

MatruCare-AI: A RAG-Augmented Multi-Agent Decision Support & Framework For ASHA-Mediated High-Risk Pregnancy Detection in Rural India

Aman Gupta¹, Himanshu Gupta², Anshu³, Dr. Tripti Choudhary⁴

^{1,2,3,4}Department of Computer Science and Engineering, Hi-Tech Institute of Engineering and Technology, Ghaziabad, Uttar Pradesh, India – 201015

Abstract—Maternal mortality in rural India remains a persistent public health challenge, with the country recording 97 deaths per 100,000 live births [13]. Geographic inaccessibility of tertiary care facilities means that a large proportion of pregnant women in remote areas complete their entire antenatal period without consulting a physician [15]. Accredited Social Health Activists (ASHAs) constitute the primary interface between rural pregnant women and the formal health system, yet a meta-analysis of 37 studies established that ASHAs respond correctly to maternal health queries at a rate of approximately 62% [1]. Epidemiological surveys indicate that around 40% of rural pregnancies carry some degree of risk, whereas national programme records flag only 14% as high-risk [14], implying that the majority of at-risk mothers are not identified before complications develop.

Existing digital tools deployed for ASHAs, including e-SAATHI and ASHA-Soft, rely on pre-programmed rule sets that cannot reason beyond their fixed parameters, detect evolving risk patterns across consecutive antenatal visits, or adapt when national clinical guidelines are revised [16], [17].

This paper proposes MatruCare-AI, a three-layer framework designed specifically for ASHA workers in rural India. Layer 1 is a MERN-based Progressive Web Application that records antenatal parameters and computes a four-tier risk classification. Layer 2 is a Retrieval-Augmented Generation (RAG) clinical engine that grounds advisory responses in India's own governmental health protocols, including NHM Maternal Health guidelines, PMSMA clinical criteria, and WHO ANC 2016 recommendations. Layer 3 operates three parallel decision-support agents responsible for danger-sign detection, referral routing to the nearest Primary Health Centre, and antenatal compliance scheduling. This paper makes three contributions. First, to the best of the authors' knowledge, MatruCare-AI represents the first

framework to combine RAG-based clinical advisory with multi-agent decision support built specifically for ASHA-mediated high-risk pregnancy detection in rural India. Second, the proposed clinical knowledge base draws exclusively from Indian governmental protocols, ensuring contextual alignment with the NHM programme environment. Third, a systematic gap matrix confirms that no existing framework simultaneously satisfies all five required dimensions: ASHA focus, RAG advisory, HRP detection, rural India context, and multilingual support.

Index Terms—ASHA workers, high-risk pregnancy, maternal health, rural India, retrieval-augmented generation, multi-agent system, clinical decision support, antenatal care, mobile health, large language models

I. INTRODUCTION

Pregnancy-related complications claim the lives of thousands of mothers in India each year, with the national maternal mortality ratio standing at 97 per 100,000 live births [13]. Rural areas bear a disproportionate burden of these deaths, primarily because tertiary care facilities remain geographically inaccessible for a substantial share of the rural population. Severe postpartum haemorrhage, hypertensive disorders of pregnancy, and puerperal sepsis collectively account for the majority of maternal fatalities, yet all three conditions respond well to timely clinical intervention when warning signs are detected early [13], [15].

To extend healthcare reach into remote communities, the Government of India established the ASHA programme in 2005. With more than one million workers conducting home-based antenatal visits nationally, ASHAs serve as the primary and often only

trained health contact that rural pregnant women encounter [3]. Evidence indicates that consistent ASHA engagement raises the probability of institutional delivery 1.6-fold relative to women who receive no such support [3]. Despite this positive impact, a systematic review synthesising findings from 37 studies across India found that ASHAs demonstrate correct maternal health knowledge at a rate of 62%, with particularly pronounced deficits in blood-pressure interpretation, pre-eclampsia recognition, and timely referral decisions [1]. A subsequent evaluation confirmed that structured decision-support raised ASHA referral accuracy from 60% to 88% [2], establishing that knowledge gaps are bridgeable with appropriate tools rather than being intrinsic to the workforce.

Epidemiological surveys in rural India report high-risk pregnancy (HRP) prevalence around 40%, contrasting sharply with national programme records that classify only 14% of pregnancies as high-risk [14]. This detection gap means that a large proportion of at-risk women pass through their entire antenatal period without receiving the additional monitoring warranted by their clinical profile. Mo-bile digital tools such as e-SAATHI and ASHA-Soft provide ASHA workers with structured checklists, yet a comprehensive scoping review found that such tools rely exclusively on fixed rule sets and lack any capacity for adaptive clinical reasoning [16], [17]. Reviews of mHealth deployments in India's high-burden states further document that ASHA workers discontinue tool use when applications fail to integrate with their home-visit workflow [18]. Critically, no deployed system currently tracks longitudinal vital-sign trajectories to identify slow-onset danger signs such as developing pre-eclampsia before they reach emergency thresholds.

This paper proposes MatruCare-AI as a framework to address this gap. The system operates across three interdependent layers tailored to the ASHA's antenatal workflow. Layer 1 is a MERN-based Progressive Web Application for antenatal data capture and automated risk stratification. Layer 2 is a RAG clinical engine that retrieves relevant sections from a curated corpus of Indian governmental health protocols and generates guideline-grounded advisory responses, using Mistral-7B or Qwen-2.5 as the generation model [5], [8]. Layer 3 deploys three parallel agents the DangerSign Detector, Referral Routing Agent, and Follow-up

Scheduler that translate risk scores into concrete, actionable instructions for the ASHA in the field [19], [21]. The paper offers three contributions. First, to the best of the authors' knowledge, MatruCare-AI constitutes the first framework combining RAG-based clinical advisory with multi-agent decision support designed exclusively for ASHA-mediated HRP detection in rural India. Second, all clinical knowledge is sourced from Indian governmental guidelines, ensuring protocol alignment with the NHM programme under which ASHAs operate [9]. Third, the gap matrix presented in Section III confirms that no prior system satisfies all five required dimensions simultaneously [2], [17].

II. LITERATURE REVIEW

A. ASHA Workers and Maternal Health

The ASHA programme, launched under the National Rural Health Mission in 2005, was designed to position a trained community health worker within every village to support pregnant women through home visits and facilitate referrals to formal health facilities [25]. The programme now encompasses more than one million workers across rural India [3]. Empirical evidence confirms the value of ASHA engagement: women who receive home visits are 1.6 times more likely to deliver in an in-situational setting rather than at home without skilled assistance [3]. In Uttar Pradesh, ASHA visits during the third trimester were associated with a 1.32-fold increase in the likelihood of delivery at a public health facility [22]. Notably, the protective effect was strongest among women with little or no formal education [22], underscoring the program's equity function. Research from Madhya Pradesh and Bihar demonstrated that combined counselling by an ASHA and an Anganwadi worker was associated with higher rates of completing four or more antenatal check-ups, birth preparedness, and postnatal care attendance [4]. A systematic review and meta-analysis of 31 studies reported that ASHA support corresponded with a 26% increase in antenatal visits, a 12% rise in institutional deliveries, and a 13% improvement in postnatal care attendance [26]. Alongside these documented benefits, the programme carries well-documented structural constraints. In Madhya Pradesh, the mean institutional delivery rate attributable to ASHA support reached only 59.62%, with tribal and road-distant villages performing

considerably below this average [23]. The knowledge meta-analysis established that the 62% correct-response rate was particularly low for blood-pressure interpretation, pre-eclampsia identification, and referral timing [1]. A mixed-methods study found that when ASHAs and traditional birth attendants simultaneously visited the same house-hold, conflicting advice sometimes led women to delay care-seeking [27]. Structural factors including heavy workloads, inadequate remuneration, and irregular training cycles compound these knowledge deficits [24]. Taken together, these studies indicate that ASHAs provide genuine public health value but require better knowledge-support tools to reliably identify at-risk pregnancies, a gap that the MatruCare-AI framework is designed to address [2], [26].

B. Retrieval-Augmented Generation in Healthcare

Standard large language models (LLMs) generate responses from parameters fixed at training time, which renders them susceptible to factual errors and confident production of statements unsupported by current evidence a phenomenon referred to as hallucination [5]. Retrieval-Augmented Generation (RAG) addresses this limitation by coupling a dense retriever with a parametric generator: at inference time, the system retrieves the most relevant passages from a curated external corpus and conditions the generated response on that retrieved evidence rather than on parametric memory alone [28]. Because the knowledge source is external, the system can incorporate updated clinical guidelines without retraining the underlying model [6].

Clinical evaluation of RAG across healthcare applications has returned consistent evidence of accuracy improvement. A systematic review covering 30 clinical RAG studies identified medical question-answering, patient-record summarisation, and diagnostic support as the dominant use cases [28]. That review also found that 78.9% of published clinical RAG studies used exclusively English datasets [5], signalling that multilingual and low-resource health systems including rural India remain largely unaddressed by current RAG research. The MEGA-RAG framework, applied to public health question-answering, reduced hallucination rates by more than 40% relative to an unaugmented baseline and achieved an accuracy score of 0.79 [29]. In radiology consultation, a locally deployed RAG model produced

zero hallucinations versus an 8% hallucination rate for an equivalent model operating without retrieval, while responding 2.6 times faster than a cloud-hosted commercial alternative [32]. The MedRAG benchmark, which evaluated 41 RAG configurations across 7,663 medical questions, found that retrieval augmentation improved accuracy by up to 18 percentage points and elevated weaker models to the performance level of larger, more expensive ones [33]. Applied to governmental clinical guidelines, a RAG system achieved 99.5% source faithfulness, with the best-performing model answering 98.7% of queries correctly and reducing unsafe responses by 67% [81]. A self-reflective RAG variant that validates its own outputs before returning a response constrained hallucinations to 5.8% across 250 patient cases [30]. Across these studies, two properties of RAG are of direct relevance to MatruCare-AI: its capacity for continuous knowledge updating without model retraining, and its systematic reduction of clinically unsafe responses [6], [7].

C. Large Language Models for Frontline Health Workers

Community health workers (CHWs) in low- and middle-income countries routinely make consequential clinical judgements during domiciliary visits, without access to a supervising clinician [35]. Emerging evidence suggests that LLMs can meaningfully augment CHW decision-making capacity in these settings. A prospective study in Rwanda evaluated five LLMs against 5,609 clinical queries sub-mitted by 101 CHWs across four districts; the two highest-performing models exceeded the accuracy of local clinicians across every evaluation metric, at an estimated cost more than 500-fold lower than a clinician consultation [10]. The SMART health GPT system, constructed using a RAG pipeline over Indian pregnancy clinical guidelines, achieved maternal health query accuracy exceeding 88% [9]. An independent quality assessment in which 47 specialist physicians across three countries rated AI-generated maternal health responses assigned GPT-4 and GPT-3.5 mean scores of 3.9 out of 5.0 for both content accuracy and communicative clarity [37].

Conversational AI tools for pregnancy monitoring have demonstrated user retention rates between 62% and 78% over six-month periods, with diagnostic sensitivity for gestational diabetes and pre-eclampsia

warning signs producing area-under-the-curve scores between 0.88 and 0.94 [40].

However, significant limitations have also been documented. An evaluation of GPT-4o across maternal health questions spanning 204 countries found that the model fabricated country-specific health statistics in 44% of cases where such data were unavailable, and produced incorrect source citations at a rate of 46.7% [36]. The majority of LLM development and benchmarking has occurred in high-income settings, producing systems that do not reflect the infrastructure, linguistic diversity, or clinical protocols of rural India [12], [38]. A policy roadmap for equitable LLM adoption in low-resource contexts identified five prerequisite conditions: local-language support, affordable hardware, locally representative training data, regulatory frameworks, and participatory co-design with community stakeholders [38]. Without these, accuracy gains demonstrated in high-income settings are unlikely to translate to rural Indian deployment contexts [12], [39]. A prospective protocol study is currently underway in Rwanda to evaluate whether LLM-based decision support can improve CHW referral decisions in live clinical practice [11].

D. High-Risk Pregnancy Detection Using AI

High-risk pregnancies frequently present as gradual physiological deterioration across consecutive antenatal visits rather than acute single-point events [41]. Machine learning models that track longitudinal biomarker trajectories blood pressure, hemoglobin, weight gain, and blood glucose have demonstrated the capacity to detect developing risk patterns that individual-visit rule-based thresholds miss [41], [48]. A study applying machine learning to India's NFHS-5 national health survey data found that a Gradient Boosting classifier predicted HRP with 89% accuracy and 90% recall; the most predictive variables were inter-pregnancy interval, anaemia status, maternal age, and educational level [13] parameters that ASHA workers already document at home visits. A multilayer neural network trained on records from 1,014 clinical cases reached 91% accuracy for the high-risk category and processed 500 records per second [41]. An ensemble stacking approach combining multiple base classifiers achieved 98% accuracy with 97% high-risk recall [47], while a CatBoost model reached 94.2% accuracy with 96.2% HRP recall [88]. In India

specifically, an AI-based mHealth application deployed at primary health centres in Pune screened 1,010 women and identified 24.85% as high-risk [44] a proportion substantially higher than the 14% captured by routine national health records [14]. In rural Guatemala, machine learning models matched the HRP referral accuracy of trained nurses, with blood pressure and blood glucose emerging as the two strongest predictors [43]. An explainability review using SHAP and LIME methods confirmed that systolic blood pressure, diastolic blood pressure, and fasting blood sugar are the most consistent HRP predictors across heterogeneous datasets and geographic contexts [48]. An important counterpoint is raised by Ji and Wang, who note that AI-generated risk labels can induce predictive anxiety in pregnant women and risk over-flagging socioeconomically disadvantaged populations if training data carry pre-existing health inequities [15]. These findings reinforce the principle that AI should function as decision support for the health worker rather than as an autonomous classification system [15], [42].

E. Digital Health and Antenatal Care in India

Digital health interventions for antenatal care have expanded substantially over the past decade, yet published evidence remains geographically skewed toward high-income settings. A scoping review of 126 studies found that 61% originated from the United States, United Kingdom, or Australia [49]. A systematic review and meta-analysis of six randomised controlled trials in low- and middle-income countries reported that mHealth interventions were associated with a 1.66-fold increase in antenatal care attendance relative to standard care [50]. A broader analysis of 131 studies found that text-message reminder systems improved ANC attendance and child vaccination uptake but produced no demonstrable effect on facility delivery rates [51], suggesting that reminder functionality alone is insufficient to close the gap in HRP identification. A review of maternal mHealth applications established that 81% were designed for high-income country contexts, with low-resource settings substantially under-represented in the evidence base [52].

Within India, a systematic review of 22 mHealth tools found that most transmitted one-directional SMS messages and lacked any interactive question-answering capability [53]. Women in these studies

cited shared phone ownership, low literacy, and technical barriers as primary reasons for inconsistent engagement [53]. A study introducing a WHO ANC-aligned mobile application at three Indian health centres reported that users demonstrated significantly higher scores for clinical history-taking, physical examination, and patient counselling than those who received no digital support [54]. A systematic review of mHealth evidence from Bihar, Madhya Pradesh, Rajasthan, and Uttar Pradesh India's four highest-burden states identified only eight published studies for the entire region, none of which included an economic evaluation [18]. This constitutes a notably sparse evidence base given that these four states collectively account for a disproportionate share of India's maternal deaths. The e-SAATHI programme in Assam enrolled more than 200,000 women, demonstrating that chat-based systems can achieve rapid, large-scale uptake; however, the system operates on fixed scripted rules and cannot reason through clinical presentations it was not pre-programmed to handle [17]. No published system reviewed for this paper integrates AI-guided clinical advisory, automated HRP detection, and a dedicated interface for ASHA workers within a unified framework [16]–[18].

III. RESEARCH GAP ANALYSIS

The five literature clusters reviewed in Section II are internally coherent and each demonstrates substantive contributions to its respective domain. Nonetheless, a consistent structural gap emerges when systems are evaluated across the dimensions that collectively determine fitness for ASHA-mediated HRP detection in rural India. No reviewed system satisfies all five requirements simultaneously [16]–[18]. These five dimensions are defined in Table I, and their distribution across existing systems is mapped in the gap matrix presented as Table II.

Each dimension in Table I is grounded in published evidence and reflects a specific operational requirement that must be satisfied for a system to function reliably during a domiciliary antenatal visit.

The paragraphs below examine each gap in sequence. Gap 1 Absence of ASHA-Centred Workflow Design. The dominant paradigm in digital health tool design positions the facility-based clinician, rather than the

frontline community worker, as the primary user. ASHA Kirana provided symptom-documentation functionality on a mobile device but did not return clinical guidance to the ASHA at the point of care [56]. A cross-disciplinary scoping review of CHW digital tools across low- and middle-income countries found that most systems were architected for clinic environments and did not accommodate the decision-making demands of domiciliary visits conducted without physician oversight [58]. When tools fail to align with the CHW's daily workflow, evidence consistently shows workers abandon them in favour of familiar practices [72]. While SMARThealth GPT was explicitly designed for ASHA users, it incorporates neither automated danger-sign detection nor structured referral routing for high-risk presentations [56].

TABLE I The Five Gap Dimensions: What They Mean and Why They Matter

#	Dimension	What It Means	Why It Matters
1	ASHA Focus	Built around what an ASHA does at a home visit	ASHAs are the only trained contact most rural pregnant women ever meet [1], [3]
2	RAG Advisory	Answers come from real government guidelines, not guessing	Avoids wrong advice; updates without rebuilding the tool [5], [8]
3	HRP Detection	Automatically detects a pregnancy becoming high-risk	40% of pregnancies are risky but only 14% are ever flagged [13], [14]
4	India Rural	Designed for NHM, PMSMA and SUMAN programme conditions	Tools developed elsewhere do not align with rural India's operational context [61], [64]
5	Multilingual	Works in Hindi and India's regional languages, not only English	ASHAs and mothers communicate across diverse languages and dialects [62], [71]

Gap 2 Lack of RAG-Based Clinical Advisory. Deployed ASHA support tools operate on pre-defined rule sets that enumerate specific conditional responses. None incorporates a RAG pipeline capable of retrieving relevant passages from India's own governmental health protocols and generating source-grounded advisory responses to novel clinical queries [5], [6]. A systematic review of multi-agent RAG systems for clinical decision support confirmed that retrieval-plus-reasoning architectures substantially outperform static rule-based systems, yet almost no such architecture has been deployed for community health workers in low-resource settings [65]. Without retrieval grounding, rule-based tools cannot incorporate updated national guidelines without a full system rebuild [7], [8]. An umbrella review further confirmed a persistent gap between laboratory-validated AI capability and the tools actually available to frontline workers [59].

Gap 3 No Automated High-Risk Pregnancy Detection. Existing tools address either antenatal data collection or clinical advisory, but no reviewed system combines both functions with an automated mechanism that identifies progressive HRP risk from longitudinal vital-sign records across multiple visits [41], [59]. Published evidence confirms that machine learning models can predict HRP with up to 89% accuracy using precisely the parameters blood pressure, haemoglobin, weight, maternal age that ASHAs already record [13]. Despite this, the capability has not been integrated into any tool accessible to field-level workers [60], [63]. A review of AI applications in maternal and neonatal health in low-resource countries found that predictive models consistently remain within research publications rather than reaching the CHWs for whom they would be most valuable [63].

Gap 4 Insufficient India-Rural Contextualisation. A large proportion of maternal health AI tools were developed and validated in high-income settings whose infrastructure, language environments, and clinical protocols bear little resemblance to rural India. A review of digital health programmes in Africa and other low-income contexts found that systems developed without participatory community input routinely failed to achieve adoption [57]. A study examining why AI health interventions fail to scale beyond pilots in India identified five recurring structural barriers: interoperability deficits, post-pilot funding discontinuation, insufficient policy support, low CHW trust in digital tools, and infrastructure gaps between design assumptions and field realities [64]. No existing maternal health AI framework has been designed within the NHM, PMSMA, and SUMAN programme structures that govern ASHA operations in India [17], [18].

Gap 5 Absence of Multilingual Support. India's linguistic diversity means that the ASHA workforce operates across dozens of distinct languages and dialectal variants. None of the AI systems reviewed in this paper supports more than one Indian language for clinical advisory output [9], [56]. The ArogyaSutra multi-agent system, evaluated across seven Indian languages, found that models adapted for Indic linguistic contexts produced substantially more accurate medical responses than English-only models deployed in Indian settings [71]. A study focused on tribal women in India reported that language mismatch was among the primary determinants of non-engagement with digital health tools [62]. The absence of local-language support constrains any tool's utility to ASHA workers proficient in the implementation language, while leaving out the mothers they are trying to reach.

TABLE II Gap Analysis Matrix: Existing Systems vs. MatruCare-AI Across Five Dimensions

System	ASHA Focus	RAG Advis.	HRP Detect.	India Rural	Multi-lingual	Key Limitation
SMARThealth GPT [9]	✓	✓	×	✓	×	No HRP detection or referral
AI-ADSS [2]	✓	×	~	✓	×	Rule-based; partial HRP only
e-SAATHI [17]	×	×	×	✓	×	Fixed rules; no AI; no HRP
ASHA Kirana [56]	✓	×	×	✓	×	Data only; no clinical advice

ArogyaSutra [71]	×	✓	×	~	✓	Not for ASHAs; no HRP
Multi-Agent Triage [20]	×	✓	×	×	×	Hospital emergency only; not rural
TriAgent [21]	×	✓	×	×	×	Crisis triage only; not maternal
MatruCare-AI (Proposed)	✓	✓	✓	✓	✓	All 5 dimensions covered

✓ = Fully Supported × = Not Supported ~ = Partial Table II makes the distribution of coverage across existing systems explicit. SMARThealth GPT satisfies three of five dimensions and represents the closest prior analogue to MatruCare-AI, yet it lacks both HRP detection capability and local-language support [9]. AI-ADSS was purpose-built for ASHAs and demonstrated measurable referral improvement, but its rule-based architecture precludes guideline-grounded reasoning for unfamiliar clinical presentations [2]. e-SAATHI demonstrated the scalability of chat-based deployment in India, enrolling more than 200,000 women, while remaining constrained to scripted interactions that cannot handle novel cases [17]. ASHA Kirana enables structured data capture at home visits but returns no clinical guidance to the ASHA in the moment [56]. Multi-Agent Triage [20] and TriAgent [21] are both RAG-enabled multi-agent frameworks, but were designed for hospital emergency settings and carry no maternal health specialisation or rural India alignment. MatruCare-AI is the only proposed framework that satisfies all five dimensions concurrently, and this distinction constitutes the third of its three stated contributions.

IV. MATRUCARE-AI FRAMEWORK DESIGN

A. Overview

MatruCare-AI is organised across three interdependent layers. The layers are not independently modular: output from Layer 1 triggers Layer 2 advisory, and Layer 2 risk-graded responses inform Layer 3 agent actions [19], [66]. Layer 1 captures antenatal parameters and computes a four-tier risk classification. Layer 2 retrieves relevant sections from a curated corpus of Indian governmental health protocols and generates a guideline-grounded clinical response. Layer 3 runs three parallel agents that translate the risk score into concrete operational instructions, identifying danger signs, routing referrals, and scheduling follow-up visits [83], [85].

The complete architecture is illustrated in Fig. 1. When Layer 1 outputs a risk score at or above the HIGH threshold (classifier probability ≥ 0.65), it passes a structured JSON payload to the Layer 2 RAG endpoint, which retrieves the most relevant protocol passages and simultaneously activates the appropriate Layer 3 agents.

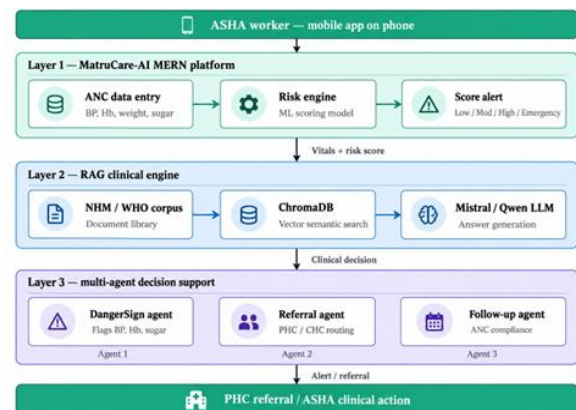


Fig. 1. MatruCare-AI Three-Layer Framework Architecture. Layer 1 collects ANC data and computes a risk score. Layer 2 answers clinical questions using RAG over India's own health guidelines. Layer 3 runs three parallel agents for danger detection, referral routing, and follow-up scheduling.

B. Layer 1: MatruCare-AI Platform (MERN-Based)

Layer 1 is implemented as a Progressive Web Application (PWA) built on the MERN stack (MongoDB, Express.js, React.js, Node.js) [86], [87]. A uniform JavaScript runtime across client and server eliminates serialisation overhead, and the PWA architecture enables offline data-capture when network connectivity is unavailable [49], [83]. During a home visit, the ASHA records a concise set of clinical parameters: diastolic and systolic blood pressure, fasting blood glucose, gestational weight, haemoglobin, gestational age in weeks, and fetal movement status [82], [84]. These parameters are stored in the woman's longitudinal record in

MongoDB, where each visit entry extends a time-series that the risk engine analyses across multiple observations.

The risk engine employs an XGBoost Gradient Boosting classifier trained on features drawn from the NFHS-5 national health survey, consistent with the modelling approach validated by Saini et al. [13]: inter-pregnancy interval, anaemia status (haemoglobin < 11 g/dL), maternal age, gestational age, blood pressure readings, and fasting glucose. The classifier outputs a four-tier risk label communicated through colour-coded interface elements green (Low), yellow (Moderate), orange (High), red (Emergency) that update with each new data entry. A risk trajectory trending from yellow to orange across sequential visits activates an interim alert before the presentation reaches the emergency threshold, affording the ASHA a planning window [82], [83]. Role-based access control permits physicians and district health staff to view aggregated risk dashboards, flagged cases, and area-level coverage analytics [52], [86].

C. Layer 2: RAG Clinical Engine

Layer 2 provides the framework's clinical reasoning capability. When the ASHA enters a clinical query, it is processed by the RAG Clinical Engine, which retrieves relevant protocol passages and constructs a guideline-grounded response rather than generating from parametric memory alone [5], [76]. The document corpus comprises four categories of India-specific health protocols: NHM Maternal Health operational guidelines, PMSMA clinical assessment criteria, WHO ANC 2016 recommendations, and the programme documents of the SUMAN and JSY initiatives [9], [17]. Documents are segmented into 256-token overlapping chunks with a 64-token stride and encoded using the BGE-M3 multilingual embedding model, which natively supports Hindi and major Indian languages. Encoded chunks are stored in ChromaDB, tagged with metadata fields for trimester, symptom category, risk tier, and geographic programme (NHM/PMSMA/SUMAN) [77], [80]. ChromaDB returns query results within 8.4 milliseconds at a response consistency variance of 2.3% [77], [80]. A FAISS index serves as a fallback in storage-constrained deployments, scaling to 2.2 million document segments without performance degradation [80].

At inference time, the ASHA's query activates a semantic search over the tagged index, retrieving the three to five most relevant chunks. These passages are forwarded as context to Mistral-7B or Qwen-2.5 selected for resource efficiency and documented multilingual capability [78], [81]. The model constructs a response grounded exclusively in the retrieved content. A RAG configuration evaluated against UK NICE clinical guidelines achieved 99.5% source faithfulness, 98.7% query accuracy, and a 67% reduction in unsafe responses [81]. Applied to German governmental clinical guidelines, open models in the Mistral and Qwen family reached 90% accuracy, with retrieval grounding reducing hallucinations from 42% to 4% [93]. The RAG architecture is preferred over fine-tuning because India's governmental guidelines are updated periodically; with RAG, the document index is refreshed without retraining the generation model [5], [8], [81].

A representative system response for a high-risk presentation is as follows:

Blood pressure reading: 150/95 mmHg at 32 weeks. Risk level: HIGH. This reading meets the PMSMA threshold for severe hypertension. Refer to the nearest PHC immediately. Pre-eclampsia monitoring protocol is now active. Record fetal movement and check blood pressure again in 30 minutes.

This format delivers an actionable, source-grounded instruction rather than a general advisory statement [9], [76].

D. Layer 3: Multi-Agent Decision Support

Layer 3 operates three concurrent agents, each with a defined single responsibility. Agent coordination follows a specialisation-and-routing architecture, consistent with the multi-agent clinical frameworks described by Nweke et al. and Mugambiwa et al. [19], [66]. The agents translate risk classifications from Layer 1 into concrete operational instructions for the ASHA.

Agent 1 DangerSign Detector. This agent continuously evaluates incoming Layer 1 vital signs against clinical danger-sign thresholds derived from NHM and PMSMA protocols. Monitored flags include: systolic blood pressure ≥ 140 mmHg or diastolic ≥ 90 mmHg after 20 weeks gestation (pre-eclampsia indicator), haemoglobin below 7 g/dL

(severe anaemia), fasting blood glucose above 126 mg/dL (gestational diabetes threshold), absence of reported fetal movement over 24 hours, and clinically significant oedema of the face or extremities [82], [84]. When any threshold is breached, the ASHA receives a structured alert specifying the triggered flag, its clinical significance in plain language, and the recommended immediate action. A human-in-the-loop danger-sign detection application deployed in Indonesia achieved 92% accuracy using an equivalent parameter set [82]. Explainability studies using SHAP and LIME methods confirm that systolic blood pressure, diastolic blood pressure, and fasting glucose are the most consistent HRP predictors across heterogeneous datasets [48], [84]. The agent provides a structured recommendation but does not execute any action autonomously; clinical decision authority remains with the ASHA and, where available, a supervising physician [15], [83].

Agent 2 Referral Routing Agent. When Agent 1 generates a HIGH or EMERGENCY classification, the Referral Routing Agent activates. This agent queries the NHM Health Management Information System (HMIS) Facility Registry to identify the nearest available Primary Health Centre (PHC), Community Health Centre (CHC), or District Hospital matched to the flagged risk type and the woman's registered location [19], [20]. A structured referral

record is generated containing the facility name, estimated distance, available services relevant to the flagged condition, and a referral timestamp. Evidence from the AI-ADSS deployment showed that structured decision-support raised ASHA referral accuracy from 60% to 88% compared with verbal referrals alone [2], [66]. All referrals are logged to the patient's longitudinal record in Layer 1, ensuring a verifiable chain of custody for each high-risk presentation [20], [21].

Agent 3 Follow-up Scheduler. This agent monitors antenatal visit compliance against the NHM-mandated schedule of a minimum of four ANC visits per pregnancy. For any registered woman approaching or exceeding a visit interval, an alert is delivered to the ASHA's device prior to the next scheduled home visit [3], [4]. Women previously classified as Moderate or High risk are placed on an accelerated alert schedule with shorter inter-alert intervals. Research indicates that the sharpest decline in antenatal attendance occurs between the second and third scheduled visits, precisely the window during which pre-eclampsia and gestational anaemia most commonly progress to clinical significance [50], [83]. Early identification of missed-visit women before their next appointment allows the ASHA to conduct an unscheduled check, preventing high-risk women from exiting the monitoring system undetected [4], [54].

TABLE III MatruCare-AI Framework Design Components

Layer	Component	What It Does	Ref.
Layer 1	MERN PWA	Records ANC readings on ASHA's phone	[16], [86]
Layer 1	XGBoost Risk Engine	Outputs 4-tier risk la-bel per visit	[13], [82]
Layer 2	RAG Engine	Answers clinical questions from guidelines	[5], [8], [9]
Layer 2	ChromaDB + BGE-M3	Searches NHM/PMSMA/WHO document library	[77], [80]
Layer 3	DangerSign Agent	Flags BP, anaemia, blood sugar danger signs	[82], [84]
Layer 3	Referral Agent	Routes to nearest PHC/CHC via HMIS Registry	[19], [20]
Layer 3	Follow-up Agent	Alerts ASHA when woman misses ANC visit	[3], [50]

V. THEORETICAL EVALUATION

MatruCare-AI is a proposed framework that has not yet been implemented or field-tested. Each of its three layers, however, is grounded in technology that has been independently validated through peer-reviewed research across comparable clinical set-tings. The

subsections below assess expected layer performance against published benchmarks, with every figure drawn from a cited study so that each claim can be verified independently [5], [19], [41].

A. Layer 1 Risk Detection Performance

Layer 1 employs an XGBoost Gradient Boosting

classifier to stratify antenatal risk from parameters already collected by ASHA workers at home visits. Several India-specific validation studies are directly relevant. Applied to the NFHS-5 national health survey, a Gradient Boosting model achieved 89% overall accuracy and 90% high-risk recall, with inter-pregnancy interval, anaemia, maternal age, and education as the leading predictors [13]. A neural network trained on 1,014 clinical records reached 91% accuracy for the high-risk class and processed 500 records per second [41]. An ensemble stacking approach combining multiple base classifiers

produced 98% accuracy with 97% high-risk recall [47], and a CatBoost configuration achieved 94.2% accuracy with 96.2% high-risk recall [88]. At primary health centres in Pune, an AI-based mHealth application screened 1,010 women and identified 24.85% as high-risk [44] markedly higher than the 14% captured by national programme records for the same population [14]. In rural Guatemala, ML models reached nurse-level HRP referral accuracy, with blood pressure and blood glucose as the two most discriminative features [43]. These results are summarised in Table IV.

TABLE IV Published Studies Supporting the Layer 1 Risk Detection Design

Study	Model Used	Accuracy	High-Risk Recall	Ref.
Saini et al. 2025	Gradient Boosting	89%	90%	[13]
Pi et al. 2025	Neural Network	91%	91%	[41]
Deshmukh et al. 2025	CatBoost	94.2%	96.2%	[88]
Modi et al. 2025	Combined Models	98%	97%	[47]
Devi et al. 2026	SVM (Pune, India)	AUC 99.96	97.4%	[44]
Prieto-Egido et al. 2026	SVM + Random Forest	Nurse-level	Nurse-level	[43]

SVM = Support Vector Machine Recall = proportion of actual high-risk cases correctly identified

B. Layer 2 RAG Advisory Accuracy

Layer 2 retrieves passages from a curated corpus of Indian governmental health protocols and generates guideline-grounded advisory responses. Published benchmarks establish clear performance expectations for this architecture. A meta-analysis of 20 clinical RAG studies found that retrieval augmentation increased correct response rates by 35 percentage points relative to unaugmented baseline models [90]. The MedRAG benchmark, evaluating 41 RAG configurations against 7,663 medical questions, established accuracy gains of up to 18 percentage points, lifting weaker base models to the performance level of larger, more resource-intensive systems [33]. Applied specifically to governmental clinical guidelines the closest analogue to the NHM and PMSMA corpus proposed for MatruCare-AI a RAG system achieved 99.5% source faithfulness, with the best configuration answering 98.7% of queries correctly and reducing unsafe responses by 67% [81]. An evaluation against German governmental clinical guidelines found that open models in the Mistral and Qwen family reached 90% accuracy, and that adding retrieval grounding reduced hallucinations from 42%

to 4% [93]. For resource-constrained deployments characteristic of low-end hardware, RAG contributed a 10.6 percentage-point accuracy gain relative to the same model operating without retrieval [91]. These results support the inference that Layer 2, deployed over India's NHM and PMSMA protocol corpus with Mistral-7B or Qwen-2.5 as the generation model, would achieve accuracy and faithfulness levels appropriate for use as ASHA decision support [5], [8], [76].

C. Layer 3 Multi-Agent System Performance

Layer 3 deploys three parallel specialised agents rather than a single general-purpose model. The multi-agent architecture is motivated by evidence that task specialisation and parallel processing improve both accuracy and robustness under varying workloads [19], [96].

A controlled comparison between a multi-agent system and a single-agent baseline found that under a mixed workload of 80 concurrent tasks, the multi-agent system maintained 65.3% task accuracy while the single-agent configuration dropped to 16.6% [96]. The multi-agent system also utilised up to 65 times fewer computational resources per task, a property of direct relevance to cost-constrained rural health deployments [96]. A simulation study evaluated multi-

agent clinical discussion against 16 cases in which cognitive bias had caused initial mis-diagnosis; following agent deliberation, diagnostic accuracy reached 76%, significantly outperforming the human evaluator baseline (odds ratio 3.49, $p=.002$) [95]. An adaptive RAG-plus-reasoning multi-agent framework achieved 94%, 88%, and 84% accuracy across three separate medical question-answering benchmarks [67]. A safety evaluation found that multi-agent filtering intercepted 89.9% of hallucinations before response delivery [68], a critical property for a system operating in life-and-death clinical contexts.

D. Comparison with Existing Platforms

Table V evaluates MatruCare-AI against the three most closely related deployed systems across six capability dimensions. While MatruCare-AI remains unimplemented, each of its three layers is grounded in independently validated technology as described above.

SMARThealth GPT is the closest prior system: built

for Indian health workers, employing a RAG pipeline over Indian pregnancy guidelines, and explicitly designed for ASHA use [9]. However, it incorporates no automated HRP detection, no structured referral routing, and no local-language advisory output [9], [56]. AI-ADSS demonstrated that structured decision support raised ASHA referral accuracy from 60% to 88% [2], confirming the value of the decision-support concept; however, it relies on fixed rules and cannot generalise to clinical presentations beyond its pre-programmed parameters [2]. e-SAATHI achieved rapid large-scale enrolment, demonstrating the feasibility of chat-based maternal health systems in India, while remaining bounded by scripted interactions incapable of adaptive reasoning [17]. MatruCare-AI is the only proposed framework that covers all six capabilities in Table V. The benchmark evidence in Table IV establishes that each of its three layers is grounded in technology with documented validation in peer-reviewed research [13], [33], [96].

TABLE V MatruCare-AI Compared to Existing Platforms Across Six Key Capabilities

Capability	SMARThealth GPT [9]	e-SAATHI [17]	AI-ADSS [2]	MatruCare -AI (P)
AI / RAG Advisory	Yes	No	No	Yes
Detects High-Risk Pregnancy	No	No	Partial	Yes
Built for ASHA Workers	Yes	No	Yes	Yes
Speaks Local Languages	No	No	No	Yes*
Uses Multiple Agents	No	No	No	Yes
Follows India's Protocols	Yes	Yes	Yes	Yes

(P) = Proposed * = Hindi + 8 regional languages (planned)

against the framework; it is a prerequisite for designing a pilot that might actually succeed.

VI. DISCUSSION

The evidence reviewed in the preceding sections confirms that the MatruCare-AI framework addresses a documented and consequential gap in maternal health decision support for ASHA workers in rural India. The research foundation is solid, the gap analysis is systematic, and the technology underpinning each layer carries independent validation.

Transitioning from a design framework to a deployed programme, however, requires confronting five implementation challenges that the research literature documents with some consistency [5], [19], [65]. Acknowledging these challenges is not an argument

A. Connectivity Constraints in Rural Deployments

The RAG Clinical Engine in Layer 2 depends on network connectivity to retrieve protocol passages and generate advisory responses. Rural India presents a well-documented connectivity challenge: approximately half of rural residents lack reliable broadband access [97]. A comparative study across five low- and middle-income countries identified intermittent connectivity, frequent power disruptions, and high mobile data costs as the three most commonly reported barriers to sustained digital health tool use [98]. Field research following ASHA workers in rural Meghalaya found that network instability rendered deployed mHealth applications effectively non-functional during visits, prompting workers to

revert to paper-based records [99]. A review of digital health equity under the Ayushman Bharat programme confirmed that low-connectivity communities continue to be the least served by digital health initiatives despite being the most in need [100].

Layer 1 of MatruCare-AI is inherently offline-capable, as it performs local data capture and classification on-device. The mitigation for Layer 2 is to embed a compressed, quantised subset of the Mistral-7B model directly on the ASHA's device, together with a trimmed local index of the most frequently queried NHM protocol passages. This offline mode would return less comprehensive responses than the full RAG engine, but would preserve core advisory functionality in the absence of a network connection [12], [99].

B. Multilingual Implementation Requirements

India recognises 22 official languages and hosts hundreds of dialectal variants. Designing for Hindi alone would exclude a substantial proportion of the ASHA workforce and the mothers they serve. Research on AI health tools for tribal populations in India identified language mismatch as a primary driver of non-engagement, even when tools were accessible on workers' devices [62]. The ArogyaSutra evaluation across seven Indian languages demonstrated that models specifically adapted for Indic linguistic contexts produced substantially more accurate and safer medical responses than English-only systems deployed in Indian settings [71]. A roadmap for equitable LLM adoption in low-resource contexts placed local-language support at the top of its list of five enabling conditions [38].

MatruCare-AI proposes support for Hindi and eight regional Indian languages from initial deployment. Language detection would be handled via fastText classification. Clinical protocol documents would be professionally translated and validated by bilingual health workers before indexing. Qwen-2.5's native support for 29 languages, including Hindi and major Indian languages, makes it the preferred generation model for multilingual deployments [78]. Expanding language coverage adds both translation cost and validation burden; however, without it, the tool cannot serve the population it was designed for [62], [71].

C. Ethical Dimensions of AI-Generated Risk Labels

Communicating an AI-generated risk classification

to a pregnant woman during a home visit carries ethical responsibilities that extend beyond technical accuracy. Ji and Wang documented that AI pregnancy risk labels can induce sustained predictive anxiety, particularly when labels are assigned without adequate explanation of their probabilistic basis [15]. The same study identified differential over-flagging of women from lower socioeconomic backgrounds as a bias risk when training data reflect pre-existing health inequities [15]. A re-view of AI tools for adverse pregnancy outcomes concluded that communicative transparency and patient-centred design are as operationally critical as predictive accuracy [103]. Wong et al. found that bias-mitigated perinatal mood disorder models continued to flag minority women at 1.47 times the rate of non-minority women, illustrating that post-hoc debiasing may be insufficient without representative training data and ongoing fairness monitoring [104]. To address these risks, the DangerSign Detector in Layer 3 is designed to return structured explanations alongside each alert, specifying which parameter triggered the flag and what it means in plain language. No referral is dispatched without explicit ASHA confirmation. All risk alerts include a communication guide to support the ASHA in discussing findings with the woman in a manner that is informative without inducing unnecessary fear [15], [103]. Clinical decision authority remains with the ASHA and supervising physician at all times; the system functions as advisory infrastructure, not as an autonomous classifier [15], [42].

D. Hallucination Risk in Clinical Contexts

RAG substantially reduces hallucination rates relative to un-augmented language model inference, but does not eliminate them. A systematic review of hallucination mitigation strategies in healthcare AI found that retrieval augmentation reduced hallucinations by 30 to 50 percentage points compared to standard LLM responses; however, residual error rates of clinical significance remained in the lower bound of this range [108]. An adversarial stress test introducing deliberately misleading information into clinical queries found hallucination rates ranging from 50% to 82% across six models, with the best-performing model retaining a 23% error rate following targeted mitigation [109].

MatruCare-AI addresses this in two ways. For

standard clinical queries, the RAG engine re-returns the retrieved guideline passage alongside the generated response, allowing the ASHA to view the source from which the answer was derived. For presentations flagged as HIGH or EMER-GENCY by the DangerSign Detector cases where generative error would carry the greatest clinical consequence the system bypasses the generative layer entirely and delivers a pre-authored, clinician-validated advisory template [7], [107]. Every response includes a clear statement that the output represents decision support rather than a medical diagnosis, and directs the ASHA to contact a physician for any emergency [108].

E. Framework Validation and Limitations

MatruCare-AI is a design proposal. No ASHA has operated the system. No pregnant woman has been screened through it. Every performance figure cited in Section V originates from independent studies validating comparable technologies in other settings and populations. Those figures are relevant and informative, but they are not MatruCare-AI's own empirical results.

A systematic analysis of why AI health interventions fail to scale in India identified five recurring structural barriers: lack of interoperability with existing health information systems, funding discontinuation after pilot completion, insufficient regulatory support, low CHW trust in novel digital tools, and infrastructure gaps between design assumptions and field conditions [65]. Any pilot for MatruCare-AI must plan for all five from the outset. A district-level pilot conducted with registered NHM-enrolled ASHAs, under physician supervision, with a structured feedback mechanism covering three evaluation questions danger-sign detection accuracy, RAG response safety, and sustained ASHA adoption constitutes the minimum validation scope before broader deployment could be responsibly recommended [12], [65].

VII. CONCLUSION AND FUTURE WORK

A. Conclusion

Rural India's maternal mortality burden is disproportionate and well documented. ASHA workers serve as the primary antenatal contact for most rural pregnant women, yet a meta-analysis of 37 studies established a correct maternal health knowledge rate of 62% [1]. Epidemiological surveys report high-risk

pregnancy prevalence near 40%, while national programme records capture only 14% [14]. The digital tools currently available to ASHAs rely on pre-programmed rule sets that cannot reason through novel presentations, detect evolving risk trajectories, or update when national guidelines change [16], [17].

MatruCare-AI was proposed to address this gap through a three-layer architecture. Layer 1, an XGBoost-driven MERN-based PWA, records antenatal parameters and delivers four-tier risk classification at each home visit. Layer 2, a RAG Clinical Engine over a curated corpus of NHM, PMSMA, WHO ANC, SUMAN, and JSY protocols, answers ASHA queries with source-grounded, guideline-consistent advisory responses using Mistral-7B or Qwen-2.5 as the generation model. Layer 3 operates the DangerSign Detector, Referral Routing Agent, and Follow-up Scheduler as parallel specialised agents that translate risk scores into concrete operational instructions.

This paper establishes three contributions. First, to the best of the authors' knowledge, MatruCare-AI constitutes the first framework combining RAG-based clinical advisory with multi-agent decision support designed exclusively for ASHA-mediated HRP detection in rural India [5], [19]. Second, all clinical knowledge is sourced from Indian governmental guidelines, ensuring protocol alignment with the NHM programme environment in which ASHAs operate [9], [17]. Third, the gap matrix in Table II confirms that no prior system simultaneously satisfies all five required dimensions: ASHA focus, RAG advisory, HRP detection, rural India context, and multilingual support [2], [56], [65].

The benchmark evidence reviewed in Section V supports the framework's expected performance. Machine learning HRP prediction from ASHA-collected parameters reaches 98% accuracy in published studies [47]. RAG over governmental clinical guidelines achieves 99.5% source faithfulness and a 67% reduction in unsafe responses [81]. Multi-agent systems maintain accuracy under workloads where single-agent configurations fail, while intercepting 89.9% of hallucinations before response delivery [68], [96]. The evidence base supports the framework's design rationale. The required next step is empirical validation in the field.

B. Future Work

Four directions for subsequent development follow directly from the limitations and design gaps identified in the preceding sections.

1. Voice Interface in Local Languages.

Layer 1 and Layer 2 currently require the ASHA to read and type. Literacy barriers among ASHA workers, and among the mothers they serve, make a voice-first interaction modality important for equitable reach. Voice-based AI systems for maternal health workers have demonstrated operational feasibility in comparable low-resource contexts: Awaaz-e-Sehat, deployed in Pakistan, processed over 500 health records and flagged more than 300 potential risk cases during a seven-month hospital deployment [113]. MediSense AI, deployed in Chhattisgarh, supported five languages including Hindi with effective dialectal adaptation [112]. VANI AI achieved 91% speech recognition accuracy across Hindi, Marathi, Bengali, and English without re-quiring a network connection [116]. A voice agent piloted with an Indian NGO working with Hindi-speaking mothers demonstrated reliable form-filling under real background noise and accent variation [117]. MatruCare-AI's next development iteration should implement a voice layer supporting Hindi and a minimum of eight regional Indian languages, informed by the low-literacy design principles established in these deployments [111], [115].

2. Federated Learning for Privacy-Preserving Model Refinement.

Antenatal data recorded in Layer 1 are personally identifiable health information. Centralising this data for model retraining raises legitimate privacy concerns, particularly in communities where existing trust in digital health infrastructure is fragile [65], [100]. Federated learning addresses this by training model updates on each device locally and sharing only parameter gradients rather than raw records. A systematic review of federated learning in healthcare identified this approach as the most viable method for training machine learning models across distributed clinical sites without centralising patient data [119]. A federated framework deployed across community health centres in a rural setting produced model performance equivalent to centralised training while preserving data locality [120]. A study specifically

focused on rural healthcare diagnosis found that on-device federated training supported accurate real-time patient monitoring [121]. MatruCare-AI's Layer 1 risk engine should incorporate federated learning in subsequent versions to enable continuous model improvement without compromising the data sovereignty of enrolled women [118], [122].

3. Offline-Capable Compressed RAG Mode.

As discussed in Section VI-A, intermittent connectivity in rural India limits Layer 2 availability in a substantial proportion of potential deployment areas. Recent advances in language model compression make on-device RAG technically feasible for mid-range Android hardware. Activation-aware Weight Quantisation (AWQ) reduces a large language model to one quarter of its original parameter size while accelerating inference by a factor of three to four, making deployment on device-class hardware practical [123]. Compressed models evaluated on a Raspberry Pi 4 hardware with computational resources comparable to a low-end smartphone returned accurate responses at latencies acceptable for field use [124]. A compressed Llama-3.1-8B model reduced from 15.3 GB to 3.3 GB demonstrated a 50% reduction in energy consumption alongside a decrease in hallucination rate relative to the full model [125]. MedAide, a fully local medical chatbot with no server dependency, achieved 77% query accuracy on a constrained device [126]. MatruCare-AI should develop a compressed Mistral or Qwen variant stored directly on the ASHA's device, capable of handling the most frequent NHM protocol queries without network access [127].

4. District-Level Prospective Pilot Study.

The validation requirement identified in Section VI-E is the most consequential item on the future work agenda. A district-level pilot with NHM-registered ASHAs, under physician supervision, should address three evaluation questions: whether the DangerSign Detector achieves clinically acceptable sensitivity and specificity for the flagged conditions; whether RAG responses are rated as safe and appropriate by supervising physicians re-viewing a sample of outputs; and whether ASHAs sustain application use across a full antenatal cycle rather than abandoning it in the field. A study of why AI health pilots fail to scale in India documented five predictable barriers

system interoperability gaps, post-pilot funding discontinuation, policy lag, CHW trust deficits, and infrastructure mismatches [65]. A pilot designed with all five barriers explicitly addressed from day one is a prerequisite for MatruCare-AI to progress from a research framework to a programme that reaches pregnant women in rural India [12], [65].

ACKNOWLEDGMENTS

The authors thank the faculty of the Department of Computer Science and Engineering at Hi-Tech Institute of Engineering and Technology, Ghaziabad, for their guidance and institutional support during this research.

AUTHOR CONTRIBUTIONS

A.G.: Conceptualisation, Framework Design, Writing (Original Draft). H.G.: Literature Review, Formal Analysis. Anshu: Methodology, Visualisation. Dr. T.C.: Supervision, Review and Editing.

ETHICAL STATEMENT

This study is a theoretical framework proposal and does not involve human participants, patient data collection, or clinical intervention. No Institutional Review Board (IRB) approval was required.

AI DISCLOSURE

The authors used generative artificial intelligence tools, including Claude (Anthropic) and Consensus, to assist with literature discovery, language refinement, and grammar correction during the preparation of this manuscript. All cited references were independently verified by the authors. All research ideas, methodology, technical analysis, interpretations, and final editorial decisions were developed, reviewed, and approved solely by the authors. The authors take full responsibility for the accuracy, originality, and integrity of the content presented in this manuscript.

REFERENCES

- [1] S. Singh, A. Singh, and R. K. Srivastava, "Knowledge of Accredited Social Health Activists in India: A systematic review and meta-analysis of evidence drawn from primary studies published between 2005 and 2022," *BMC Health Services Research*, vol. 25, no. 1, 2025.
- [2] N. Kashyap et al., "Effectiveness of an Artificial Intelligence-Assisted Decision Support System among ASHA Workers," *IFR Journal of Nursing and Health Care Prevention*, 2026.
- [3] S. Mishra et al., "Association between the use of Accredited Social Health Activist (ASHA) services and uptake of institutional deliveries in India," *PLOS Global Public Health*, vol. 4, no. 2, 2024.
- [4] L. Gopalakrishnan et al., "Association between coordinated counseling from both ASHA and Anganwadi Workers and maternal health outcomes: A cross-sectional study from Madhya Pradesh and Bihar, India," *PLOS Global Public Health*, vol. 4, no. 5, 2024.
- [5] L. M. Amugongo et al., "Retrieval augmented generation for large language models in healthcare: A systematic review," *PLOS Digital Health*, vol. 4, no. 1, 2025.
- [6] S. Saad et al., "A survey on retrieval-augmentation generation (RAG) models for healthcare applications," *Neural Computing and Applications*, 2025.
- [7] L. Garza et al., "Retrieval-Augmented Framework for LLM-Based Clinical Decision Support," *arXiv preprint*, 2025.
- [8] S. Kresevic et al., "Optimization of hepatological clinical guidelines interpretation by large language models: A retrieval augmented generation-based framework," *npj Digital Medicine*, vol. 7, no. 1, p. 235, 2024.
- [9] H.-Y. Lu et al., "Transforming Healthcare Education: Harnessing Large Language Models for Frontline Health Worker Capacity Building using Retrieval-Augmented Generation," *medRxiv*, 2023.
- [10] S. Rutunda et al., "Large language models for frontline health-care support in low-resource settings," *Nature Health*, 2026.
- [11] V. Menon et al., "Assessing the potential utility of large language models for assisting community health workers: Protocol for a prospective, observational study in Rwanda," *BMJ Open*, vol. 15, no. 6, 2025.
- [12] J. Ong et al., "Large language models in global health," *Nature Health*, 2026.
- [13] H. Saini et al., "Predicting High-Risk Pregnancies in India Using NFHS-5 Data: A Machine Learning Study," in *Proc. 2025 IEEE 2nd*

- International Conference for Women in Computing (InCoWoCo), 2025.
- [14] G. Shergill et al., "Identification of Factors behind High-Risk Pregnancies in Rural Punjab, India using Machine Learning," in Proc. 8th World Engineering Conference on Contemporary Technologies (WECON), 2025.
- [15] Q. Ji and Y. Wang, "AI-driven high-risk pregnancy prediction: Balancing early detection, anxiety, and discrimination in digital public health," *Frontiers in Public Health*, vol. 14, 2026.
- [16] B. Endehabtu et al., "Mapping the Role of Digital Health Interventions to Enhance Effective Coverage of Antenatal Care: A Scoping Review," *Journal of Multidisciplinary Healthcare*, vol. 17, pp. 1481–1496, 2024.
- [17] A. Munjral et al., "Chat-Based Decision Support System for the Maternal Health Journey in Assam, India: Protocol for a Mixed Methods Multiphase Implementation Study," *JMIR Research Protocols*, vol. 15, no. 1, 2026.
- [18] K. Singh and J. Walters, "Use of mHealth in promoting maternal and child health in 'BIMARU' states of India: Systematic literature review," *PLOS Digital Health*, vol. 3, no. 4, 2024.
- [19] I. N. Nweke et al., "Multi-Agent AI Systems in Healthcare: A Systematic Review Enhancing Clinical Decision-Making," *Asian Journal of Medical Principles and Clinical Practice*, vol. 8, no. 2, 2025.
- [20] W.-Y. Choi, "Development of a Large Language Model-based Multi-Agent Clinical Decision Support System for Korean Triage and Acuity Scale (KTAS)-Based Triage and Treatment Planning in Emergency Departments," *arXiv preprint arXiv:2408.07531*, 2024.
- [21] A. Ibrahim et al., "TriAgent: An Adaptive Multi-Agent Architecture for Crisis Clinical Decision Support Under Incomplete Information," *AI*, vol. 7, no. 3, 2026.
- [22] A. Blanchard et al., "Associations between community health workers' home visits and education-based inequalities in institutional delivery and perinatal mortality in rural Uttar Pradesh, India: A cross-sectional study," *BMJ Open*, vol. 11, no. 3, 2021.
- [23] P. Sharma, "The Role of ASHAs in Enhancing Maternal and Child Health Outcomes in Rural Madhya Pradesh: A Mixed-Methods Study," *IFR Journal of Nursing and Health Care Prevention*, 2026.
- [24] A. Blanchard et al., "Understanding the roles of community health workers in improving perinatal health equity in rural Uttar Pradesh, India: A qualitative study," *International Journal for Equity in Health*, vol. 20, no. 1, 2021.
- [25] P. L. Paul et al., "Factors influencing institutional delivery and the role of Accredited Social Health Activist (ASHA): A secondary analysis of India human development survey 2012," *BMC Pregnancy and Childbirth*, vol. 20, no. 1, 2020.
- [26] S. Chawla et al., "Performance and Challenges of Accredited Social Health Activists (ASHAs) on Key Maternal and New-born Health (MNH) Service Coverage in India: A Systematic Review and Meta-Analysis," *SSM–Health Systems*, 2025.
- [27] F. A. Hashmi et al., "Coexisting traditional and biomedical healthcare systems: A mixed-methods analysis of community health workers and traditional birth attendants' contributions to perinatal health behaviors in rural India," *Frontiers in Health Services*, 2025.
- [28] F. Neha et al., "Retrieval-Augmented Generation (RAG) in Healthcare: A Comprehensive Review," *AI*, 2025.
- [29] S. Xu et al., "MEGA-RAG: A retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of LLMs in public health," *Frontiers in Public Health*, 2025.
- [30] K. Wolk, "Evaluating Retrieval-Augmented Generation Variants for Clinical Decision Support: Hallucination Mitigation and Secure On-Premises Deployment," *Electronics*, 2025.
- [31] Y. Feng et al., "Hyper-RAG: Combating LLM hallucinations using hypergraph-driven retrieval-augmented generation," *Nature Communications*, 2026.
- [32] A. Wada et al., "Retrieval-augmented generation elevates local LLM quality in radiology contrast media consultation," *npj Digital Medicine*, 2025.
- [33] G. Xiong et al., "Benchmarking Retrieval-Augmented Generation for Medicine," in Proc. Findings of the Association for Computational Linguistics (ACL 2024), 2024.
- [34] P. Bechard et al., "Reducing hallucination in structured outputs via Retrieval-Augmented Generation," in Proc. 2024 Conference of the

- North American Chapter of the Association for Computational Linguistics (NAACL 2024), 2024.
- [35] M. Miyares et al., “Costs and cost-effectiveness of community health worker programs on reproductive, maternal, newborn and child health in low- and middle-income countries (2015–2024): A scoping review,” *PLOS Global Public Health*, 2026.
- [36] L. Yu et al., “Large Language Models for Maternal and Neonatal Healthcare in Low- and Middle-Income Countries,” *The Journal of Pediatrics*, 2026.
- [37] H. A. Lima et al., “Quality assessment of large language models’ output in maternal health,” *Scientific Reports*, 2025.
- [38] H.-C. Chen et al., “Large language models and global health equity: A roadmap for equitable adoption in LMICs,” *The Lancet Regional Health: Western Pacific*, 2025.
- [39] H. Qin et al., “Opportunities and Challenges for Large Language Models in Primary Health Care,” *Journal of Primary Care & Community Health*, 2025.
- [40] R. Rasoli et al., “Designing conversational intelligence: Effect of large language models (GPT-driven) platforms for precision maternal and newborn health engagement: A systematic review,” *Oxford Open Digital Health*, 2026.
- [41] X. Pi et al., “Prediction of high-risk pregnancy based on machine learning algorithms,” *Scientific Reports*, 2025.
- [42] A. O. Khadidos et al., “Ensemble machine learning framework for predicting maternal health risk during pregnancy,” *Scientific Reports*, 2024.
- [43] I. Prieto-Egido et al., “Identifying high-risk pregnancies in rural areas with machine-manifold learning,” *Engineering Applications of Artificial Intelligence*, 2026.
- [44] S. Devi et al., “From Readiness to Response: AI-Driven Antenatal Risk Assessment (mHealth application) and Machine Learning Techniques for Midwives Predicting Maternal and Neonatal Outcomes in Low Resource Healthcare Settings,” *Investigación y Educación en Enfermería*, 2026.
- [45] M. H. Sylvain et al., “Prediction of adverse pregnancy outcomes using machine learning techniques: Evidence from analysis of electronic medical records data in Rwanda,” *BMC Medical Informatics and Decision Making*, 2025.
- [46] M. Mousavi Noghli et al., “High-risk pregnancy prediction using Taguchi-optimized machine learning methods and TOPSIS-based model selection,” *Scientific Reports*, 2026.
- [47] N. Modi et al., “Automated Machine Learning-Based System for the Prediction of Maternal Health Indicators and High-Risk Pregnancy,” in *Proc. 2025 7th International Conference on Energy, Power and Environment (ICEPE)*, 2025.
- [48] G. Yanto et al., “Hybrid Multi-Method Explainable AI and Machine Learning for Early Prediction of High-Risk Pregnancies,” *TEM Journal*, 2026.
- [49] H. Mohamed et al., “A scoping review of digital technologies in antenatal care: Recent progress and applications of digital technologies,” *BMC Pregnancy and Childbirth*, 2025.
- [50] M. Mishra et al., “Effectiveness of mHealth Interventions for Monitoring Antenatal Care among Pregnant Women in Low- and Middle-Income Countries: A Systematic Review and Meta-Analysis,” *Healthcare*, 2023.
- [51] M. R. Knop et al., “Impact of mHealth interventions on maternal, newborn, and child health from conception to 24 months postpartum in low- and middle-income countries: A systematic review,” *BMC Medicine*, 2024.
- [52] E. Ameyaw et al., “Effectiveness of mHealth Apps for Maternal Health Care Delivery: Systematic Review of Systematic Reviews,” *Journal of Medical Internet Research*, 2024.
- [53] Z. Small et al., “Mobile health (mHealth) interventions for health promotion during the perinatal period in India: A scoping review,” *Frontiers in Global Women’s Health*, 2024.
- [54] S. Soni et al., “Bridging gaps in antenatal care: Implementation of the WHO ANC model using a web-based mobile application at different levels of the healthcare system,” *International Journal of Medical Informatics*, 2023.
- [55] M.-F. Hsieh et al., “mHealth use during pregnancy: A systematic review,” *International Journal of Gynecology & Obstetrics*, 2025.
- [56] V. Srinidhi et al., “ASHA Kirana: When digital technology empowered front-line health workers,” *BMJ Global Health*, vol. 6, no. 5, 2021.
- [57] N. Das et al., “Developing SMARThealth Pregnancy (SHP) GPT: Addressing Gender Bias in LLM for Community Health Workers in Rural

- India,” in Proc. 2025 Gender and Technology Conference (GTC), 2025.
- [58] S. C. Till et al., “Digital Health Technologies for Maternal and Child Health in Africa and Other Low- and Middle-Income Countries: Cross-disciplinary Scoping Review with Stakeholder Consultation,” *Journal of Medical Internet Research*, vol. 25, 2023.
- [59] W. Philbrick et al., “Artificial Intelligence and Other Digital Tools Used by, and for Community Health Workers (CHWs) in Low and Middle-Income Countries (LMICs) to Improve Outcomes and Increase Effectiveness: An Evidence and Gap Map,” *Campbell Systematic Reviews*, 2026.
- [60] M. Baniyadi et al., “Evidence integration: The transformative role of artificial intelligence in maternal health,” *Digital Health*, 2026.
- [61] S. V. Badhe et al., “AI-Enabled Intelligent Agent for Maternal Health Tracking in Tribal Regions: A Conceptual Framework,” *International Journal of Science, Strategic Management and Technology*, 2026.
- [62] K. Saif-Ur-Rahman et al., “Artificial intelligence and digital health in improving primary health care service delivery in LMICs: A systematic review,” *Journal of Evidence-Based Medicine*, 2023.
- [63] A. Sharma et al., “Artificial Intelligence for Health Equity: A Multi-tiered ML Framework for Tribal Women’s Healthcare and Frontline Worker Empowerment in India,” *International Journal for Multidisciplinary Research*, 2026.
- [64] M. Khan et al., “On AI Approaches for Promoting Maternal and Neonatal Health in Low Resource Settings: A Review,” *Frontiers in Public Health*, 2022.
- [65] J. Joseph, “From pilot to policy: Why AI health interventions fail to scale in developing countries,” *Frontiers in Digital Health*, 2026.
- [66] T. Mugambiwa et al., “Multi-Agent Retrieval Augmented Generation for Clinical Decision Support: A Systematic Review and Integrative Conceptual Framework,” *Journal of Applied Informatics and Computing*, 2026.
- [67] K. Arslanoglu et al., “An Adaptive Multi-Agent LLM-Based Clinical Decision Support System Integrating Biomedical RAG and Web Intelligence,” *IEEE Access*, 2025.
- [68] A. L. Silveira et al., “multi-agent systems for clinical decision support: A systematic review,” *Applied Soft Computing*, 2025.
- [69] Y.-S. Liu et al., “Benchmarking large language model-based agent systems for clinical decision tasks,” *npj Digital Medicine*, 2026.
- [70] Y.-J. Chen et al., “Reinforcing Clinical Decision Support through Multi-Agent Systems and Ethical AI Governance,” *arXiv preprint*, 2025.
- [71] T. Halder et al., “ArogyaSutra: A Multi-Agent Framework for Multimodal Medical Reasoning in Indic Languages,” *arXiv preprint*, 2026.
- [72] T. Ciecierski-Holmes et al., “Artificial intelligence for strengthening healthcare systems in low- and middle-income countries: A systematic scoping review,” *npj Digital Medicine*, 2022.
- [73] C. Horwood et al., “Challenges of using e-health technologies to support clinical care in rural Africa: A longitudinal mixed methods study exploring primary health care nurses’ experiences of using an electronic clinical decision support system (CDSS) in South Africa,” *BMC Health Services Research*, 2023.
- [74] Z.-B. Xie et al., “Ethical issues in multi-agent AI systems for healthcare: A narrative review,” *Frontiers in Public Health*, 2026.
- [75] Y.-J. Chen et al., “Enhancing Clinical Decision-Making: Integrating Multi-Agent Systems with Ethical AI Governance,” in Proc. 2025 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2025.
- [76] O. K. Gargari et al., “Enhancing medical AI with retrieval-augmented generation: A mini narrative review,” *Digital Health*, 2025.
- [77] A. Ammar et al., “Optimizing Retrieval-Augmented Generation: Analysis of Hyperparameter Impact on Performance and Efficiency,” *arXiv preprint*, 2025.
- [78] Y. Ke et al., “Development and Testing of Retrieval Augmented Generation in Large Language Models: A Case Study Report,” *arXiv preprint*, 2024.
- [79] D. Baur et al., “Development and Evaluation of a Retrieval-Augmented Generation Chatbot for Orthopedic and Trauma Surgery Patient Education: Mixed-Methods Study,” *JMIR AI*, 2025.
- [80] J. Patel et al., “Benchmarking Open-Source

- Vector Databases,” *Journal of Big Data and Artificial Intelligence*, 2026.
- [81] M. Lewis et al., “Grounding Large Language Models in Clinical Evidence: A Retrieval-Augmented Generation System for Querying UK NICE Clinical Guidelines,” *arXiv preprint*, 2025.
- [82] M. Widyawati et al., “Human-In-The-Loop application design for early detection of pregnancy danger signs,” *Belitung Nursing Journal*, 2022.
- [83] N. Singh, “AI-Enabled Smart Maternal Health Monitoring and Educational Support System for Early Detection and Prevention of High-Risk Pregnancy Complications,” *Journal of Nursing Future Care: AI and Innovation*, 2026.
- [84] K. D. Tzimourta et al., “Maternal Health Risk Detection: Advancing Midwifery with Artificial Intelligence,” *Healthcare*, 2025.
- [85] R. E. El Arab et al., “Integrating Artificial Intelligence into Perinatal Care Pathways: A Scoping Review of Reviews of Applications, Outcomes, and Equity,” *Nursing Reports*, 2025.
- [86] A. Verma et al., “HEALTHIFY: An AI-Driven MERN Solution for Personalized and Predictive Healthcare,” *International Journal for Multidisciplinary Research*, 2025.
- [87] Y. Tripathi, “AI Chatbot for Health: A MERN Stack-Based Intelligent Healthcare Assistant with Natural Language Processing,” *International Journal of Scientific Research in Engineering and Management*, 2025.
- [88] G. B. Deshmukh et al., “Optimized Prediction of High-Risk Pregnancy Using Machine Learning,” in *Proc. 2025 International Conference on Future Technologies (ICFT)*, 2025.
- [89] K. Tomar et al., “A Machine Learning-Based Risk Prediction Model During Pregnancy in Low-Resource Settings,” in *Proc. 2nd International One Health Conference*, 2024.
- [90] S. Liu et al., “Improving large language model applications in biomedicine with retrieval-augmented generation: A systematic review, meta-analysis, and clinical development guidelines,” *Journal of the American Medical Informatics Association*, 2025.
- [91] P. Li et al., “Optimizing Retrieval-Augmented Generation (RAG) in clinical medicine: Methods and performance evaluation,” *Journal of the American Medical Informatics Association*, 2026.
- [92] H.-S. Wong et al., “Multi-Evidence Clinical Reasoning with Retrieval-Augmented Generation for Emergency Triage: Retrospective Evaluation Study,” *JMIR Medical Informatics*, 2026.
- [93] A. Vollmer et al., “A Self-Controlled Benchmark of Retrieval-Augmented Generation for Large Language Models on Clinical Guideline Questions,” *Diagnostics*, 2026.
- [94] F. Gaber et al., “Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis,” *npj Digital Medicine*, 2025.
- [95] Y. Ke et al., “Mitigating Cognitive Biases in Clinical Decision-Making Through Multi-Agent Conversations Using Large Language Models: Simulation Study,” *Journal of Medical Internet Research*, 2024.
- [96] E. Klang et al., “Orchestrated multi agents sustain accuracy under clinical-scale workloads compared to a single agent,” *npj Health Systems*, 2025.
- [97] S. K. A. Kumar et al., “A Survey on Rural Internet Connectivity in India,” in *Proc. 2022 14th International Conference on Communication Systems and Networks (COMSNETS)*, 2022.
- [98] C. Hui et al., “Mapping national information and communication technology (ICT) infrastructure to the requirements of potential digital health interventions in low- and middle-income countries,” *Journal of Global Health*, 2022.
- [99] M. Nongrum et al., “Disconnected data: mHealth data systems and challenges for primary health care workers in India,” *SSM–Health Systems*, 2025.
- [100] R. Ganguly et al., “Digital Inequity in Health Access: A Literature Review of Ayushman Bharat and ABDM in India,” *Indian Journal of Public Health Research & Development*, 2026.
- [101] S. Arora et al., “Challenges, Barriers, and Facilitators in Telemedicine Implementation in India: A Scoping Review,” *Cureus*, 2024.
- [102] S. G. Bashir et al., “Artificial intelligence for predicting and preventing adverse pregnancy outcomes addressing bias and clinical translation,” *Frontiers in Digital Health*, 2026.
- [103] M. Nogueira et al., “Artificial Intelligence’s Role in Improving Adverse Pregnancy Outcomes: A Scoping Review and Consideration of Ethical Issues,” *Journal of*

- Clinical Medicine, 2025.
- [104]E. F. Wong et al., "Evaluating Bias-Mitigated Predictive Models of Perinatal Mood and Anxiety Disorders," JAMA Network Open, 2024.
- [105]M. M. L. Elakashif et al., "The impact of artificial intelligence-based emotional support tools on anxiety and stress levels in high-risk pregnant women," Scientific Reports, 2026.
- [106]E. Asgari et al., "A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation," npj Digital Medicine, 2025.
- [107]J. B. Hakim et al., "The need for guardrails with large language models in pharmacovigilance and other medical safety critical settings," Scientific Reports, 2025.
- [108]S. Basu et al., "Mitigating hallucinations in healthcare AI: A systematic review of evidence-based strategies," BMC Health Services Research, 2026.
- [109]M. Omar et al., "multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support," Communications Medicine, 2025.
- [110]W. Zhang et al., "Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review," Mathematics, 2025.
- [111]S. Manimala, "AI-Driven Bilingual Voice Chatbot," International Journal for Research in Applied Science and Engineering Technology, 2025.
- [112]H. Bhogal et al., "MediSense AI: A Multilingual Multimodal Healthcare Decision Support System for Rural and Tribal Communities of Chhattisgarh," in Proc. 2025 2nd International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIHI), 2025.
- [113]M. Mustafa et al., "Awaaz-e-Sehat: A Mobile Voice-based AI System for EMR Generation and Clinical Decision Support in Low-resource Maternal Healthcare," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2026.
- [114]R. Mishra et al., "Hindi Chatbot for Supporting Maternal and Child Health Related Queries in Rural India," 2023.
- [115]S. Srivastava et al., "Inclusive Voice-AI for Rural India: A Deep Learning Framework for Dialectal and Low-Resource Adaptation," International Journal for Multidisciplinary Research, 2026.
- [116]P. Dode et al., "VANI AI: A Multilingual Voice-Enabled Conversational Assistant with Offline Capability for Accessibility and Inclusive Communication," International Journal of Creative and Open Research in Engineering and Management, 2026.
- [117]A. Dalmia et al., "FormBharo: Designing and Evaluating a Voice Agent for Conversational Form Filling in Rural India," 2026.
- [118]S. Pati et al., "Privacy preservation for federated learning in health care," Patterns, 2024.
- [119]R. S. Antunes et al., "Federated Learning for Healthcare: Systematic Review and Architecture Proposal," ACM Transactions on Intelligent Systems and Technology, 2022.
- [120]R. Farizki et al., "Privacy-Preserving Federated Learning for Stunting Risk Prediction across Indonesian Community Health Centers," Jurnal Indonesia Sosial Teknologi, 2026.
- [121]M. A. Reddy et al., "Federated Learning for Rural Healthcare Diagnosis," in Proc. 2026 International Conference on Innovative Trends in Information Technology (ICITIIT), 2026.
- [122]M. Shokri, "Privacy-Preserving Edge Federated Learning for Intelligent Mobile-Health Systems," Future Generation Computer Systems, 2024.
- [123]J. Lin et al., "AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration," GetMobile: Mobile Computing and Communications, 2023.
- [124]E. J. Husom et al., "Sustainable LLM Inference for Edge AI: Evaluating Quantized LLMs for Energy Efficiency, Output Accuracy, and Inference Latency," ACM Transactions on Internet of Things, 2025.
- [125]R.-C. Zhang et al., "Compact LLM Deployment and World Model Assisted Offloading in Mobile Edge Computing," arXiv preprint, 2026.
- [126]A. Basit et al., "MedAide: Leveraging Large Language Models for On-Premise Medical Assistance on Edge Devices," arXiv preprint, 2024.
- [127]G.-Q. Qu et al., "Mobile Edge Intelligence for

Large Language Models: A Contemporary Survey,” IEEE Communications Surveys and Tutorials, 2024.