

ASHABot: An LLM-Powered Chatbot to Support the Informational Needs of Community Health Workers

Pragnya Ramjee*
 Microsoft Research
 Bangalore, India
 t-pramjee@microsoft.com

Mahendra Meena
 Khushi Baby
 Udaipur, India
 mahendra@khushibaby.org

Ruchit Nagar
 Khushi Baby
 Udaipur, India
 ruchit@khushibaby.org

Mehak Chhokar*
 Khushi Baby
 Udaipur, India
 mehak@khushibaby.org

Hamid Abdullah
 Khushi Baby
 Udaipur, India
 hamid@khushibaby.org

Bhuvan Sachdeva
 Microsoft Research
 Bangalore, India
 b-bsachdeva@microsoft.com

Aditya Vashistha
 Cornell University
 Ithaca, USA
 adityav@cornell.edu

Mohit Jain
 Microsoft Research
 Bangalore, India
 mohja@microsoft.com

Abstract

Community health workers (CHWs) provide last-mile healthcare services but face challenges due to limited medical knowledge and training. This paper describes the design, deployment, and evaluation of *ASHABot*, an LLM-powered, experts-in-the-loop, WhatsApp-based chatbot to address the information needs of CHWs in India. Through interviews with CHWs and their supervisors and log analysis, we examine factors affecting their engagement with *ASHABot*, and *ASHABot*'s role in addressing CHWs' informational needs. We found that *ASHABot* provided a private channel for CHWs to ask rudimentary and sensitive questions they hesitated to ask supervisors. CHWs trusted the information they received on *ASHABot* and treated it as an authoritative resource. CHWs' supervisors expanded their knowledge by contributing answers to questions *ASHABot* failed to answer, but were concerned about demands on their workload and increased accountability. We emphasize positioning LLMs as supplemental fallible resources within the community healthcare ecosystem, instead of as replacements for supervisor support.

CCS Concepts

- **Human-centered computing** → Ubiquitous and mobile computing systems and tools; *Field studies*; • **Applied computing** → **Health care information systems**.

Keywords

Chatbot, GPT-4, Experts-in-the-loop, Medical, Frontline Healthcare, HCI4D, ASHA, India

1 Introduction

India's healthcare ecosystem relies on nearly a million Accredited Social Health Activist (ASHA) workers [71]. ASHAs are volunteer community health workers who receive 3-4 weeks of initial training and are compensated through task-based incentives provided

by the Government of India [92]. They play a vital role in delivering essential health services, particularly in maternal and child health, within their communities [71, 92]. However, ASHAs often encounter challenges in effectively delivering these services due to their limited medical knowledge [109]. The reference material available to them are frequently insufficient for addressing complex queries of their care recipients [109]. Additionally, ASHAs have limited opportunities for further training and upskilling [109]. For instance, although ASHAs attend monthly meetings at local hospitals where they can interact with doctors and participate in group training sessions [38, 65], these sessions often lack adequate time to address ASHAs' individual queries comprehensively [38]. Overall, there is a lack of skilled teachers, funds, space, and materials to effectively train and support ASHAs [70, 107, 110].

ASHAs work under the supervision of Auxiliary Nurse-Midwives (ANMs), senior health workers who are formally employed and given 18 months of training by the government [41, 49, 92]. ASHAs often seek assistance from ANMs to perform their tasks and resolve doubts [41, 107, 110]. However, since a single ANM manages ASHAs across multiple villages, coordinating phone calls or in-person visits can be challenging [38]. Furthermore, ANMs, with their limited formal education [56, 107], may not always have the necessary answers [46]. ASHAs also hesitate to discuss sensitive topics such as sexual health [38, 84]. As a result, many ASHAs' questions remain unanswered, and misunderstandings persist [107].

HCI researchers have investigated various technologies to address the informational needs of ASHA workers, including voice messaging [20], IVRS [110], feature phone apps [85], and smartphone apps [60, 79, 88, 107]. However, ASHAs often struggle with these tools due to a steep learning curve, largely stemming from their limited familiarity with digital technologies [38, 39]. Recently, chatbots have emerged as a more user-friendly alternative [53, 66, 80, 108]. However, despite their promise, the impact of previous chatbots has been constrained by their rule-based design, which limits natural language interaction, requires a steep learning curve, and often yields inaccurate responses.

*Both authors contributed equally to this research.

In 2023, the release of LLM (Large Language Model)-powered chatbots marked a turning point, addressing these limitations and achieving higher adoption rates. This trend has sparked renewed interest in using LLMs in healthcare [86, 111]. However, applying LLMs to support healthcare in frontline settings remains largely unexplored and is still in early stages of design and evaluation [26, 61].

In this work, we examine the potential of an LLM-powered chatbot to address the informational needs of ASHAs. In collaboration with Khushi Baby, a non-governmental organization in Rajasthan, India, we designed and developed *ASHABot* using the open-source ‘Build Your Own expert Bot’ framework [86]. Key features of *ASHABot* include its integration with WhatsApp—a platform already familiar to ASHAs and ANMs [38]—and its ability to provide context-specific answers based on a custom knowledge base that incorporates government guidelines and ASHA handbooks [30]. *ASHABot* values ANMs for their expertise, involving them as experts-in-the-loop to guide consensus-driven decisions on unfamiliar questions. The ANMs’ answers are used to update the bot’s knowledge base. *ASHABot* supports Hindi and speech-based input and output, making it accessible to ASHAs and ANMs, who have varying literacy levels and digital skills [38, 39]. Given the application of LLMs in a high-stakes setting, we conducted iterative pilot testing to ensure the bot’s usability and the accuracy of its responses. We then conducted a field deployment study with 20 ASHAs and 15 ANMs in the Salumbar district of Rajasthan, India. After two months of usage, we conducted interviews with ASHAs and ANMs to examine three key research questions: **(RQ1)** How did ASHAs and ANMs use *ASHABot*? **(RQ2)** What factors influenced their engagement with *ASHABot*? **(RQ3)** How effective is *ASHABot* in addressing ASHAs’ information needs?

Our analysis of interviews and log data showed that ASHAs valued *ASHABot* for convenient, anytime access to detailed information, allowing them to learn at their own pace and save time. The bot provided a private channel for ASHAs to ask rudimentary or sensitive questions that they might hesitate to ask ANMs or doctors. ASHAs generally trusted the bot’s answers, especially as *ASHABot* admitted when it did not know an answer. Over time, *ASHABot* became an authoritative resource, influencing the decisions of both ASHAs and their care recipients. Although most ANMs expanded their knowledge by providing answers to the bot, they expressed concerns about the added workload and being held accountable for their responses. We conclude the paper by discussing design opportunities and challenges, such as personalization for ASHAs, incentivization for ANMs, and the importance of continuous training and support to address misconceptions in high-stakes frontline health settings. Our work makes the following contributions:

- We design, build, and deploy an LLM-powered, experts-in-the-loop chatbot on WhatsApp to support informational needs of low-skilled, low-literate ASHAs.
- Through a field deployment and interviews with ASHAs and ANMs, we evaluate the efficacy of the bot and identify factors that shape ASHAs’ and ANMs’ engagement with it.
- We discuss the limits and opportunities of using emergent AI technologies like LLMs to strengthen community health infrastructure.

2 Background and Related Work

ASHAs play a crucial role in supporting the well-being of their communities, particularly in providing maternal and child health services, with each ASHA serving a population of ~1,000 people [29, 98]. Their responsibilities include visiting pregnant women, assisting with checkups and deliveries, offering post-natal care, monitoring newborn health, and providing guidance on immunizations, family planning, and anemia prevention [108]. While our research focuses on ASHAs in India, it is situated within the broader context of literature on technology supporting community health workers worldwide. Additionally, our work explores the use of AI in healthcare. We summarize these topics below.

2.1 Technology to Support Community Health Workers

The HCI4D community has developed and deployed various solutions to improve the engagement of community health workers with care recipients. These include tools for planning home visits [21, 114], data collection [78], clinical decision-making [61], health-related education through text messages [53, 83, 99] and mobile videos [55, 68], and feedback gathering [69, 74]. Researchers have also designed technologies to improve the performance of community health workers, for example by enabling supervisors to provide personalized feedback [21, 35, 52, 105] and facilitating peer-comparison opportunities [20, 22]. These technological interventions have not only helped towards legitimizing the roles of community health workers as healthcare providers—as they are often disregarded or unpaid [41, 55, 85]—but have also motivated them to improve their digital skills [39, 62].

In particular, the use of technology to address community health workers’ information needs has been extensively studied [109]. Prior research has explored Interactive Voice Response (IVR) systems for delivering automated refresher courses and quizzes [6, 13] and enabling ASHAs to participate in live audio training sessions with peers and trainers remotely [85, 110]. However, IVR systems have significant limitations. Sequential navigation requires users to listen to all options, leading to long wait times [32]. Their menu structures are restrictive, particularly when navigating back to the root or previous menus [54]. Limited support for local language speech technologies further forces reliance on Dual Tone Multi-Frequency (DTMF) inputs, which are unintuitive and slow [7]. Additionally, the lack of persistent information reduces usability [32]. These factors make IVR systems cost-prohibitive and inefficient [103]. Visual solutions, which facilitate quick referencing, can overcome some of these limitations. Researchers have proposed various alternatives, including non-interactive videos [46, 102], smartphone-based games [60], custom platforms that leverage peer and supervisor knowledge for collaborative learning and telemonitoring [79, 88, 107], and even virtual and augmented reality applications [12, 59]. Despite the innovative nature of these technologies, their adoption has been limited due to high costs, as well as the steep learning curve for community health workers, who have limited education and poor digital skills [38, 39].

Recently, chatbots have emerged as an effective tool to meet the information needs of community health workers [53, 66, 108]. With their natural language interface, 24x7 availability, support

for multiple languages, and speech-based interactions, chatbots minimize the learning curve, making them accessible even to those with limited digital skills [43, 44, 86]. Chatbots can be easily deployed on familiar platforms such as WhatsApp [35, 38]. Before the advent of LLMs, researchers explored various rule-based chatbots. For instance, Yadav et al. [108] used a participatory Wizard-of-Oz approach to examine the potential of chatbots for breastfeeding-related questions, highlighting that ASHAs saw it in the possible role of trainer to enhance their knowledge, Karusala et al. [53] developed a WhatsApp-based service that allows nurses to manually respond to antenatal and postnatal care questions, and Mishra et al. [66] created a rule-based chatbot using an FAQ to answer maternal and child health questions. These non-LLM chatbots faced significant limitations [42], including (1) a limited scope that reduced their overall utility, and (2) difficulties in understanding natural language queries, leading to inaccurate responses. In contrast, *ASHABot* uses the state-of-the-art LLM model, GPT-4, to provide instant and accurate responses to a wide range of queries from community health workers via WhatsApp, addressing these previous shortcomings. We now describe the scholarly work on designing, building, and evaluating AI technologies in community health context.

2.2 AI and Community Health

Researchers have extensively explored the potential of AI to enhance the healthcare ecosystem [15, 33, 93]. This research spans a variety of applications, including disease diagnosis [24, 25, 104], clinical decision support [91, 113], improving communication between patients and healthcare providers [67, 86, 111], and optimizing hospital logistics [8, 14, 112]. In the context of community healthcare, researchers have developed AI tools to help ASHAs and ANMs diagnose diseases [9, 16], analyze rapid diagnostic tests [18, 81], and manage patient-care [72, 89]. Studies have also examined ASHAs' knowledge of AI and their perception of AI-based diagnostic applications, and found that ASHAs possess low levels of AI understanding and a tendency towards overreliance on AI [76]. To ensure community health workers use AI technologies safely, researchers have explored methods to make AI more understandable to them [75] and designed novel tools, such as Explorable AI explanations, that allow low-skilled ASHAs to test their expectations of AI behavior against its actual behaviors, thereby reducing AI overreliance [94].

With the advent of LLMs, a growing body of work has explored how these emerging technologies can address the informational needs of community health workers [3, 5, 26, 61, 82, 96]. For instance, Al Ghadban et al. [3] developed an LLM that applies retrieval-augmented generation over a custom knowledge base, and Gangavarapu [26] proposed a fine-tuned LLM. Unlike our study, both these works focus on technical development and benchmarking rather than deployment with end-users. Similarly, Taylor et al. [96] developed a custom LLM-powered application to provide on-demand guidance to community health workers in Malawi. They evaluated the tool by conducting pilot focus group sessions, which indicated its potential for delivering point-of-care assistance in low-resource settings, but without a field deployment study. McPeak et al. [61] explored the use of LLMs to aid Nigerian community health workers in decision-making by offering a 'second opinion'. This system relies solely on notes from patient interactions and

does not address additional questions that may arise. Our work contributes to this emerging body of research, by conducting a deployment study with 35 community health workers to assess the factors influencing their interaction with *ASHABot*, an LLM-powered, WhatsApp-based expert-in-the-loop chatbot designed to meet their information needs.

Finally, our field study was informed by design guidelines for AI in frontline health [40, 75, 76]. For example, Ismail and Kumar [40] advocate for strong partnerships with community organizations and Okolo et al. [75] emphasize the need for training and ongoing technical support during deployment. Throughout the process, we collaborated closely with our partner organization Khushi Baby, who played a crucial role in onboarding community health workers onto *ASHABot* and providing them with continuous support. In the following sections, we describe the design of *ASHABot* and the methodology used to evaluate its efficacy in addressing the informational needs of ASHAs.

3 System Design

ASHABot is an LLM-powered, expert-in-the-loop, WhatsApp-based chatbot system. It is built on an open-sourced framework called BYOeB¹, which was originally designed, developed, and deployed in a hospital setting for cataract surgery patients by Ramjee et al. [86]. *ASHABot* (Figure 1) responds to ASHAs' questions by searching a doctor-curated knowledge base and using GPT-4 to generate an answer from the retrieved information. If no answer is found, it forwards the question to ANMs, gathers their responses, generates a consensus answer, and sends that back to the ASHA. Additions to the knowledge base are based on the ANMs' answers and approved by medical doctors on our research team, enabling the bot to improve over time.

The *ASHABot* system incorporates a few key modifications from the original BYOeB framework. First, in BYOeB, each query is sent to a single expert for verification, which is well-suited to hospital settings where doctors have extensive medical training and expertise. In the *ASHABot* context, ANMs serve as the experts. They typically have limited formal education, supplemented by only 18 months of training without a professional degree [56, 107]. To ensure a more reliable and comprehensive verification process, *ASHABot* uses a crowdsourcing approach for answers that cannot be found in its knowledge base, aggregating responses from multiple ANMs (Figure 2). The final answer shared with the ASHA is determined by adopting the most common response among the ANMs. In cases of conflicting opinions, majority voting is employed. If fewer than three relevant responses are received, or if a tie occurs during majority voting, the system automatically forwards the query to additional ANMs. The detailed prompt used in this consensus-obtaining process is available in Appendix A.1. This prompt was refined through multiple iterations over a synthetic dataset that includes questions posed by ASHAs and answers provided by ANMs.

Second, in the BYOeB framework, every LLM-generated response is sent to an expert for verification. In *ASHABot*, however, only questions that result in an 'I don't know' response—indicating the chatbot cannot generate an answer from the curated knowledge

¹Build Your Own expert Bot: <https://github.com/microsoft/byoeb>

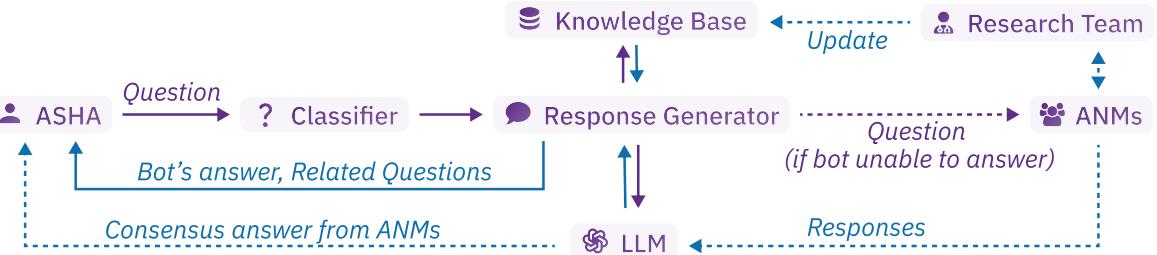


Figure 1: Flow diagram of the ASHABot system.

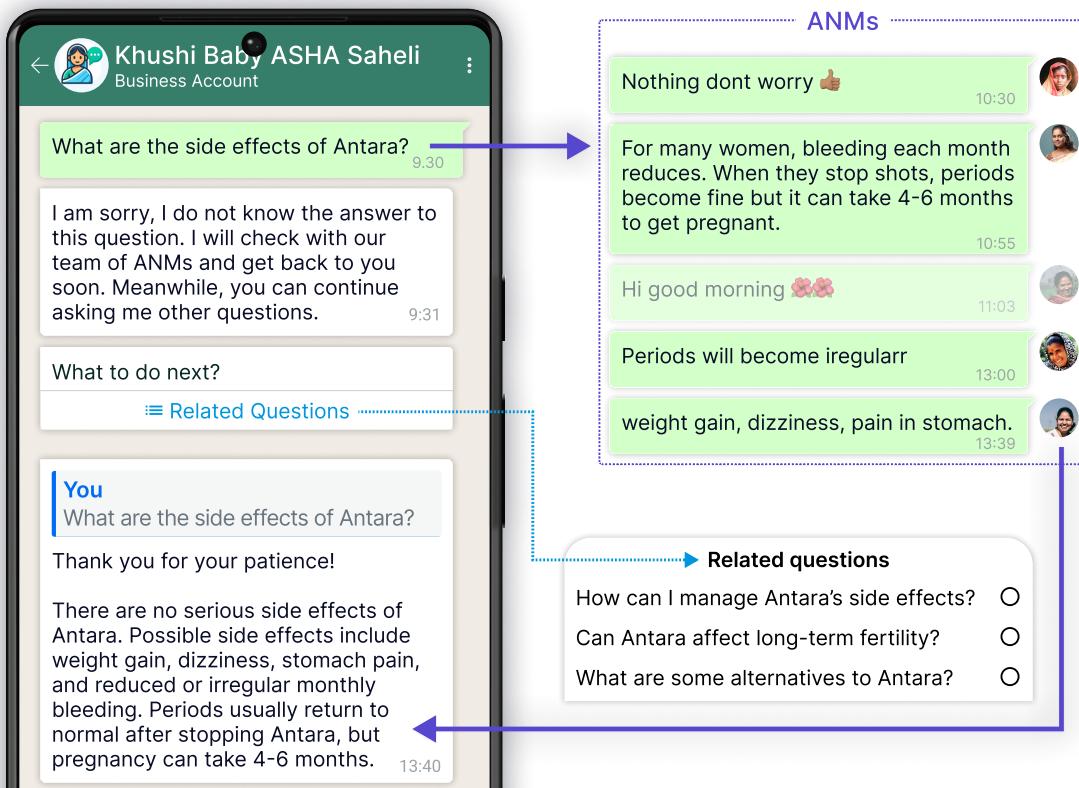


Figure 2: When ASHABot cannot answer a question using its knowledge base, it sends that question to multiple ANMs. It identifies the relevant information from their responses and generates a consensus answer, which it sends back to the ASHA. While we present this example in English, we note that all ASHAs and ANMs interacted with ASHABot in Hindi.

base—are forwarded to ANM experts (Figure 2). This design choice was made to balance scalability with the demanding schedules of ANMs, who are often busy and overstretched [110]. To further reduce ‘I don’t know’ responses and minimize the workload of ANMs, we iteratively improved the custom knowledge base. Initially, the knowledge base was prepared using the seven-part ASHA training module [30]. Khushi Baby then collected 197 questions frequently asked by ASHAs and generated responses from ASHABot for each question. ASHABot was unable to answer 76 of these questions, primarily due to gaps in the knowledge base. To address this, we

supplemented ASHABot’s knowledge base with additional documents specific to those topics and iterated the process twice.

Third, for questions asked by ASHAs using the speech modality, ASHABot provides ANMs with both the original audio recording and the transcribed text version. This approach addresses limitations identified by Ramjee et al. [86] where only the text version was shared with experts, leading to a few instances of misunderstandings due to transcription errors. To accommodate varying levels of (technological) literacy, ANMs can provide their answers via either text or voice messages, unlike the text-only approach used for experts by Ramjee et al. [86].

Lastly, we intentionally defined a female gender for *ASHABot*. In deference to Indian cultural sensitivities about conversations related to pregnancy and childcare [101], we avoided an explicit male presence in our system. As shown in Figure 2, *ASHABot*'s display name on WhatsApp was '*Khushi Baby ASHA Saheli*', which translates to '*a female friend of ASHAs*' in English.

3.1 Pilot Deployment and Response Evaluation

In March-May 2024, we deployed *ASHABot* with test users from Khushi Baby and our research team with 19 onboarded as ASHAs and 10 as ANMs. Test users from Khushi Baby included field workers, who serve as direct liaisons with ASHAs and ANMs. These field workers asked 374 questions, reviewed *ASHABot*'s responses, and provided feedback on their accuracy and completion. Of these, 273 responses were marked as accurate and complete, 17 as accurate but incomplete, 9 as inaccurate, and 75 were 'I don't know' responses. Of the inaccurate responses, 3 were due to discrepancies (e.g., dosage inconsistencies) in the government-approved documents. We addressed these by manually updating the knowledge base with doctor-verified answers, and instructing the bot to prioritize recent updates while generating responses. The 'I don't know' responses were mainly due to translation and transcription errors. For example, *ASHABot* mistranslated 'ASHA' in the Hindi question to 'hope' in English, resulting in an 'I don't know' response. To resolve this issue, we experimented with different query inputs to the GPT-4 model: (a) the original Hindi query, (b) the English-translated query, and (c) both the English-translated query and the Hindi source query. The number of 'I don't know' responses for these methods were 32, 35, and 32, respectively. We decided to input both Hindi and English queries into the system. This approach further distinguishes our method from that of Ramjee et al. [86], who only used English queries.

To address transcription failures, we developed a dictionary, initially containing 4 word pairings. For instance, the term 'Antara', referring to an injectable contraceptive (medically known as Depot Medroxy Progesterone Acetate [87]), was often misheard by the bot as 'Anna'. Second, to assist ASHAs when their audio query resulted in an 'I don't know' response, we included a transcription of the audio processed by *ASHABot* alongside the 'I don't know' message. This allowed ASHAs to understand how their question was interpreted and to refine their question accordingly. Pilot feedback also indicated that the Hindi used by the bot was perceived as "*too pure*". As a result, instead of using Azure for English-to-Hindi translation, we experimented using GPT-4 for translation. We requested Khushi Baby to evaluate the understandability of 26 bot responses translated by both Azure and GPT, using a 5-point Likert scale, where 1 represented 'Very easy' and 5 represented 'Very hard'. The average rating for Azure was 1.92 ± 1.29 , while for GPT, it was 1.58 ± 0.90 . Hence, we decided to use GPT for translation of the output.

3.2 Evaluation of *ASHABot*'s Responses by Medical Professionals

Given the deployment of *ASHABot* in a high-stakes healthcare context, we rigorously evaluated its responses using multiple approaches (Table 1).

First, we randomly selected 50 questions posed by test users and compared *ASHABot*'s responses against responses independently generated by three expert groups: doctors, ANMs, and Community Health Officers (CHOs)². These expert-generated responses served as the ground truth for our analysis. For this evaluation, we used a prompt (Appendix A.2) which included six metrics based on those identified in prior literature [1]. The metrics were (1) 'Accuracy', the correctness of the information in the answer, compared to the ground truth, (2) 'Subset', the degree to which the information in the answer is contained within the ground truth, (3) 'Completeness', the degree to which the information in the ground truth is contained within the answer, (4) 'Conciseness', the degree of repetition in the answer, (5) 'Clarity', the simplicity of medical information in the answer, and (6) 'Structure', the logical arrangement of information in the answer compared to the ground truth. Each metric used a 3-point scale, except for 'Accuracy', which used a 2-point scale but included an 'NA' category for indeterminate answers which included information not present in the ground truth. We refined the evaluation prompt iteratively by using it to evaluate *ASHABot*'s answers to randomly chosen questions. Each iteration involved comparing the LLM's evaluation with a manual evaluation, conducted by the first author using the same six metrics.

In addition to the automated evaluation, two doctors also manually assessed the responses generated by *ASHABot* based on three metrics: 'Accuracy', 'Subset', and 'Completeness'. Doctor 1 evaluated 63 responses, and Doctor 2 reviewed 100. The results of these evaluations are also summarized in Table 1. Our evaluations found that *ASHABot* achieved an accuracy of 85% or higher in all cases, which provided us with the confidence to deploy *ASHABot* among ASHAs and ANMs in the field, which we describe next.

4 Field Deployment and Evaluation

With the help of Khushi Baby, we deployed *ASHABot* with ASHAs and ANMs, and conducted an IRB-approved mixed-methods user study to understand their engagement with *ASHABot*. Below, we detail our deployment approach and evaluation methodology, including participant recruitment, demographic information, interviews, and data analysis methods.

4.1 Field Deployment

We deployed *ASHABot* between May-August 2024 to ASHAs and ANMs within the Salumbar block of Salumbar district in Rajasthan, India's largest state. The Salumbar district was selected due to Khushi Baby's established presence there, allowing them to obtain the necessary government permissions for deployment and identify a pool of potential study participants. In Salumbar, ASHAs usually work for 6 hours daily, visiting ~10 households in their village. On the other hand, ANMs work in local clinics (subcenters), with 1-2 ANMs providing basic healthcare and family planning services to 4-5 villages. ASHAs typically meet their ANM supervisor once a week at the subcenter. At the end of each month, members from all subcenters within a sector—usually 25-30 ASHAs, 5-7 ANMs, and 2-3 CHOs—gather at the primary health center for a review meeting led by a doctor. Each sector also maintains an official WhatsApp

²CHOs are frontline health workers with a bachelor's degree in nursing, who supervise ANMs in rural areas [31].

Table 1: Evaluation of ASHABot’s answers. First, we used an LLM prompt with six metrics to compare the bot’s answers to Ground Truth (GT) answers independently generated by three human experts (Doctor, CHO, and ANM). Then, two doctors manually evaluated the bot’s answers as per their knowledge, using three metrics.

	Accuracy	Subset	Completeness	Conciseness	Clarity	Structure
GT = Doctor	87.88%±26.32%	72.67%±16.26%	67.33%±10.74%	100.00%±0.00%	98.00%±8.00%	74.00%±24.09%
GT = CHO	93.65%±20.05%	65.99%±22.27%	63.27%±19.74%	100.00%±0.00%	98.67%±6.60%	65.97%±29.89%
GT = ANM	96.08%±16.17%	63.95%±22.18%	59.86%±22.45%	99.33%±4.76%	96.00%±11.04%	61.90%±30.46%
Doctor 1	93.65%±19.73%	96.30%±10.56%	89.42%±15.64%			
Doctor 2	85.00%±19.75%	82.00%±17.36%	94.67%±12.28%			

group, primarily for announcements, with ASHAs typically contributing field photos as proof of their work or raising issues like delayed payments. Beyond this, ASHAs actively participate in informal WhatsApp groups for emotional support, where they discuss common challenges such as limited incentives or overwhelming workloads.

A field worker from Khushi Baby, who had prior experience working with ASHAs and ANMs, facilitated participant recruitment. The field worker introduced the participants to ASHABot and explained the study protocol, which involved using ASHABot for three months and participating in an interview with our research team during that period. A total of 20 ASHAs and 15 ANMs agreed to participate in the study. Their phone numbers were added to ASHABot’s database, and the bot sent them ‘welcome messages’ with a ‘Yes’ option as consent for opt-in. The field worker then conducted brief training sessions (~15 minutes) with groups of 2-5 participants. For the ASHAs, the training included an overview of the bot’s purpose, demonstrations of how to ask questions using text, voice, and related-question inputs, and a reminder that the bot’s responses might occasionally be inaccurate. ASHAs were encouraged to report inaccuracies and consult their ANMs if unsure. For the ANMs, the training highlighted their role in addressing ASHAs’ questions the bot could not answer. The field worker demonstrated input methods and explained how their responses would be relayed anonymously to ASHAs after reaching a group consensus. Throughout the 3.5-month deployment, the field worker maintained regular phone contact with all participants, addressing any issues or questions and offering encouragement. One of the authors also visited five participants in-person during the first two weeks of deployment to gather early feedback and address any questions or concerns.

We logged participants interactions with ASHABot for research analysis. To promote open communication, ASHAs’ conversations with ASHABot were not shared with their supervisors, ensuring privacy and avoiding additional burdens on supervisors to review these logs. Participants were not compensated for this study, although they received financial support, including covering their mobile data cost, from Khushi Baby as part of a separate project deployment.

4.2 Semi-structured Interviews

After ASHAs and ANMs had used the bot for two months, we interviewed them over a one-week period. Following Khushi Baby’s recommendation, we focused on active ASHABot users. Specifically,

we included ASHAs who had asked at least 10 questions and ANMs who had answered at least 2 questions. As a result, we interviewed 18 out of the 20 ASHAs and 10 out of the 15 ANMs.

The first two authors conducted semi-structured interviews in Hindi. One researcher led the interviews, while the other took extensive notes. To ensure participants’ convenience, the interviews were held in-person at local nursery schools, health centers, hospitals, or participants’ homes. The interviews began with chit-chat to create a welcoming environment and then explored the participants’ overall experience with ASHABot. We focused on specific instances of why, when, and how they used the bot, encouraging comparisons with their existing methods for seeking or disseminating information. Discussions included specific features they liked and disliked, suggestions for improvement, and their views on ASHABot’s trustworthiness and accuracy. At the end of the interview, any misconceptions about the chatbot were clarified. Each interview lasted ~30 minutes. With participants’ consent, we audio-recorded 25 interviews; three participants did not consent to recording. After the interviews, participants continued using ASHABot for an additional month. Table 2 presents the demographic details of the interviewed participants.

4.3 Data Analysis

The first two authors translated and transcribed the 25 interview recordings, totaling 12 hours and 10 minutes, into English. They also digitized their respective field notes. The first author then conducted a thematic analysis of these transcripts and field notes, applying open-coding on a line-by-line basis. Throughout the analysis, four authors regularly met to review, merge, or eliminate codes. We identified 124 unique codes in the ASHA transcripts and 96 in the ANM transcripts. Through peer debriefing, we organized our findings around seven key themes, such as ‘Learning’, ‘Authoritativeness’, and ‘Motivators for Usage’. We supplemented these qualitative insights with quantitative data derived from the interaction logs, including the count, type, and frequency of ASHAs’ questions, as well as responses from ASHABot and ANMs.

4.4 Ethics

Community health care presents a sensitive landscape within which, without proper safeguards, research activities could lead to unintended negative consequences. We took a number of steps to design, deploy, and evaluate ASHABot ethically and responsibly. To begin with, our team included HCI researchers, medical doctors, and experienced staff from Khushi Baby, with over a decade of experience

Table 2: Demographic details of interviewed ASHAs and ANMs

	ASHAs (n = 18)	ANMs (n = 10)
Gender	Female	Female
Age (yrs)	34.7 ± 6.21	33.3 ± 8.69
Education	<10th grade: 3, 12th grade: 9, Bachelors: 4, Masters: 2	<10th grade: 1, 12th grade: 4, Bachelors: 2, Masters: 3
Work experience (yrs)	10.7 ± 4.81	7.90 ± 9.13
Smartphone usage (yrs)	5.75 ± 3.03	5.15 ± 3.49
Languages known	Hindi: 18, Mewari: 16, Wagdi: 1, Dawadi: 1, Gujarati: 1, Sindhi: 1	Hindi: 10, Mewari: 9, Wagdi: 2, Marwari: 1, Gujarati: 1, English: 1

working directly with ASHAs and ANMs. This interdisciplinary expertise, coupled with a commitment to community-engaged research, guided the design and evaluation of *ASHABot* to meet the specific needs of ASHAs and ANMs.

Recognizing the potential for LLMs to hallucinate [28], we conducted iterative pilot testing to refine the bot’s knowledge base and periodically evaluated its performance. Collaborating with medical doctors and field workers, we incorporated their feedback to ensure that the bot’s responses were medically accurate, complete, and contextually and linguistically appropriate. We also systematically evaluated *ASHABot*’s performance via manual reviews of its responses by field staff and two doctors on our team, and also by comparing its answers to those provided independently by doctors, ANMs, and CHOs. It was only after receiving quantitative reassurance of the bot’s accuracy as well as approval from the partner organization and local government officials that we deployed *ASHABot* to ASHAs and ANMs in the field. To further enhance reliability, we employed conservative confidence thresholds for *ASHABot*’s responses. Questions beyond the bot’s knowledge were forwarded to ANMs, whose collective responses were synthesized and relayed to ASHAs. Considering prior evidence of high AI over-reliance among ASHAs [75, 76], participants underwent training before deployment. During these sessions, ASHAs were cautioned about potential inaccuracies in the bot’s responses and advised to consult their supervisors when in doubt.

Finally, throughout the deployment, our research team and Khushi Baby staff actively monitored the questions posed by ASHAs and the responses provided by both *ASHABot* and ANMs. This continuous oversight ensured the system’s effectiveness and maintained its alignment with the needs of the community.

4.5 Positionality

All the authors are from India, and four have extensive experience conducting fieldwork in disadvantaged Indian communities. Three authors have been part of the partner organization, Khushi Baby, for several years, and one is a field worker who lives and works closely with community health workers on a daily basis. We recognize, however, that our socioeconomic status creates a power imbalance between us and our participants, who are women working in low-income, highly patriarchal contexts. A male field worker with strong community connections attended the beginning of the interviews to build rapport but then stepped away to allow participants to discuss sensitive topics freely. To create a comfortable and respectful environment for ASHAs and ANMs, two female researchers conducted

the interviews in Hindi, addressing participants with respectful terms like “*ji*”, and “*ma’am*”, mirroring the participants’ own use of these terms. This approach elicited nuanced responses. In this work, we aim to use our relative privilege to amplify the voices of our participants, highlight their information needs, and critically examine the potential of emerging LLM-powered technologies to strengthen the community healthcare ecosystem.

5 Findings

Our analysis revealed that *ASHABot* effectively met the information needs of ASHAs and served as an authoritative resource during their fieldwork. Section 5.1 presents usage statistics extracted from the interaction logs to describe how ASHAs and ANMs engaged with the bot (RQ1). Section 5.2 describes key factors like convenience, human infrastructure limitations, privacy, and accountability that affected ASHAs’ and ANMs’ engagement with the bot (RQ1 and RQ2). Finally, Section 5.3 describes how *ASHABot* supported ASHAs’ learning and examines issues of trust, expertise, and authority (RQ2 and RQ3).

5.1 Bot Usage by ASHAs and ANMs

Over the 107 days of deployment, ASHAs and ANMs used *ASHABot* extensively (Figure 3). ASHAs sent 1,761 messages, averaging 88.05 ± 123.01 messages and 0.83 ± 1.16 daily messages, per ASHA. The high standard deviations are due to two ASHAs who sent no messages, six who sent 100+, and one (ASHA1) who sent 555 messages. ASHAs sent 37.0% of messages between 5–9 AM, using the early hours to prepare for their workday or reflect on the previous one. ASHAs prioritized ease of use in their interactions. They sent 1,040 messages by tapping on the ‘Related Questions’ that the bot provided them with, recorded 470 audio messages, and typed just 251 text messages. Table 3 shows question topics and corresponding examples of questions asked by the ASHAs. *ASHABot* classified 1,612 (91.5%) of ASHAs’ messages as clinical queries, covering a broad range of topics beyond pregnancy and childcare. Of the remaining messages, 72 were smalltalk, such as ‘Hii’, ‘Thank you 🙏’, and ‘आपसे प्रश्न पूछकर बहुत अच्छा’ [‘It’s great to ask you questions’] (ASHA4), and 77 messages were ignored due to a system bug (Figure 3). *ASHABot* responded with ‘I don’t know’ to 264 clinical questions, forwarding them to ANMs.

When *ASHABot* prompted ANMs to answer a question, they frequently ignored that. Out of the 6915 prompts sent (461 ± 4.40 prompts/ANM), ANMs responded ‘Yes’ to only 1491, therefore receiving 99.4 ± 99.78 questions/ANM. In response to the questions,

Table 3: Question Topics and Examples

Question Topic	Example (Original Question)	Example (English Translation)
Pregnancy	'अगर एक गमेवती महिला एचाएइवी पॉजिटिव हो तो उसके लिए क्या करें?' 'गमेवती के कितनी बीपी कम है तो माना जाता।'	'What to do if a pregnant woman is HIV-positive?' 'For a pregnant woman, what BP level is considered low?'
Childcare	'इस बच्चे को जन्म के तुरंत पीलिया होता है वो कैसे हो जाता है?' 'एक बच्चा दो किलो का है तो उसकी देखभाल किस तरह करनी चाहिए? घर?'	'This child has jaundice immediately after birth, how does it happen?' 'If a child weighs two kilos, how should he be cared for? At home?'
Family planning	'अंतरा इंजेक्शन लगवाने के बाद महिला की महावारी कूँ रुक जाती है'	'Why does a woman's period stop after getting Antara injection?'
Societal issues	'प्रियंका और प्रियंका के पति का झगड़ा हो जाता है। क्या करना पड़ता है प्रियंका?' 'बाल विवाह होना चाहिए या नहीं?'	'Priyanka and Priyanka's husband get into a fight. What does Priyanka have to do?' 'Should there be child marriage or not?'
General healthcare	'एसिडिटी होती है तो क्या करना पड़ता है?' 'Kya aayushman yojna pure bharat me lagu hai'	'What to do if acidity occurs?' 'Is Ayushman Yojana applicable to the whole of India?'
General knowledge	'साइबर मशीन क्या है?'	'What is a cyber machine?'

ANMs sent 788 messages overall (52.53 ± 68.93 responses/ANM, and 0.46 ± 0.63 daily responses/ANM). Similar to ASHAs, three ANMs sent no messages, three sent 100+, while one (ANM1) sent 244. ANMs sent 37.9% of their responses before their work day began at 9AM, with activity peaking at 1PM during lunch breaks. After 5PM, ANMs did not send any responses. Unlike ASHAs, only the minority of ANM responses used audio modality rather than text (3.2%, or 1.67 ± 3.06 responses/ANM).

To generate a consensus answer, *ASHABot* required agreement from at least three ANMs. As ANMs averaged 2.28 ± 1.76 responses per question, consensus was reached for just 19 questions. Individual ANMs took 30.75 ± 66.46 hours to respond, and the consensus answer was shared with the ASHA within 58.59 ± 99.18 hours. By that time, ASHAs typically had already found answers elsewhere and often overlooked *ASHABot*'s consensus answers. This suggests that ASHAs primarily relied on *ASHABot*'s automated answers, which accounted for 83.6% of responses to clinical questions. A poll was sent along with 20% of the bot's answers, asking 'Is this answer helpful?'. 95.7% of ASHAs' responses were 'Yes', suggesting its effectiveness. To more deeply understand the ASHAs' and ANMs' lived experiences with *ASHABot*, we proceed to examine factors that shaped their engagement with it.

5.2 Factors Impacting Users' Engagement with the Bot

5.2.1 Motivators for Bot Usage. ASHAs described several motivations to use *ASHABot*.

Convenience. ASHAs found the bot highly convenient, allowing them to ask questions anytime, anywhere, without waiting for or traveling to meetings with ANMs and doctors. Previously, ASHAs might forget their questions that they intended to ask their ANMs later, but *ASHABot* enabled them to ask questions as they had them, even "*at night also*" (ASHA3). The bot provided instant answers, addressing their doubts immediately. ASHAs and ANMs reported that they often used *ASHABot* during their free time, such as while travelling on the bus. Some ASHAs, especially those with difficulties typing or reading in Hindi, found the audio features for input and

output quick and easy to use. ASHA12 noted a preference for voice messages: "*It is... difficult to text in Hindi, particularly half letters... There are chances of spelling mistakes.*" Overall, 26.7% of ASHAs' questions were in audio format, suggesting the convenience that the audio feature brought to them. On the system end, *ASHABot* was robust enough to recognize a few translation and transcription errors while addressing these audio questions, as it received both Hindi and English versions of the question. For instance, when ASHAs' audio question was incorrectly transcribed as asking about a baby's 'slow mother-in-law' (सास, or *sas*) instead of a baby's slow breathing rate (साँस, or *saans*), *ASHABot* began its response with, 'I believe there may be a translation error in your question. If you are asking about what to do if the newborn baby's breathing is slow...'. On the other hand, 96.8% of ANMs' responses were text-based. They preferred text because it allowed them to carefully revise each part of their response before sending it out, unlike audio messages. We also observed that both ASHAs and ANMs used Google Keyboard's speech-to-text feature for composing WhatsApp text messages, preferring it over WhatsApp's built-in 'voice message' feature due to the latter's less familiar press-and-hold interaction. Thus, some text messages sent to *ASHABot* were likely generated from speech inputs.

ASHAs also appreciated the persistent nature of communication on *ASHABot*, as it allowed them to reread messages when needed. However, despite the value of being able to access past interactions with the bot anytime, many ASHAs and ANMs had to regularly use the 'clear chat' feature on WhatsApp to delete messages across their personal and group chats. This was because their phones—provided by Khushi Baby over two years ago—often ran out of storage and became slow, especially when logging field data.

Overworked ANMs and Doctors. ASHAs shared that they usually seek answers from ANMs or occasionally from doctors during monthly training sessions. However, because ANMs and doctors are often busy, ASHAs reported prioritizing their questions and feeling hesitant in asking "*simple questions*" (ASHA11). ASHAs feared being judged, as they had been scolded in the past for not knowing basic

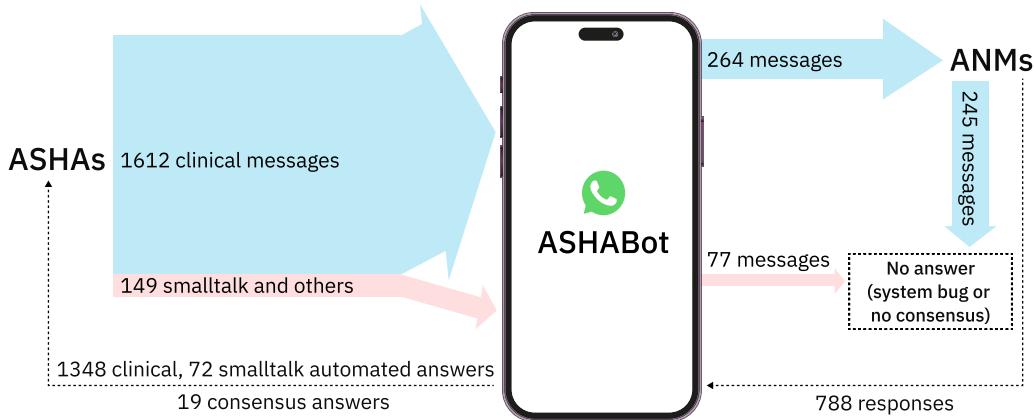


Figure 3: ASHABot deployment statistics.

information. In contrast, *ASHABot* provided a non-judgemental platform, allowing many ASHAs to ask anything without hesitation.

“Some questions stay in my mind, I do not ask ANMs or doctors... but with the bot, I can ask anything... (e.g., Once) there was a crowd, it was chaotic, and I mistakenly gave the vaccine (intended) for a 3.5-month-old child to a 2.5 month old and vice versa. I asked ASHABot... It said that ‘You can give the missed vaccine next month... so the babies can get vaccinated within 6 months.’ I was relaxed... Otherwise I would have visited those children’s homes again and again to make sure that they were okay.”
– ASHA10

With multiple responsibilities and limited time, ANMs and doctors often provide “*short*” and “*cut to cut*” (ASHA10) answers to ASHAs. In contrast, *ASHABot* offered detailed, nuanced responses, equipping ASHAs to make informed decisions. ASHAs felt that they could ask *ASHABot* as many follow-up questions as needed without any concerns for time constraints or expecting impatience, like they had to experience sometimes with ANMs and doctors. This freedom allowed them to explore topics in depth and revisit information until they fully understood it.

“I asked about haemoglobin deficiency... The bot said ‘If you eat this vegetable, this will happen, eat daal [lentils], eat this, eat that.’ It’s a detailed explanation, so I understand it well. But ma’am [ANM] only tells me the main point: ‘The haemoglobin needs to be at this level... take the medicines, done.’ The (bot) is better, because the patient might not even need to take medicines.” – ASHA3

Older ASHAs found *ASHABot* especially useful for staying up-to-date on new information on vaccinations, medicines, and contraceptive methods. The bot helped them avoid the embarrassment they sometimes felt when asking questions to ANMs, who were often younger than them. Reflecting on their own initial struggles, these older ASHAs also acknowledged *ASHABot*’s importance for newer ASHAs, who often lack experience and confidence, and feel hesitant to approach ANMs and doctors.

We note that despite ASHAs’ shift to using *ASHABot*, it was not seen as a replacement for the feedback and support provided by ANMs, nor did it undermine the ANMs’ perceived authority. Instead, ANMs recognized *ASHABot* as a time-saving tool that could potentially reduce their workload by offering ASHAs the ability to independently resolve their questions. Many ASHAs noted that their ANMs encouraged them to use the bot not only to address their immediate doubts, but also to expand their knowledge beyond the basics.

Asking Sensitive Questions. ASHAs were often hesitant to discuss sensitive topics, such as women’s sexual health, with male doctors and, at times, even with ANMs. *ASHABot* provided a private channel to address these queries, which many ASHAs found beneficial.

“The bot is better... I cannot openly ask the Doctor sir about white discharge, about itching and burning...”
– ASHA3

However, a few ASHAs refrained from asking the bot questions they considered too sensitive. When asked why, they expressed concerns that Khushi Baby might be monitoring their queries, extrapolating from previous experiences where Khushi Baby was involved in monitoring their use of other data collection applications. Misunderstandings about how *ASHABot* operated exacerbated this issue. A few ASHAs mistakenly believed that a human controlled the system, making them hesitant to ask such questions. For example, ASHA3 typed and then deleted a question about menstruation, assuming that the male field worker who introduced her to *ASHABot* was the one responding. ASHA17 believed a woman acted as *ASHABot* due to the female voice in the audio messages she received from *ASHABot*. Furthermore, because ASHAs and ANMs typically use WhatsApp groups for communication with each other and with Khushi Baby, some assumed *ASHABot* could identify them and relay their identity along with the questions to ANMs. For instance, ASHA16 hesitated to ask about domestic violence, fearing her questions could be traced back to her.

“Even though I want to ask, I cannot, because it will take a toll on the woman’s family... I was scared that

if I ask the bot, you will have all the details that (this) ASHA has asked certain questions... It's in the records... If that gets leaked, then I will be asked more questions." – ASHA16

5.2.2 Barriers to ANMs' Usage. Most ANMs felt that there were no incentives, financial or otherwise, for them to use *ASHABot*, making it just another online task that added to their effort and screen time. Some ANMs felt overwhelmed by the number of questions asked to them. They also felt that some of those questions were too basic and increased their workload unnecessarily.

"Sometimes there are so many silly questions about pregnancy and vaccinations... ASHAs have so many years of experience, they should already know these things. Why do they ask such questions?" – ANM4

This reinforces the ASHAs' concern that they hesitate to ask questions because they fear judgment from their supervisors.

A fear of accountability also influenced ANMs' use of *ASHABot*. Some ANMs were reluctant to admit their lack of knowledge of certain answers on record, fearing negative consequences. As ANM8 expressed, "*Everyone will know that I don't know the answer... What if it is a simple question? They will judge me.*" Consequently, most ANMs chose to ignore the questions they could not answer rather than respond with 'I don't know' on the bot.

ANMs also struggled to answer questions that were ambiguous or not clearly phrased. This was sometimes due to a lack of context, such as when ASHAs asked about medicine dosage without specifying whether the patient was a mother or a newborn (ANM5 and ANM9). At other times, the lack of clarity was due to transcription and translation errors. 38.3% of the questions sent to ANMs were audio messages, as regional dialects and background noise sometimes hindered the bot's ability to directly answer those. For instance, when the vaccine 'BCG' (Bacillus Calmette-Guérin, a tuberculosis prevention vaccine used in India [73]) was mistakenly transcribed as 'BCC', ANM6 chose to ignore the question, assuming it referred to an unfamiliar vaccine. ASHAs' original audio messages, sent along with the transcribed text, occasionally helped clarify these issues. When 'वजन' [vajan, meaning weight] was transcribed as 'भजन' [bhajan, meaning devotional song], ANMs answered the question correctly by listening to the corresponding audio. However, the effectiveness of audio was limited, as ANMs were often reluctant to listen to the audio messages, especially in public spaces where playing it could be "awkward" (ANM8). Such scenarios where questions were incomplete or incorrectly formatted often required back-and-forth message exchanges between ASHAs and ANMs, which was not supported by *ASHABot*.

"There was a question, 'For pregnant women, how much time is safe for walking?' It depends on the pregnancy. If it is a high risk case, bed rest is advised... totally depends on the prescription of doctors. I didn't answer... If I answer anything wrong, the chances of risk would rise." – ANM8

As a result, ANMs preferred answering ASHAs' questions in person rather than using the bot. In-person interactions allowed ANMs to better understand ASHAs' queries and address follow-up

questions, enabling them to provide more detailed answers, compared to the "brief or one-word" (ANM8) responses given via the bot. A minority even felt the bot reduced their individual liability for providing accurate and complete answers, as it aggregated responses from multiple ANMs.

Finally, when responding to questions, ANMs found *ASHABot*'s default reply, 'Thank you for your answer 🙏' unsatisfactory, as it did not offer meaningful feedback. They complained that the bot thanked them regardless of the content, even when they responded with 'I do not know' or provided incorrect information. ANMs wanted to know whether their answers were correct, and if not, what the correct answer was. They desired constructive feedback to improve their knowledge. When we clarified that *ASHABot* sought their input because it lacked the correct answers, ANMs suggested that they should at least be able to view the responses provided by other ANMs.

5.3 Bot's Role in Addressing the Information Needs of ASHAs

5.3.1 Learning. *ASHABot* contributed to ASHAs' learning experience by addressing their immediate knowledge needs and helping them confirm familiar information. Older ASHAs reported that it allowed them to revisit and reinforce knowledge gained from previous training sessions. ASHAs also noted that the bot was particularly useful for confirming specific numerical details. For instance, the Centchroman contraceptive tablet [47] has a complex dosage schedule that changes weekly, making it difficult to recall.

Beyond routine work, ASHAs used the bot to explore advanced topics, equipping themselves with knowledge that might be useful in future. For example, ASHA2 consulted *ASHABot* on 'home-based care' for babies with low birth weight. She recalled feeling unprepared in the past when families could not afford hospital treatment for their child and wanted to be ready "*just in case*" similar situations arise again. ASHAs also valued the bot's ability to address questions beyond public health, such as about government schemes or financial planning. With *ASHABot*, they could provide immediate information to care recipients, without redirecting them to other sources of information, such as banks or post offices. Even outside of work, ASHAs sometimes sought information on *ASHABot* prompted by their personal experiences. For instance, after witnessing a heart attack at a local temple, ASHA3 used the bot to gain clarity on the incident: "*What could have happened? He died so suddenly... it was on my mind. So, when I came back... I asked questions about heart attack.*"

The Related Questions feature proved beneficial for ASHA workers, enabling them to ask follow-up questions and deepen their understanding. Usage logs revealed that 59.1% of the questions asked by ASHAs were Related Questions, which they found challenging to formulate on their own. This feature also introduced ASHAs to medical terminology they were otherwise unfamiliar with, such as "*toxoplasmosis*" (ASHA13).

The resulting increase in knowledge improved ASHAs' practical readiness for monthly training sessions, making them more active and confident participants. As ASHA16 noted:

"[In these training sessions] We get asked a lot of questions, related to HBNC, geriatrics, etc., and I could

answer them... It (The bot) increased my confidence."
– ASHA16

ASHAs' Strategies to Handle ASHABot's 'I don't know' Responses. ASHABot responded with 'I don't know' to 16.4% of ASHAs' questions. In 92.8% of such cases, ASHAs did not receive a subsequent consensus answer from ANMs too. They then employed alternative strategies based on their existing knowledge and the urgency of the question. For critical questions, ASHAs consulted other sources, such as ASHA handbooks, online resources, ANMs, or doctors.

"I asked the bot about accidentally giving too much iron syrup to infants... It is hard to give the correct dosage when they move their hands... The bot could not answer, so I asked the CHO [doctor], who reassured me that... an accidental extra drop or two wouldn't cause harm." – ASHA16

When ASHAs had partial knowledge, they relied on their existing information to continue their tasks, often retrying the bot later. For instance, ASHA4 thrice asked for a list of hospitals offering free treatment under a specific government scheme. When ASHABot could not provide an answer, she directed her care recipients to the two hospitals she knew. Finally, for questions prompted by curiosity rather than immediate need, ASHAs typically accepted 'I don't know' responses without further action. For example, ASHA9, a yoga practitioner, inquired about appropriate levels of exercise for pregnant women. When ASHABot could not provide an answer, she did not attempt to seek the information elsewhere.

ANMs' Learning. ASHABot also contributed to learning for ANMs. ANM4 even regarded ASHABot as a productive use of her smartphone, remarking that it "reduces my YouTube time." Although ANMs were generally familiar with the answers to most of the questions raised on ASHABot, when ASHAs initiated detailed follow-ups and advanced medical queries, often through Related Questions, ANMs were encouraged to reflect deeply, expand their knowledge and "stay updated" (ANM8). These complex questions often led ANMs to consult senior experts or conduct online research, which they then shared. For instance, when ASHA6 asked 'What precautions should be taken while donating eyes?', ANM1, who viewed ASHABot as a "challenge" to test both her knowledge and resourcefulness, consulted a doctor at the local hospital. She conveyed her learnings via ASHABot: 'The donation must be completed within four hours of death'.

Despite this proactive approach, the new knowledge gained by ANMs was not always communicated back to the bot. For example, when ANM3 researched 'What are the symptoms of ectopic pregnancy?' on YouTube, she understood the answer but did not add it to ASHABot because the information was too lengthy to summarize, and the video format did not allow for easy referencing, requiring her to draft an answer from scratch, for which she had little time.

Collective Sensemaking. ASHABot facilitated collective understanding among both ASHAs and ANMs. ASHAs reported that their peers, who were not directly participating in the study, would relay their questions through them to the bot. For instance, ASHA3,

while discussing how to protect newborns from heatwaves with a group of ASHAs, used ASHABot to inquire about the topic and played the audio response aloud:

"I just raised the question, but all 16-17 ASHAs around me learned from the answer." – ASHA3

Among the ANMs, ANM5 and ANM9, who were neighbours, often collaborated by discussing and answering ASHABot questions using each other's phones. For instance, when ANM5 received a question about the purpose of folic acid, they collectively researched the topic on Google and then provided a response. Through this exercise, ANM9 learned that folic acid helps increase red blood cell counts, which she then added as a response on ANM5's phone. Overall, most ANMs indicated that they did not hesitate to ask for help when responding to ASHABot questions.

5.3.2 Trust, Expertise, and Authoritativeness. ASHAs considered the answers provided by ASHABot to be trustworthy and dependable. This confidence developed as ASHAs initially tested the bot with questions they already knew the answers to, and compared the bot's responses with their own knowledge. Sometimes when uncertain, they would verify ASHABot's answers by consulting ANMs, doctors, ASHA handbooks, or online sources, and consistently found the bot's answers to be "correct" (ASHA2) and "complete" (ASHA6).

As a result, ASHAs gradually reduced the frequency with which they double-checked the bot's answers and began to rely on it directly. For example, ASHA13 used ASHABot while visiting the home of an anaemic pregnant woman. When the woman asked for specific dietary recommendations, the ASHA asked the bot, which suggested foods such as '*daal* [lentils], rice, egg, and fish'. She promptly shared this information with the woman.

ASHAs and ANMs also appreciated the quality and precision of the information that ASHABot shared. They contrasted the bot with other tools they used to find answers such as "Google" or "YouTube". As ANM1 noted, "*If I Google a question, it does not give me a short answer... It gives the whole Ramayana [a lengthy Indian epic] of the topic, which is not useful to me.*" ANM9 observed that ASHAs often chose to click on the simplest options during Internet searches, and consumed information with little regard for its reliability. In contrast, she felt that ASHABot would provide more trustworthy and appropriate responses.

Interestingly, ASHABot's 'I don't know' responses also contributed to building trust. ASHAs appreciated that the bot either provided accurate information or admitted when it did not know the answer. However, ASHABot's transparency was a double-edged sword. 'I don't know' responses to ASHAs' audio messages included the text transcription of their corresponding audios. Reading this, the ASHAs recognized that the bot sometimes misunderstood them. For some ASHAs, this lowered their confidence in the bot's ability to meet their information needs. A few even avoided asking critical questions via audio, fearing misinterpretation. For example, ASHA6, who predominantly used the audio modality, preferred asking her ANM directly about which vaccination to use rather than consulting ASHABot. She feared that a misunderstanding on behalf of the bot could lead to her administering the wrong vaccination.

ASHAs noted that if ASHABot and their ANM provided different answers, they were more inclined to rely on the bot. ASHABot's

onboarding messages emphasised its partnership with Khushi Baby and the Ministry of Health and Family Welfare, Government of Rajasthan. This reinforced ASHAs' trust, as they believed that "*whatever the government says is correct.*" (ASHA17). Additionally, ASHAs perceived ANMs as having limited knowledge and only slightly more formal education than themselves. ASHAs had previously encountered situations where ANMs provided incorrect information, leading to this reduced credibility for ANMs compared to *ASHABot*. An ASHA recounted:

"Albendazole is generally given to pregnant women when they are in the 6th month... but that ANM gave it to one woman when she was only 3 months pregnant... I raised this issue but she said, 'It is okay, nothing will happen.' I reported this to [doctor] sir during a sector meeting." – ASHA16

In practice, there were two reported instances, both mentioned by ASHA1—the most active ASHA among our participants—in which *ASHABot*'s answers were inconsistent with the ASHAs' existing knowledge. First, the ASHA believed that iron tablets should be taken at night after food. When the bot instead recommended taking them in the morning on an empty stomach, she was unconvinced. To verify, she asked her ANM, who supported her original understanding. Second, when asked about the responsibilities of an ASHA worker, *ASHABot* mentioned '*... assisting victims of violence in accessing medical care, counselling, and legal aid.*' ASHA1 disagreed, believing her role was limited to health issues, and did not pursue the topic further with the bot or her ANM. In both cases, ASHA1 relied on the existing (though incorrect) human knowledge over *ASHABot*'s factually correct answers.

Authoritativeness. Barring ASHA1, all other ASHAs reported no inconsistencies in *ASHABot*'s answers. In fact, when a doctor provided unsatisfactory guidance to ASHA10, she turned to the bot for verification, treating it as a supplemental authority. Similarly, *ASHABot* influenced the decisions of care recipients by providing clear and reliable information. For example, ASHA3 used the bot to address concerns from two women in her village who were hesitant to receive the Antara injectable contraceptive. The women were worried that this injection would cause permanent infertility, as they had heard from others about experiencing irregular menstrual cycles after receiving it. *ASHABot*'s clarification enabled ASHA3 to effectively convince them.

"Because most women don't take Antara... and they [the ones who have taken Antara] ask us about their [irregular] periods... How to make them understand? I asked about this yesterday to the bot. It told me, 'Irregularities are just the side effects of the injection... it is temporary and does not cause infertility.' In the evening I explained to the two women about this, even showed them the bot's answer... and this morning both of them came to get the injection." – ASHA3

6 Discussion

Our study demonstrates the potential of an LLM-powered, ANMs-in-the-loop chatbot to meet the information needs of ASHA workers. Through a thorough examination of the factors that shaped the participation of ANMs and ASHAs with *ASHABot*, we now discuss opportunities and challenges in designing LLM-powered chatbots for low-wage, low-literate frontline workers like ASHAs in high-stakes settings.

6.1 Accounting for Technodeterminism

Our findings show that ASHAs used the bot for its accurate, complete, and detailed responses, which helped address their queries, support care decisions, and persuade care recipients. They were not skeptical about *ASHABot*'s responses; instead, they expressed trust and gratitude, viewing it as an authority. Prior work shows that community health workers often lack familiarity with AI-driven tools and tend to rely heavily on them [76]. Even outside of healthcare, individuals with low digital literacy often over-rely on new technologies [51] like conversational agents [43]. This trust and reliance on AI are particularly pronounced in non-Western settings like ours [2, 57]. However, such over-reliance can cause serious harm in high-stakes frontline health contexts, particularly since LLMs are prone to generating misleading information [28] and culturally inappropriate content [2]. While researchers have proposed developer-centric tools like design playbooks [36] and documentation protocols [10, 37] to enhance AI safety, the effectiveness of these tools remains underexplored [11], and AI safety continues to be a complex, evolving challenge, particularly in high-stakes, non-Western contexts where formal benchmarks are often absent.

To promote a realistic user understanding of AI capabilities and limitations, it is essential to incorporate AI competency training into user onboarding to ensure safe engagement with AI systems [50, 76, 86]. Similar to our approach, such training should emphasize the potential for errors in LLM-generated responses and their ability to produce synthetic data. It should provide specific examples of possible hallucinations and encourage critical assessment of AI outputs, and prompt users to question AI authority and verify critical information with experts when needed. This aligns with broader human-AI design guidelines, emphasizing the need to "*help the user understand how often the AI system may make mistakes*" [4].

Our findings, however, revealed that even with training, some ASHAs tended to overrely on AI-generated outputs, often overlooking the need to verify critical information. Addressing this issue requires strategies that help novice users better understand AI's limitations—an area that still poses open research challenges. Approaches like adding accuracy indicators, making data sources more transparent, or designing prompts that push users to question and validate AI responses need further exploration and validation.

Our findings also revealed that *ASHABot* was not perceived as undermining the authority of ASHAs in their interactions with care recipients, nor did it diminish the role of ANMs in supporting ASHAs. Instead, participants regarded *ASHABot* as a resource to improve ASHAs' effectiveness and reduce ANMs' workload. As LLM-powered tools become more integrated into everyday services, it is critical that these technologies are viewed as complements, not replacements, for knowledge workers. For instance in the *ASHABot*

deployment, the support ANMs offer to ASHAs extends well beyond answering questions—it includes mentorship, emotional support, and guidance in navigating complex healthcare situations in messy, resource-constrained environments. The designers and developers of AI technologies must recognize that technology cannot replace human judgment and care, a concern emphasized in prior HCI4D scholarship [97, 108]. Moving forward, LLM-powered bots should be framed as tools that assist end users, working to “*amplify human intent and capacity*” [76, 97] rather than as solutions that can independently address complex societal problems. By positioning these bots as supplemental, fallible tools, we can maintain trust in human elements while still benefiting from the efficiencies that AI offers.

6.2 Multimodality and Personalization

Prior work has demonstrated that speech can be an effective medium for users with low levels of digital literacy such as ASHAs and ANMs [23, 43, 86, 103]. While *ASHABot* allowed ANMs and ASHAs to interact using audio and textual modality, our deployment revealed that they faced several challenges in using audio inputs and outputs. To begin with, errors in translation and transcription significantly impacted usability and usefulness, with some ASHAs being so concerned about potential misinterpretation by the system that they even avoided asking critical questions via audio. Furthermore, the inability to review and edit audio messages before sending them affected ANMs’ use of the audio features. Also, ASHAs and ANMs hesitated to speak (or play audio responses) out loud and had concerns about being overheard. These privacy concerns were evident during interviews, where ASHAs and ANMs avoided sensitive topics or spoke in hushed tones, mirroring patterns noted by Ismail and Kumar [38].

While ASHAs and ANMs struggled with the audio, there might be unique opportunities to promote other forms of multi-modal interactions with the bot. For example, research has shown that visual aids can be highly effective for users with limited literacy [63], offering an intuitive way to interact. Recent advancements in multimodal LLMs [77] can make it feasible for bots to process and respond to image-based inputs, which ASHAs can use to submit photos and videos of growth charts, breastfeeding positions, medicine dosages, or even the data they collect. Once *ASHABot* has access to such information, it could analyze images to provide relevant answers or additional visual resources. Furthermore, by querying the data that knowledge workers collect, a multi-modal bot could provide them with timely, relevant, and personalized information in an accessible manner. For instance, the bot could integrate with users’ work calendars—which, in the case of ASHAs, include key dates like pregnancy checkups, childbirth, vaccinations, and training meetings—and send reminders at critical moments.

However, as the bot gains visibility within the frontline healthcare ecosystem, it could amplify invisible work [64] as well as tensions between stakeholders. For instance, the government may seek to leverage the bot as a tool for monitoring ASHAs’ learning and performance, in order to decide incentives and inform training programs. This objective could directly conflict with ASHAs’ desire to ask the bot sensitive questions which they feel less comfortable asking to their supervisors. If ASHAs perceive that their queries or

mistakes are under scrutiny, it could deter them from fully engaging with—and benefiting from—the system. This tension between users’ privacy needs and authorities’ monitoring goals has been noted in prior work in multi-stakeholder healthcare [48] and educational settings [100]. Hence while incorporating personalization features, it is essential to ensure that users have both awareness and control over how much of their collected data and past conversations are shared with an LLM-powered system and other stakeholders. Balancing personalization with ethical considerations in AI for health remains an ongoing discussion [58, 76, 108], which becomes even more complex when working with low-literate frontline health workers [38, 39].

6.3 Value Alignment

While facilitating collaboration between LLM-powered bots and healthcare workers, it is crucial to ensure that the underlying models have a deep understanding of the sociocultural context in which they are used. For instance, we found that some ASHAs used *ASHABot* not only to ask health-related questions but also to seek guidance on social issues, such as domestic violence and child marriage. In such critical scenarios, it is essential that the values embedded in the LLMs align with those of the ASHAs and ANMs, ensuring that the responses are culturally appropriate. Without this value alignment, there is a risk of LLMs imposing Western values and norms [2], potentially causing harm, especially when addressing complex social issues such as patriarchy, as highlighted by Sultana et al. [95].

The use of technology for frontline workers has predominantly focused on tracking, monitoring, and evaluating them for rewards and compensation [21, 35, 52, 105]. In our study, some ASHAs mistakenly assumed that their questions were being monitored, causing hesitation in asking sensitive or basic questions. Likewise, ANMs were concerned about being held accountable for providing incorrect answers and were reluctant to admit knowledge gaps. A few ASHAs asked as many questions as possible, while ANMs provided numerous answers, both assuming their interactions would be used to evaluate them. While we clarified these misconceptions during the interviews, the ways users engaged with the bot was influenced by the values and apprehensions around surveillance and tracking, instead of using the bot as a learning tool.

These findings highlight the importance of aligning AI systems with the values of frontline health workers to ensure that the technology fosters trust, rather than fear or anxiety. To achieve this alignment, future deployments should clearly communicate the purpose of the technology from the outset, ensuring that frontline workers understand it is designed to support learning, not to track or monitor performance. This proactive approach will help mitigate misconceptions and create a more supportive environment for the growth and development of frontline workers.

6.4 Governance and Accountability on Scale

Our findings show that ASHAs derived significant value from using the *ASHABot*, highlighting its potential for broader deployment. However, there are several open questions before scaling up such LLM-powered chatbots. For example: *Who governs what data should be fed into the bot and keep accountability in check? Whose voices and*

values should be prioritized in a multi-stakeholder environment? Who is accountable for potential mistakes and value misalignment? How can we prevent additional burdens on already overstretched ASHAs and ANMs?

Xiao et al. [106] emphasizes that domain expertise is essential for identifying and curating reliable and clear content for information portals. In our pilot study, medical doctors from our research team managed and updated the bot’s knowledge base, the field team ensured that the bot’s responses matched the linguistic styles of ASHAs and ANMs, and the development team ensured that the bot prioritized local values instead of enforcing Western norms. While this approach ensured that the bot was deployed in a safe and controlled environment, new governance models, fairness frameworks, and safety approaches are needed to scale such bots at a population level, especially in a country like India with over 1 million frontline health workers operating in diverse cultural and socioeconomic contexts.

If not done with care, scaling could introduce tensions between government priorities and the information needs of ASHAs and their care recipients, particularly on sensitive topics such as family planning. As noted in prior work, decisions about what constitutes ‘appropriate’ LLM responses can inadvertently privilege certain perspectives over others [45]. For instance, government-issued healthcare guidelines might emphasize the benefits and ease of contraceptive methods while downplaying potential side effects [27, 34], creating a risk of biased or incomplete information. Moreover, ASHAs’ need for detailed, regionally specific information could conflict with the government’s preference for consistent, standardized messaging. Hence, for large-scale deployment, it is crucial to center the voices and values of community health workers and use participatory approaches to design and develop emerging AI technologies *effectively*. In doing so, it is essential to be mindful of whose values go into the design and development of AI technologies, ensure that the designers and builders account for power imbalances (e.g., between ASHAs and their supervisors) and avoid tokenism, where input is sought but not genuinely considered in decision-making [17].

Additionally, preventing catastrophic failures, such as hallucinations, requires leveraging the “expert-in-the-loop” approach. In our deployment, this did not work as well as expected—many ANMs hesitated to provide answers to *ASHABot*’s queries due to time constraints, lack of knowledge, or concerns about accountability. This raises concerns about the burden placed on an already overstretched workforce and underscores the need to align ASHAs’ needs with their supervisors’ capacities realistically. While previous research has compensated experts financially for their contributions to healthcare chatbots [53], more work is needed to explore the sustainability of this approach and examine alternative strategies to keep experts motivated, such as peer comparison, which has proven effective in frontline health contexts [20]. Finally, it is crucial to limit the demands on experts, for example, by setting daily limits on how much they can contribute to be respectful of their existing workload. Reducing unnecessary expert involvement is equally important. For instance, to address referrals caused by translation and transcription errors, we recommend offering users the flexibility to rephrase and resubmit unclear questions before forwarding them to experts. It is also important to incorporate expert input into the knowledge base, enabling the bot to answer

more questions independently over time. As seen in prior work on expert-in-the-loop LLM chatbots [90], this approach is likely to progressively reduce supervisors’ bot-related workload.

6.5 Limitations

Our exploratory work has some limitations. First, we employed a small sample size and interview-based methods to assess the feasibility of *ASHABot* and understand the individual experiences of ASHAs and ANMs. While our findings suggest that *ASHABot* effectively met the information needs of these ASHAs, participant bias [19] may have influenced the results. Second, the study is limited to a specific geographic area within a single state in India, which may restrict the generalizability of the findings. The nuances of user experiences may vary if the bot is scaled across community workers in other parts of India or in other Global South contexts. Future research should involve a larger, geographically diverse group of community workers and examine their usage over an extended period to develop more robust quantitative insights, such as understanding how ASHAs’ questions evolve over time and the subsequent impact on ANMs’ workload and attitudes.

7 Conclusion

ASHAs are essential to India’s healthcare system, yet the infrastructure to improve their limited medical knowledge and skill is inadequate. To tackle this, we used a LLM-powered experts-in-the-loop open-source framework, to design and develop *ASHABot*. *ASHABot* is a WhatsApp-based chatbot that uses a doctor-curated knowledge base to provide ASHAs with instant answers to their questions. When uncertain, the bot consults multiple ANMs and generates a consensus response, which is also used to enhance the knowledge base. After extensive pilot testing to ensure our system’s accuracy and appropriateness, we conducted a field study involving 20 ASHAs and 15 ANMs. Our interviews with them, and analysis of their interaction log data, highlighted the effectiveness of *ASHABot* in addressing ASHAs’ information needs. ASHAs valued the bot as a convenient and private channel for them to ask rudimentary or sensitive questions without fear of judgement. They generally trusted the bot, especially as it admitted when it did not know an answer. Eventually, *ASHABot* played an authoritative role, influencing the decisions of both ASHAs and their care recipients. On the ANM side, they appreciated the chance to review and expand their knowledge by providing answers on *ASHABot*, but felt overwhelmed by the workload and accountability, suggesting they may need incentives to contribute regularly. Overall, we find that an LLM-powered chatbot can significantly improve access to information for those with limited literacy and technological exposure in healthcare roles. Finally, we caution against over-reliance and emphasize the need to position these resources for community health workers as a supplement, and not a replacement, for their supervisors.

Acknowledgments

We express our sincere gratitude to the participants for their time and insights. This work is supported by Microsoft and Khushi Baby. Aditya Vashistha’s contributions to this project are supported by the President’s Council of Cornell Women.

References

- [1] Alaa Abd-Alrazaq, Zeineb Safi, Mohannad Alajlani, Jim Warren, Mowafa Househ, Kerstin Denecke, et al. 2020. Technical metrics used to evaluate health care chatbots: scoping review. *Journal of medical Internet research* 22, 6 (2020), e18301.
- [2] Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. 2024. AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances. <https://doi.org/10.48550/arXiv.2409.11360> [cs].
- [3] Yasmina Al Ghadban, Huiqi (Yvonne) Lu, Uday Adavi, Ankita Sharma, Sridevi Gara, Neelanjana Das, Bhaskar Kumar, Reny John, Praveen Devarsetty, and Jane E. Hirst. 2023. Transforming Healthcare Education: Harnessing Large Language Models for Frontline Health Worker Capacity Building using Retrieval-Augmented Generation. *medRxiv* N/A, N/A (2023), 574–579. <https://doi.org/10.1101/2023.12.15.23300009>
- [4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collison, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Tveean, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3290605.3300233>
- [5] Maria Antoniak, Aankanksa Naik, Carla S. Alvarado, Lucy Lu Wang, and Irene Y. Chen. 2024. NLP for Maternal Healthcare: Perspectives and Guiding Principles in the Age of LLMs. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) (FAccT '24). Association for Computing Machinery, New York, NY, USA, 1446–1463. <https://doi.org/10.1145/3630106.3658982>
- [6] ARMMAN. 2024. Advancing Reduction In Mortality And Morbidity Of Mothers, Children And Neonates (ARMMAN) Mobile Academy. <https://armman.org/mobile-academy/> Accessed: 2024-08-20.
- [7] Siddhartha Asthana, Pushpendra Singh, and Amarjeet Singh. 2013. Design and evaluation of adaptive interfaces for IVR systems. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (Paris, France) (CHI EA '13). Association for Computing Machinery, New York, NY, USA, 1713–1718. <https://doi.org/10.1145/2468356.2468663>
- [8] Carolin Bauerhenne, Rainer Kolisch, and Andreas S. Schulz. 2024. Robust Appointment Scheduling with Waiting Time Guarantees. [arXiv:2402.12561 \[econ.GN\]](https://arxiv.org/abs/2402.12561) <https://arxiv.org/abs/2402.12561>
- [9] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3313831.3376718>
- [10] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604. https://doi.org/10.1162/tacl_a_00041
- [11] Glen Berman, Nitesh Goyal, and Michael Madaio. 2024. A Scoping Study of Evaluation Practices for Responsible AI Tools: Steps Towards Effectiveness Evaluations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 294, 24 pages. <https://doi.org/10.1145/3613904.3642398>
- [12] Shimmila Bhowmick, Rajkumar Darbar, and Keyur Sorathia. 2018. Pragati: design and evaluation of a mobile phone-based head mounted virtual reality interface to train community health workers in rural India. In *Proceedings of the 10th Nordic Conference on Human-Computer Interaction* (Oslo, Norway) (NordiCHI '18). Association for Computing Machinery, New York, NY, USA, 299–310. <https://doi.org/10.1145/3240167.3240201>
- [13] CapacityPlus. 2024. Senegal. <https://www.capacityplus.org/senegal.html> Accessed: 2024-08-20.
- [14] Shraddha Chakradhar. 2017. Predictable response: finding optimal drugs and doses using artificial intelligence. *Nature medicine* 23, 11 (2017), 1244–1248.
- [15] Liang-Kung Chen. 2018. Artificial intelligence in medicine and healthcare. *Journal of Clinical Gerontology and Geriatrics* 9, 3 (2018), 77–78.
- [16] Lilian de Greef, Mayank Goel, Min Joon Seo, Eric C. Larson, James W. Stout, James A. Taylor, and Shweta N. Patel. 2014. Bilicam: using mobile phones to monitor newborn jaundice. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Seattle, Washington) (UbiComp '14). Association for Computing Machinery, New York, NY, USA, 331–342. <https://doi.org/10.1145/2632048.2632076>
- [17] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (EAAMO '23). Association for Computing Machinery, New York, NY, USA, 1–23. <https://doi.org/10.1145/3617694.3623261>
- [18] Nicola Dell, Ian Francis, Haynes Sheppard, Raiva Simbi, and Gaetano Borriello. 2014. Field evaluation of a camera-based mobile health system in low-resource settings. In *Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices & Services* (Toronto, ON, Canada) (MobileHCI '14). Association for Computing Machinery, New York, NY, USA, 33–42. <https://doi.org/10.1145/2628363.2628366>
- [19] Nicola Dell, Vidya Vaideyanathan, Indrani Medhi, Edward Cutrell, and William Thies. 2012. "Yours is better!": participant response bias in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) (CHI '12). Association for Computing Machinery, New York, NY, USA, 1321–1330. <https://doi.org/10.1145/2207676.2208589>
- [20] Brian DeRenzi, Nicola Dell, Jeremy Wacksman, Scott Lee, and Neal Lesh. 2017. Supporting Community Health Workers in India through Voice- and Web-Based Feedback. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) (CHI '17). Association for Computing Machinery, New York, NY, USA, 2770–2781. <https://doi.org/10.1145/3025453.3025514>
- [21] Brian DeRenzi, Leah Findlater, Jonathan Payne, Benjamin Birnbaum, Joachim Mangilima, Tapas Parikh, Gaetano Borriello, and Neal Lesh. 2012. Improving community health worker performance through automated SMS. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development* (Atlanta, Georgia, USA) (ICTD '12). Association for Computing Machinery, New York, NY, USA, 25–34. <https://doi.org/10.1145/2160673.2160677>
- [22] Brian DeRenzi, Jeremy Wacksman, Nicola Dell, Scott Lee, Neal Lesh, Gaetano Borriello, and Andrew Ellner. 2016. Closing the Feedback Loop: A 12-month Evaluation of ASTA, a Self-Tracking Application for ASHAs. In *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development* (Ann Arbor, MI, USA) (ICTD '16). Association for Computing Machinery, New York, NY, USA, Article 22, 10 pages. <https://doi.org/10.1145/2909609.2909652>
- [23] Leah Findlater, Ravin Balakrishnan, and Kentaro Toyama. 2009. Comparing semiliterate and illiterate users' ability to transition from audio+text to text-only interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) (CHI '09). Association for Computing Machinery, New York, NY, USA, 1751–1760. <https://doi.org/10.1145/1518701.1518971>
- [24] Siddhartha Gairola, Murtuza Bohra, Nadeem Shaheer, Navya Jayaprakash, Pallavi Joshi, Anand Balasubramaniam, Kaushil Murali, Nipun Kwatra, and Mohit Jain. 2022. SmartKC: Smartphone-based Corneal Topographer for Keratoconus Detection. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 155 (dec 2022), 27 pages. <https://doi.org/10.1145/3494982>
- [25] Siddhartha Gairola, Francis Tom, Nipun Kwatra, and Mohit Jain. 2021. RespirNet: A Deep Neural Network for Accurately Detecting Abnormal Lung Sounds in Limited Data Setting. , 527–530 pages. <https://doi.org/10.1109/EMBC46164.2021.9630091>
- [26] Agasthya Gangavarapu. 2024. Introducing L2M3, A Multilingual Medical Large Language Model to Advance Health Equity in Low-Resource Regions.
- [27] Alina Geampana. 2016. Pregnancy is more dangerous than the pill: A critical analysis of professional responses to the Yaz/Yasmin controversy. *Social Science & Medicine* 166 (2016), 9–16. <https://doi.org/10.1016/j.socscimed.2016.08.005>
- [28] Sandy J. J. Gould, Duncan P. Brumby, and Anna L. Cox. 2024. ChatTL;DR – You Really Ought to Check What the LLM Said on Your Behalf. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems* (CHI EA '24). Association for Computing Machinery, New York, NY, USA, Article 552, 7 pages. <https://doi.org/10.1145/3613905.3644062>
- [29] National Health Mission Government of India. 2014. Guidelines for Community Processes 2014. https://nhm.gov.in/images/pdf/communityisation/asha/Orders-Guidelines/Guidelines_for_Community_Processes_2014_English.pdf
- [30] National Health Mission Government of India. 2024. ASHA Training Modules. <https://nhm.gov.in/index1.php?lang=1&level=3&sublinkid=184&lid=257> Accessed: 2024-08-24.
- [31] National Health Systems Resource Centre Government of India. 2021. Induction Training Module for CHO. <https://nhsrcindia.org/sites/default/files/2021-12/Induction%20Training%20Module%20for%20CHO%20at%20AB-HWC%28English%29.pdf> Accessed: 2024-09-03.
- [32] Nielsen Norman Group and Tanner Kohler. 2023. The UX of Phone-Tree Systems: 16 Usability Guidelines. <https://www.nngroup.com/articles/phone-tree-guidelines/> Accessed: 2024-08-20.
- [33] Abid Haleem, Mohd Javaid, and Ibrahim Haleem Khan. 2019. Current status and applications of Artificial Intelligence (AI) in medical field: An overview. *Current Medicine Research and Practice* 9, 6 (2019), 231–237. <https://doi.org/10.1016/j.cmrp.2019.11.005>
- [34] BETSY HARTMANN and MOHAN RAO. 2015. India's Population Programme: Obstacles and Opportunities. *Economic and Political Weekly* 50, 44 (2015), 10–13. <http://www.jstor.org/stable/44002793>
- [35] Jade Vu Henry, Niall Winters, Alice Lakati, Martin Oliver, Anne Geniets, Simon M Mbae, and Hannah Wanjiru. 2016. Enhancing the Supervision of

- Community Health Workers With WhatsApp Mobile Messaging: Qualitative Findings From 2 Low-Resource Settings in Kenya. *Global Health: Science and Practice* 4, 2 (2016), 311–325. <https://doi.org/10.9745/GHSP-D-15-00386> arXiv:<https://www.ghspjournal.org/content/4/2/311.full.pdf>
- [36] Matthew K. Hong, Adam Fourney, Derek DeBellis, and Saleema Amershi. 2021. Planning for Natural Language Failures with the AI Playbook. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 386, 11 pages. <https://doi.org/10.1145/3411764.3445735>
- [37] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) (FAccT '21). Association for Computing Machinery, New York, NY, USA, 560–575. <https://doi.org/10.1145/3442188.3445918>
- [38] Azra Ismail and Neha Kumar. 2018. Engaging Solidarity in Data Collection Practices for Community Health. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 76 (nov 2018), 24 pages. <https://doi.org/10.1145/3274345>
- [39] Azra Ismail and Neha Kumar. 2019. Empowerment on the Margins: The Online Experiences of Community Health Workers. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–15. <https://doi.org/10.1145/3290605.3300329>
- [40] Azra Ismail and Neha Kumar. 2021. AI in Global Health: The View from the Front Lines. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 598, 21 pages. <https://doi.org/10.1145/3411764.3445130>
- [41] Azra Ismail, Deepika Yadav, Meghna Gupta, Kirti Dabas, Pushpendra Singh, and Neha Kumar. 2022. Imagining Caring Futures for Frontline Health Work. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 480 (nov 2022), 30 pages. <https://doi.org/10.1145/3555581>
- [42] Mohit Jain, Ramachandra Kota, Pratyush Kumar, and Shwetak N. Patel. 2018. Convey: Exploring the Use of a Context View for Chatbots. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–6. <https://doi.org/10.1145/3173574.3174042>
- [43] Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q. Vera Liao, Khai Truong, and Shwetak Patel. 2018. FarmChat: A Conversational Agent to Answer Farmer Queries. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 170 (dec 2018), 22 pages. <https://doi.org/10.1145/3287048>
- [44] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N. Patel. 2018. Evaluating and Informing the Design of Chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) (DIS '18). Association for Computing Machinery, New York, NY, USA, 895–906. <https://doi.org/10.1145/3196709.3196735>
- [45] JiWoong Jang, Sanika Moharana, Patrick Carrington, and Andrew Begel. 2024. "It's the only thing I can trust": Envisioning Large Language Model Use by Autistic Workers for Communication Assistance. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 77, 18 pages. <https://doi.org/10.1145/3613904.3642894>
- [46] Maham Javaid, Beenish Fatima, and Amna Batool. 2017. Bridging the Knowledge Gaps in Lady Health Visitors through Video Based Learning Tool. In *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development* (Lahore, Pakistan) (ICTD '17). Association for Computing Machinery, New York, NY, USA, Article 38, 4 pages. <https://doi.org/10.1145/3136560.3136603>
- [47] Rajesh Kumar Jha, Arun K. Trivedi, and Sanjeev Yadav. 2024. World's First Non-Steroidal Birth Control Pill. Council of Scientific & Industrial Research, Ministry of Science & Technology, Govt. of India. <https://www.csir.res.in/csr-success-stories/worlds-first-non-steroidal-birth-control-pill> Accessed: 2024-11-19.
- [48] Eunkyoung Jo, Young-Ho Kim, Yuin Jeong, SoHyun Park, and Daniel Epstein. 2024. Incorporating Multi-Stakeholder Perspectives in Evaluating and Auditing of Health Chatbots.
- [49] Poonam S. Kalne, Pooja S. Kalne, and Ashok M. Mehendale. 2022. Acknowledging the Role of Community Health Workers in Providing Essential Healthcare Services in Rural India-A Review. *Cureus* 14, 9 (20 Sep 2022), e29372. <https://doi.org/10.7759/cureus.29372>
- [50] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. "Because AI is 100% right and safe": User Attitudes and Sources of AI Authority in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 158, 18 pages. <https://doi.org/10.1145/3491102.3517533>
- [51] Nimisha Karnatak, Brooke Loughrin, Tiffany Amy Kuo, Odeline Mateu-Silvernail, Indrani Medhi Thies, William Thies, and Mohit Jain. 2023. "Is it Even Giving the Correct Reading or Not?": How Trust and Relationships Mediate Blood Pressure Management in India. *ACM Trans. Comput.-Hum. Interact.* 30, 6, Article 90 (Sept. 2023), 27 pages. <https://doi.org/10.1145/3609327>
- [52] Ramesha Karunasena, Mohammad Sarparajul Ambiya, Arunesh Sinha, Ruchit Nagar, Saachi Dalal, Hamid Abdulla, Divy Thakkar, Dhyanesh Narayanan, and Milind Tambe. 2021. Measuring Data Collection Diligence for Community Healthcare. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (-, NY, USA) (EAAMO '21). Association for Computing Machinery, New York, NY, USA, Article 10, 12 pages. <https://doi.org/10.1145/3465416.3483292>
- [53] Naveena Karusala, Shirley Yan, Nupoor Rajkumar, Victoria G, and Richard Anderson. 2023. Speculating with Care: Worker-centered Perspectives on Scale in a Chat-based Health Information Service. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 361 (oct 2023), 26 pages. <https://doi.org/10.1145/3610210>
- [54] Hee-Cheol Kim. 2012. An Experimental Study to Explore Usability Problems of Interactive Voice Response Systems. In *Intelligent Information and Database Systems*. Jeng-Shyang Pan, Shyi-Ming Chen, and Ngoc Thanh Nguyen (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 169–177.
- [55] Neha Kumar, Trevor Perrier, Michelle Desmond, Kiersten Israel-Ballard, Vikrant Kumar, Sudip Mahapatra, Anil Mishra, Shreya Agarwal, Rikin Gandhi, Pallavi Lal, and Richard Anderson. 2015. Projecting health community-led video education for maternal health. In *Proceedings of the Seventh International Conference on Information and Communication Technologies and Development* (Singapore, Singapore) (ICTD '15). Association for Computing Machinery, New York, NY, USA, Article 17, 10 pages. <https://doi.org/10.1145/2737856.2738023>
- [56] Simon Lewin, Judy Dick, Philip Pond, Merrick Zwarenstein, Godwin N Aja, Brian E van Wyk, Xavier Bosch-Capblanch, and Mary Patrick. 2010. Lay health workers in primary and community health care for maternal and child health and the management of infectious diseases. *Cochrane Database of Systematic Reviews* N/A, N/A (2010), N/A. Issue 3. <https://doi.org/10.1002/14651858.CD004015.pub3>
- [57] Zihan Liu, Han Li, Anfan Chen, Renwen Zhang, and Yi-Chieh Lee. 2024. Understanding Public Perceptions of AI Conversational Agents: A Cross-Cultural Analysis. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (CHI '24). Association for Computing Machinery, New York, NY, USA, 1–17. <https://doi.org/10.1145/3613904.3642840>
- [58] Els Maekelbergh, Kinga Zdunek, Sara Marceglia, Bobbie Farsides, and Michael Rigby. 2023. The ethical challenges of personalized digital health. *Frontiers in Medicine* 10 (2023), 1123863.
- [59] Arka Majhi, Satish Agnihotri, and Aparajita Mondal. 2023. Physical and Augmented Reality based Playful Activities for Refresher Training of ASHA Workers in India. In *Proceedings of the Asian HCI Symposium 2022* (New Orleans, LA, USA) (Asian HCI '22). Association for Computing Machinery, New York, NY, USA, 16–22. <https://doi.org/10.1145/3516492.3558788>
- [60] Arka Majhi, Aparajita Mondal, and Satish B. Agnihotri. 2024. Replay, Revise, and Refresh: Smartphone-Based Refresher Training for Community Healthcare Workers in India. In *HCI International 2024 Posters*, Constantine Stephanidis, Margherita Antonia, Stavroula Ntao, and Gavriel Salvendy (Eds.). Springer Nature Switzerland, Cham, 310–320.
- [61] Grady McPeak, Anja Sautmann, Ohia George, Adham Hallal, Eduardo Arancón Simal, Aaron L Schwartz, Jason Abaluck, Nirmal Ravi, and Robert Pless. 2024. An LLM's Medical Testing Recommendations in a Nigerian Clinic: Potential and Limits of Prompt Engineering for Clinical Decision Support. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*. IEEE, N/A, 586–591.
- [62] Indrani Medhi, Mohit Jain, Anuj Tewari, Mohini Bhavsar, Michael Matheke-Fischer, and Edward Cutrell. 2012. Combating rural child malnutrition through inexpensive mobile phones. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design* (Copenhagen, Denmark) (NordiCHI '12). Association for Computing Machinery, New York, NY, USA, 635–644. <https://doi.org/10.1145/2399016.2399113>
- [63] Indrani Medhi, Aman Sagar, and Kentaro Toyama. 2006. Text-Free User Interfaces for Illiterate and Semi-Literate Users. In *2006 International Conference on Information and Communication Technologies and Development*. IEEE, N/A, 72–82. <https://doi.org/10.1109/ICTD.2006.301841>
- [64] Joy Ming, Srijana Kamath, Elizabeth Kuo, Madeline Sterling, Nicola Dell, and Aditya Vashistha. 2022. Invisible Work in Two Frontline Health Contexts. In *Proceedings of the 5th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS '22)*. Association for Computing Machinery, New York, NY, USA, 139–151. <https://doi.org/10.1145/3530190.3534814>
- [65] Government of India Ministry of Health & Family Welfare. 2022. Indian Public Health Standards (IPHS) Guidelines for Primary Healthcare Centres (PHCs). https://nhm.gov.in/images/pdf/guidelines/iphhs-revised-guidelines-2022/03_PHC_IPHS_Guidelines-2022.pdf Accessed: 2024-08-08.
- [66] Ritwik Mishra, Simranjeet Singh, Jasmeet Kaur, Pushpendra Singh, and Rajiv Shah. 2023. Hindi Chatbot for Supporting Maternal and Child Health Related Queries in Rural India. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, Tristan Naumann, Asma Ben Abacha, Steven Bethard, Kirk Roberts, and Anna Rumshisky (Eds.). Association for Computational Linguistics,

- Toronto, Canada, 69–77. <https://doi.org/10.18653/v1/2023.clinicalnlp-1.9>
- [67] Mamta Mittal, Gopi Battineni, Dharmendra Singh, Thakursingh Nagarwal, and Prabhakar Yadav. 2021. Web-based chatbot for Frequently Asked Queries (FAQ) in Hospitals. *Journal of Taibah University Medical Sciences* 16, 5 (2021), 740–746. <https://doi.org/10.1016/j.jtumed.2021.06.002>
- [68] Maletsabisa Molapo, Melissa Densmore, and Brian DeRenzi. 2017. Video Consumption Patterns for First Time Smartphone Users: Community Health Workers in Lesotho. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) (CHI '17). Association for Computing Machinery, New York, NY, USA, 6159–6170. <https://doi.org/10.1145/3025453.3025616>
- [69] Maletsabisa Molapo, Melissa Densmore, and Limpopo Morie. 2016. Designing with Community Health Workers: Enabling Productive Participation Through Exploration. In *Proceedings of the First African Conference on Human Computer Interaction* (Nairobi, Kenya) (AfriCHI '16). Association for Computing Machinery, New York, NY, USA, 58–68. <https://doi.org/10.1145/2998581.2998589>
- [70] Ministry of Health National Health Mission and Government of India Family Welfare. 2013. Update on ASHA Program: July 2013. https://nhm.gov.in/images/pdf/communitisation/asha/Reports/Update_on_ASHA_Program_July_2013.pdf Accessed: 2024-08-05.
- [71] National Health Systems Resource Centre (NHSRC). 2021. Annual ASHA Update 2020-21. <https://nhsrcindia.org/sites/default/files/2022-07/Annual%20ASHA%20Update%202020-21.pdf>
- [72] Siddharth Nishthala, Harshavardhan Kamarthi, Divy Thakkar, Dhyanesh Narayanan, Anirudh Grama, Ramesh Padmanabhan, Neha Madhiwalla, Suresh Chaudhary, Balaraman Ravindra, and Milind Tambe. 2020. Missed calls, Automated Calls and Health Support: Using AI to improve maternal health outcomes by increasing program engagement.
- [73] Chika N Okafor, Ayesan Rewane, and Ifeanyi I Momodu. 2024. *Bacillus Calmette Guerin*. StatPearls Publishing, Treasure Island (FL). <https://www.ncbi.nlm.nih.gov/books/NBK538185/> [Updated 2023 Jul 3].
- [74] Fabian Okeke, Lucas Nene, Anne Muthee, Stephen Odindo, Dianna Kane, Isaac Holman, and Nicola Dell. 2019. Opportunities and challenges in connecting care recipients to the community health feedback loop. In *Proceedings of the Tenth International Conference on Information and Communication Technologies and Development* (Ahmedabad, India) (ICTD '19). Association for Computing Machinery, New York, NY, USA, Article 13, 11 pages. <https://doi.org/10.1145/3287098.3287111>
- [75] Chinasa T. Okolo, Dhruv Agarwal, Nicola Dell, and Aditya Vashistha. 2024. "If it is easy to understand then it will have value": Examining Perceptions of Explainable AI with Community Health Workers in Rural India. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 71 (apr 2024), 28 pages. <https://doi.org/10.1145/3637348>
- [76] Chinasa T. Okolo, Srujana Kamath, Nicola Dell, and Aditya Vashistha. 2021. "It cannot do all of my work": Community Health Worker Perceptions of AI-Enabled Mobile Health Applications in Rural India. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 701, 20 pages. <https://doi.org/10.1145/3411764.3445420>
- [77] OpenAI. 2024. Hello GPT-4O! <https://openai.com/index/hello-gpt-4o/> Accessed: 2024-09-04.
- [78] Joyojeet Pal, Anjuli Dasika, Ahmad Hasan, Jackie Wolf, Nick Reid, Vaishnav Kameswaran, Purva Yardi, Allyson Mackay, Abram Wagner, Bhramar Mukherjee, Sucheta Joshi, Sujay Santra, and Priyamvada Pandey. 2017. Changing data practices for community health workers: Introducing digital data collection in West Bengal, India. In *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development* (Lahore, Pakistan) (ICTD '17). Association for Computing Machinery, New York, NY, USA, Article 17, 12 pages. <https://doi.org/10.1145/3136560.3136582>
- [79] Rajmohan Panda, Supriya Lahoti, Nivedita Mishra, Rajath R Prabhu, Kalpana Singh, Apoorva Karan Rai, and Kumud Rai. 2024. A mixed methods evaluation of the impact of ECHO® telementoring model for capacity building of community health workers in India. *Human Resources for Health* 22, 1 (2024), 26.
- [80] Ankur Pandey, Inshita Murtreja, Saru Brar, and Pushpendra Singh. 2020. Exploring Automated Q&A Support System for Maternal and Child Health in Rural India. In *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies* (Ecuador) (COMPASS '20). Association for Computing Machinery, New York, NY, USA, 349–350. <https://doi.org/10.1145/3378393.3402281>
- [81] Chunjong Park, Alex Mariakakis, Jane Yang, Diego Lassala, Yasamba Djiguba, Youssouf Keita, Hawa Diarra, Beatrice Wasunna, Fatou Fall, Marème Soda Gaye, Bara Ndiaye, Ari Johnson, Isaac Holman, and Shwetak Patel. 2020. Supporting Smartphone-Based Image Capture of Rapid Diagnostic Tests in Low-Resource Settings. In *Proceedings of the 2020 International Conference on Information and Communication Technologies and Development* (Guayaquil, Ecuador) (ICTD '20). Association for Computing Machinery, New York, NY, USA, Article 2, 11 pages. <https://doi.org/10.1145/3392561.3394630>
- [82] África Periáñez, Ana Fernandez Del Río, Ivan Nazarov, Enric Jané, Moi Hassan, Aditya Rastogi, and Dezian Tang. 2024. The Digital Transformation in Health: How AI Can Improve the Performance of Health Systems. *Health Systems & Reform* 10, 2 (2024), 2387138. <https://doi.org/10.1080/23288604.2024.2387138> arXiv:<https://doi.org/10.1080/23288604.2024.2387138> PMID: 39437247.
- [83] Trevor Perrier, Nicola Dell, Brian DeRenzi, Richard Anderson, John Kinuthia, Jennifer Unger, and Grace John-Stewart. 2015. Engaging Pregnant Women in Kenya with a Hybrid Computer-Human SMS Communication System. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) (CHI '15). Association for Computing Machinery, New York, NY, USA, 1429–1438. <https://doi.org/10.1145/2702123.2702124>
- [84] Rifat Rahman, Md. Rishadur Rahman, Nafis Irtiza Tripto, Mohammed Eunus Ali, Sajid Hasan Apón, and Rifat Shahriyar. 2021. AdolescentBot: Understanding Opportunities for Chatbots in Combating Adolescent Sexual and Reproductive Health Problems in Bangladesh. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 710, 15 pages. <https://doi.org/10.1145/3411764.3445694>
- [85] Divya Ramachandran, John Canny, Prabhu Dutta Das, and Edward Cutrell. 2010. Mobile-izing health workers in rural India. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) (CHI '10). Association for Computing Machinery, New York, NY, USA, 1889–1898. <https://doi.org/10.1145/175326.1753610>
- [86] Pragnya Ramjee, Bhuvan Sachdeva, Satvik Golechha, Shreyas Kulkarni, Geeta Fulari, Kaushik Murali, and Mohit Jain. 2024. CataractBot: An LLM-Powered Expert-in-the-Loop Chatbot for Cataract Patients. arXiv:2402.04620 [cs.HC] <https://arxiv.org/abs/2402.04620>
- [87] Sabyasachi Ray, Kinkar Sing, Titol Biswas, Siva Tejaswi Manepalli, and Akanksha Chaturvedi. 2024. Experiences of "Antara": The Injectable Contraceptive in Rural Indian Women Presenting to a Tertiary Care Hospital of Eastern India. *The Journal of Obstetrics and Gynecology of India* 74, 3 (01 Jun 2024), 243–249. <https://doi.org/10.1007/s13224-023-01863-8>
- [88] Sarah M Rodrigues, Anil Kanduri, Adeline Nyamathi, Nikil Dutt, Pramod Khargonekar, and Amir M Rahmani. 2022. Digital Health-Enabled Community-Centered Care: Scalable Model to Empower Future Community Health Workers Using Human-in-the-Loop Artificial Intelligence. *JMIR formative research* 6, 4 (2022), e29553.
- [89] Sarah M Rodrigues, Anil Kanduri, Adeline Nyamathi, Nikil Dutt, Pramod Khargonekar, and Amir M Rahmani. 2022. Digital Health-Enabled Community-Centered Care: Scalable Model to Empower Future Community Health Workers Using Human-in-the-Loop Artificial Intelligence. *JMIR Formative Research* 6, 4 (April 2022), e29553. <https://doi.org/10.2196/29553>
- [90] Bhuvan Sachdeva, Pragnya Ramjee, Geeta Fulari, Kaushik Murali, and Mohit Jain. 2024. Learnings from a Large-Scale Deployment of an LLM-Powered Expert-in-the-Loop Healthcare Chatbot. arXiv:2409.10354 [cs.HC] <https://arxiv.org/abs/2409.10354>
- [91] Mike Schaeckermann, Graeme Beaton, Elaheh Sanoubari, Andrew Lim, Kate Larson, and Edith Law. 2020. Ambiguity-aware AI Assistants for Medical Data Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3313831.3376506>
- [92] Kerry Scott, Dena Javadi, and Jessica Gergen. 2024. India's Auxiliary Nurse-Midwife, Anganwadi Worker, Accredited Social Health Activist, Multipurpose Worker, and Lady Health Visitor Programs. <https://chwcentral.org/indiias-auxiliary-nurse-midwife-anganwadi-worker-accredited-social-health-activist-multipurpose-worker-and-lady-health-visitor-programs/> Accessed: 2024-08-12.
- [93] Silvana Secinaro, Davide Calandra, Aurelio Secinaro, Vivek Muthurangam, and Paolo Biancone. 2021. The role of artificial intelligence in healthcare: a structured literature review. *BMC medical informatics and decision making* 21 (2021), 1–23.
- [94] Ian René Solano-Kamaiko, Dibyendu Mishra, Nicola Dell, and Aditya Vashistha. 2024. Explorable Explainable AI: Improving AI Understanding for Community Health Workers in India. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (CHI '24). Association for Computing Machinery, New York, NY, USA, 1–21. <https://doi.org/10.1145/3613904.3642733>
- [95] Sharifa Sultana, François Guimbretière, Phoebe Sengers, and Nicola Dell. 2018. Design Within a Patriarchal Society: Opportunities and Challenges in Designing for Rural Women in Bangladesh. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3173574.3174110>
- [96] Amelia Taylor, Macphail Magwira, Chimwemwe Chamangwana, Evelyn Chaphuma, Thokozani Liwewe, and Chisomo Kankhwali. 2024. Self-Directed Learning for Community Health Workers in Malawi Through Generative AI. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, Vol. N/A. IEEE, N/A, 574–579. <https://doi.org/10.1109/ICHI61247.2024.00092>
- [97] Kentaro Toyama. 2017. Geek Heresy: Rescuing Social Change from the Cult of Technology.

- [98] UNAIDS, WHO, and PEPFAR. 2008. Task shifting: Global recommendations and guidelines.
- [99] United Nations Population Fund (UNFPA) India. 2024. Just Ask Chatbot. <https://india.unfpa.org/en/just-ask-chatbot> Accessed: 2024-08-24.
- [100] Rama Adithya Varanasi, Aditya Vashistha, Rene F. Kizilcec, and Nicola Dell. 2021. Investigating Technostress Among Teachers in Low-Income Indian Schools. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (Oct. 2021), 340:1–340:29. <https://doi.org/10.1145/3476081>
- [101] Leila Caleb Varkey, Anurag Mishra, Anjana Das, Emma Ottolenghi, Dale Huntington, Susan Adamchak, ME Khan, and Frederick Homan. 2004. Involving men in maternity care in India. , 62 pages.
- [102] Aditya Vashistha, Neha Kumar, Anil Mishra, and Richard Anderson. 2017. Examining Localization Approaches for Community Health. In *Proceedings of the 2017 Conference on Designing Interactive Systems (DIS '17)*. ACM, New York, NY, USA, 357–368. <https://doi.org/10.1145/3064663.3064754>
- [103] Aditya Vashistha and Agha Ali Raza. 2023. Voice Interfaces for Underserved Communities. In *Introduction to Development Engineering: A Framework with Applications from the Field*, Temima Madon, Ashok J. Gadgil, Richard Anderson, Lorenzo Casaburi, Kenneth Lee, and Arman Rezaee (Eds.). Springer International Publishing, Cham, 589–611. https://doi.org/10.1007/978-3-030-86065-3_22
- [104] Edward Jay Wang, Junyi Zhu, Mohit Jain, Tien-Jui Lee, Elliot Saba, Lama Nachman, and Shwetak N. Patel. 2018. Seismo: Blood Pressure Monitoring using Built-in Smartphone Accelerometer and Camera. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–9. <https://doi.org/10.1145/3173574.3173999>
- [105] Caroline Whidden, Kassoum Kayentao, Jenny X Liu, Scott Lee, Youssouf Keita, Djoumé Diakité, Alexander Keita, Samba Diarra, Jacqueline Edwards, Amanda Yembrik, Isaac Holman, Salif Samaké, Bouréima Plea, Mama Coumaré, and Ari D Johnson. 2018. Improving Community Health Worker performance by using a personalised feedback dashboard for supervision: a randomised controlled trial. *J Glob Health* 8, 2 (Dec. 2018), 020418.
- [106] Ziang Xiao, Q. Vera Liao, Michelle Zhou, Tyrone Grandison, and Yunyao Li. 2023. Powering an AI Chatbot with Expert Sourcing to Support Credible Health Information Access. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) (IUI '23). Association for Computing Machinery, New York, NY, USA, 2–18. <https://doi.org/10.1145/3581641.3584031>
- [107] Deepika Yadav, Anushka Bhandari, and Pushpendra Singh. 2019. LEAP: Scaffolding Collaborative Learning of Community Health Workers in India. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 169 (nov 2019), 27 pages. <https://doi.org/10.1145/3359271>
- [108] Deepika Yadav, Preerna Malik, Kirti Dabas, and Pushpendra Singh. 2019. Feedpal: Understanding Opportunities for Chatbots in Breastfeeding Education of Women in India. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 170 (nov 2019), 30 pages. <https://doi.org/10.1145/3359272>
- [109] Deepika Yadav, Preerna Malik, Kirti Dabas, and Pushpendra Singh. 2021. Illustrating the Gaps and Needs in the Training Support of Community Health Workers in India. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 231, 16 pages. <https://doi.org/10.1145/3411764.3445111>
- [110] Deepika Yadav, Pushpendra Singh, Kyle Montague, Vijay Kumar, Deepak Sood, Madeline Balaam, Drishti Sharma, Móna Duggal, Tom Bartindale, Delvin Varghese, and Patrick Olivier. 2017. Sangoshti: Empowering Community Health Workers through Peer Learning in Rural India. In *Proceedings of the 26th International Conference on World Wide Web* (Perth, Australia) (WWW '17). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 499–508. <https://doi.org/10.1145/3038912.3052624>
- [111] Ziqi Yang, Xuhai Xu, Bingsheng Yao, Ethan Rogers, Shaor Zhang, Stephen Intille, Nawar Shara, Guodong Gordon Gao, and Dakuo Wang. 2024. Talk2Care: An LLM-based Voice Assistant for Communication between Healthcare Providers and Older Adults. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 73 (may 2024), 35 pages. <https://doi.org/10.1145/3659625>
- [112] Nur Yildirim, Susanna Zlotnikov, Deniz Sayar, Jeremy M. Kahn, Leigh A Bukowski, Sher Shah Amin, Kathryn A. Riman, Billie S. Davis, John S. Minturn, Andrew J. King, Dan Ricketts, Lu Tang, Venkatesh Sivaraman, Adam Perer, Sarah M. Preum, James McCann, and John Zimmerman. 2024. Sketching AI Concepts with Capabilities and Examples: AI Innovation in the Intensive Care Unit. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 451, 18 pages. <https://doi.org/10.1145/3613904.3641896>
- [113] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) (FAT* '20). Association for Computing Machinery, New York, NY, USA, 295–305. <https://doi.org/10.1145/3351095.3372852>
- [114] Dejan Zurovac, Raymond K Sudoi, Willis S Akhwale, Moses Ndiritu, Davidson H Hamer, Alexander K Rowe, and Robert W Snow. 2011. The effect of mobile phone text-message reminders on Kenyan health workers' adherence to malaria treatment guidelines: a cluster randomised trial. *The Lancet* 378, 9793 (2011), 795–803.

A Appendix

A.1 ANM Consensus Prompt

A.1.1 System Prompt. ###Task Description:

A question ("q") asked by an Accredited Social Health Activist (ASHA) has been answered by multiple Auxiliary Nurse Midwives (ANMs) ("anm_answers"). Your task is to synthesise facts and clarifications in anm_answers into a simple and comprehensive answer ("consensus_answer"), by first identifying any conflicting information in anm_answers and providing a count ("anm_votes"), and second generating a precise explanation ("consensus_explanation").

###Steps:

1. Read q and anm_answers carefully.
 2. Identify information within anm_answers relevant to q. Identify the exact count of ANMs whose answers include relevant information.
 - 3a. If the count is less than 3: (A) Provide an empty string for anm_votes. (B) Provide only "Consensus not reached." as consensus_answer.
 - 3b. Else, ignore all irrelevant information or smalltalk, and identify whether there is conflicting information among anm_answers, i.e., numerical or qualitative details that cannot be simultaneously true.
 - 3b.i. If there are no conflicts: (A) Provide an empty string for anm_votes. (B) Provide consensus_explanation. (C) Provide consensus_answer.
 - 3b.ii. Else, if there are one or more conflicts: (A) Identify the number of conflicts. (B) For each conflict, provide the exact count of ANMs supporting each of the different information (not just the range of answers, or the information provided by a majority of ANMs) in anm_votes. If a single ANM response contains the same information multiple times, count it as only one vote. (C) Use this Python function 'majority_voting(anm_votes)' to count if the number of votes among ANMs resulted in a voting majority or not.
- ```
def majority_voting(anm_votes): results = {} for conflict, votes in anm_votes.items(): # Convert the vote counts from strings to integers votes = key: int(value) for key, value in votes.items() # Find the maximum number of votes received max_votes = max(votes.values()) # Check how many pieces of information have the maximum vote count max_vote_keys = [key for key, value in votes.items() if value == max_votes] if len(max_vote_keys) == 1: # If one piece of information has the highest count, it's the majority results[conflict] = "Voting majority" else: # If there's a tie, indicate no voting majority results[conflict] = "No voting majority" return results
```
- 3b.ii.1. If there is 'No voting majority' (i.e., if there is a tie in votes) for one or more conflicts: (A) Provide consensus\_explanation. (B) Provide only "Consensus not reached." as consensus\_answer.
  - 3b.ii.2. Else, if there is 'Voting majority' for all conflicts: (A) For each conflict, identify the information provided by a majority of ANMs. (B) Provide consensus\_explanation. (C) For each conflict, integrate only that majority information into consensus\_answer. Do not mention minority information or conflicts in consensus\_answer.

#### ###Instructions:

1. Only use information in anm\_answers to generate consensus\_answer.

Do not use any other source.

2. Strictly follow the JSON output format in the examples. Do not generate any other (opening or closing) explanations or code.
3. Your output (anm\_votes, consensus\_explanation, and consensus\_answer) must only be in English. The input (q and anm\_answers) can be in Hindi, English, or Hinglisch.
4. consensus\_answer must explain information in simple terms without using medical jargon or uncommon words.
5. consensus\_answer must be as short as possible.
6. consensus\_answer must be framed as an answer for ASHA workers, who are not patients themselves.
7. consensus\_explanation must only be 1-2 sentences long.
8. Do not allow the length of the ANM answers to influence your output.
9. Be as objective as possible.
10. Make sure you read and understand these instructions carefully.
11. Keep this document open while reviewing, and refer to it as needed.
12. Think step-by-step.

###Examples:

##Example 1:

#Input:

```
{
 "q": "माला एन की टेबलेट कैसे इस्तेमाल करें ?",
 "anm_answers": ["यहां गर्भ निरोधक गोलियां होती हैं जिसमें 21 गोली सफेद रंग की एवं 7 गोली काले रंग की होती हैं टोटल 28 गोली होती है प्रथम बार शुरू करते समय महावारी के पांचवें दिन से स्टार्ट करनी है और बिना किसी रुकावट के प्रत्येक दिन निर्धारित समय पर ही गोली लेनी है अगर किसी दिन गोली लेना भूल जाए तो जैसे ही याद आता है तुरंत गोली लेना है स्टार्ट की तरफ से गोली स्टार्ट करनी है और एक पता खत्म होते ही दूसरा स्टार्ट करना है गोली को बच्चों से दूर रखना है", "यह एक सुरक्षित गर्भ निरोधक साधन है, पिरयड आने के पांचवें दिन से शुरू करते हैं हर रोज एक गोली खानी ह", "haa", "Daily ek goli", "Mala N ke tab.mahila ke period ke 5ve din par lene h, Ansar shi h kiya", "Mala and tablet MC ke paanchvein din se tablet Ke Piche Teer ka Nishan se chalu karni hai Lal tablet MC ke Samay per leni hai", "Per Day 1 tablet"]
}
```

#Output:

```
{
 "anm_votes": "",
 "consensus_explanation": "This answer synthesises the unanimous guidance provided by ANMs on the correct usage of Mala N tablets, focusing on starting the cycle, daily intake, and handling missed doses. As the information given by ANMs was qualitatively different but can be simultaneously true, there was no conflicting information and counting votes and identifying the majority was not required."
}
```

"consensus\_answer": "Mala N tablets should be started on the fifth day of the menstrual cycle. The pack contains 28 pills, with 21 white and 7 black pills. One pill should be taken daily at the same time without any interruption. If a pill is missed, take it as soon as you remember. After finishing one pack, start the next pack immediately. Keep the pills out of reach of children."

}

##Example 2:

#Input:

```
{
 "q": "HIV walo ko pension milti h kya?",
 "anm_answers": ["No", "Yes", "haa", "5 लाख तक का इलाज सरकारी या प्राइवेट हॉस्पिटलों में निशुल्क किया जायेगा", "Han ji"]
}
```

#Output:

```
{
 "anm_votes": "Pension": "No": "1", "Yes": "3", "consensus_explanation": "The majority of responses indicate that individuals living with HIV do receive pensions. The information about free treatment provides additional support and is included as part of the comprehensive answer.",
 "consensus_answer": "People living with HIV are eligible for pensions and can receive free treatment up to Rs 5 lakh in government or private hospitals."
}
```

##Example 3:

#Input:

```
{
 "q": "How much money will we get for nasbandi and when we will get it?",
 "anm_answers": ["Under the Janani Suraksha Yojana, ASHA workers will receive Rs 600 for transport and other costs when they help women at the hospital. In 2 months you will get.", "1 महीने में सात सौ रुपये सर ", "haan 600 in 2 months", "ASHA will get 600 rs for ur travel and other costs when you stay with women at the hospital.", "600 In 1 month"]
}
```

#Output:

```
{
 "anm_votes": "Money for sterilisation": "Rs 700": "1", "Rs 600": "4", "When they would get it": "1 month": "2", "2 months": "2",
 "consensus_explanation": "There was confusion among ANM responses regarding the amount and timing of payments for sterilisation support. Although a majority of ANMs agreed that ASH workers will receive support of Rs 600 for their travel and other costs, however variation in the timeframe for receiving the payment (1 month vs. 2 months) prevented a unified answer.",
 "consensus_answer": "Consensus not reached."
}
```

#### A.1.2 Query Prompt. XX

## A.2 Medical Answer Evaluation Prompt

###Task Description:

1. Evaluate the quality of a student answer ("Ans") on a scale of 1-3, for each of the below six distinct "Evaluation Metrics", by comparing it with a teacher's reference answer ("Ref") that scores "3" on all metrics.
2. While evaluating, begin with a short and precise one-sentence "Explanation" of your rating for each metric, strictly based on the definitions given below. Using the Explanation, output your final "Rating" for each metric.

###Instructions:

1. Strictly follow the JSON output format in the Examples below.

- Do not generate any other (opening or closing) explanations or code.
2. For each metric that involves comparison, only compare Ans with Ref. Do not use any other source.
  3. Do not allow the length of the answers to influence your evaluation.
  4. Be as objective as possible.
  5. Make sure you read and understand these instructions carefully.
  6. Keep this document open while reviewing, and refer to it as needed.
  7. Think step-by-step.
- ###Evaluation Metrics:
1. Accuracy: Correctness of information in Ans compared to Ref. Note: if Ans is missing information that is present in Ref, or if Ans repeats information, do not consider this while evaluating Accuracy. Also note: this scale uses "NA" instead of "2".
    - 1a. "1": Ans contains some information directly contradicted by Ref.
    - 1b. "NA": Ans includes information not present in Ref, making their accuracy indeterminate.
    - 1c. "3": All information in Ans is directly supported by Ref.
  2. Subset: Degree to which the set of information in Ans is contained within the set of information in Ref. Note: if Ans is missing some information that is present in Ref, or if Ans repeats information, do not consider this while evaluating Subset.
    - 2a. "1": None of the information in Ans is present in Ref. Ans and Ref are disjoint sets of information.
    - 2b. "2": Ans includes some information not present in Ref. Either Ans and Ref are intersecting sets of information, or the information in Ref is a subset of the information in Ans.
    - 2c. "3": All information in Ans is present in Ref. Ans is a subset of Ref.
  3. Completeness: Degree to which the set of information in Ref is contained within the set of information in Ans. Note: if Ref is missing some information that is present in Ans, or if Ans repeats information, do not consider this while evaluating Completeness.
    - 3a. "1": None of the information in Ref is present in Ans. Ref and Ans are disjoint sets of information.
    - 3b. "2": Ref includes some information not present in Ans. Either Ref and Ans are intersecting sets of information, or the information in Ans is a subset of the information in Ref.
    - 3c. "3": All information in Ref is present in Ans. Ref is a subset of Ans.
  4. Conciseness: Degree of repetition in Ans. Note: This metric does not require comparison with Ref.
    - 4a. "1": All information in Ans is mentioned twice or more.
    - 4b. "2": Some information in Ans is mentioned twice or more.
    - 4c. "3": All information in Ans is mentioned only once.
  5. Clarity: Degree to which Ans explains (medical) information in simple terms without using jargon or uncommon words, making it understandable to those without a medical background. Note: This metric does not require comparison with Ref.
    - 5a. "1": Ans relies entirely on medical jargon or uncommon words without providing explanations in layman's terms.
    - 5b. "2": Ans primarily uses simple language but includes some medical jargon or uncommon words without providing explanations in layman's terms.

5c. "3": Ans primarily uses simple language. Medical jargon or uncommon words are either not included at all, or accompanied with explanations in layman's terms.

6. Structure: Degree of logical arrangement of information in Ans compared to Ref. Note: If Ans is missing some information that is present in Ref, or if Ref is missing some information that is present in Ans, do not consider this while evaluating Structure. Only consider the set of information that is present in both Ref and Ans.
  - 6a. "1": Ans does not align with the logical arrangement of any information present in Ref. This is also applicable when Ans and Ref are disjoint sets of information.
  - 6b. "2": Ans aligns with the logical arrangement of some information present in Ref. Some information in Ans are logically arranged in the same way as in Ref, but others are not.
  - 6c. "3": Ans aligns with the logical arrangement of all information in Ref.

###Examples:

##Example 1:

```
#Input:
{
 "Ref": "The eye drops prescribed are to be applied only in the eye that has undergone surgery, in the lower fornix, i.e., the fold between the back of the lower eyelid and the eyeball. However, if you have any doubts, it's always best to consult with your doctor. If you already have been advised eye drops for the other eye, it is best to continue the same as per advice.",
 "Ans": "The eye drops prescribed are typically to be applied in the eye that has undergone surgery, in the lower fornix. However, it's always best to follow the specific instructions given by your doctor. If you're unsure, please consult with your doctor."
}
```

#Output:
{
 "Accuracy": {
 "Explanation": "All information in Ans (the use of eye drops in the operated eye) is directly supported by Ref; although Ans is missing some information that is present in Ref (whether prescribed drops should continue in the other eye), omissions in Ans do not affect the evaluation of Accuracy.",
 "Rating": "3"
 },
 "Subset": {
 "Explanation": "All information in Ans is present in Ref; Ans is a subset of Ref; although Ans is missing some information that is present in Ref (about existing prescriptions), omissions in Ans do not affect the evaluation of Subset.",
 "Rating": "3"
 },
 "Completeness": {
 "Explanation": "Ref includes some intersecting information with Ans (the use of eye drops in the operated eye) but Ref also includes some information not present in Ans (continuation of existing treatment for the other eye and an explanation of the term 'fornix').",
 "Rating": "2"
 },
 "Conciseness": {
 "Explanation": "All information in Ans (the use of eye drops in the operated eye) is mentioned only once, making it concise.",
 "Rating": "3"
 }
}

```

the operated eye and the suggestion to consult with a doctor) is
mentioned only once.",
"Rating": "3"
},
"Clarity": {
"Explanation": "Ans primarily uses simple language, but includes a
medical jargon 'fornix' without providing any explanation.",
"Rating": "2"
},
"Structure": {
"Explanation": "Ans aligns with the logical arrangement of all in-
formation in Ref that is present in both Ref and Ans (instruction
followed by the suggestion to consult with a doctor).",
"Rating": "3"
}
}
##Example 2:
#Input:
{
"Ref": "We are in the process of enabling online scheduling. In the
meanwhile please schedule your appointment by calling the num-
ber 9739270477 or writing in to pc.blr@sankaraeye.com.",
"Ans": "You can schedule your appointment by contacting the hos-
pital at sankaraeye.com or writing in to info@sankaraeye.com. You
could also call during working hours on +9197399270477."
}
#Output:
{
"Accuracy": {
"Explanation": "The phone number in Ans is directly contradicted
by the phone number in Ref, as Ans includes an extra digit '9' which
might be a typing mistake.
"Rating": "1"
},
"Subset": {
"Explanation": "Ans includes some intersecting information with
Ref (calling and email as contact methods); but Ans also includes
some information not present in Ref (website, different email, wrong
phone number, and the advice to call during working hours).",
"Rating": "2"
},
"Completeness": {
"Explanation": "Ref includes some intersecting information with
Ans (calling and email as contact methods); but Ref also includes
some information not present in Ans (correct phone number,email
ID, and the fact that online scheduling is yet to be enabled.)",
"Rating": "2"
},
"Conciseness": {
"Explanation": "All information in Ans (website, email ID, and phone
number) is mentioned only once",
"Rating": "3"
},
"Clarity": {
"Explanation": "Ans primarily uses simple language to explain the
contact methods with no medical jargon or uncommon words",
"Rating": "3"
}
},
"Structure": {
"Explanation": "Ans mentions the email option first, and the phone
call option second, while Ref is ordered conversely.",
"Rating": "1"
}
}
##Example 3:
#Input:
{
"Ref": "For a cataract surgery under topical or local anaesthesia,
you can have a light breakfast at your usual time. It's ideal to en-
sure around a 45-minute gap between your breakfast and being
wheeled into the operation theatre, especially if your surgery is
scheduled early. However, if the surgery is under general anaesthe-
sia, an empty stomach is recommended.
This would be advised by the anaesthesia team.", "Ans": "If the
surgery is planned under general anaesthesia, the patient should
check with the attending staff or anaesthetist on when the last meal
could be had. It is always best to consult with them when under
general anaesthesia. It's generally recommended to have a light
meal before coming for the surgery."
}
#Output:
{
"Accuracy": {
"Explanation": "Ans directly contradicts Ref by saying that it's rec-
ommended to have a light meal before surgery under general anaes-
thesia, while Ref states that an empty stomach is recommended for
the same.",
"Rating": "1"
},
"Subset": {
"Explanation": "None of the information in Ans about food intake
prior to general anaesthesia is present in Ref",
"Rating": "1"
},
"Completeness": {
"Explanation": "None of the information in Ref about food intake
recommendations for either anaesthesia prior to cataract surgery
is present in Ans",
"Rating": "1"
},
"Conciseness": {
"Explanation": "Ans twice mentions the advice to consult with med-
ical staff when under general anaesthesia",
"Rating": "2"
},
"Clarity": {
"Explanation": "Ans primarily uses simple language while explain-
ing food intake recommendations and does not include any medical
jargon or uncommon words",
"Rating": "3"
}
}

```

while Ref makes different recommendations based on the type of anaesthesia.",  
"Rating": "1"