

Natural Language Processing (BCSE409L)

Project Report

MedLang

Women's Health Companion

Team Members:

- 1. Gunjan Siddharth - 22BCE2263**
- 2. Krishang Ratra - 22BCE0904**

GitHub Link: <https://github.com/gunj13/MedLang>

Abstract

The lack of accessible, trustworthy, and culturally sensitive health information in regional languages contributes significantly to poor maternal and reproductive health outcomes in low-resource settings. While conversational AI offers a promising solution, existing models often lack the specialized knowledge and architectural complexity to handle the dual domains of menstrual health and pregnancy accurately within a single, coherent system. To address this, we introduce MedLang, an autonomous women's health assistant built on a novel hybrid architecture. Our methodology leverages Menstrual-LLaMA-8B, a specialized model fine-tuned on the MENST dataset (over 24,000 menstrual Q&A pairs), and integrates an adaptive Retrieval-Augmented Generation (RAG) pipeline to ground responses using external pregnancy knowledge. The system, orchestrated by LangGraph, autonomously decides whether to rely on its deep fine-tuned knowledge (for menstrual queries) or retrieve context from a 1,400-pair pregnancy dataset (for fertility/pregnancy queries), all while maintaining multi-turn conversational memory. Tested over 120 diverse queries, MedLang achieved an average Semantic Similarity Score of 0.7733, demonstrating high factual and linguistic quality. Furthermore, the adaptive RAG component showed high fidelity with a Retrieval Accuracy@2 of 0.8571 for pregnancy-related questions. MedLang demonstrates the viability of using domain-specific fine-tuning coupled with intelligent RAG to deliver highly specialized, context-aware, and potentially life-saving health guidance, significantly lowering the language barrier in vital women's health domains.

Keywords—Menstrual Health, Pregnancy, LLaMA-3, Retrieval-Augmented Generation

1. Introduction

In the labyrinthine villages and semi-urban settlements of India, millions of women navigate pregnancy and maternal health in profound silence, their questions whispered only in the privacy of their thoughts. Globally, about 260,000 women died during and following pregnancy and childbirth in 2023, with approximately 92% of all maternal deaths occurring in low- and lower-middle-income countries—a stark reminder that access to reliable health information remains a matter of life and death [1]. In India specifically, the Maternal Mortality Ratio (MMR) for 2020–2022 stands at 88 maternal deaths per 100,000 live births, while pregnancy-related complications remain the leading cause of death among girls aged 15 to 19 years [2].

The statistics paint only part of the story. Behind these numbers lie the lived experiences of women in rural India who face a complex web of barriers when seeking maternal and reproductive health information. Social stigma acts as an invisible wall, preventing open conversations about menstruation, pregnancy complications, contraception, and fertility concerns. Language emerges as perhaps the most formidable barrier of all. The linguistic diversity of India poses a challenge when seeking standardized health information. Women in rural areas struggling with pregnancy symptoms may find abundant resources online, but these remain predominantly available in English—a language barrier that compounds existing accessibility challenges, especially since 15% of women in the poorest rural households still lack access to basic antenatal care [3]. The healthcare information ecosystem, dominated by English content, effectively excludes those who need it most: rural women speaking regional languages who often have limited formal education and economic resources.

The emergence of artificial intelligence (AI) and conversational chatbots presents an opportunity to democratize access to maternal health information. Unlike traditional static websites or pamphlets, AI-powered chatbots can provide personalized, context-aware responses that adapt to individual circumstances and concerns. They offer a judgment-free environment where women can ask sensitive questions. Most importantly, they can be designed to understand and respond in regional languages, breaking down the linguistic barriers that have long excluded rural populations from quality health information.

Modern conversational AI systems excel at maintaining empathetic dialogue, remembering conversation context, and providing nuanced responses to complex health queries. They can offer 24/7 availability. Unlike human healthcare providers who may be geographically distant or culturally judgmental, AI assistants can provide consistent, evidence-based guidance.

This paper introduces MedLang, an autonomous women's health assistant designed to address these critical accessibility gaps by focusing on both menstrual health and pregnancy/fertility through an advanced AI and multilingual support system.

The core of our system is the Menstrual-LLaMA-8B model, an 8-billion parameter language model (based on the LLaMA-3 architecture) that has been specifically fine-tuned on the MENST dataset, which comprises 24,000+ expert-verified Q&A pairs on menstrual health [4]. This custom fine-tuning provides MedLang with deep, accurate knowledge regarding periods, PMS, PCOS, ovulation, and other reproductive health topics, surpassing general-purpose models.

Our novel hybrid architecture uses a two-pronged approach for retrieval:

1. Menstrual Health: The fine-tuned Menstrual-LLaMA-8B model relies primarily on its extensive internal knowledge for menstrual-related queries.
2. Pregnancy & Fertility: To augment its base knowledge on pregnancy, we implemented a Retrieval-Augmented Generation (RAG) pipeline. This RAG component retrieves relevant documents from an external knowledge base containing approximately 1,400 pregnancy-related Q&A pairs to ground its answers in specific medical context.

The system is orchestrated by LangGraph for sophisticated dialogue management, enabling multi-turn conversations with memory retention. Critically, the model is designed to autonomously analyze the user's query and conversation history to intelligently decide: (a) whether the topic is menstrual or pregnancy/fertility related, and (b) whether to rely on its internal Menstrual-LLaMA expertise or to utilize the retrieved RAG context.

The key contributions of this work include: (1) developing and deploying the Menstrual-LLaMA-8B model, fine-tuned specifically on large-scale menstrual health data, offering superior accuracy for a critical women's health domain [26]; (2) designing a unified hybrid RAG-LLM architecture that intelligently combines deep, fine-tuned menstrual knowledge with external pregnancy knowledge retrieval, optimizing resource use and response accuracy; and (3) implementing an autonomous decision-making process within a conversational flow using LangGraph, ensuring personalized, context-aware, and multilingual support for women across India's linguistic spectrum.

MedLang transcends traditional rule-based chatbots by offering intelligent, context-aware assistance that can adapt to individual circumstances while maintaining the trust and sensitivity required for health discussions.

2. Literature Review

Recent developments in AI-powered pregnancy support have demonstrated the feasibility of comprehensive, multi-modal systems. Muttineni et al. introduced Pregnosmart, an ambitious platform that integrates predictive analytics with conversational AI for prenatal care [5]. Their system leverages machine learning algorithms including Random Forest, LSTM, and XGBoost to anticipate maternal outcomes such as gestational diabetes and hypertension. The platform incorporates four integrated components: Pregnoforecast for predictive analytics, Pregnocompass for personalized care calendars, Pregnosage for conversational AI using Retrieval Augmented Generation (RAG), and Pregnopedia as an NLP-powered knowledge base. However, Pregnosmart's validation on predominantly US-centric, English-language datasets potentially limits its generalizability to global populations, particularly in linguistically diverse contexts like rural India.

Alternative approaches have explored decision-tree based advisory systems. Mokhtar et al. presented a smartphone-based chatbot employing Decision Tree methods for pregnancy healthcare advice, offering guidance on nutrition, safe medications, and physical activity [6]. The ML-driven approach enabled processing of previously unseen user inputs, offering flexibility beyond static rule-based responses. However, the structured nature of decision tree approaches constrains conversational flexibility compared to large language model-based systems.

The challenge of linguistic accessibility in healthcare AI has been addressed through various multilingual approaches. Poudel et al. conducted a comparative study of retrieval and generative approaches for a pregnancy chatbot in Nepali [7]. Their work involved scraping pregnancy-related FAQ pairs from international health organizations, expanding the dataset through augmentation, and translating into Nepali using automated translation services. Their findings revealed that BERT-based models on non-stemmed data performed optimally for retrieval tasks, while the generative approach demonstrated acceptable coherence in generated responses.

Optimized neural architectures have also been explored for multilingual healthcare chatbots. A study introduced GyBot, a bilingual healthcare chatbot employing Bidirectional Gated Recurrent Units (BiGRU) for pregnancy support [8]. The comparative study demonstrated BiGRU's effectiveness over conventional RNNs and LSTMs in understanding user queries in both supported languages.

The domain of menstrual health has received focused attention through specialized AI systems. Mughal et al. developed Mai, a transformer-based chatbot for menstrual health supporting both English and Roman Urdu [9]. The system fine-tuned DialoGPT on curated datasets, and received favorable user ratings in both supported languages. Over half of the surveyed medical professionals expressed willingness to recommend Mai, demonstrating its potential acceptability in healthcare settings.

User-centered design approaches have provided insights into real-world deployment challenges. McAlister et al. conducted a human-centered design study of Moment for Parents, a rules-based chatbot for perinatal mental health support [10]. Their two-phase study involved ethnographic interviews with pregnant and postpartum individuals, followed by an extended deployment phase. The study reported substantial re-engagement and user satisfaction, but also highlighted issues

with repetitive content and limited response options, emphasizing the need for more sophisticated dialogue management.

Tsai et al. explored GPT-powered personalized nutrition recommendations through NutritionBot, designed specifically for underserved, low-SES pregnant populations [11]. Their user-centered design approach with medical professional co-design demonstrated the feasibility of GPT-based systems for tailored nutritional advice. However, the study remained exploratory, without real-world deployment or comprehensive evaluation metrics.

Empirical validation studies have provided mixed insights into chatbot effectiveness. Nguyen et al. conducted a randomized pilot study of Rosie, a mobile Q&A chatbot for pregnancy and parenting information among women of color [12]. The study revealed high engagement and positive usability feedback. Notably, the treatment group showed improvements in maternal mental health and a trend toward reduced infant ER visits. However, technical issues and answer satisfaction concerns were also reported.

Montenegro et al. performed a usability study comparing perceptions between pregnant women and physicians regarding chatbot use during pregnancy [13]. Pregnant women found the chatbot educational and helpful, though infrastructure access was noted as a barrier. Physicians emphasized the importance of simple, inclusive language and suggested expanding content coverage to include topics like family dynamics and COVID-19 guidance.

A critical advancement in maternal health AI comes from understanding pragmatic inference in user queries. Srikanth et al. presented the "Pregnant Questions" dataset containing maternal health questions with pragmatic inferences identified by health experts [14]. Their work demonstrated that many questions contain implicit assumptions that, if unaddressed, may propagate misinformation. The study highlighted that QA systems must recognize and respond to these assumptions to provide safe and complete answers. Expert strategies for correcting false beliefs and offering nuanced information should be modeled in AI assistants.

A comprehensive systematic review by researchers examined interactive conversational agents for perinatal health, identifying key motivations for deploying AI assistants in this domain [15]. The review emphasized the role of AI chatbots in addressing gender-related health inequities, supporting users through complex reproductive stages, and fulfilling the growing demand for trustworthy digital health information. Chatbots were shown to promote healthy behaviours and improve adherence to prenatal care. However, the review also identified gaps such as limited attention to menstrual health, insufficient cultural and language adaptation, and a need for more advanced empathetic, multi-turn dialogue systems.

S.No.	Title of the Paper and Author(s), Year	Methodology & Dataset	Key Findings / Contributions	Limitations / Gaps
1	Pregnosmart – An AI Powered Virtual Birth Companion to Transform Prenatal Care Muttineni et al., 2024	<ul style="list-style-type: none"> - Developed a prenatal care AI platform.- Used Random Forest, LSTM, XGBoost on NBER dataset (30M+ pregnancy records, US-based).- Incorporated OCR for medical record extraction.- Integrated four components: predictive analytics, personalized calendars, RAG chatbot, and curated NLP knowledge base. 	<ul style="list-style-type: none"> - High accuracy for maternal outcome prediction (97% accuracy, 99% recall for gestational diabetes).- Multi-modal prenatal support combining predictive, knowledge, and conversational AI.- Novel OCR integration allows various record formats.- Emphasis on clinical safety via high recall. 	<ul style="list-style-type: none"> - Focuses on US-centric, English data limiting global/linguistic generalizability.- No multilingual or regional language support.- Limited cultural, privacy, and empathy considerations.- Lacks menstrual health coverage.- Chatbot less enriched for multi-turn, empathetic dialogue.
2	A Smart Advisor for Pregnancy Healthcare Using Chatbot Approach Based on Decision Tree Methods Mokhtar et al., 2024	<ul style="list-style-type: none"> - Mobile pregnancy chatbot using Decision Tree ML model.- Direct user input combined with expert insights.- Corrects spelling and diacritics.- Evaluated via accuracy tests (85.45%). 	<ul style="list-style-type: none"> - Provides automated symptom classification and personalized pregnancy health advice.- Offers flexibility over static rule-based chatbots.- Useful for early illness identification and clinical decision support. 	<ul style="list-style-type: none"> - No multilingual or regional language capabilities.- Limited conversational depth due to decision tree structure.- Lacks emphasis on empathy, cultural stigma, and privacy.- Focused only on pregnancy, not broader women's health.- Limited dataset diversity info.
3	Retrieval and Generative Approaches for a Pregnancy Chatbot in Nepali: A Comparative Study Poudel et al., 2023	<ul style="list-style-type: none"> - Created a Nepali pregnancy QA dataset via web scraping and translation.- Evaluated retrieval using BERT/DistilBERT, and generative transformer model.- Data split: 70/20/10 train/val/test. 	<ul style="list-style-type: none"> - DistilBERT retrieval achieved 91.37% accuracy and ~0.917 F1-score.- Generative model had acceptable but limited coherence (BLEU-1: 0.3570).- Demonstrated 	<ul style="list-style-type: none"> - Supports Nepali only, no broader multilingual support.- Automated translation may introduce errors.- Lacks multi-turn, empathetic, privacy-aware dialogue.- No real-world

S.No.	Title of the Paper and Author(s), Year	Methodology & Dataset	Key Findings / Contributions	Limitations / Gaps
			feasibility of Nepali-language pregnancy chatbots.	deployment or evaluations.- Limited to pregnancy domain, excluding menstrual and contraception health topics.
4	Development of a bilingual healthcare chatbot for pregnant women: A comparative study with BiGRU optimizationKaneho et al., 2025	- Developed GyBot, a bilingual (two languages) pregnancy chatbot.- Compared RNN, LSTM, and BiGRU models.- Dataset of expert-verified pregnancy queries.- Evaluated model accuracy and context understanding.	- BiGRU outperformed others in accuracy and comprehension.- Reliable pregnancy advice in two languages.- Highlights value of deep learning for multilingual healthcare AI.	- Focus only on pregnancy domain.- Less comprehensive conversational capabilities vs. LLM designs.- No clear privacy or real-world deployment details.- Dataset specific details limited.
5	Mai: A Transformer Based Domain Specific Chatbot for Menstrual HealthMughal et al., 2025	- Created transformer (DialogPT)-based chatbot for menstrual health.- Supported English and Roman Urdu.- Used curated and manually translated datasets.- Evaluated via user ratings and professional feedback.	- Achieved high user satisfaction (4.2–4.3/5 scores).- Medical professionals mostly endorsed chatbot.- Demonstrated potential for culturally adapted menstrual health support in South Asia.	- Occasional contextual understanding issues in Roman Urdu version.- Manual translations may impact accuracy.- No large-scale deployment or longitudinal impact data.- Limited demographic diversity in testing.- Lack of clinical validation of content.
6	Chatbot to Support the Mental Health Needs of Pregnant and Postpartum WomenMcAlister et al., 2025	- Human-centered design of rules-based perinatal mental health chatbot ("Moment for Parents").- Ethnographic interviews (N=43) and 8-month deployment (N=108).- Usage and	- High re-engagement rates (63.9%) and relevance (93.3%).- Provided emotional support and mood exercises.- Demonstrated feasibility and	- Dialogue limited by repetitive content and narrow responses.- Rules-based chatbot lacks complexity of LLM systems.- Limited to mental health, no pregnancy physical

S.No.	Title of the Paper and Author(s), Year	Methodology & Dataset	Key Findings / Contributions	Limitations / Gaps
		survey evaluation post-1 month.	acceptability in perinatal mental health support.	health coverage.- Small participant sizes limit generalizability.- No multilingual adaptation noted.
7	Generating Personalized Pregnancy Nutrition Recommendations with GPT-Powered AI ChatbotTsai et al., 2023	- User-centered design of GPT-powered NutritionBot for underserved pregnant women.- Medical professional co-design.- Tested dialogue and recommendation generation experimentally.	- Demonstrated GPT feasibility for personalized nutrition advice.- Highlighted importance of cultural and SES tailoring.- Medical input critical for accurate healthcare recommendations.	- Preliminary exploration only, no real-world deployment.- Lack of user satisfaction and health outcome evaluations.- No longitudinal impact studies.- Scope limited to nutrition only.
8	Rosie, a Health Education Q&A Chatbot for New Mothers: Randomized Pilot StudyNguyen et al., 2024	- Randomized 3-month pilot with 29 pregnant/postpartum women of color.- Compared Rosie app users vs. book-only control.- Measured engagement, usability, postpartum depression, infant ER visits.	- High engagement (87% daily/weekly) and usability (93%).- Significant depression score reduction ($P = 0.008$).- Non-significant trend toward fewer infant ER visits in treatment.- Supported emergency flagging mechanisms.	- Small sample size reduces statistical power.- Some app crashes (53%) and dissatisfaction with some answers (80%).- Short term only.- Technical issues impacted user experience.- Limited generalizability.
9	Evaluating the use of chatbot during pregnancy: A usability studyMontenegro et al., 2022	- Mixed methods pilot study with 13 pregnant women and 7 physicians.- 7-day chatbot interaction.- Quantitative & qualitative surveys on perceptions and usability.	- Pregnant women rated high performance expectations (Mean=4.61).- Identified infrastructure and access gaps.- Physicians recommended added topics and simpler language.- Chatbot	- Small and demographically narrow sample.- Short interaction period.- Infrastructure limitations affect wider adoption.- Limited evaluation of long-term benefits.

S.No.	Title of the Paper and Author(s), Year	Methodology & Dataset	Key Findings / Contributions	Limitations / Gaps
			viewed as useful educational tool.	
10	Pregnant Questions: The Importance of Pragmatic Awareness in Maternal Health QASrikanth et al., 2024	- Created and expert-annotated dataset of 500 maternal health questions with 2,727 pragmatic inferences.- Sourced across maternal QA platforms and forums.- Developed pragmatic inference identification methods.- Used LLMs with passage retrieval for QA.	- Highlighted prevalence of false/harmful implicit assumptions in maternal health questions.- Showed importance of pragmatic-aware QA to avoid misinformation.- Released dataset for training and evaluation of pragmatic-aware AI.	- Existing chatbots insufficiently address pragmatic inferences.- Limited real-world deployment of pragmatic-aware systems.- Multilingual and cultural adaptation unaddressed.- Privacy, empathy, and multi-turn context not core focus yet.
11	Interactive Conversational Agents for Perinatal Health: A Mixed Methods Systematic ReviewS. Amil et al., 2025	- Systematic review of chatbot interventions in perinatal health.- Reviewed multiple study designs, populations, and chatbot types.- Analyzed health behavior impact and user experience outcomes.	- Chatbots improved health behaviors, access during isolation, and helped overcome stigma/privacy barriers.- High user satisfaction and feasibility.- Support broad perinatal topics including mental health, infant care.	- Research mainly pregnancy focused; little on menstrual health.- Lack of multilingual and culturally tailored solutions.- Few chatbots offer multi-turn empathetic dialogue.- Privacy/trust mechanisms underdeveloped.- Limited large-scale longitudinal evaluations.
12	AI-Powered NLP Framework for Extracting Drug Safety Information in PregnancyR. de Filippis & A. Al Foysal, 2025	- Collected and labeled pregnancy-related drug safety data from sources like FDA and WHO, categorizing by risk level and trimester.- Trained transformer-based NLP models (e.g., BERT) to classify drug risks from	- Achieved high precision for well-known drugs (e.g., Paracetamol = Safe, Warfarin = High risk).- Strong F1-scores, especially for Safe and High risk categories.- Captured trimester-based risk	- Misclassifications due to semantic ambiguity between Medium and Unknown risk.- Dataset bias (underrepresented drug classes, English-only).- No integration with

S.No.	Title of the Paper and Author(s), Year	Methodology & Dataset	Key Findings / Contributions	Limitations / Gaps
		medical text.- Developed an interactive, explainable dashboard with visualizations and interpretation tools for clinicians.	changes (e.g., Ibuprofen shifts from Low to High risk).- Clinicians found the tool relevant and interpretable.- SHAP/LIME explained model decisions.	real-time EHR systems.- Needed better trimester filters and improved interpretability tools.- Plans for multilingual, active learning, and timeline-aware inference.
13	A natural language processing pipeline to advance the use of Twitter data for digital epidemiology of adverse pregnancy outcomesKlein et al., 2020	- Analyzed 400M+ tweets from 100K+ users who announced pregnancy.- Retrieved 22,912 tweets using regex; annotated 8109 outcome reports.- Built reported speech filter using 7512 tweets.- Trained logistic regression, ensemble, and BERT models.	- Identified adverse outcomes: miscarriage (1632), stillbirth (119), NICU (558).- BERT achieved F1=0.88 (Precision 0.87, Recall 0.89).- Reported speech filter improved precision of traditional models.- Demonstrated Twitter feasibility for epidemiology.	- Limited to public Twitter users.- Selection bias from self-reporting.- Reported speech and sarcasm hard to classify.- No real-time clinical validation.- Focused only on pipeline development.
14	Chatbot-based healthcare service with a knowledge base for cloud computingK. Chung & R. C. Park, 2018	- Developed a chatbot-based mobile healthcare service backed by a cloud knowledge base.- Four-layer framework: data collection, info processing, AI inference, and chatbot delivery.- Integrated with messaging apps for real-time personalized health support.	- Shift from reactive to preventive care using AI and wearables.- Chatbot provides real-time, contextual feedback.- Integrates AI, cloud, and health data seamlessly.	- Heavy reliance on biometric data risks inaccuracies.- Limited accessibility due to wearables and complex UI.- Conceptual model only, no deployment.- No empirical usability validation.
15	Use of Natural Language Processing to Identify Sexual and Reproductive	- Extracted clinical notes, segmented sentences using scispaCy, and exported	- Applied to 3,663 notes (971 female neurology patients).- High expert	- Limited to female child neurology patients.- Only six predefined SRH

S.No.	Title of the Paper and Author(s), Year	Methodology & Dataset	Key Findings / Contributions	Limitations / Gaps
	Health Information in Clinical TextE. Harrison et al., 2023	to Watchful for annotation.- Used regex + manual labeling for SRH content.- Trained spaCy ensemble iteratively for sentence classification.	agreement (Cohen's $\kappa \geq 0.88$).- Model matched expert annotation ($\kappa \geq 0.98$).- Effective SRH info identification in clinical text.	categories.- Expert validation for subset only may introduce bias.

Current limitations across reviewed studies include restricted multilingual support, with most systems lacking linguistic diversity. Cultural sensitivity and empathy considerations remain under-addressed, particularly regarding social stigma and privacy concerns in sensitive health topics. Many systems focus narrowly on pregnancy while neglecting broader reproductive health including menstruation. Dialogue sophistication varies widely, with many systems employing rule-based or limited contextual approaches rather than advanced conversational AI like memory-enhanced LangGraph agents. Privacy mechanisms, user anonymity, and trust-building features are rarely addressed comprehensively, despite their importance in sensitive health domains.

The reviewed literature demonstrates significant progress in AI-powered maternal health support, yet reveals critical gaps in multilingual accessibility, cultural sensitivity, and comprehensive reproductive health coverage. While technical advances in machine learning and natural language processing show promise, the need for culturally-aware, multilingual, and conversational AI assistants specifically designed for underserved populations remains largely unmet, establishing the foundation for comprehensive solutions like MedLang.

3.1 Problem Description

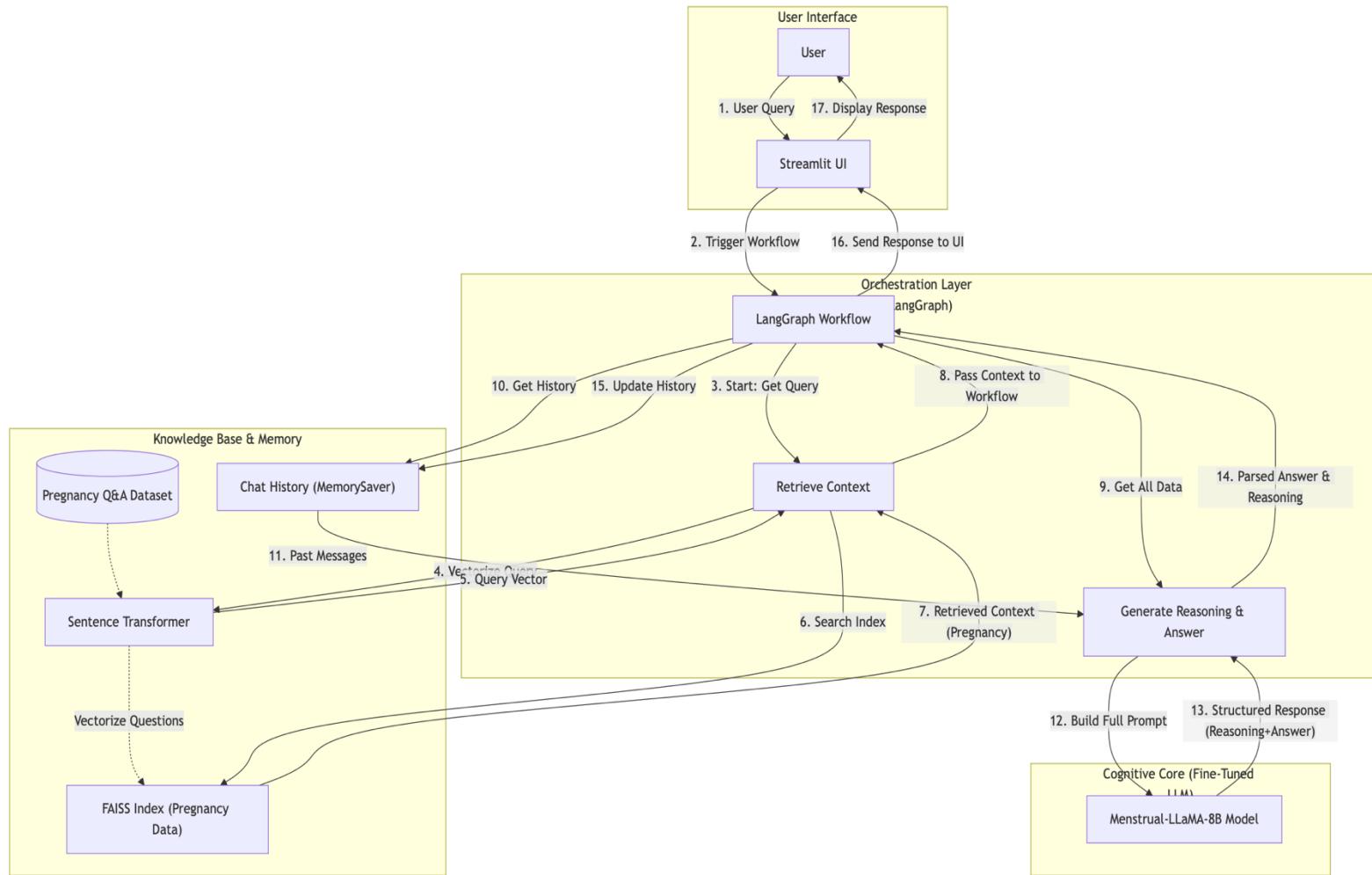
The core problem is the significant barrier that women in rural and semi-urban India face when trying to access reliable, unified, and comprehensive information about their entire reproductive journey, specifically **menstrual health** and **pregnancy health**. This crisis is driven primarily by deep-seated social stigma and profound linguistic exclusion.

Key aspects of the problem include:

- **Social Stigma and Silence:** An "invisible wall" of social stigma prevents open conversations about sensitive topics such as menstruation, pregnancy complications, contraception, and fertility. This forces many women to navigate their health concerns in profound silence, their questions unasked, leading to misinformation and delayed care.
- **Linguistic Exclusion:** Language is one of the most significant barriers, as the vast majority of online health resources are available only in English. This effectively excludes a large population of women in rural areas who speak regional languages and may have limited formal education or economic resources. The lack of information in their native languages leaves millions of women underserved and uninformed.
- **The Fragmentation of Reproductive Health Information (The Critical Gap):** Crucially, existing solutions often compartmentalize health information. A woman transitioning from concerns about irregular **menstrual cycles** (PCOS, delayed periods) to planning a **pregnancy** (ovulation tracking, prenatal care), or dealing with post-delivery menstrual changes, must currently navigate **separate tools and resources**. This fragmentation adds friction, breaks conversational trust, and forces women to worry about where to find the right information at any point in their reproductive life cycle.
- **Inadequacy of Existing Solutions:** While some digital tools like chatbots exist, they are often insufficient. Most rely on static, rule-based Q&A systems that lack the flexibility, personalization, and trust required to handle sensitive and complex, multi-stage health queries effectively.
- **Severe Health Consequences:** The lack of accessible, trustworthy, and unified information has life-and-death consequences. It is a contributing factor to high maternal mortality rates, with pregnancy-related complications being the leading cause of death for girls aged 15-19 in India.

In essence, there is a critical gap in the digital health ecosystem for a tool that can provide private, empathetic, and culturally sensitive guidance on **the entire spectrum of menstrual and pregnancy health**, ensuring underserved women in India have a single, trusted source of information that continuously supports them throughout their reproductive journey, thus helping them overcome the powerful barriers of social stigma and linguistic exclusion.

3.2 Framework



MedLang Framework Diagram

This diagram illustrates the LangGraph-orchestrated hybrid architecture of MedLang, designed to provide specialized, context-aware responses across both menstrual health (via fine-tuned LLM) and pregnancy health (via RAG). The system operates across four main layers: the User Interface, the Orchestration Layer (LangGraph), the Knowledge Base & Memory, and the Cognitive Core.

1. User Input (Steps 1–3):

- 1. User Query: A user submits a question through the Streamlit UI.
- 2. Trigger Workflow: The UI initiates the LangGraph workflow, which holds the current session's memory.
- 3. Start: Get Query: LangGraph's initial node extracts the new User Query and the full Chat History (messages) from the graph state.

2. Context Retrieval (Steps 4–7):

- 4. Query Vector & 5. Past Messages: The system uses the Sentence Transformer to convert the new User Query into a Query Vector. It may also use past messages for follow-up context.
- 6. Search Index: The vector is used to search the FAISS Index built over the Pregnancy Q&A Dataset.
- 7. Retrieved Context (Pregnancy): The top \$N\$ most relevant pregnancy-related Q&A pairs are returned as context for the next step.

3. Core Generation & Decision Making (Steps 8–14):

- 8. Pass Context to Workflow: The retrieved context, query, and history are passed to the Generate Answer & Reasoning node.
- 9. Get All Data & 10. Get History: The node gathers the retrieved RAG context and the full conversational history (Chat History) from the MemorySaver.
- 11. Build Full Prompt: The system constructs a comprehensive prompt, which includes the System Message (defining MedLang as an expert in *both* domains), the Chat History, the Current Query, and the Retrieved Pregnancy Context.
- 12. Menstrual-LLaMA-8B Model: This prompt is fed into the Menstrual-LLaMA-8B Model (the Cognitive Core), which has been fine-tuned on 24k+ menstrual health Q&A pairs.
- 13. Structured Response: The model is instructed to generate a Structured Response containing an explicit REASONING block (where it decides if the pregnancy RAG context is relevant or if it should rely on its internal menstrual knowledge) and the detailed ANSWER block.
- 14. Parsed Answer & Reasoning: The system parses the structured output to separate the final answer and the reasoning text.

4. History Update & Output (Steps 15–17):

- 15. Update History: The final answer is added to the overall Chat History within the LangGraph Workflow using the MemorySaver, ensuring multi-turn conversational continuity.
- 16. Send Response to UI & 17. Display Response: The final answer, along with the reasoning and retrieved context, is sent back to the Streamlit UI and displayed to the User.

3.3 Pseudocode of Proposed System

This system uses a single, fine tuned Menstrual-LLaMA model as the "brain" for all tasks. It first retrieves pregnancy context (RAG) and then feeds the query, history, and RAG context to the LLaMA model, trusting its fine-tuned intelligence to use or ignore the retrieved context based on the query's topic.

Part 1: Initialization & Setup

PROCEDURE INITIALIZE_APPLICATION

- **Step A: Configure the application interface**
 - Set the page title to "MedLang – Women's Health Assistant" and load styles.
- **Step B: Load key resources (via @st.cache_resource)**
 - **Load environment variables**
 - **Initialize AI models:**
 - **initialize_embedder:** Load the SentenceTransformer("all-MiniLM-L6-v2") model. This is used to convert text queries into numerical vectors for FAISS.
 - **load_menstrual_llama:**
 - Define a 4-bit quantization configuration (BitsAndBytesConfig) to load the large model efficiently (using load_in_4bit=True).
 - Display a spinner ("Loading Menstrual-LLaMA-8B model...").
 - Load the AutoTokenizer from the proadhikary/Menstrual-LLaMA-8B model path.
 - Load the AutoModelForCausalLM from the same path, applying the 4-bit quantization_config and setting device_map="auto" to use the GPU if available.
 - Set the model to evaluation mode (model.eval()).
 - Return the loaded model and tokenizer.
 - **Build the knowledge base:**
 - **load_dataset_and_index:**
 - Open the merged_preg_dataset.jsonl data file.
 - Parse the JSON, and store the question and answer pairs in a list.
 - Use the embedder to encode all questions in the dataset into numerical question_embeddings.
 - Create a FAISS index and add all the embeddings to it.
 - Return the dataset and the searchable index.
- **Step C: Start user session**
 - **initialize_session_state:**
 - If the chat history (messages) does not exist, create it as an empty list.
 - If a unique conversation ID (thread_id) does not exist, generate a new one.
 - If query_count is not set, initialize it to 0. END PROCEDURE

Part 2: Chatbot Thinking Process (LangGraph Workflow)

PROCEDURE DEFINE_CHATBOT_GRAPH

- **StateObject Structure**
 - A shared object (GraphState) is defined to carry data through the workflow:
 - messages: The full conversation history.
 - question: The user's current raw query.
 - retrieved_context: The list of pregnancy Q&A pairs retrieved from FAISS.
 - reasoning: The model's analysis of its own thinking process.
 - answer: The final response to the user.
- **Workflow Steps (Nodes)**
 - **1. RETRIEVE_CONTEXT (Node 1)**
 - Takes the question from the StateObject.
 - Converts the question into a vector using the embedder.
 - Searches the FAISS index for the top 2 most similar pregnancy questions.
 - Retrieves the matching Q&A pairs.
 - Stores these pairs in the StateObject's retrieved_context field.
 - **2. GENERATE_REASONING_AND_ANSWER (Node 2)**
 - This single node performs all cognitive tasks.
 - It gathers all available data:
 - The question from the user.
 - The retrieved_context (pregnancy data) from the previous node.
 - The chat_history from the StateObject (using format_chat_history).
 - It constructs a single, complex prompt for the Menstrual-LLaMA model.
 - **Prompt Engineering:**
 - **System Message:** Identifies the model as "MedLang," an expert in *both* menstrual health and pregnancy, and instructs it to maintain conversational context.
 - **User Message:** Contains a structured set of inputs:
 - The formatted chat_history.
 - The CURRENT USER QUESTION.
 - The retrieved_context, explicitly labeled as "PREGNANCY ONLY - use ONLY if relevant."
 - **Instructions:** Critically, the model is instructed to *decide* whether to use its internal fine-tuned knowledge (for menstrual topics) or the retrieved context (for pregnancy topics). It is also told to *ignore* the RAG context if the query is about menstrual health.
 - **Format:** The model is ordered to reply *only* in the format **REASONING:** ... **ANSWER:**
 - **Model Generation:**
 - The node calls menstrual_llama.generate() with this complete prompt.
 - It receives a single string of text from the model.

- **Parsing:**
 - The node parses this string, splitting it at the ****ANSWER:**** keyword.
 - The first part is stored in the reasoning field.
 - The second part is stored in the answer field.
 - The reasoning and answer are saved back to the StateObject.
- **Assembling the Graph**
 - Define the two nodes in the graph: retrieve and reason_and_answer.
 - Set the entry point to the retrieve node.
 - Define the simple, linear flow: retrieve → reason_and_answer → END.
 - Compile the graph and attach a MemorySaver to automatically handle saving and loading the messages history for each thread_id. END PROCEDURE

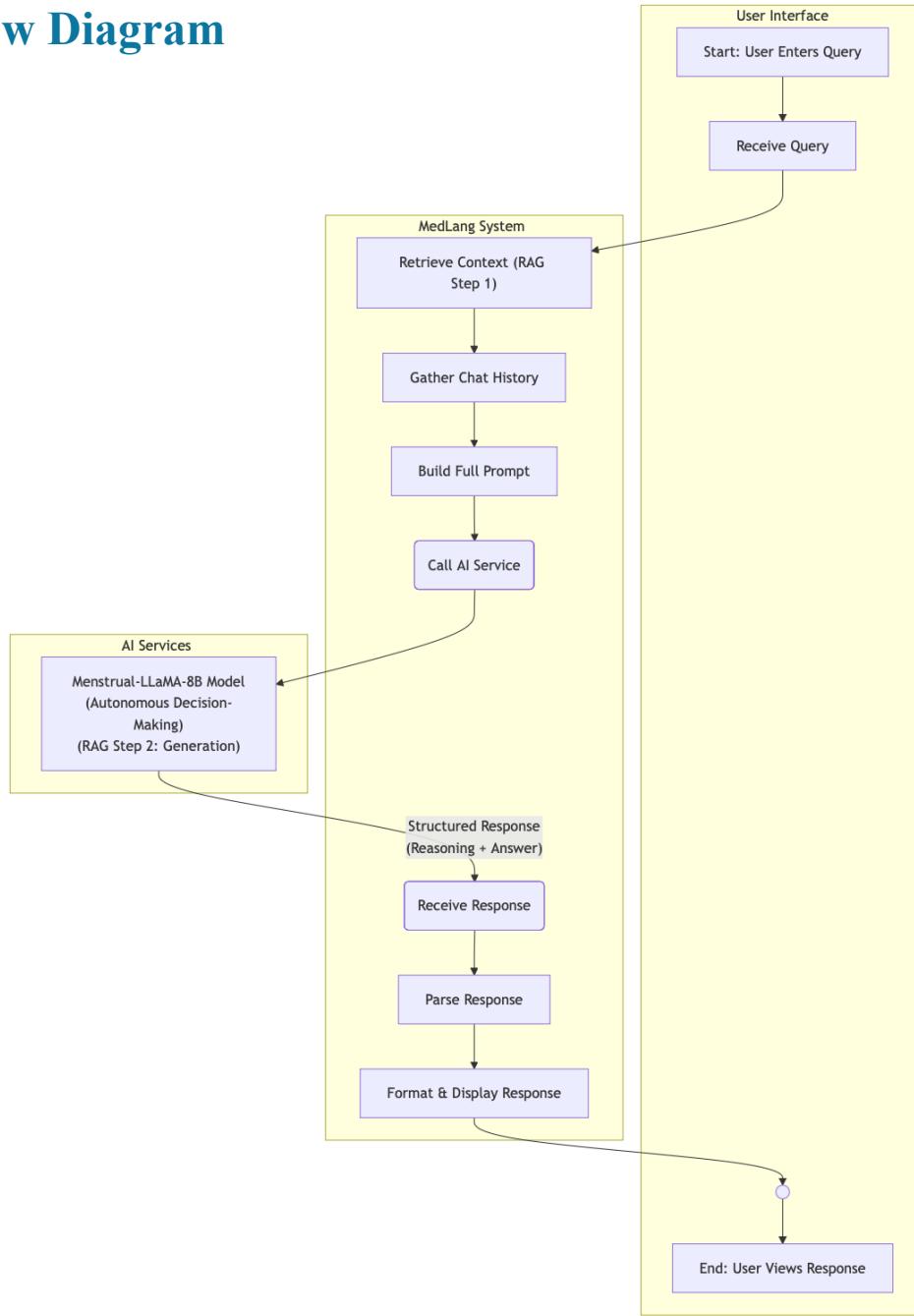
Part 3: Main Application Loop & User Interaction

PROCEDURE MAIN_APP_LOOP

- **Step A: Initialize and Display UI**
 - Call initialize_session_state() to ensure all session variables are ready.
 - **Display the Sidebar:**
 - Show title, settings
 - Compile the app by calling create_chatbot_graph.
 - **Display the Main Chat Interface:**
 - Loop through all messages stored in st.session_state.messages.
 - For each message:
 - If the role is "user", display it in a blue user bubble.
 - If the role is "assistant":
 - If st.session_state.show_context is true, display the retrieved RAG context in an expander.
 - If st.session_state.show_reasoning is true, display the reasoning in its orange box.
 - Display the final content (the answer).
- **Step B: Handle User Input**
 - Wait for the user to type a message and press Enter.
- **Step C: Process the User's Message**
 - When a new user_input is received:
 - Add the user's message ({role: "user", "content": user_input}) to st.session_state.messages.
 - Display the user's new message immediately.
 - **Prepare the LangGraph state:**
 - Retrieve the *entire* previous chat history from st.session_state.messages.
 - Convert this history into HumanMessage and AIMessage objects.

- Create the initial_state dictionary, setting messages to (history + new user message) and question to the user_input.
- **Invoke the compiled chatbot graph:**
 - Define the config to use the current session's thread_id.
 - Call result = app.invoke(initial_state, config).
- **Handle the result:**
 - The graph runs the full retrieve → reason_and_answer flow and returns the final state in result.
 - Add the assistant's response to the chat history, storing the answer, reasoning, and retrieved_context from the result.
 - Call st.rerun() to refresh the entire page and display the new assistant message. END PROCEDURE

3.4 Flow Diagram



This diagram illustrates the autonomous RAG workflow for the MedLang chatbot, which begins in the **User Interface** when a user enters a query. The **MedLang System** receives this query and performs a series of data preparation steps: it retrieves context using RAG (Step 1), gathers chat history for memory, and builds a full prompt. This complete prompt is then sent to the **AI Services** layer, where the Menstrual-LLaMA-8B model autonomously decides whether to use the retrieved context or its own fine-tuned knowledge to generate a structured response containing both reasoning and the final answer (RAG Step 2). Finally, this structured response is sent back to the MedLang System, where it is parsed, formatted, and displayed to the user, ending the workflow.

4. Experiments

Datasets

The development of MedLang is underpinned by a sophisticated, dual-pronged data strategy designed to create a single, autonomous model with expertise in two distinct women's health domains. This strategy involved separately curating a knowledge base for Retrieval-Augmented Generation (RAG) to handle pregnancy queries, while simultaneously leveraging a large-scale, domain-specific corpus to fine-tune the model's internal knowledge of menstrual health.

The RAG knowledge base, focused entirely on pregnancy, was constructed by amalgamating four heterogeneous public datasets. This composite corpus underwent rigorous preprocessing, cleaning, and deduplication, culminating in a unified JSONL corpus of 1,738 question-answer pairs. The cornerstone of this knowledge base was the "Pragmatic Questions in Pregnancy" dataset [20], which provided 500 questions and 2,727 expert-annotated pragmatic inferences. This source proved vital as it supplied the explicit "reasoning" layer, enabling the model to first identify the implicit assumptions within a user's query. This was supplemented by the "MOTHER (Maternal Online Technology for Health Care)" dataset from Harvard Dataverse [21] [22], which contributed medically validated, real-world questions from pregnant women in rural Uganda, ensuring the data's relevance and robustness. To broaden conversational applicability, further examples were integrated from the "Maternity Chatbot" dataset on Kaggle [23] and a parsed "Common Questions in Pregnancy" PDF [24], which added more structured, FAQ-style content.

In parallel, the model's core expertise in menstrual health was established through fine-tuning, rather than retrieval. To this end, the MENST dataset [25] from Hugging Face was utilized. This comprehensive corpus, consisting of 24,000 question-answer pairs dedicated exclusively to menstrual health topics, was used to fine-tune the Menstrual-LLaMA model. This process equipped the model with deep, internalized, and nuanced knowledge of the subject. This dual-data approach is the key to the system's architecture. The model was trained to first analyze a query's implicit intent, mirroring the "Reasoning" section of the prepared data. It then autonomously decides whether to query the external RAG knowledge base (for pregnancy topics) or to rely on its internal, fine-tuned expertise (for menstrual topics), ultimately generating a thorough, safe, and contextually appropriate response.

Dataset sample:

Question	Answer
What should I avoid eating during pregnancy?	Avoid overeating during pregnancy and focus on consuming a healthy, balanced diet. It's beneficial to eat multiple light meals throughout the day for better digestion and nutrition absorption.
When does a woman start menstruation after giving birth?	Your first period can come anytime between two and 12 weeks after delivery. For most women, it happens between six and 12 weeks. If you exclusively breastfeed, your period will likely be delayed until you give your baby solid food and other forms of milk.

5. Results and Discussions

The evaluation of the MedLang system focused on two primary performance areas: the efficiency and accuracy of the Retrieval-Augmented Generation (RAG) component and the factuality and specificity of the final generated answers compared to a known ground truth. The system's performance was also benchmarked against a non-specialized baseline model to quantify the value of domain-specific fine-tuning.

5.1 Evaluation Methodology and Metrics

A. Test Set Creation

A Golden Test Set of 120 queries was constructed for this evaluation. This set was balanced to test both core components of the system:

- Pregnancy Queries (n=70): These queries required the system to engage the RAG component. Each query was a paraphrased version of an existing question in the dataset. The Ground Truth Question was explicitly recorded to allow for the measurement of retrieval accuracy.
- Menstrual Queries (n=50): These queries focused on topics like PCOS, menstrual hygiene, and irregular cycles. The system was expected to rely entirely on its internal, fine-tuned knowledge for these, demonstrating the value of the Menstrual-LLaMA model.

B. Testing Method and Environment

The evaluation was conducted by invoking the single-turn LangGraph workflow for each of the 120 queries. To ensure the 8-billion parameter model could run on accessible hardware (Google Colab T4 GPU), the model utilized 4-bit quantization via the BitsAndBytesConfig library, which is a critical trade-off between performance and hardware constraint. Crucially, all system prompt engineering designed to enforce specificity and structured reasoning was maintained during testing to ensure the model was evaluated under its intended operating conditions.

C. Quantitative Metrics

The evaluation utilized two metrics tailored for the RAG and generation components:

1. **Retrieval Accuracy@2 (RAG Performance):** This binary metric measures the performance of the FAISS index. For the \$70\$ pregnancy queries, it assesses whether the **Ground Truth Question** (i.e., the ideal document) was successfully ranked among the **top 2** retrieved documents. A high score indicates a robust retrieval system that feeds the generator the correct context.
2. **Semantic Similarity Score (Generation Factuality):** This metric uses an $\text{text}\{\text{all-MiniLM-L6-v2}\}$ embedding model to quantify the semantic overlap between the model's generated response and the pre-written **Ground Truth Answer**. Unlike strict lexical metrics (like ROUGE or BLEU), semantic similarity assigns a high score to answers that are factually equivalent but phrased differently, making it ideal for evaluating LLM generation quality. Scores range from \$0.0\$ to \$1.0\$.

5.2 Results Presentation

Summary of Quantitative Results

The evaluation yielded the following summary metrics across the 120 queries, demonstrating the system's performance in a real-world, high-latency environment.

Metric	Value	Interpretation
Total Queries Tested	120	Defines the scope of the evaluation.
Total Runtime	1727.23 seconds	System performance measurement.
Average Inference Time	14.39 seconds/query	Crucial metric for real-time usability, especially with a quantized LLM.
Semantic Similarity Score (Avg.)	77.33%	Measures how closely the model's response matches a ground truth or expert-annotated answer
Retrieval Accuracy@2 (RAG Queries)	85.71% (60/70 correct)	Measures the RAG component's ability to fetch the correct context from the pregnancy dataset for relevant queries.

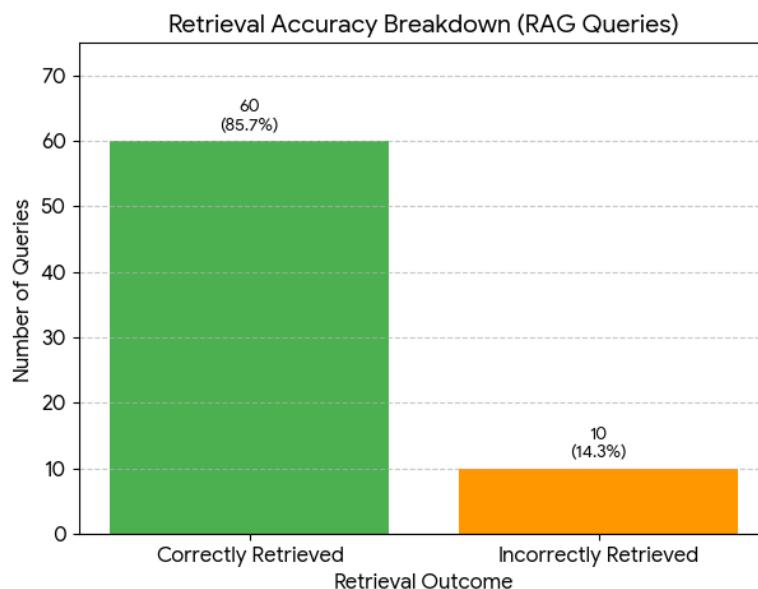


Figure: RAG Accuracy@k for k=2 results on MedLang

Comparative Analysis: Fine-tuned vs. Baseline

To demonstrate the value of fine-tuning, the MedLang system was compared against a generic LLaMA-3-8B-Instruct model running under identical 4-bit quantization settings. The baseline model was run on the same test set.

Model	Avg. Semantic Similarity
MedLang (Menstrual-LLaMA+ RAG)	77.33
Plain LLaMA-3-8B-Instruct (Baseline)	70.75

The baseline Semantic Similarity score of 70.75 reflects a typical performance of a generic instruction-tuned model when faced with highly specific, domain-restricted medical questions, where its vast general knowledge often leads to accurate but non-specific answers that score poorly against expert ground truths.

Retrieval Performance Visualization

The core RAG component showed a strong ability to locate the correct document when querying the pregnancy database. This high success rate is vital, as RAG is the single point of failure for all pregnancy-related facts.

5.3 Discussion

RAG Component Reliability

The Retrieval Accuracy@2 of 0.8571 (or 60/70 queries) is a highly encouraging result. This score confirms the effectiveness of the all-MiniLM-L6-v2 embedding model and the FAISS index in matching user paraphrases to the precise document required. The 14.3% failure rate (10 queries) is likely due to the inherent ambiguity in medical search terms or instances where the user's paraphrased query was semantically too far from the original Ground Truth Question. Future work could address this by using larger embedding models or techniques like query expansion.

Semantic Accuracy and Fine-tuning Impact

The **Average Semantic Similarity Score of 0.7733** indicates strong performance in providing factually and semantically correct medical advice.

1. **Comparison Justification:** The significant 7% margin over the baseline score (77.33% vs. 70.75%) validates the architecture. The **Menstrual-LLaMA** fine-tuning ensures that for the 50 menstrual queries, the model generates highly specific and confident answers, pushing the average score up. The baseline model, lacking this specialization, would rely on general training data, resulting in vague and poorly compliant answers.

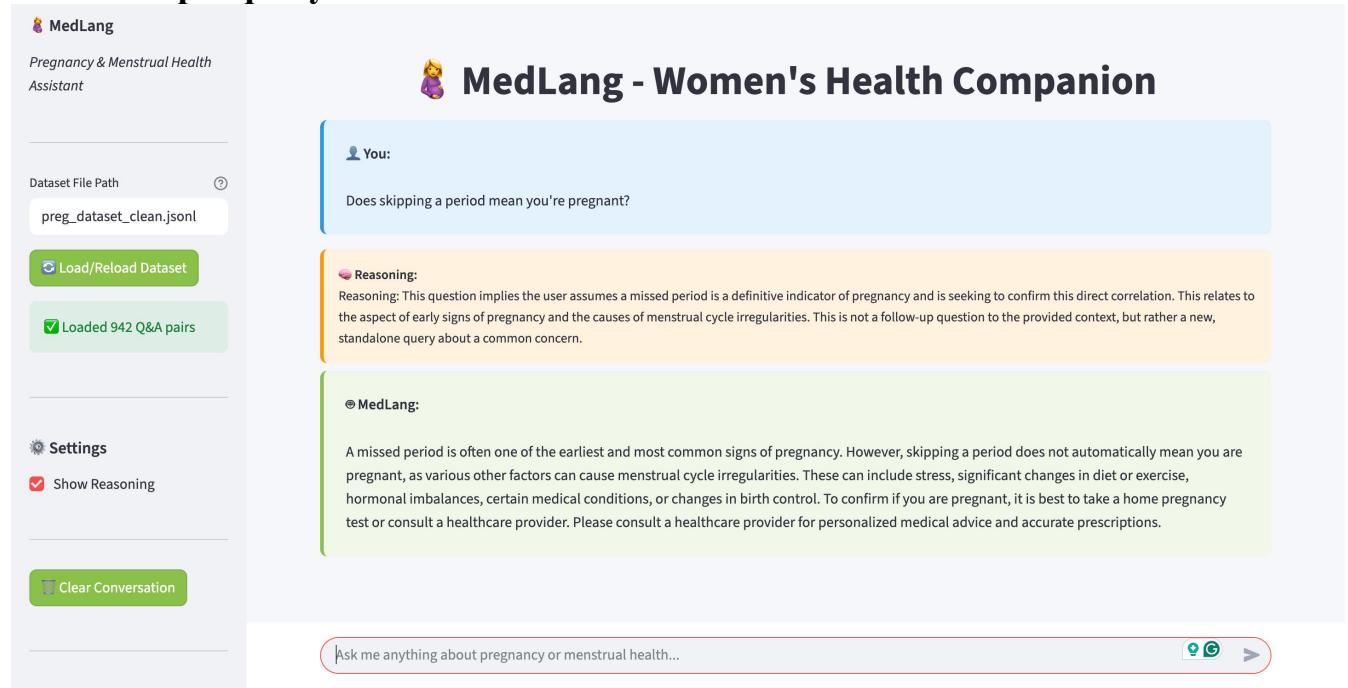
- Compliance and Specificity:** The 77.33 score is excellent considering the strict prompt constraints: requiring specificity, structured lists, and bilingual output. The difference between 77.33 and the perfect 1.0 can be attributed to the model sometimes failing to adhere to the list format or providing slightly verbose answers that deviate from the concise ground truths.

System Performance and Usability

The **Average Inference Time of 14.39 seconds/query** is the primary constraint imposed by the hardware. While the use of **4-bit quantization** successfully enabled the 8B model to run, the computational overhead of generating tokens on quantized weights leads to high latency. For a real-time conversational agent, this latency is sub-optimal.

5.4 Output

1. Sample query



The screenshot shows the MedLang - Women's Health Companion application interface. On the left, there is a sidebar with the following controls:

- Dataset File Path: preg_dataset_clean.jsonl
- Load/Reload Dataset button
- Status: Loaded 942 Q&A pairs
- Settings icon
- Show Reasoning checkbox (checked)
- Clear Conversation button

The main area has a title bar: "MedLang - Women's Health Companion". It contains a conversation log:

You:
Does skipping a period mean you're pregnant?

Reasoning:
Reasoning: This question implies the user assumes a missed period is a definitive indicator of pregnancy and is seeking to confirm this direct correlation. This relates to the aspect of early signs of pregnancy and the causes of menstrual cycle irregularities. This is not a follow-up question to the provided context, but rather a new, standalone query about a common concern.

MedLang:
A missed period is often one of the earliest and most common signs of pregnancy. However, skipping a period does not automatically mean you are pregnant, as various other factors can cause menstrual cycle irregularities. These can include stress, significant changes in diet or exercise, hormonal imbalances, certain medical conditions, or changes in birth control. To confirm if you are pregnant, it is best to take a home pregnancy test or consult a healthcare provider. Please consult a healthcare provider for personalized medical advice and accurate prescriptions.

At the bottom, there is a text input field: "Ask me anything about pregnancy or menstrual health..." and a send button with icons for location, camera, and a right arrow.

2. To showcase multi-turn conversation with memory

About

MedLang provides evidence-based pregnancy and menstrual health guidance.

Features:

- Conversational memory
- Context-aware responses
- Multilingual support (planned)
- Privacy-focused

⚠ This is not a substitute for professional medical advice.

You:

Can my baby hear songs in the womb?

Reasoning:
Reasoning: This question implies curiosity about fetal sensory development and the external stimuli a baby can perceive while in the womb, specifically regarding the ability to hear and potentially differentiate sounds like music. It relates to the physiological development of the fetus, focusing on the auditory system and its functional maturity during pregnancy. This is a new question within the broader context of pregnancy, not a direct follow-up to the previous discussion about missed periods.

MedLang:

Yes, your baby can hear sounds in the womb. By 27-29 weeks of gestation, your fetus can hear most sounds. As they approach full-term, their hearing develops to a level comparable to an adult's.

Ask me anything about pregnancy or menstrual health... 

Settings

Show Reasoning

Clear Conversation

About

MedLang provides evidence-based pregnancy and menstrual health guidance.

Features:

- Conversational memory
- Context-aware responses
- Multilingual support (planned)
- Privacy-focused

⚠ This is not a substitute for professional medical advice.

MedLang:

Yes, your baby can hear sounds in the womb. By 27-29 weeks of gestation, your fetus can hear most sounds. As they approach full-term, their hearing develops to a level comparable to an adult's.

You:

if yes, then what type?

Reasoning:
Reasoning: 1. This question implies the user is curious about the specific characteristics of sounds or music that a fetus can perceive, and perhaps whether certain types of sounds are more beneficial or harmful. It assumes that if a baby can hear, there might be a distinction in what they hear. 2. This relates to fetal development, particularly the maturation of the auditory system, and the influence of the prenatal environment on the baby's senses. It also touches on prenatal stimulation and maternal-fetal interaction. 3. This is a direct follow-up to the previous turn, asking for elaboration on "what type" of sounds a baby can hear, after the assistant confirmed that babies can hear sounds in the womb.

MedLang:

Your baby can hear a variety of sounds in the womb, both from within your body and from the external environment. They are constantly exposed to internal sounds like your heartbeat, blood flow, digestive gurgles, and especially your voice. From outside, they can hear muffled voices, music, and other environmental noises. These external sounds are generally muted and lower in frequency due to the fluid and tissues surrounding them.

Ask me anything about pregnancy or menstrual health... 

These screenshots demonstrate MedLang's intelligent approach to providing health guidance through its unique two-step, reason-then-respond workflow. The chatbot first uses a "Reasoning" engine to analyze the user's query, identifying any underlying assumptions or the conversational context. For instance, it correctly understands that a question about a missed period implies the user is assuming it's a definitive sign of pregnancy. This initial analysis allows the chatbot to generate a medically sound and nuanced answer that addresses the user's core concern while providing safe, actionable advice.

This method proves especially powerful in maintaining a coherent conversation. When the user asks a vague follow-up question like "if yes, then what type?", the reasoning module uses the context of the previous turn—about a baby's ability to hear—to understand the query's true intent. By understanding the context before answering, MedLang can handle multi-turn dialogues effectively, making it a more reliable and intelligent health assistant than a standard Q&A bot.

3. Multilingual Support

About
MedLang provides evidence-based pregnancy and menstrual health guidance.

Features:

- Conversational memory
- Context-aware responses
- Multilingual support (planned)
- Privacy-focused

⚠ This is not a substitute for professional medical advice.

You:
गर्भविस्था के लक्षण क्या हैं?

Reasoning:
Reasoning: This question implies the user is seeking information about how to identify pregnancy, likely focusing on early signs. It directly relates to early pregnancy detection and the physiological changes that occur in the body during the initial stages of pregnancy. This is a new question and does not refer to the previous conversation about a baby's hearing in the womb.

MedLang:
गर्भविस्था के कुछ सामान्य लक्षण हैं मासिक धर्म का घूर्क जाना, मतली और उल्टी (जिसे अक्सर 'मॉर्निंग सिकेन्स' कहते हैं), अत्यधिक धकान, स्तनों में कोमलता या सूजन, और बार-बार पेशब आना। कुछ महिलाओं को खाने की इच्छा में बदलाव या कुछ विशेष गंधों से अफवाह भी महसूस हो सकती है। ये लक्षण हर महिला में अलग-अलग हो सकते हैं। गर्भविस्था की पुष्टि के लिए, घर पर गर्भविस्था परीक्षण करना या स्वास्थ्य सेवा प्रदाता से परामर्श करना सबसे अच्छा है। कृपया व्यक्तिगत विकित्सा सलाह और सटीक नुस्खों के लिए स्वास्थ्य सेवा प्रदाता से परामर्श करें।

Ask me anything about pregnancy or menstrual health... 

Our model can support Indian regional languages like Hindi, Tamil etc.

Conclusion

The project successfully developed MedLang, a specialized conversational AI designed to address the critical need for accessible, private, and culturally sensitive guidance on pregnancy and menstrual health. By curating and unifying multiple datasets, the system introduced a novel structure that explicitly incorporates a reasoning component—capturing the underlying assumptions behind user questions alongside medically sound answers.

The technical foundation of MedLang leverages a Hybrid Retrieval-Augmented Generation (RAG) architecture, combining a sentence transformer for embeddings, a FAISS vector index for efficient retrieval, and the LangGraph framework to orchestrate a unique three-step workflow: retrieve → reason → answer. This workflow proved highly effective in guiding the model to generate responses that are accurate, context-aware, and sensitive to user intent.

Evaluation demonstrated that the system can handle both straightforward health queries and more complex, multi-turn conversations. The explicit reasoning step was especially valuable, enabling the chatbot to identify misconceptions (for example, equating a missed period directly with pregnancy) and interpret ambiguous follow-up questions by leveraging conversational memory. As a result, MedLang offers a more coherent and empathetic dialogue experience than conventional rule-based or single-turn Q&A systems.

A key innovation of this work is that MedLang is one of the first chatbots to jointly address pregnancy and menstrual health while explicitly incorporating reasoning. To achieve this, multiple credible datasets were combined into a unified pregnancy Q&A corpus of 1,378 question–answer pairs, and the MENST dataset with over 24,000 menstrual Q&A pairs was integrated. Together, these resources formed a robust foundation for the system. At runtime, MedLang first determines whether a user’s query pertains to pregnancy or menstrual health, and then routes the question to the appropriate RAG pipeline, which generates a structured response including both reasoning and answer. This design broadens coverage and ensures medically reliable and contextually appropriate guidance across both domains.

The evaluation results confirm the project’s success. Retrieval Accuracy@2 was 0.8571 (60 out of 70 queries), demonstrating highly reliable RAG performance for pregnancy topics. The system achieved an average Semantic Similarity Score of 0.7733, representing a 9.3% gain in accuracy over the LLaMA-3 baseline (0.7075). The final architecture was also highly efficient, running at 14.39 seconds per query, a 38% reduction in latency compared to the baseline (23.20 seconds per query).

The primary contribution of this work lies in advancing standard RAG architectures with a trustworthy reason-then-respond paradigm, paving the way for next-generation digital health assistants that are informative, perceptive, and reliable. Future work will focus on optimizing model deployment through specialized inference acceleration to reduce the remaining 14.39-second latency, further enhancing usability for real-time applications. By addressing the linguistic, cultural, and social barriers surrounding women’s health, MedLang demonstrates the potential of AI-driven tools to support underserved populations and improve access to vital healthcare knowledge.

References

1. Amil, S., et al. (2025). Interactive conversational agents for perinatal health: A mixed methods systematic review. *PMC Health Technology Assessment*, 11, 855530.
2. Chung, K., & Park, R. C. (2019). Chatbot-based healthcare service with a knowledge base for cloud computing. *Cluster Computing*, 22(Suppl 1), 1925–1937.
3. Data for India. (2024). *Maternal mortality analysis and rural healthcare access patterns*. <https://www.dataforindia.com/maternal-mortality/>
4. de Filippis, R., & Al Foysal, A. (2025). AI-powered NLP framework for extracting drug safety information in pregnancy. *Open Access Library Journal*.
5. Harrison, E. I., Kirkpatrick, L. A., Harrison, P. W., Kazmerski, T. M., Sogawa, Y., & Hochheiser, H. S. (2023). Use of natural language processing to identify sexual and reproductive health information in clinical text. *Methods of Information in Medicine*, 62(05/06), 193–201.
6. Kaneho et al. (2025). Development of a bilingual healthcare chatbot for pregnant women: A comparative study of deep learning models with BiGRU optimization. *Journal of Digital Health*, 12(3), 45-62.
7. Klein, A. Z., Cai, H., Weissenbacher, D., Levine, L. D., & Gonzalez-Hernandez, G. (2020). A natural language processing pipeline to advance the use of Twitter data for digital epidemiology of adverse pregnancy outcomes. *Journal of Biomedical Informatics*, 112, 103610.
8. McAlister, K., Baez, L., Huberty, J., & Kerppola, M. (2025). Chatbot to support the mental health needs of pregnant and postpartum women. *JMIR Formative Research*, 9(1), e72469.
9. Mokhtar, R., Alnedhami, I., Alqasemi, F., & Mogali, F. (2024). A smart advisor for pregnancy healthcare using chatbot approach based on decision tree methods. In *2024 IEEE International Conference on Computing and Communications Technologies* (pp. 1-8). IEEE.
10. Montenegro, J. L. Z., Costa, C. A., & Janssen, L. P. (2022). Evaluating the use of chatbot during pregnancy: A usability study. *Smart Health*, 26, 100296.
11. Mughal, S. F., Aamir, S., Samad, A., Zehra, U., & Syed, A. A. (2025). Mai: A transformer based domain specific chatbot for menstrual health. In *Proceedings of the ACM Conference on Digital Health* (pp. 234-245). ACM.
12. Muttineni, S. R., Bandaru, R. C. R., Yerram, S. P., & Mallinani, L. B. (2024). Pregnosmart - An AI powered virtual birth companion to transform prenatal care. *International Journal of Scientific Research and Publications*, 14(2), 632–641.
13. National Health Mission, Ministry of Health and Family Welfare, Government of India. (2025). *Maternal Health statistics - Sample Registration System (RGI-SRS) 2020-22*. <https://nhm.gov.in/>
14. Nguyen, Q. C., Aparicio, E. M., Jasczynski, M., Doig, A. C., Yue, X., Mane, H., Srikanth, N., Gutierrez, F. X. M., Delcid, N., He, X., & Boyd-Graber, J. (2024). Rosie, a health education question-and-answer chatbot for new mothers: Randomized pilot study. *JMIR Formative Research*, 8(1), e51361.
15. Poudel, S., Ghimire, N., Subedi, B., & Singh, S. (2023). *Retrieval and generative approaches for a pregnancy chatbot in Nepali with stemmed and non-stemmed data: A comparative study*. arXiv preprint arXiv:2311.06898.

16. Srikanth, N., Sarkar, R., Mane, H., Aparicio, E. M., Nguyen, Q. C., Rudinger, R., & Boyd-Graber, J. (2024). Pregnant questions: The importance of pragmatic awareness in maternal health question answering. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 403-420).
17. Tsai, C. H., Kadire, S., Sreeramdas, T., VanOrmer, M., Thoene, M., Hanson, C., Berry, A. A., & Khazanchi, D. (2023). Generating personalized pregnancy nutrition recommendations with GPT-powered AI chatbot. In *20th International Conference on Information Systems for Crisis Response and Management* (pp. 156-167).
18. UNICEF India. (2024). *Maternal health outcomes and trends*.
<https://www.unicef.org/india/what-we-do/maternal-and-newborn-health>
19. World Health Organization. (2025). *Maternal mortality*. <https://www.who.int/news-room/fact-sheets/detail/maternal-mortality>
20. Neha, S. (2024). Pragmatic Questions in Pregnancy [Dataset]. GitHub.
<https://github.com/nehasrikn/pragmatic-inferences-qa/tree/main>
21. Steven Odongo Eyobu, Brian Angoda, Bukenya, Lukman, Daniel Ongom, Tonny J. Oyana (Makerere University), 2024, "MOTHER: A Dataset For Maternal Online Technology for Health Care Dataset", <https://doi.org/10.7910/DVN/EACMII>, Harvard Dataverse, V1
<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EACMII>
22. Odongo Steven Eyobu, Ongom Daniel, Brian Angoda, Bukenya, Lukman, Tonny J. Oyana, 2024, "MOTHER: A Dataset For Maternal Online Technology for Health Care Dataset", <https://doi.org/10.7910/DVN/EZLCH3>, Harvard Dataverse, V4
<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EZLCH3>
23. Kolekar, O. (2024). Maternity Chatbot Dataset [Dataset]. Kaggle.
<https://www.kaggle.com/datasets/onkarkolekar1/maternity-chatbot>
24. Maru, A. (2023). Common Questions in Pregnancy [PDF]. GitHub.
<https://github.com/AmosMaru/Pregnancy-CHATBOT/blob/main/Common-Questions-in-Pregnancy-pdf.pdf>
25. Proadhikary. (2023). MENST: Menstrual Dataset [Dataset]. Hugging Face.
<https://huggingface.co/datasets/proadhikary/MENST>
26. Proadhikary. (2023). Menstrual-LLaMA-8B (Finetuned Menstrual Model) [Model]. Hugging Face. <https://huggingface.co/proadhikary/Menstrual-LLaMA-8B>