. By referencing prior messages, the bot can tailor follow-up answers, avoid repetition, and sustain conversational coherence. From a relational standpoint, continuity fosters trust and empathy: users feel acknowledged as individuals rather than treated as isolated inputs. At the same time, the choice to restrict memory to the session balances this benefit against privacy risks, ensuring that sensitive disclosures are not stored beyond the immediate interaction.

5.2.4 Transparency as Emotional and Epistemic Resource

Across all output modes—narrative text, tabular displays, and session memory—the system treats transparency not merely as a technical safeguard but as an affective resource. By exposing retrieved sources, showing tool calls, and maintaining visible continuity, the bot positions itself as accountable and trustworthy. For women navigating health anxieties often compounded by stigma, this visible honesty becomes an emotional anchor. In this way, technical mechanisms such as retrieval pipelines and memory management are mobilized as design resources for empathy, trust, and inclusion, ensuring that the *output interface* is not only functional but also relational.

Figure 5.5: Flow of messaging in a session; context of previous messages is preserved

Chapter 6

Technical Evaluation

This section analyzes the performance of the Retrieval-Augmented Generation (RAG) application across three different language models: Qwen 2.5:14B, Gemini 2.5 Flash - Lite, and Gemini 2.5 Flash. The evaluation considers various metrics including Conciseness (C), Contextual Accuracy (CoT Acc.), Chain of Thought Contextual Accuracy (CoT Acc.), Hallucination (H), Relevance (R), and latency at the 50th (p50) and 99th (p99) percentiles, with comparisons made for two different values of K (5 and 10), representing the number of retrieved documents.

Table 6.1: Performance metrics by models at different retrieved context window (K)

K	Model	CoT Acc.	Contextual Acc	R	C	H
5	Qwen 2.5:14B	0.733	0.466	0.63	0.466	0.51
	Gemini 2.5 flash - lite	0.85	0.75	0.8	0.75	0.18
	Gemini 2.5 flash	0.933	0.9	0.91	0.933	0.11
10	Qwen 2.5:14B	0.8	0.57	0.65	0.52	0.43
	Gemini 2.5 flash - lite	0.85	0.84	0.83	0.78	0.18
	Gemini 2.5 flash	0.95	0.94	0.89	0.9	0.08

Table 6.2: P50 and P99 Latency by models at different retrieved context window (K)

K	Model	p50 latency (s)	p99 latency (s)
5	Qwen 2.5:14B	15.58	27.12
	Gemini 2.5 flash - lite	5.1	8.5
	Gemini 2.5 flash	7.46	10.57
10	Qwen 2.5:14B	20.1	40.4
	Gemini 2.5 flash - lite	7.42	12.82
	Gemini 2.5 flash	13.56	20.7

6.1 Conciseness

The Conciseness metric, as shown in Fig. 6.1, generally improves with more advanced models. Qwen 2.5:14B exhibits the lowest conciseness (0.47 for K=5, 0.52 for K=10). Gemini 2.5 Flash - Lite shows a significant improvement (0.75 for K=5, 0.78 for K=10), while Gemini 2.5 Flash achieves the highest scores (0.93 for K=5, 0.90 for K=10). Interestingly, for Gemini 2.5 Flash, K=5 slightly outperforms K=10, suggesting that beyond a certain point, increasing the number of retrieved documents might not proportionally enhance conciseness, or could even introduce extraneous information.

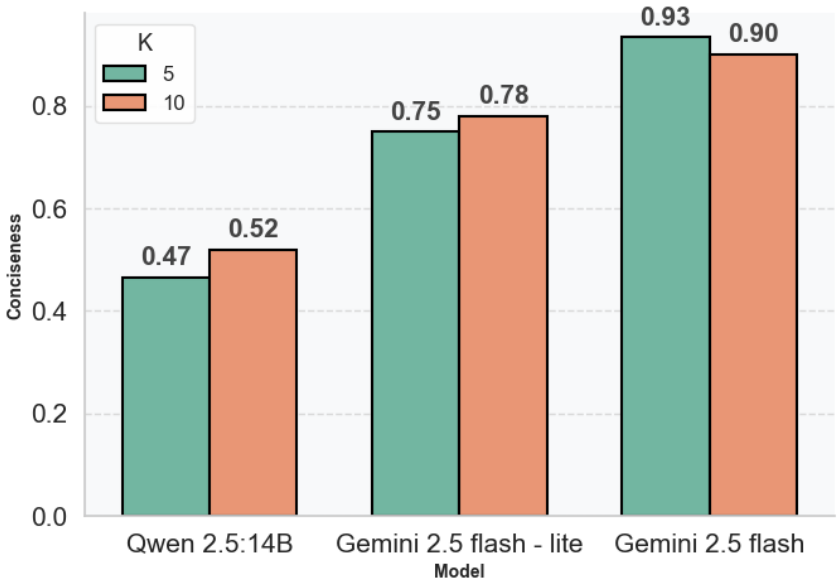


Figure 6.1: Conciseness

6.2 Contextual Accuracy

Similar to conciseness, Contextual Accuracy (fig. 6.2) demonstrates a clear upward trend with model sophistication. Qwen 2.5:14B again lags with scores of 0.47 (K=5) and 0.57 (K=10). Gemini 2.5 Flash - Lite shows good performance (0.75 for K=5, 0.84 for K=10), and Gemini 2.5 Flash achieves the highest contextual accuracy (0.90 for K=5, 0.94 for K=10). Across all models, increasing K from 5 to 10 consistently leads to an improvement in contextual accuracy, indicating that more retrieved context generally aids in generating accurate responses.

6.3 Chain of Thought Contextual Accuracy

Chain-of-Thought (CoT) Contextual Accuracy in fig. 6.3 also follows the trend of improving with model capability. Qwen 2.5:14B shows 0.73 (K=5) and 0.80 (K=10). Both Gemini 2.5 Flash - Lite and Gemini 2.5 Flash demonstrate high and very similar CoT contextual accuracy scores. Gemini 2.5 Flash - Lite achieves 0.85 for both K=5 and K=10, while Gemini 2.5 Flash reaches 0.93 (K=5) and 0.95

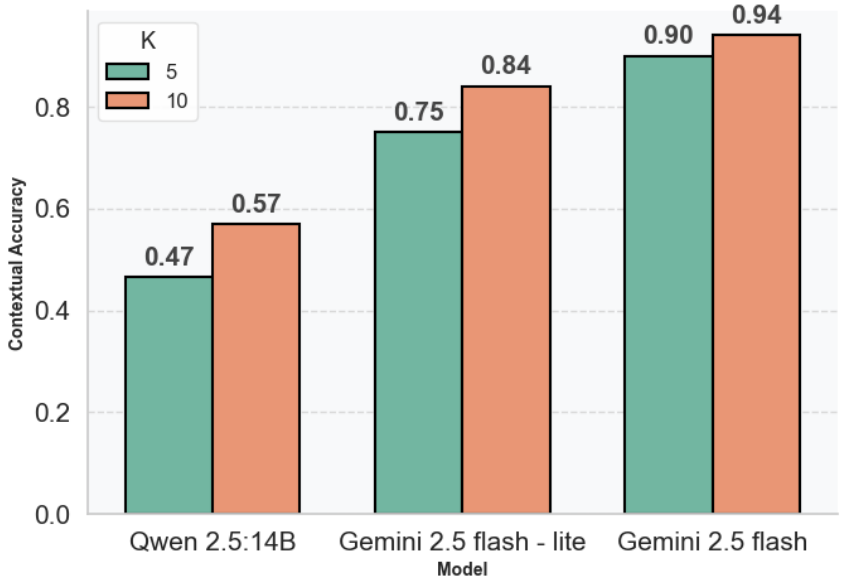


Figure 6.2: Contextual Accuracy

(K=10). This suggests that for CoT-based reasoning, the more advanced Gemini models are highly effective, and the impact of K might be less pronounced once a sufficient amount of relevant context is provided.

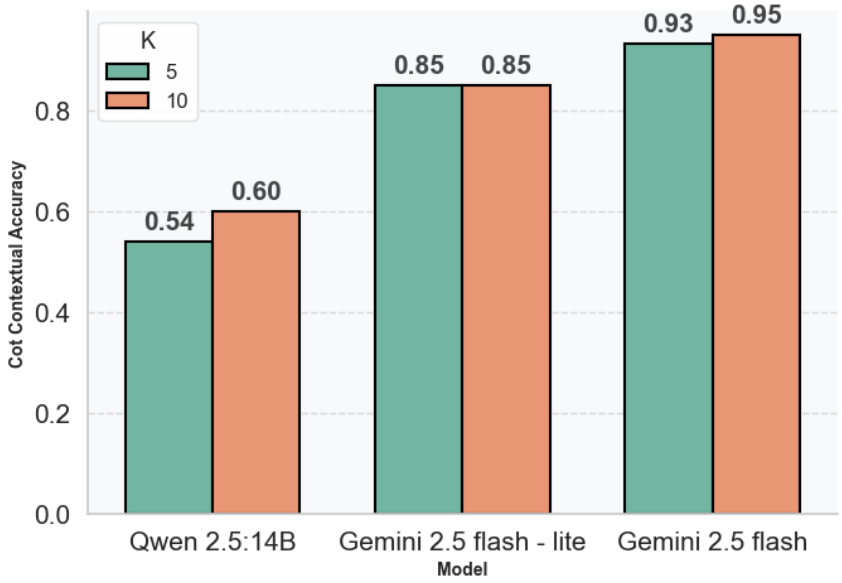


Figure 6.3: Chain of Thought Contextual Accuracy

6.4 Hallucination

The Hallucination metric (fig. 6.4) shows a reverse trend, with lower values indicating better performance (less hallucination). Qwen 2.5:14B exhibits the highest hallucination rates (0.51 for K=5, 0.43 for K=10). Both Gemini 2.5 Flash - Lite and Gemini 2.5 Flash demonstrate significantly lower hallucination, with Gemini 2.5

Flash - Lite at 0.18 for both K values, and Gemini 2.5 Flash at 0.11 (K=5) and 0.08 (K=10). Note that these values are likely representing the inverse of hallucination (i.e., accuracy or non-hallucination rate), as higher values are better. Assuming this interpretation, increasing K seems to increase the non-hallucination slightly rate for Gemini 2.5 Flash, while for the 'Lite' version, K has no discernible effect.

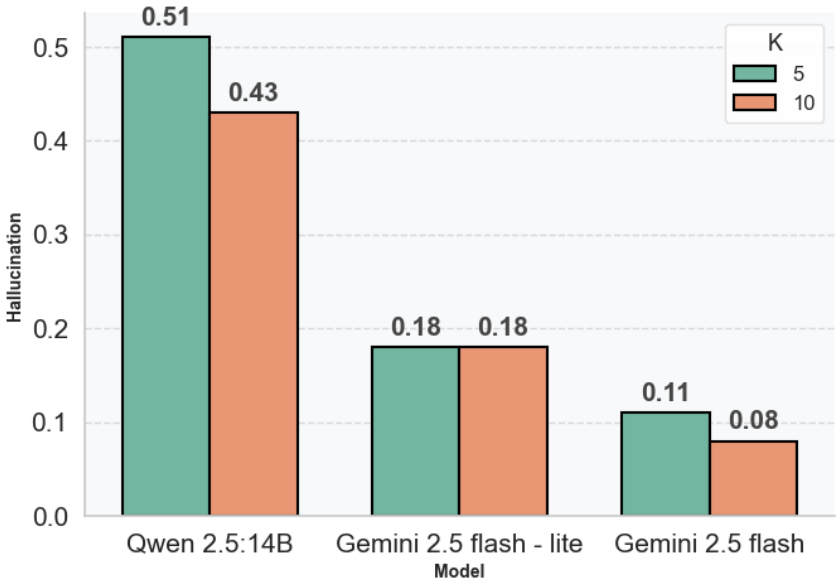


Figure 6.4: Hallucination

6.5 Relevance

Relevance indicates how closely the generated output aligns with the query. All models show an improvement in relevance as K increases from 5 to 10, though the effect is modest for the Gemini models 6.5. Qwen 2.5:14B has relevance scores of 0.63 (K=5) and 0.65 (K=10). Gemini 2.5 Flash - Lite performs better at 0.80 (K=5) and 0.83 (K=10). Gemini 2.5 Flash achieves the highest relevance, 0.91 (K=5), though it slightly decreases to 0.89 for K=10, similar to the conciseness observation, suggesting potential saturation or noise introduction with excessive context.

6.6 Latency

Latency measurements at fig. 6.6 and fig. 6.7 reveal a trade-off between model complexity, K value, and response time. Qwen 2.5:14B consistently exhibits the highest latencies, both at p50 (15.58s for K=5, 20.10s for K=10) and p99 (27.12s for K=5, 40.40s for K=10). Notably, increasing K from 5 to 10 significantly increases latency for Qwen 2.5:14B, which is expected due to the processing of more retrieved documents.

The Gemini models demonstrate substantially lower latencies. Gemini 2.5 Flash - Lite is the fastest, with p50 latencies of 5.10s (K=5) and 7.42s (K=10), and p99 latencies of 8.50s (K=5) and 12.82s (K=10). Gemini 2.5 Flash, while performing

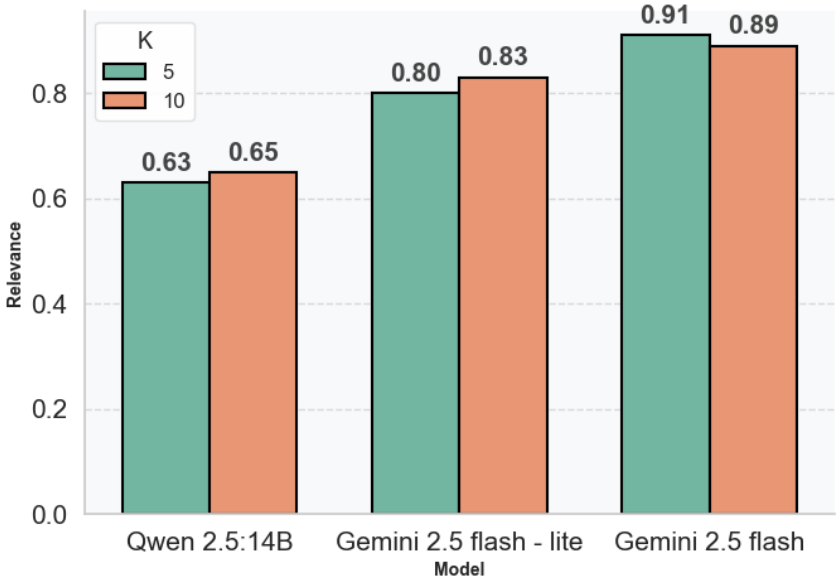


Figure 6.5: Relevance

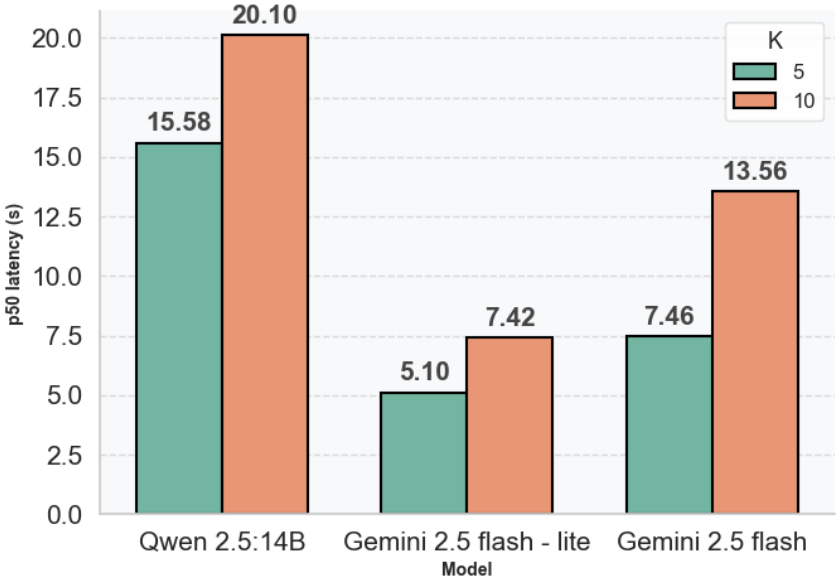


Figure 6.6: P50 Latency

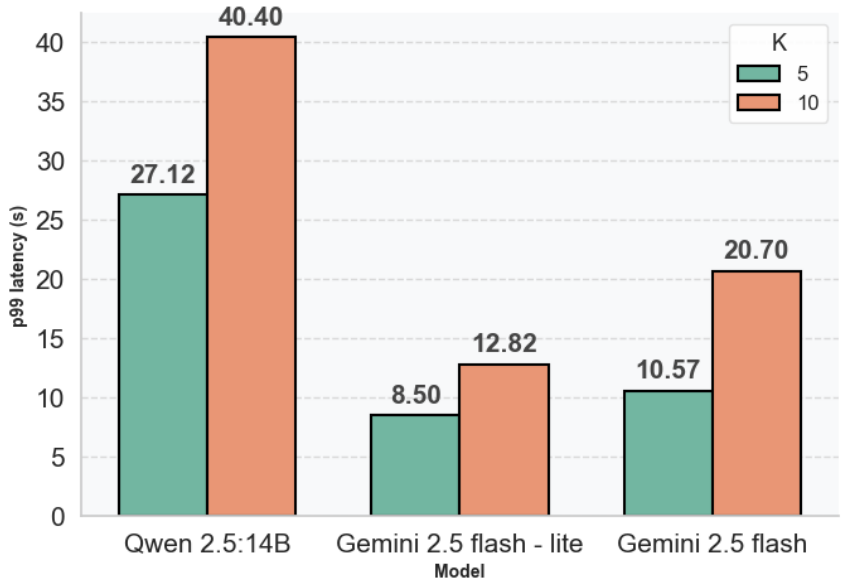


Figure 6.7: P99 Latency

very well on accuracy metrics, is slightly slower than its 'Lite' counterpart, with p50 latencies of 7.46s (K=5) and 13.56s (K=10), and p99 latencies of 10.57s (K=5) and 20.70s (K=10). For all models, increasing K predictably increases latency, reflecting the computational overhead of processing more information.

The performance metrics across all models indicate that the system meets acceptable standards for real-world deployment. While Qwen 2.5:14B shows moderate accuracy and higher latency, the Gemini 2.5 Flash models achieve high conciseness, contextual and chain-of-thought accuracy, relevance, and very low hallucination rates, with response times suitable for interactive use. The results also show that increasing the number of retrieved documents (K) improves contextual accuracy, though for conciseness and relevance, there is a point of diminishing returns. Overall, the Gemini models strike an effective balance between accuracy, reliability, and speed, making the system appropriate for maternal health guidance in Bangla.

Moreover, the impact of K (number of retrieved documents) is nuanced. For most accuracy-related metrics (Contextual Accuracy, CoT Contextual Accuracy, non-Hallucination), increasing K from 5 to 10 generally improves performance, suggesting that more context is beneficial. However, for Conciseness and Relevance in the Gemini 2.5 Flash model, K=5 occasionally outperforms K=10, indicating a potential saturation point where additional context may introduce redundancy or minor inaccuracies, impacting conciseness or the direct relevance perception. Unsurprisingly, increasing K consistently leads to higher latencies across all models, reflecting the increased computational load.

Chapter 7

End User Feedback Evaluation

7.1 Online Survey

This section presents a quantitative analysis of user feedback regarding the application's performance and user experience, collected from 24 participants. The feedback was solicited across five key dimensions: ease of use, clarity and understandability of answers, perceived caring/empathy in responses, alignment of answers with user concerns, and trustworthiness of the provided information. Participants' ages were also recorded.
Baby and Me bot was deployed and sent among pregnant and postpartum women with a feedback survey to provide illustrative depth and contextual insights. To validate the effectiveness and usability of the chatbot, we formulated three primary research questions (RQs):

1. RQ1 (Process): Can women use the chatbot easily? Any friction?
2. RQ2 (Results): How do they interpret and trust the answers?
3. RQ3 (Influence): Does it change how they think/act about health?

Participants: The app was sent to the women who initially took part in the survey. Later, the app was opened to a wider audience for a week. The final user group consisted of 24 individuals with an average age of 29.79 years (SD = 5.51). The age range was broad, from 20 to 47 years, indicating a diverse user base within the young to middle adult demographic.

7.1.1 Likert Scale Question Answer Analysis

The user group consisted of 24 participants with an average age of 29.79 years (SD = 5.51, range 20–47). This distribution reflects a diverse base spanning younger to middle-aged adults. Figure 7.1 presents the Likert-scale evaluation of the chatbot across five core user experience dimensions (1–5 scale, 5 = most positive).

1. **Ease of Use:** Usability emerged as the strongest aspect of the system (M = 4.79, SD = 0.41). Three-quarters of participants rated the application a perfect 5, with no ratings below 4. This indicates that the interface design and navigation affordances were intuitive and well-aligned with user expectations.

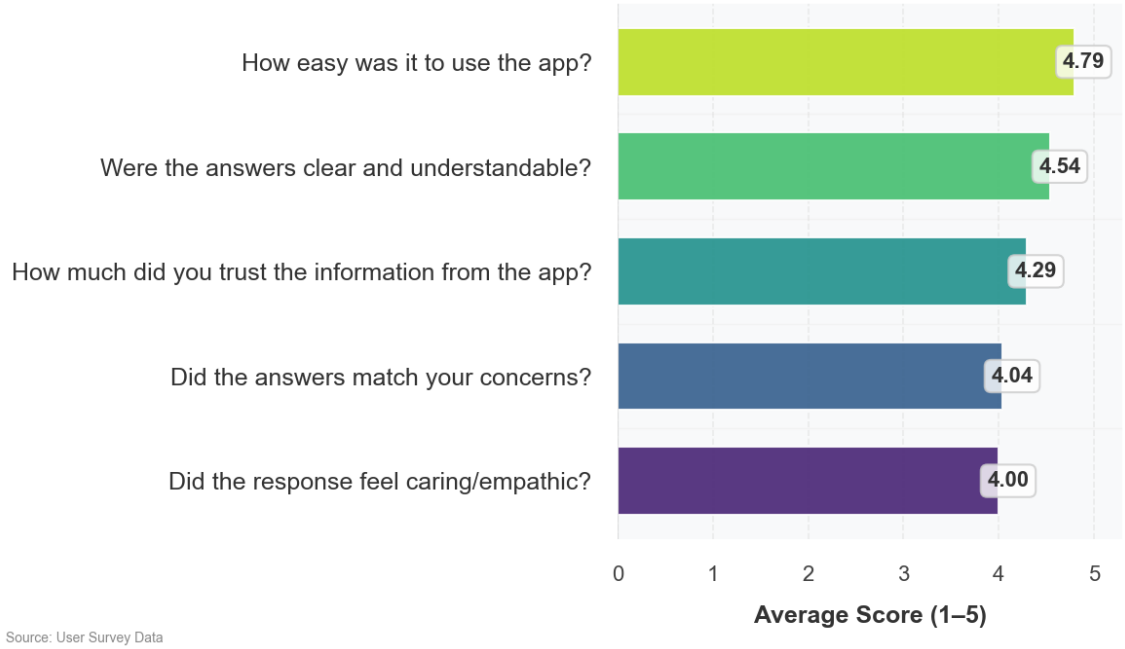


Figure 7.1: Average user feedback scores in different categories

2. **Clarity and Understandability of Answers:** Clarity scored consistently high (M = 4.54, SD = 0.59). While most participants rated clarity positively (4 or 5), the lowest rating was 3, signaling that some responses could be refined for better interpretability. Overall, this suggests that the information was communicated effectively, though occasional ambiguities remained.

3. **Perceived Caring/Empathy in Responses:** Empathy received a moderately positive score (M = 4.00, SD = 0.72). This dimension exhibited the widest variability, ranging from 2 to 5. While many participants felt the responses conveyed care, others reported a lack of warmth or emotional resonance. This highlights the challenge of consistently embedding empathy in automated responses.

4. **Alignment of Answers with User Concerns:** Relevance of the answers achieved an average of 4.04 (SD = 0.69). Most users rated alignment positively, though a minimum rating of 2 revealed that some users felt their specific concerns were not sufficiently addressed. The median remained at 4, showing that the majority found the answers generally helpful.

5. **Trustworthiness of Information:** Trust in the chatbot's responses scored relatively high (M = 4.29, SD = 0.91). While most users rated trustworthiness positively (4 or 5), variability was substantial, with ratings as low as 2. This suggests that although the system was perceived as reliable overall, credibility cues must be further reinforced.

This shows the system achieved strong usability outcomes. Ease of use and answers with clarity are two strong points of the system, though a small number of users noted occasional ambiguities. For example, P2 and P7 said that *"it took frustratingly long to get answers sometimes."* Whereas P3 mentioned, *"The app misunderstood*

my questions and gave incorrect answers." More variable results emerged in perceived empathy and alignment with user concerns. While most users found the responses supportive and relevant, a minority rated them lower, suggesting uneven emotional resonance and contextual coverage. P17 noted, *"The app sometimes repeats answers or gives very generic responses. More personalized advice would feel a lot nicer."* Trustworthiness of information scored moderately high (M = 4.29) but with the widest spread, indicating that while most users found the system reliable, a few expressed doubts.

These results highlight two key design priorities: (1) strengthening the perceived emotional tone and personalization of responses, and (2) reinforcing credibility cues to build trust in sensitive health contexts. Also, users were more likely to recommend the system to others than to use it as a full replacement for Google. This suggests strong positive sentiment, but also some hesitation in displacing an entrenched tool.

Notably, uncertainty was higher for "Use Instead of Google," indicating that trust and reliability remain barriers to complete adoption. The survey also showed a strong preference for personalization by storing user data, with a smaller but notable segment prioritizing data privacy. P10 wrote *"If I could save my questions and answers or revisit previous conversations, that would be very convenient."* Some participants wanted more than a chat platform, but a sharing aspect as well. P8 added *"Sometimes I just want to feel like I'm not alone. If the app shared real experiences from other mothers, it would be emotionally comforting."*

Overall, the survey results indicate that ease of use scored highest, empathy and personalization lagged for *Baby and Me*. This gap points to a deeper design challenge: technical clarity does not guarantee emotional resonance. Women may interpret a technically accurate but emotionally flat response as unhelpful, reminding us that trust in health AI is relational, not purely informational

7.1.2 Trichotomous Questions Analysis

Participants also responded to trichotomous questions (Yes/No/Not Sure) on recommendation and replacement preferences (see Figure 7.2). Results indicate that 70.8% of participants were willing to recommend the system to others, while 62.5% were open to using it instead of Google. Negative responses were identical across both questions (12.5%), suggesting a consistent skeptical minority. Uncertainty was higher regarding replacement (25.0%) than recommendation (16.7%), implying that while participants valued the system enough to endorse it, hesitations remained in viewing it as a primary information source. It should be noted that the comparison question was framed against Google, as it remains the dominant health information tool among Bangladeshi mothers. Although emerging AI assistants such as ChatGPT are increasingly relevant, our survey did not explicitly measure substitution against them. Future research could extend this comparison to evaluate whether mothers would prefer the proposed system over general-purpose AI assistants, offering a more current benchmark of adoption.

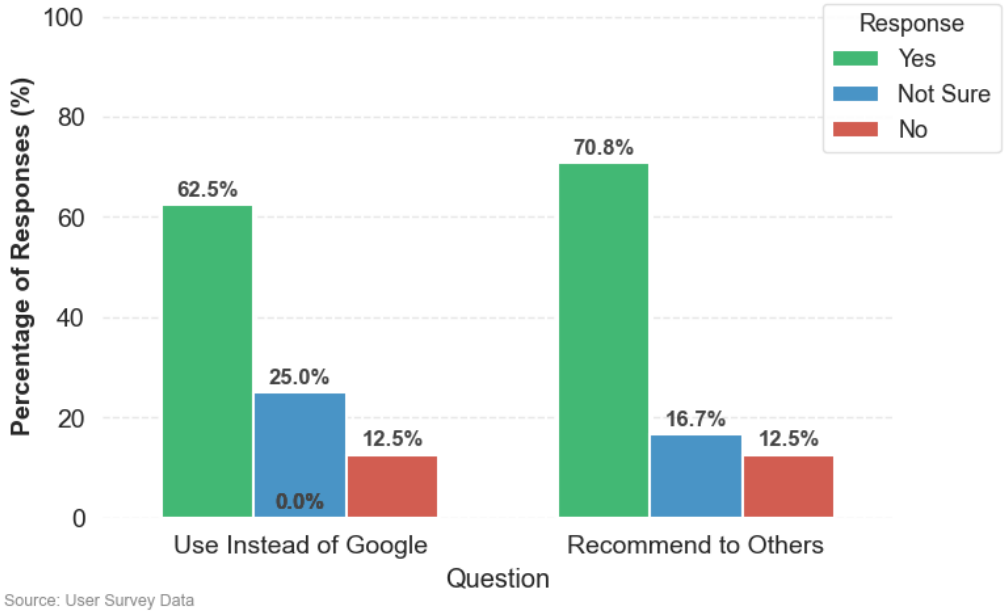


Figure 7.2: User Responses to Trichotomous Questions

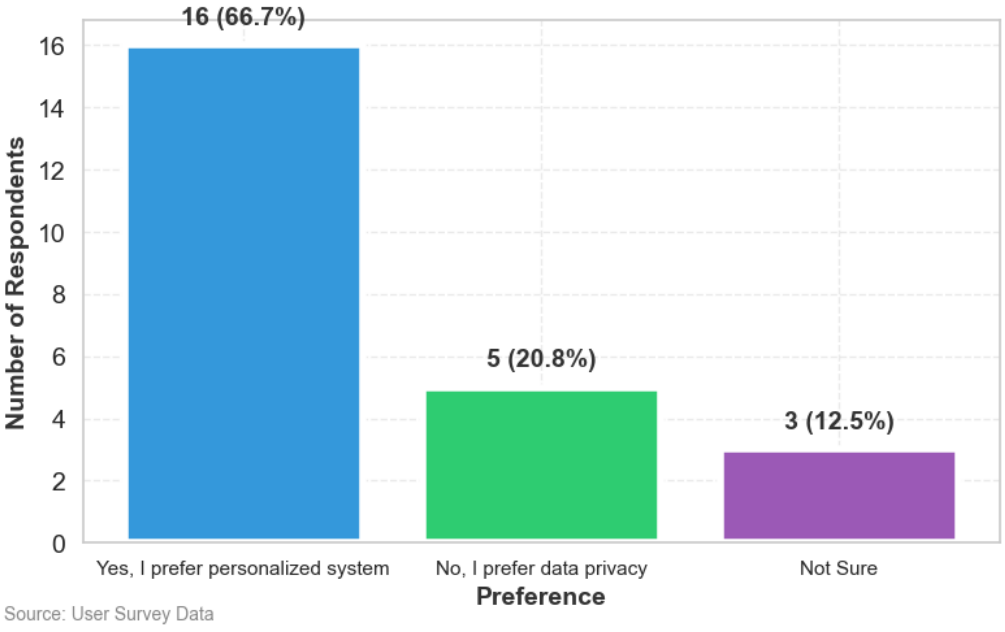


Figure 7.3: Privacy Preference of Users in Chatbot Interactions

7.1.3 Privacy Preferences

Figure 7.3 illustrates participants' preferences in the trade-off between personalization and privacy. Two-thirds (66.7%) preferred a personalized system, suggesting a willingness to share data in exchange for tailored support. A significant minority (20.8%) prioritized privacy, while 12.5% remained undecided. This distribution reveals a general inclination towards personalization, but also underscores the importance of safeguarding users' autonomy and offering explicit control over data usage. The qualitative feedback provided by participants revealed several recurring themes that reflect both functional and emotional needs. These insights offer valuable direction for enhancing the app's usability, credibility, and emotional resonance with its target users.

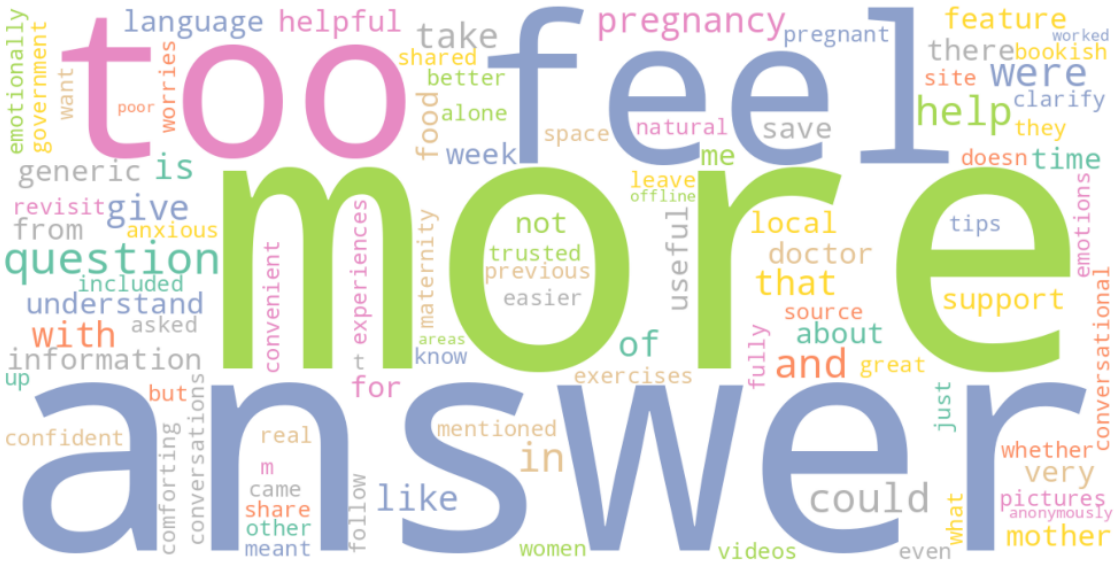


Figure 7.4: User Feedback wordcloud

7.2 Insights from In person interviews

The semi-structured interviews revealed not only practical feature requests but also deeper insights into how women situated the chatbot in their everyday lives. While all three participants engaged with the system for several days, their reflections emphasized the intersection of embodied experience, trust, and the scope of care that AI should provide.

Voice as embodied interaction: Participants expressed frustration with typing, suggesting that audio input would be more natural during daily routines. As 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 framing positions voice not as a convenience feature but as an accessibility requirement rooted in the physical realities of pregnancy and postpartum life. For

HCI, this highlights how modality choices are shaped by bodily circumstances, not just by efficiency metrics.

Transparency as reassurance: Women also emphasized the importance of knowing the source of medical guidance. P2 noted,

> "If the app mentioned whether the answer came from a doctor or a trusted site, I would feel more confident."

Here, transparency was less about verifying factual correctness and more about cultivating emotional reassurance in moments of uncertainty. This suggests that trust in health AI may hinge not only on technical accuracy but on visible signals of accountability and legitimacy.

Reminders as relational support: Finally, participants envisioned the chatbot evolving beyond a Q&A tool into a partner in managing daily health routines. Requests for features like prenatal vitamin reminders or appointment notifications reframed the system as an active participant in care, as both P2 and P3 said,

> "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."

This shift blurs the boundary between information service and caregiver, raising critical design questions about the extent to which AI should assume responsibility in sensitive health contexts.

Offline access: A recurring concern was the fragility of internet connectivity, particularly for women in semi-urban and rural areas. All the participants explained,

> "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 highlights how connectivity gaps can exclude precisely those most in need of support. For HCI, the challenge is to envision hybrid AI systems that can gracefully degrade into offline modes—providing cached guidance and stored conversations—while resuming richer interactivity when online again. Designing for the absence, not just the presence, of infrastructure is thus central to ensuring equitable access.
These interviews illustrate how user expectations exceeded technical functionality. They position maternal health chatbots not merely as information retrieval systems but as relational technologies, expected to adapt to the embodied, emotional, and infrastructural rhythms of pregnancy.

7.3 Expert Feedback

The initial and feedback survey was designed and reviewed by a licensed physician on our research team to ensure medical appropriateness and cultural sensitivity.

The physician provided input on question wording, response options, and overall structure to ensure clarity and relevance for the target population. To ensure medical appropriateness and safety, we also conducted an expert review of the chatbot's outputs. The doctor examined a representative sample of user queries and system responses, focusing on the accuracy, clarity, and potential risks of the information provided. Feedback confirmed that the chatbot generally produced clinically sound guidance, while also highlighting the importance of reinforcing disclaimers that the system is not a substitute for professional medical advice. This expert validation provides an initial check on clinical reliability, though future work should include more systematic evaluations with a broader set of healthcare professionals.

Chapter 8

Discussion

This chapter integrates insights from the survey, system design, technical evaluation, and user feedback to critically reflect on the contributions of the *Baby and Me* system. The discussion highlights how the work advances both maternal health support in Bangladesh and broader understandings of human–AI interaction. We also address our research questions in the discussion.

8.1 Relational and Contextual Design

The survey findings showed that maternal concerns were not limited to physical complications such as miscarriage or labor pain; psychosocial issues like anxiety, depression, and mood swings were equally pressing. This duality revealed that any technology limited to medical facts would miss an essential dimension of care. The design, therefore, combined retrieval-augmented generation for factual reliability with empathetic response generation for emotional reassurance.

This design choice was not cosmetic. In practice, mothers frequently reported that empathetic tones and culturally resonant language mattered as much as factual correctness. Several participants described the chatbot as "someone listening to me," suggesting relational value beyond information delivery. Such interpretations echo prior research on systems like Woebot and Wysa, which emphasize empathy as a therapeutic element.

These findings demonstrate that multilingual interaction in Bangla and Banglish, paired with culturally contextualized content, significantly increased accessibility and trust. Women who might otherwise have disengaged with English-only or generic systems perceived the bot as "authentic" and approachable. The result was higher engagement and stronger willingness to share sensitive concerns, confirming that linguistic inclusivity and relational empathy are not optional add-ons but prerequisites for meaningful maternal health technologies in Bangladesh (RQ1).

8.2 Transparency, Trust, and Personalization

A deliberate feature of the system was partial transparency—users could see the retrieved documents and tool outputs that informed the chatbot's responses. This

feature functioned as an "epistemic resource," reassuring women that the system's knowledge was evidence-based rather than arbitrary. Feedback suggested that women equated such transparency with honesty, which directly bolstered trust.

However, transparency alone did not satisfy all expectations. Many participants expressed a desire for deeper personalization: advice adapted to their trimester, daily routines, or specific complications. This tension mirrors broader challenges in health AI, where static content undermines perceived relevance. While the RAG framework succeeded in reducing hallucinations and grounding answers, personalization remained a visible gap.

Transparency combined with multilingual support increased user confidence, reinforcing the system's perceived trustworthiness (RQ1). However, trust was closely tied to personalization; women felt reassured only when responses reflected their stage of pregnancy or specific symptoms. This indicates that cultural contextualization must be coupled with personalization to fully realize accessibility and engagement.

The use of conversational memory and adaptive decision-making already distinguishes the system from rule-based chatbots. Unlike static systems, *Baby and Me* could recall prior queries within a session and build continuity. Users highlighted this as a sign of attentiveness—something traditional rule-based bots cannot achieve (RQ2). This finding confirms that agent-based architectures improve the quality and relational depth of support.

8.3 Technical Performance in Context

The technical evaluation revealed a decisive factor for adoption in low-resource contexts: the ability to generate fast, accurate Bangla responses. Among the tested models, Gemini 2.5 Flash outperformed Qwen 2.5 in both contextual accuracy and latency. Faster responses preserved conversational flow, while subtle improvements in Bangla phrasing prevented breakdowns in trust.

In practice, these seemingly technical benchmarks translated into lived usability. Delays of more than a few seconds were seen as frustrating, and awkward or overly literal Bangla translations felt "foreign" to participants. Thus, performance optimization for both language capability and responsiveness was not simply about engineering efficiency but about ensuring social accessibility.

The integration of Bangla/Banglish language capability ensured that users could interact naturally without switching to English, which directly improved accessibility (RQ1). Trust was particularly fragile when phrasing felt unnatural; smooth multilingual performance, therefore, became a trust signal in itself.

By combining real-time retrieval with efficient orchestration, the agent outperformed rule-based alternatives that would either fail to adapt or deliver delayed responses (RQ2). This demonstrates that technical optimization is not peripheral but central to the superiority of agent-based systems in maternal health contexts.

8.4 Agentic AI as Socio-Technical Mediation

The *Baby and Me* system did not aim to replace doctors or midwives. Instead, it mediated between structural gaps in healthcare and the everyday needs of women. In Bangladesh, where over 70% of births occur at home and perinatal mental health remains stigmatized, such mediation is critical. Mothers often lack access to both professional expertise and safe spaces for disclosure.

By offering a readily available, stigma-free environment, the chatbot created an alternative channel for emotional and informational support. Its agentic design—drawing on retrieval, memory, and decision-making—allowed it to respond dynamically to user concerns. In doing so, it complemented overburdened healthcare systems rather than competing with them.

This reframes the contribution of agentic AI. Unlike rule-based chatbots that deliver static advice, *Baby and Me* functioned as a socio-technical bridge, empowering women with informational equity and emotional reassurance (RQ2). This demonstrates how tool orchestration and memory can transform AI from a transactional interface into a relational companion that mediates structural inequities.

8.5 Implications for HCI and Health AI

The study contributes to ongoing HCI debates about the roles of conversational agents. Findings show that users increasingly perceive such agents not only as tools but as companions. In a low-resource Bangla-first context, this relational framing is especially significant because it compensates for fragile infrastructures and cultural stigma.

Three broader implications emerge:

1. **Interdependence of properties:** Transparency, empathy, and contextual accuracy must be treated as interconnected drivers of trust, not isolated design features.

2. **Localized language as trust signal:** Multilingual support, especially Bangla-first design, is more than usability—it is a marker of cultural legitimacy.

3. **Promise of agentic frameworks:** Protocols like the Model Context Protocol demonstrate how modular, extensible architectures can sustain both technical performance and cultural sensitivity.

The integration of multilingual, culturally grounded knowledgebases improved accessibility and trust (RQ1), while the shift from rule-based to agentic design—through tool orchestration, memory, and real-time decision-making—significantly enhanced personalization and quality of support (RQ2). Together, these results illustrate how socio-technical design choices determine whether maternal health AI is adopted, trusted, and sustained in real-world contexts.

Table 8.1: Comparison of Model Options for Bangla Maternal Health Agent

Feature	Qwen 2.5 (14B)	Gemini 2.5 Flash-Lite	Gemini 2.5 Flash
Open Source	yes	no	no
Self-Hosting	yes	no	no
Bangla Capabilities	Mediocre	Good	State-of-the-Art
Response Time	Slowest	Fastest	Fast
Cost	Depends on hosting platform	Free tier available (0.5 USD / million tokens)	Free tier available (3 USD / million tokens)

8.6 Cost Analysis and Scalability

While the technical evaluation demonstrates feasibility, long-term deployment hinges on cost-effectiveness and scalability. Large language models differ not only in accuracy and latency but also in recurring operational costs. As shown in Table 8.1, commercial offerings like Gemini 2.5 Flash deliver state-of-the-art Bangla performance but are priced at approximately **$3 per million tokens**, whereas open-source models such as Qwen 2.5 can be self-hosted at significantly lower rates (as low as **$0.5 per million tokens**), albeit with infrastructure overheads.
For maternal health applications in Bangladesh, where budgets are constrained, these differences are decisive. If an average user session involves 1,000 tokens, then supporting 10,000 monthly users would generate around **10 million tokens**. At scale, this translates to:

- **Gemini 2.5 Flash:** ˜$30 per month
- **Gemini 2.5 Flash-Lite:** ˜$5 per month
- **Qwen (self-hosted):** ˜$5 per month in token-equivalent costs, plus server maintenance (estimated $50–200 per month depending on GPU/CPU configuration).

Thus, while Gemini APIs provide predictable pricing and lower engineering burden, self-hosting Qwen trades higher upfront costs (servers, DevOps, monitoring) for long-term affordability. A hybrid model—using self-hosted Qwen for routine queries and commercial APIs for complex reasoning—emerges as the most sustainable path for LMIC settings.

8.6.1 Requirements for Scaling the Application

To scale beyond a prototype into a national-level maternal health support system, several components are required:

1. **Infrastructure:** Cloud or hybrid deployment with caching, load balancing, and low-bandwidth mobile clients for rural users.

2. **Model Optimization:** Distillation and quantization of large models, along with session-level token management to minimize usage.

3. **Operations:** Continuous monitoring for accuracy, secure data handling, and regular updates to maternal health guidelines.

4. **User-Centered Expansion:** Multilingual support (regional dialects), community health worker integration, and personalization through semantic memory.

8.6.2 Funding Pathways

Given the public health relevance, funding strategies must align with sustainability and equity:

- **Government and Health Ministry Programs:** Integration with Bangladesh's digital health initiatives, supported by national maternal health budgets.
- **NGO and Development Partners:** Collaboration with organizations such as BRAC, UNFPA, or UNICEF, who already fund maternal and reproductive health programs.
- **Research and Innovation Grants:** Applications to global AI and health funding calls (e.g., Gates Foundation, Wellcome Trust, Grand Challenges Canada).
- **Public-Private Partnerships:** Telecom operators and mobile service providers can subsidize access, given the high mobile penetration in Bangladesh.
- **Sponsorship and CSR:** Local technology companies and hospitals can provide infrastructure support as part of corporate social responsibility.

These diversified funding streams would not only cover computational costs but also support outreach, training, and integration with existing maternal health services. Ultimately, financial sustainability is as central to inclusivity as linguistic or cultural adaptation, since prohibitive costs risk excluding precisely the low-resource communities most in need.
In sum, the research demonstrates that agentic, culturally grounded AI can act as both an informational and relational companion. By integrating evidence-based guidance with empathetic support, the *Baby and Me* system provides design insights for equitable health technologies that are scalable, trustworthy, and grounded in the everyday realities of pregnant and postpartum women in Bangladesh.

Chapter 9

Limitations and Future Works

9.1 Limitations of The Study

No research is without constraints, and acknowledging these helps to contextualize the contributions and guide future work. The development and evaluation of the *Baby and Me* system faced several methodological, technical, and contextual limitations that shaped the scope and interpretation of the findings.

9.1.1 Survey and Data Collection Constraints

The initial survey of 72 women provided valuable insights into maternal health concerns in Bangladesh, but the sample size and distribution limit generalizability. While participants came from both urban and rural settings, urban respondents were somewhat overrepresented. Furthermore, recruiting participants in hospitals and clinics may have excluded the most marginalized women who do not access formal care at all. Responses to sensitive topics such as mental health may also have been underreported due to cultural stigma and hesitancy in disclosure. As such, while the survey data grounded the design in lived experiences, it cannot fully represent the diversity of maternal health contexts across Bangladesh. We have not got a single sensitive issue (abortion, early motherhood etc.) in our surveys, showing that we haven't reached every spectrum.

9.1.2 Knowledgebase Development

The knowledgebase was curated from publicly available Bangla health content and cross-checked against authoritative guidelines. However, reliable and comprehensive Bangla-language resources remain scarce. This meant that certain maternal health topics, particularly mental health, were underrepresented. Additionally, the static nature of the curated knowledgebase means that it cannot automatically adapt to newly emerging guidelines or updated clinical recommendations without manual intervention.

9.1.3 Language and Model Limitations

Although the system supported both Bangla and Banglish, limitations in large language model performance remain a challenge. Even the best-performing models oc-

casionally produced awkward phrasing, grammatical errors, or overly literal translations that reduced naturalness. Moreover, Bangla-specific colloquialisms or regional dialects (sylheti, chatgayyian, etc.) were not handled effectively. These limitations may have constrained inclusivity for women outside major urban or educated populations. Finally, while hallucination rates were relatively low, the potential for incorrect outputs remains inherent to generative models, raising concerns for clinical reliability. In our testing, we found different answers for the same questions almost every time. Though they were similar, this goes to show the randomness in answer generation by the LLM's.

9.1.4 System Design and Technical Boundaries

The technical design prioritized accessibility, but constraints remain. The system requires internet connectivity, which limits use in low-bandwidth or offline contexts where many rural women live. Latency, though minimized in testing, can still pose challenges for prolonged interactions. Furthermore, the scope of personalization was limited: while the system could adapt responses to user queries, it lacked persistent, individualized profiles that could account for long-term pregnancy stages or health histories. The absence of multimodal input (e.g., images, voice) further constrained the richness of interactions, particularly for women with low literacy.

9.1.5 Evaluation and Feedback Study

The evaluation combined technical metrics with user feedback, but both carried constraints. The technical review was conducted via LLM as a Judge, which is widely adopted but has its own caveats. Some research has questioned the authenticity of this method of evaluation [69], [75]. Similarly, while user testing revealed strong satisfaction, the study population was relatively small and biased toward digitally literate mothers who were willing to experiment with an AI system. Feedback may therefore reflect the perspectives of early adopters rather than the broad population of Bangladeshi mothers. Finally, testing was conducted over short interactions, leaving longer-term engagement and sustained trust unexamined.

9.1.6 Socio-Cultural and Ethical Constraints

Maternal health is shaped not only by access to information but also by socio-economic structures, family dynamics, and cultural practices. An AI agent, no matter how empathetic, cannot fully address systemic barriers such as a lack of skilled birth attendants or stigma surrounding maternal mental health. Ethical challenges also remain: the absence of professional oversight in chatbot interactions raises risks if users treat the system as a substitute for medical advice. While the design explicitly mentions that it's never a substitute for a doctor and redirects users to healthcare professionals for urgent concerns, there is no guarantee that users always interpret or follow such guidance.

These limitations highlight the boundaries of the *Baby and Me* system as an exploratory, proof-of-concept intervention. The study demonstrates the potential of agentic AI for maternal health support in Bangladesh, but its survey data, curated knowledgebase, technical design, and evaluation scope are not exhaustive. These

constraints must be considered when interpreting the findings and in designing future iterations of the system.

9.2 Future Works

Our study demonstrates the feasibility of an MCP-enabled, Bangla-first maternal health chatbot, but it also surfaces broader questions for the design of empathetic AI in resource-constrained contexts. Based on the comments from feedback survey and interviews we take into account on how to improve the later versions of *Baby and Me* bot.

9.2.1 Multimodal Input for Richer Context

Almost all the participants mentioned the need for voice input option. Some imagined a future where the chatbot could interpret ultrasound reports, test results or prescriptions. This desire highlights a critical design trajectory: moving beyond text toward multimodal maternal health support. In HCI terms, multimodality offers richer grounding of conversations in lived experience, yet it also introduces new design dilemmas. Medical document parsing raises issues of standardization, translation between Bangla and English, and safeguarding sensitive health data. Thus, the push toward multimodality must be accompanied by an equally strong push for privacy-preserving design that respects women's control over their health information.

9.2.2 Personalized Health Tracking

Our current system treats each conversation as an isolated event. While many participants welcomed personalization—such as the ability to revisit prior chats or receive reminders—others voiced concerns about data storage and potential misuse. This tension illustrates a classic HCI challenge: personalization is often framed as an unquestioned good, yet in maternal health it is entangled with trust, consent, and stigma. Designing longitudinal maternal health companions therefore requires not only technical infrastructure for secure memory but also new interaction paradigms that allow women to negotiate what is remembered, forgotten, or shared.

9.2.3 AI Insights for Risk Detection

Participants frequently expressed anxiety about miscarriage, depression, or confusing bodily changes. This raises the possibility of AI not just answering questions reactively, but proactively detecting risk patterns across interactions. Early warning systems could flag depressive language or repeated mentions of severe pain. Yet proactive detection also risks over-alerting or misclassifying, especially in a context where false alarms carry social and emotional consequences. The design challenge is to build systems that recognize risk without pathologizing normal experiences of pregnancy, balancing algorithmic sensitivity with cultural nuance and human judgment.

9.2.4 Accessibility and Inclusive Design

Our evaluation surfaced persistent barriers around language preference, literacy, and digital familiarity. Future work should therefore prioritize accessibility features such as voice input/output in Bangla, simplified interfaces for women with low digital literacy, and offline functionality for rural areas with unreliable internet. Several interviewees described moments when connectivity was unavailable for hours, suggesting the need for hybrid systems that cache previous conversations and provide essential health guidance without requiring constant connectivity. Recent advances in small language models and optimized open-source architectures make this vision increasingly feasible. Models such as LLaMA, GPT-OSS, and other lightweight variants can now be pruned, quantized, and deployed directly on mobile devices, reducing dependency on always-online infrastructure. For maternal health contexts, this shift opens a design space where AI agents are not just cloud services but companions that remain accessible even in network dead zones. This reframes accessibility from a matter of *better interfaces* to a matter of infrastructural resilience. Designing agentic AI that thrives in intermittent or absent connectivity conditions is not simply a technical optimization—it is a step toward equity, ensuring that women in low-resource settings are not excluded by the very systems intended to support them.

9.2.5 Longitudinal Evaluation of Impact

While our current evaluation focused on short-term usability and satisfaction, future research should adopt longitudinal designs to track how sustained interaction with the chatbot influences maternal health behaviors and outcomes over time. This would allow for measuring changes in trust, knowledge retention, mental well-being, and care-seeking practices across pregnancy and postpartum periods. Such longitudinal studies would provide stronger evidence of real-world effectiveness and help refine personalization strategies.

These reflections suggest that empathetic, agentic AI for maternal health is not just a technical endeavor but a socio-technical endeavor of negotiation. Future systems will need to reconcile multimodality with privacy, personalization with consent, risk detection with cultural sensitivity, and Inclusivity with infrastructural realities. Our findings invite the HCI community to see maternal health AI not only as a design problem but also as a lens into how trust, equity, and empathy must be reimagined when advanced AI meets everyday healthcare in low-resource settings.

Chapter 10

Conclusion

This thesis has presented the design and development of the first MCP-enabled agentic AI system for maternal health support in the Bangla language. By integrating retrieval-augmented generation, conversational memory, and external tool orchestration within a modular agentic framework, the system demonstrates how advanced AI can be localized to meet the linguistic, cultural, and healthcare realities of Bangladeshi women. Unlike conventional rule-based chatbots, the proposed solution adapts dynamically to the user context, incorporates curated maternal health knowledge, and supports both Bangla and Banglish inputs to ensure inclusivity.

The empirical evaluation—combining survey-based user insights, system performance benchmarking, and language model judgment—highlights both the promise and challenges of deploying such technology in low-resource healthcare environments. Findings suggest that the system can effectively address common maternal queries, reduce information asymmetry, and provide supportive dialogue in contexts where formal healthcare access is limited or stigmatized, particularly for perinatal mental health.

Beyond technical innovation, this work underscores the ethical and social responsibility of AI research. Designing for maternal health in Bangladesh required not only robust engineering but also sensitivity to issues of privacy, stigma, and digital literacy. The emphasis on Bangla-language capability directly addresses structural inequities in global AI development, where non-English users are frequently excluded.

In conclusion, this thesis contributes a novel, localized, and technically advanced framework for maternal health support in LMIC contexts. While the system demonstrates feasibility and initial impact, it also opens critical avenues for future research: large-scale deployment studies, integration with national health systems, longitudinal assessments of maternal and mental health outcomes, and exploration of hybrid models combining AI with human community health workers. By pioneering an agentic, MCP-enabled, Bangla-language maternal health assistant, this work lays the groundwork for more equitable, culturally attuned applications of AI in global health.

Bibliography

[1] F. Arias, A. G. Bhide, K. Damania, S. N. Daftary, et al., *Practical Guide to High Risk Pregnancy and Delivery-E-Book: A South Asian Perspective*. Elsevier health sciences, 2008.

[2] M. A. Hossen and A. Westhues, "Rural women's access to health care in bangladesh: Swimming against the tide?" *Social work in public health*, vol. 26, no. 3, pp. 278–293, 2011.

[3] L. T. Rose and K. W. Fischer, "Garbage in, garbage out: Having useful data is everything," *Measurement: Interdisciplinary Research & Perspective*, vol. 9, no. 4, pp. 222–226, 2011.

[4] R. A. Edwards, T. Bickmore, L. Jenkins, M. Foley, and J. Manjourides, "Use of an interactive computer agent to support breastfeeding," *Maternal and child health journal*, vol. 17, no. 10, pp. 1961–1968, 2013.

[5] K. I. Pollak et al., "Weight-related sms texts promoting appropriate pregnancy weight gain: A pilot study," *Patient education and counseling*, vol. 97, no. 2, pp. 256–260, 2014.

[6] B. Jack et al., "Reducing preconception risks among african american women with conversational agent technology," *The Journal of the American Board of Family Medicine*, vol. 28, no. 4, pp. 441–451, 2015.

[7] K. E. Driscoll et al., "Mood symptoms in pregnant and postpartum women with bipolar disorder: A naturalistic study," *Bipolar disorders*, vol. 19, no. 4, pp. 295–304, 2017.

[8] K. K. Fitzpatrick, A. Darcy, and M. 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*, vol. 4, no. 2, e7785, 2017.

[9] P. M. Gardiner et al., "Engaging women with an embodied conversational agent to deliver mindfulness and lifestyle recommendations: A feasibility randomized control trial," *Patient education and counseling*, vol. 100, no. 9, pp. 1720–1729, 2017.

[10] S. Liu and L. Zou. "Premom ovulation tracker app with ovulation tests, pregnancy tests." Accessed: 2025-07-29. [Online]. Available: https://premom.com/.

[11] M. Akter, S. Yimyam, J. Chareonsanti, and S. Tiansawad, "The challenges of prenatal care for bangladeshi women: A qualitative study," *International nursing review*, vol. 65, no. 4, pp. 534–541, 2018.

[12] L. Vaira, M. A. Bochicchio, M. Conte, F. M. Casaluci, and A. 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*, 2018, pp. 273–277.

[13] A. Williams, M. Sarker, and S. T. Ferdous, "Cultural attitudes toward postpartum depression in dhaka, bangladesh," *Medical anthropology*, vol. 37, no. 3, pp. 194–205, 2018.

[14] S. Akter, J. L. Rich, K. Davies, and K. J. Inder, "Access to maternal healthcare services among indigenous women in the chittagong hill tracts, bangladesh: A cross-sectional study," *BMJ open*, vol. 9, no. 10, e033224, 2019.

[15] A. T. Hossain, *Mit solve: Tiger challenge (bangladesh-based applicants) lifebot*, 2019. [Online]. Available: https://solve.mit.edu/solutions/10934.

[16] M. Islam and N. Sultana, "Risk factors for pregnancy related complications among urban slum and non-slum women in bangladesh," *BMC pregnancy and childbirth*, vol. 19, no. 1, p. 235, 2019.

[17] M. M. Rahman, R. Amin, M. N. K. Liton, and N. Hossain, "Disha: An implementation of machine learning based bangla healthcare chatbot," in *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, IEEE, 2019, pp. 1–6.

[18] M. Rahman et al., "Maternal pregnancy intention and its association with low birthweight and pregnancy complications in bangladesh: Findings from a hospital-based study," *International health*, vol. 11, no. 6, pp. 447–454, 2019.

[19] S. Abedin and D. Arunachalam, "Maternal autonomy and high-risk pregnancy in bangladesh: The mediating influences of childbearing practices and antenatal care," *BMC Pregnancy and Childbirth*, vol. 20, no. 1, p. 555, 2020.

[20] S. Akter, K. Davies, J. L. Rich, and K. J. 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*, vol. 15, no. 8, e0237002, 2020.

[21] T. Bickmore, Z. Zhang, M. Reichert, C. Julce, and B. Jack, "Promotion of preconception care among adolescents and young adults by conversational agent," *Journal of Adolescent Health*, vol. 67, no. 2, S45–S51, 2020.

[22] A. Dosani, H. Arora, and S. 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*, vol. 17, no. 20, p. 7679, 2020.

[23] E. Maeda et al., "Promoting fertility awareness and preconception health using a chatbot: A randomized controlled trial," *Reproductive BioMedicine Online*, vol. 41, no. 6, pp. 1133–1143, 2020.

[24] R. Venkatachari, "Maternal mental health and child outcomes: A human-centered design perspective on preventive mental health care," Ph.D. dissertation, Massachusetts Institute of Technology, 2020.

[25] K. Chung, H. Y. Cho, and J. Y. Park, "A chatbot for perinatal women's and partners' obstetric and mental health care: Development and usability evaluation study," *JMIR Medical Informatics*, vol. 9, no. 3, e18607, 2021.

[26] U. M. Jui, *Maya: An ai-powered start-up that plans to become the go-to name in healthcare*, Oct. 2021. [Online]. Available: https://www.tbsnews.net/features/panorama/maya-ai-powered-start-plans-become-go-name-healthcare-314215.

[27] J. J. Prochaska et al., "A therapeutic relational agent for reducing problematic substance use (woebot): Development and usability study," *Journal of medical Internet research*, vol. 23, no. 3, e24850, 2021.

[28] G. K. Dutta et al., "Mental healthcare-seeking behavior during the perinatal period among women in rural bangladesh," *BMC health services research*, vol. 22, no. 1, p. 310, 2022.

[29] N. Insan, A. Weke, J. Rankin, and S. Forrest, "Perceptions and attitudes around perinatal mental health in bangladesh, india and pakistan: A systematic review of qualitative data," *BMC Pregnancy and Childbirth*, vol. 22, no. 1, p. 293, 2022.

[30] M. N. Khan, M. L. Harris, and D. Loxton, "Low utilisation of postnatal care among women with unwanted pregnancy: A challenge for bangladesh to achieve sustainable development goal targets to reduce maternal and newborn deaths," *Health & Social Care in the Community*, vol. 30, no. 2, e524–e536, 2022.

[31] J. L. Z. Montenegro, C. A. da Costa, and L. P. Janssen, "Evaluating the use of chatbot during pregnancy: A usability study," *Healthcare Analytics*, vol. 2, p. 100 072, 2022.

[32] P. De and M. R. 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*, vol. 23, no. 1, p. 664, 2023.

[33] E. Durden, M. C. Pirner, S. J. Rapoport, A. Williams, A. Robinson, and V. L. Forman-Hoffman, "Changes in stress, burnout, and resilience associated with an 8-week intervention with relational agent "woebot"," *Internet Interventions*, vol. 33, p. 100 637, 2023.

[34] B. Inkster, M. Kadaba, and V. 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*, vol. 4, p. 1 084 302, 2023.

[35] A. Islam, F. Begum, A. Williams, R. Basri, R. Ara, and R. Anderson, "Midwife-led pandemic telemedicine services for maternal health and gender-based violence screening in bangladesh: An implementation research case study," *Reproductive health*, vol. 20, no. 1, p. 128, 2023.

[36] S. Suharwardy et al., "Feasibility and impact of a mental health chatbot on postpartum mental health: A randomized controlled trial," *AJOG Global Reports*, vol. 3, no. 3, p. 100 165, 2023.

[37] WHO, *Perinatal mental health*, 2023. [Online]. Available: https://www.who.int/teams/mental-health-and-substance-use/promotion-prevention/maternal-mental-health.

[38] R. Afreen et al., "Enhancing mental health literacy and care through community-driven solutions in rural bangladesh," *Frontiers in Global Women's Health*, vol. 5, p. 1 478 817, 2024.

[39] A. AI, *Probahini: Ai-powered menstrual health chatbot*, 2024. [Online]. Available: https://www.acmeai.tech/product/probahini-chatbot.

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

[41] T. BEATY, *How ai health care chatbots learn from the questions of an indian women's organization*, Feb. 2024. [Online]. Available: https://apnews.com/article/ai-health-care-chatbot-india-6e820475e89dbc15ce1880695f698934.

[42] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," *arXiv preprint arXiv:2402.03216*, 2024.

[43] D. Dominguez, *A technical guide to multi-agent orchestration*, Nov. 2024. [Online]. Available: https://dominguezdaniel.medium.com/a-technical-guide-to-multi-agent-orchestration-5f979c831c0d.

[44] J. Gao, P. Garsole, R. Agarwal, and S. Liu, "Ai test modeling and analysis for intelligent chatbot mobile app-a case study on wysa," in *2024 IEEE International Conference on Artificial Intelligence Testing (AITest)*, IEEE, 2024, pp. 132–141.

[45] L. Jones-Esan, N. Somasiri, and K. 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.

[46] A. Leto, C. Aguerrebere, I. Bhati, T. Willke, M. Tepper, and V. A. Vo, "Toward optimal search and retrieval for rag," *arXiv preprint arXiv:2411.07396*, 2024.

[47] H. Li et al., "Llms-as-judges: A comprehensive survey on llm-based evaluation methods," *arXiv preprint arXiv:2412.05579*, 2024.

[48] S. A. Mapari et al., "Revolutionizing maternal health: The role of artificial intelligence in enhancing care and accessibility," *Cureus*, vol. 16, no. 9, 2024.

[49] S. N. Mohtasim, F. O. Arpita, and I. Ahmed, "Gorbhoshongi: A mhealth app for the mental well-being of bangladeshi pregnant and postpartum women," 2024.

[50] E. Olwanda et al., "Women's autonomy and maternal health decision making in kenya: Implications for service delivery reform-a qualitative study," *BMC Women's Health*, vol. 24, no. 1, p. 181, 2024.

[51] I. Ouédraogo, "Mobile technology and artificial intelligence for improving health literacy among underserved communities," Ph.D. dissertation, Université de Bordeaux; Université Nazi Boni (Bobo-Dioulasso, Burkina Faso), 2024.

[52] M. P. Priola, "Addressing hallucinations with rag and nmiss in italian healthcare llm chatbots," *arXiv preprint arXiv:2412.04235*, 2024.

[53] J. N. R. Rivera et al., "Development and refinement of a chatbot for birthing individuals and newborn caregivers: Mixed methods study," *JMIR Pediatrics and Parenting*, vol. 7, no. 1, e56807, 2024.

[54] P. Sheidu. "Maternal health in bangladesh - the borgen project. "[Online]. Available: https://borgenproject.org/maternal-health-in-bangladesh/.

[55] D. A. Siddiqi et al., "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*, vol. 181, p. 105 288, 2024.

[56] B. Siwicki, *Conversational ai improves "fourth trimester" maternal care at penn medicine*, Apr. 2024. [Online]. Available: https://www.healthcareitnews.com/news/conversational-ai-improves-fourth-trimester-maternal-care-penn-medicine.

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

[58] M. Vilarim et al., "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*, vol. 46, e20233453, 2024.

[59] X. Wang et al., "Searching for best practices in retrieval-augmented generation," *arXiv preprint arXiv:2407.01219*, 2024.

[60] S. Amil et al., "Interactive conversational agents for perinatal health: A mixed methods systematic review," in *Healthcare*, MDPI, vol. 13, 2025, p. 363.

[61] G. Comanici 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.

[62] R. Deva, D. Ramani, T. Divate, S. Jalota, and A. 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*, 2025, pp. 1–23.

[63] A. Ehtesham, A. Singh, G. K. Gupta, and S. 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] M. Finio and A. Downie, *What is ai agent orchestration? — ibm*, Feb. 2025. [Online]. Available: https://www.ibm.com/think/topics/ai-agent-orchestration.

[65] X. Hou, Y. Zhao, S. Wang, and H. Wang, "Model context protocol (mcp): Landscape, security threats, and future research directions," *arXiv preprint arXiv:2503.23278*, 2025.

[66] N. Karunanayake, "Next-generation agentic ai for transforming healthcare," *Informatics and Health*, vol. 2, no. 2, pp. 73–83, 2025.

[67] Y. Kasse, R. Riantofani, R. Rahmi, and K. Kiftiyah, "Designing ai-driven chatbots for adolescent mental health support in rural schools," *International Journal of Research in Counseling*, vol. 4, no. 1, pp. 83–92, 2025.

[68] G. Kerimoglu Yildiz, R. Turk Delibalta, and Z. Coktay, "Artificial intelligence-assisted chatbot: Impact on breastfeeding outcomes and maternal anxiety," *BMC Pregnancy and Childbirth*, vol. 25, no. 1, p. 631, 2025.

[69] D. Li et al., "Preference leakage: A contamination problem in llm-as-a-judge," *arXiv preprint arXiv:2502.01534*, 2025.

[70] K. McAlister, L. Baez, J. Huberty, M. 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*, vol. 9, no. 1, e72469, 2025.

[71] S. Nasrin, "The association of participation in a text messaging program with reduction in depression severity among pregnant women," Ph.D. dissertation, University of British Columbia, 2025.

[72] İ. Özer Aslan and M. T. Aslan, "Benchmarking ai chatbots for maternal lactation support: A cross-platform evaluation of quality, readability, and clinical accuracy," in *Healthcare*, MDPI, vol. 13, 2025, p. 1756.

[73] N. K. Sehgal, H. Kambhamettu, S. P. Matam, L. Ungar, and S. C. 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*, 2025, pp. 1–7.

[74] Supabase, *Supabase docs*, 2025. [Online]. Available: https://supabase.com/docs.

[75] A. Szymanski, N. Ziems, H. A. Eicher-Miller, T. J.-J. Li, M. Jiang, and R. 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*, 2025, pp. 952–966.

[76] B. T. Worku, M. Abdulahi, D. Amenu, and B. 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*, vol. 25, no. 1, p. 155, 2025.

[77] A. Yang et al., "Qwen2. 5-1m technical report," *arXiv preprint arXiv:2501.15383*, 2025.

[78] F. Yasmin, A. Roy, M. Z. Alam, and A. S. Alam, "Unseen obstacles: The neglect of antenatal and postnatal care in bangladesh and the underlying factors," *TAJ: Journal of Teachers Association*, vol. 38, no. 1, pp. 121–129, 2025.

[79] D. Zhu et al., "Chain-of-thought matters: Improving long-context language models with reasoning path supervision," *arXiv preprint arXiv:2502.20790*, 2025.

[80] S. Ahamed et al., "Susastho. ai: A multimodal medical copilot for adolescents using evidence-based medicine and large language models," *AI: A Multimodal Medical Copilot for Adolescents Using Evidence-Based Medicine and Large Language Models*,

Appendix A

Survey Questionnaire

Table A.1: Pregnancy Experience Survey Questions

No	Question	Options
1	Name	*Short Answer*
2	Age	• Less than 18 years • 18–25 years • 25–35 years • 35–45 years • Greater than 45 years
3	Education Level	• No Formal Education • Primary • Secondary • College • Bachelor's • Master's or higher
4	Occupation	• Homemaker • Service • Business • Student • Daily wage labour • 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 Lakh – 5 Lakh • 5 Lakh – 8 Lakh • More than 8 Lakh

Continued on next page

Table A.1: Pregnancy Experience Survey Questions (Continued)

8	Internet / Smartphone Access	• Personal Smart Phone and Internet • Shared Access • Don't have smartphone
9	Language Comfort in App	• Only Bangla • Only English • Both Bangla and English
10	Number of Pregnancies	• 1 • 2 • 3 • 4 • 5 • 5+
11	Major 5 Complications Faced During Pregnancy	• Morning sickness or nausea • Back pain or body aches • Fatigue or low energy • Sleep difficulties • Fear of miscarriage or complications • Fear of childbirth/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 healthcare services • Negative experiences with doctors/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)*

Continued on next page

Table A.1: Pregnancy Experience Survey Questions (Continued)

12	Major 5 Concerns During Pregnancy	• 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 A.1: Pregnancy Experience Survey Questions (Continued)

13	Major 5 Happy Points During Pregnancy	• 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 (clothes, room, etc.) • Bonding emotionally with the baby • Receiving gifts or blessings (baby shower, etc.) • Positive feedback from doctors or nurses • Family getting excited 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 belly grow and baby develop • 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

Table A.2: Feedback Questionnaire

No	Question	Options
1	Name	*Short Answer*
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 scale* • 1 = very hard • 5 = very easy
6	Were the answers clear and understandable?	*Likert scale* • 1 = not clear • 5 = very clear
7	Did the response feel caring/empathic?	*Likert scale* • 1 = completely robotic • 5 = very empathetic
8	Did the answers match your concerns?	*Likert scale* • 1 = not at all • 5 = fully satisfied
9	How much did you trust the information from the app?	*Likert scale* • 1 = didn't 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 personalized chatbot (share your personal information to the system)?	• Yes, I prefer personalized system • No, I prefer data privacy • Not Sure
13	Would you recommend it to others?	• Yes • No • Not Sure

Appendix B

Semi Structured Interview Protocol

Baby and Me – Exit Interview Protocol

1. Introduction & Consent (2 mins)

Thank you for taking the time to speak with me today. This interview will take about 20–30 minutes. The purpose is to understand your experiences using the maternal health chatbot. Your responses will be anonymized, and no personal identifying information will be shared. Participation is completely voluntary; you can skip any question or stop at any time. With your permission, I'd like to record the conversation for accuracy in analysis. Do you agree to participate?

2. Warm-up & Context (3–5 mins)

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?

3. Process / Usability (10 mins)

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

4. Results / Understanding & Trust (10 mins)

1. Can you recall a specific question you asked the chatbot? What did you think about the response?
2. Was the answer clear and understandable?
3. Did the chatbot ever give information you doubted or didn't trust? Why?
4. Did you feel reassured, anxious, or neutral after reading the answers?
5. Compared to asking a doctor or midwife, how much would you trust this chatbot?

5. Influence / Impact (7–10 mins)

1. Did using the chatbot change how you think about your health, pregnancy, or care?
2. Did it motivate you to take any action (e.g., looking up more info, talking to someone, changing routine)?
3. If this chatbot was widely available, do you think women in your community would use it? Why or why not?
4. How often do you think you would realistically use it in the future?

6. Improvements & Wrap-up (3–5 mins)

1. If you could change one thing about the chatbot, what would it be?
2. What's one feature that you really liked and think should stay?
3. Is there anything else about your experience that I didn't ask but you'd like to share?