Mbaza RBC: Deploying and evaluation of an LLM powered Chatbot for Community Health Workers in Rwanda

Samuel Rutunda[1], **Pierrette Mahoro Mastel**[3], **Jean De Dieu Nyandwi**[3], **Kleber Kabanda**[1], **Celestin Niyindagiriye**[1], **Boris Ishimwe Mugisha**[1], **Ali Nengo**[1], **Francis Nkurunziza**[1], **Saad Byiringiro**[1], **Mivumbi Michael Patrick**[1], **Elvis Peace Rugero Ndahayo**[1], **Emmanuel Igirimbabazi**[1], **Gilbert Nzabonimpa**[1], **Fabrice Hakuzimana**[1], **Olivier Niyomugisha**[1], **Cyprien Nshimiyimana**[2], **Alain Ndayishimiye**[2]

[1]Digital Umuganda
[2]Center for the Fourth Industrial Revolution
[3] Carnegie Mellon University
samuel@digitalumuganda.com

Abstract

The emergence of Large Language Models (LLMs) offers an opportunity to support health systems, particularly in low and middle income countries such as Rwanda where there exists limited health infrastructure. By providing information and support to front-line workers, especially community health workers (CHWs), LLMs offer to improve the quality of care by providing quick access to medical guidelines, supporting clinical decision-making, and facilitating health education in local languages. This work deploy and evaluates the performance of Large Language Model (LLM)-based chatbots to assist Community Health Workers (CHWs) in Rwanda, focusing on usability, interaction modalities, and local language processing. A total of 3,000 questions generated by Frontline workers using text and voice input methods were analyzed to determine preferences and error rates. Results indicate a strong preference for text-based queries (66%), though voice queries showed high satisfaction (97.5%) with minor transcription errors (2.47%). The most common focus areas for CHW queries were Maternal and Newborn Health, Integrated Community Case Management, and Nutrition. These findings suggest that, while voice interactions hold some potential, improvements in speech-to-text models are needed for optimal functionality in low-resource settings.

Introduction

Rwanda, like many countries in the Global South, relies heavily on community health workers (CHW) to provide essential healthcare services at the community level (Board 2017). These CHWs, often recruited from their local communities and trained in basic health services, address critical needs such as maternal and newborn care, screening for malnutrition, provision of contraception, and prevention of non-communicable diseases (NCDs) (Rwanda 2013). Established in 1995 to alleviate the shortage of healthcare providers, the CHW program has grown from 12,000 workers to nearly 60,000 by 2020, with four CHWs serving each rural village of 50–150 households [1]. Although the program has significantly improved access to reproductive, maternal, neonatal, child, and adolescent health (RMNCAH) services, CHWs performance remain limited due to insufficient medical expertise. With recent advances in Large Language Model (LLM)-based chatbots demonstrating showing strong performance on several medical tasks (Chen et al. 2023; Singhal et al. 2023), this has sparked interest in their application to support and upskill CHWs (Ramjee et al. 2024; Al Ghadban et al. 2023). However, most AI tools are designed for high-resource settings, raising questions about their effectiveness in low-resource contexts.

This study takes first step to evaluate the performance of current LLM-based chatbots in low-resource CHWs settings, focusing on their usability, including preferred modes of interaction (text vs. audio), and their ability to process local language (Kinyarwanda) and contextual nuances.

Background and related work

To support CHWs, researchers have explored various ways to integrate technology into their workflows to their healthcare delivery. These tools collaborate with CHWs to facilitate key tasks such as data collection for patient monitoring and evaluation (Pal et al. 2017), receiving performance feedback (Whidden et al. 2018), and improving their knowledge and skills through training. With the growing adoption of artificial intelligence (AI) in healthcare, these tools are increasingly incorporating AI features. For example, researchers have developed AI-powered chatbots to assist CHWs (Ramjee et al. 2024), while others have examined the feasibility and challenges of the adoption of such tools (Okolo et al. 2024). Many of these studies have focused on rural settings in India as case studies. Although AI-driven solutions such as chatbots are being designed to cater to local, resource-constrained environments (Wu et al. 2024), this work shifts the focus to the realities of CHWs in Sub-Saharan Africa, specifically eastern Africa, with Rwanda serving as the primary case study.

[1] https://documents.worldbank.org/en/publication/documents-reports/documentdetail/099090823041017390/p17558305e05c50c0b4360a03380d580b0

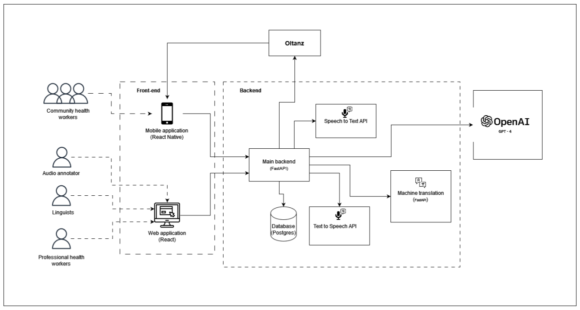

Figure 1: Component diagram of the Mbaza RBC application

System design

The system consists of a back-end and two front-ends, as illustrated in Figure 1. The backend is built using a modular architecture, with the core backend serving as the central component. It links all other modules, manages user authentication, and handles data storage. The Kinyarwanda speech-to-text module (Elamin et al. 2023) transcribes spoken questions into text, while the machine translation module converts questions from Kinyarwanda to English and translates the responses generated by the LLM from English back into Kinyarwanda. The text-to-speech module (Rutunda, Kabanda, and Stan 2023) produces audio responses in Kinyarwanda. For the LLM, GPT-4 was utilized.

The frontend consist of a mobile application and a web application. The mobile application, tailored for community health workers (CHWs), is designed for ease of use. CHWs log in using their phone numbers and a five-digit PIN, similar to those used in mobile money systems. After signing in, CHWs can pose questions either via text or voice. Responses are displayed in text format, with an option to play them as audio. The application is exclusively in Kinyarwanda to align with the language CHWs use in their daily activities.

The web application is used for annotation by the Professional health workers and linguists. The professional health workers annotate by rating conversations generated by the CHWs, the rating is from 0 to 5 and contains helpfulness, correctness and coherence metrics. The linguists annotate the audio quality and the translation. The audio quality annotation is done by annotating both the audio question and audio responses, the audio question only appears if the user asked the question using voice otherwise the annotation is done on the audio responses, the audio annotation is from 0 to 5 on the Mean opinion score metric for both the audio question and audio response.

The linguists also annotates the translation done by the machine translation annotation where they rate the question and its equivalent translation and the response and its equivalent translation, rating the adequacy and helpfulness. however during our preliminary test of GPT-4, gpt-4-0125-preview had improved compared to the previous gpt-4-1106-preview improving the accuracy rating by linguist to 86% from 8% thus we no longer saw the need to evaluate the translation.

Evaluation

For the evaluation we generated 3000 diverse questions based on 14 work packages currently covered by the Rwandan Community health workers, the users could either ask using text or voice. We conducted the analysis guided by the following research questions:

1. Interaction Modalities and Satisfaction (voice vs. text): How do CHWs interact with the AI system across different modalities ?
2. Audio Error Rates: What is the error rate in audio-based queries, and how does it affect user satisfaction and outcomes?
3. Question Categorization: What are the common focus areas of CHW queries, and how do these categories reflect their priorities and challenges?
4. Human Ratings: How does the AI system perform in terms of coherence, translation fluency, and audio quality, once rated by users?

In the sections below, we will answer these questions through comprehensive data exploration and visualizations, providing actionable insights to enhance the AI system's functionality and usability in healthcare settings. By focusing on the core elements of interaction quality, this analysis lays the groundwork for optimizing AI systems in healthcare settings.

Interaction Modalities: Voice vs Text

Community Health Workers (CHWs) exhibit a notable preference for text-based interactions, accounting for approximately 66% of the queries, while voice-based interactions constitute the remaining 34%. This inclination toward text could be attributed to several factors, including the reliability of written input in eliminating potential errors caused by speech-to-text (STT) transcription. Additionally, text allows for more deliberate and precise articulation of queries, which is essential in healthcare contexts where clarity is paramount.

Interestingly, CHWs who opted for voice interactions might have been leveraging its convenience, especially in scenarios where typing was less practical or feasible. CHWs demonstrated an ability to ask clearer questions via voice than text. This finding underscores the potential of voice as a valuable modality, particularly for users with limited typing proficiency or in environments conducive to verbal communication. However, network connectivity and response latency were noted as significant barriers, potentially affecting the general preference for voice-based interactions.

Error Rates in Voice Queries and Their Impact on CHW Results

The integration of voice queries in AI systems for Community Health Workers (CHWs) enables accessibility and ease of interaction, particularly in resource-limited settings. However, transcription errors inherent in voice-to-text processing may affect response quality and user satisfaction. This analysis investigates the error rate from voice queries in the context of CHWs interacting with an LLM, evaluates

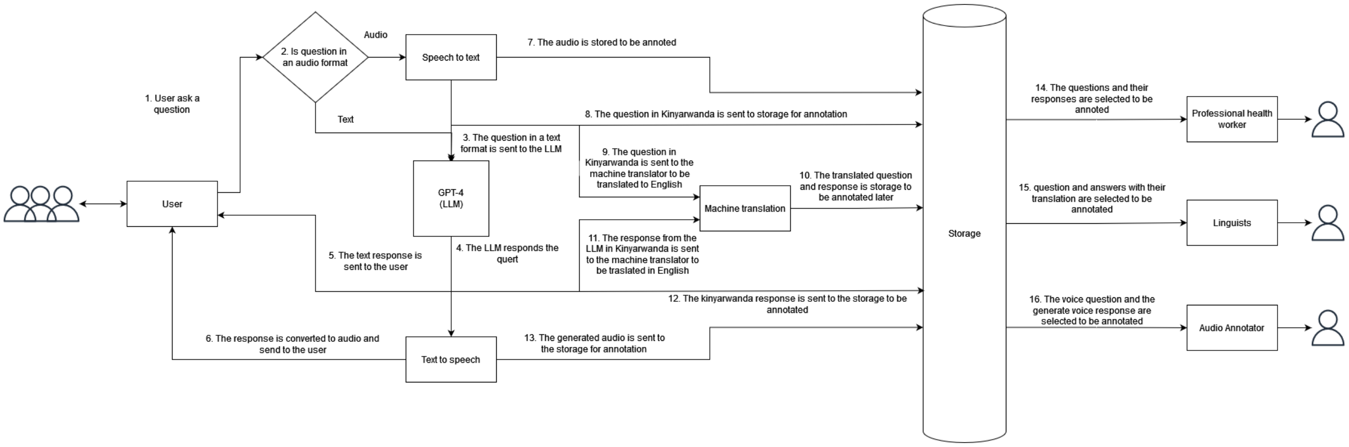
1. User ask a question
User
2. Is question in an audio format
Audio
Text
Speech to text
7. The audio is stored to be annoted
8. The question in Kinyarwanda is sent to storage for annotation
3. The question in a text format is sent to the LLM
GPT-4 (LLM)
4. The LLM responds the quert
5. The text response is sent to the user
9. The question in Kinyarwanda is sent to the machine translator to be translated to English
Machine translation
10. The translated question and response is storage to be annotated later
11. The response from the LLM in Kinyarwanda is sent to the machine translator to be traslated in English
12. The kinyarwanda response is sent to the storage to be annotated
6. The response is converted to audio and send to the user
Text to speech
13. The generated audio is sent to the storage for annotation
Storage
14. The questions and their responses are selected to be annoted
Professional health worker
15. question and answers with their translation are selected to be annotated
Linguists
16. The voice question and the generate voice response are selected to be annotated
Audio Annotator

Figure 2: Process diagram of the Mbaza RBC, a flow diagram of how the entire components are linked together from the user asking the questions with all the process it takes until the user receives the response to the rating by the annotators

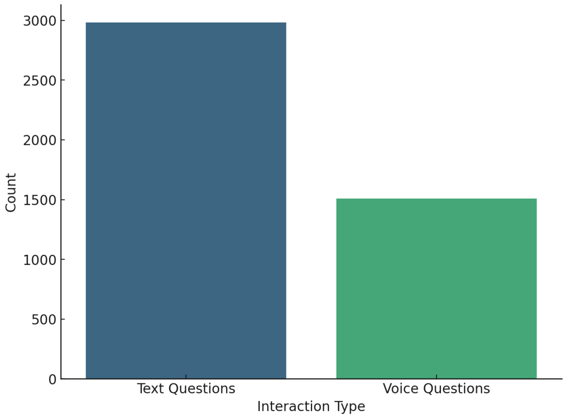

Figure 3: Comparison of Text vs. Voice Queries Asked by CHWs. The majority of queries were made through text (67%), indicating a preference for text-based interactions, while voice queries accounted for approximately 33%.

its effect on results, and compares it with text queries to gain deeper insights.

Voice queries constituted a total of 1,606 interactions, with a normalized satisfaction rate of approximately 97 5%. However, a small error rate of 2.47% was observed, predominantly due to transcription errors in the speech-to-text conversion process. These errors were especially evident in complex or nuanced questions, highlighting a need for further improvements in the handling of local dialects and specialized terminologies. Despite this, voice queries proved to be an effective and user-friendly medium for simpler interactions, demonstrating high overall satisfaction among CHWs.

Text queries remain the most reliable mode for CHWs, whereas voice queries perform well but show room for improvement. Enhancements in speech-to-text models for local dialects and addressing contextual nuances in transcription are essential to minimize errors and boost satisfaction.

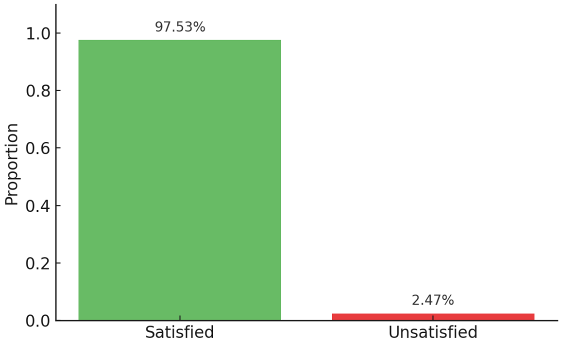

Figure 4: Normalized Satisfaction for Voice Queries. The majority of voice-based interactions (97.53%) resulted in user satisfaction, with a small proportion (2.47%) reporting dissatisfaction.

Analysis of Question By Focus Areas

Community Health Workers (CHWs) play a pivotal role in addressing the health needs of their communities, often relying on AI systems to assist in answering critical questions. By categorizing these questions into predefined focus areas—Maternal New Born Health (MNBH), Integrated Community Case Management (ICCM), Nutrition, Community-Based Provision Family Planning (CBP/FP), Mental health, Non-communicable diseases (NCDs), First Aid, Drug management, Tuberculosis, Malaria, HIV, Behavior Change Communication (BCC), Emergency response to epidemics (ERE), Water, Sanitation and Hygiene (WASH), Early Childhood Development (ECD), Adolescent Sexual and Reproductive Health (ASRH), Gender-Based Violence (GBV) —we can better understand the priorities and challenges faced by CHWs.

Using the all-MiniLM-L6-v2(Reimers and Gurevych 2020) model, questions were semantically matched to the most appropriate category. This method ensures contextual accuracy, going beyond keyword-based heuristics to capture the nuanced meaning of each question. Below, we present the distribution of questions across focus areas to identify trends and guide resource allocation.

The analysis highlights key trends in CHWs' focus areas:

Maternal Newborn and Child Health (MNBH):

- MNBH is the most dominant category, with the highest number of queries (800). This reflects its central importance in community health interventions and highlights the need for robust support systems in this area.

High-Volume Categories:

- Categories such as Integrated Community Case Management (ICCM), Nutrition, and Community-Based Provision of Family Planning (CBP/FP) show significant query volumes, underlining their relevance in day-to-day CHW activities.
- These areas represent both health promotion and logistical challenges, emphasizing the diversity of CHW responsibilities.

Emerging Focus Areas:

- Categories like Mental Health and Non-Communicable Diseases (NCDs) are gaining traction, reflecting broader health challenges in community settings.
- The increasing focus on First Aid and Drug Management also indicates growing awareness of emergency response and supply chain management.

Underrepresented Areas:

- Some categories, such as Gender-Based Violence (GBV) and Behavior Change Communication (BCC), appear less frequently. This may suggest underreporting, limited awareness, or insufficient integration of these topics into CHWs' workflows.

The semantic categorization, powered by the all-MiniLM-L6-v2 model, provided a nuanced understanding of CHWs' priorities. These insights can inform targeted resource allocation, ensuring that support aligns with the most pressing community health needs.

Comprehensive Analysis of Human Ratings

Helpfulness, Correctness, and Coherence Analyzing the ratings of AI responses across key dimensions—helpfulness, correctness, and coherence—is essential to evaluate the system's effectiveness and identify areas for improvement. These metrics reflect the quality of interactions and the ability of the AI to deliver accurate, relevant, and understandable responses. Each response is rated by users on a scale from 0 to 5, with higher scores indicating better performance. By examining the distribution of these ratings, we gain insights into the strengths and weaknesses of the AI system, guiding future enhancements.

The comparison of scores across response metrics—helpfulness, correctness, and coherence—reveals strong performance by the model. Most responses are rated at the maximum score of 5, indicating a consistent perception of quality. Correctness shows slightly more variability, with a small number of scores below 5, suggesting occasional inaccuracies or room for improvement. Overall, the system demonstrates high levels of helpfulness, correctness, and coherence, reflecting well-received responses across all metrics.

Audio Quality The Audio Mean Opinion Score (MOS) evaluates the quality of both input (question audio) and output (response audio) in AI-assisted communication. This metric reflects the clarity, naturalness, and overall user satisfaction with audio interactions, rated on a scale from 0 to 5. Understanding these scores is crucial in identifying areas where audio processing, such as speech-to-text and text-to-speech, can be improved to enhance user experience.

The analysis shows that:

- Question Audio MOS tends to score higher, with many ratings concentrated around 4 and 5, indicating generally good audio clarity for input questions.
- Response Audio MOS exhibits slightly more variability, with a notable presence of mid-range scores (2–3). This suggests opportunities to enhance the naturalness and clarity of synthesized responses.
- The differences highlight a potential need for better optimization in the text-to-speech pipeline for responses.

This analysis underscores the importance of refining audio components to ensure consistent quality across all interactions.

Evaluation: Key Insights and Takeaways

This evaluation highlights critical insights into how Community Health Workers (CHWs) interact with an AI-powered system and how the system performs across key dimensions. The findings shed light on strengths, weaknesses, and opportunities for optimization.

1. Interaction Modalities and Satisfaction:
   - CHWs demonstrated a clear preference for text-based interactions, constituting the majority of queries. Voice-based interactions accounted for 34% of the queries.
   - Voice queries achieved a high satisfaction rate of 97.53%, with a low error rate of 2.47%, primarily due to speech-to-text transcription errors in complex queries.
2. Audio Error Rates:
   - Audio-based queries showed strong input quality, reflected in higher scores for question audio clarity (MOS 3.82).
   - The response audio clarity (MOS 3.41) was slightly lower, suggesting the need for improvements in the text-to-speech system to enhance synthesized audio naturalness and clarity.
3. Question Categorization:
   - Maternal and Child Care emerged as the most frequently addressed focus area, representing over 40% of queries. This underscores its centrality in CHWs' daily operations.

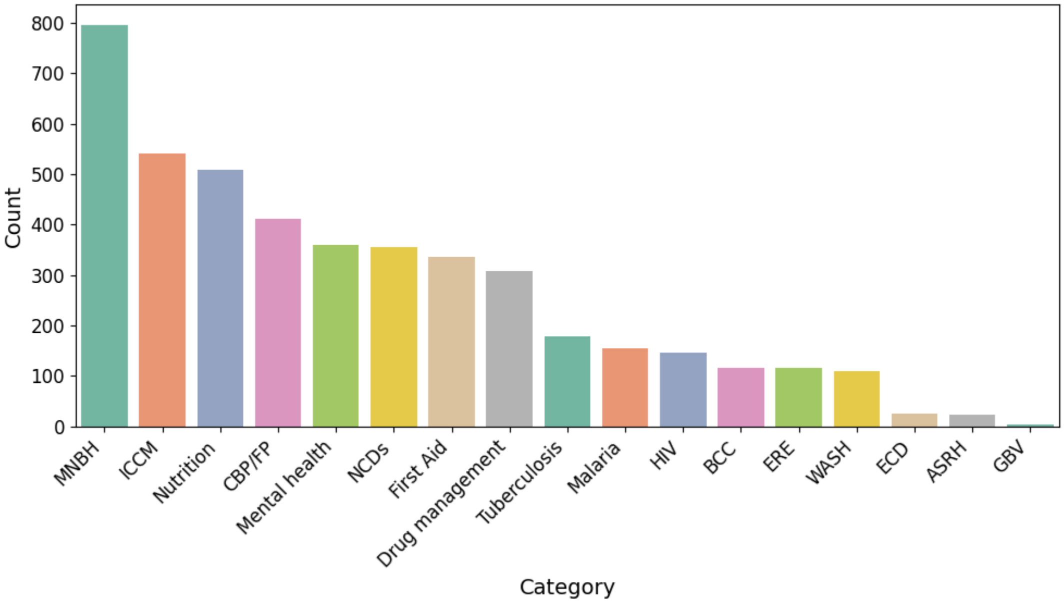

Figure 5: Distribution of Questions Across Focus Areas. The chart highlights the prevalence of different focus areas in CHW queries. Maternal Newborn and Child Health (MNBH) leads with the highest number of queries, followed by ICCM and Nutrition. Categories like Gender-Based Violence (GBV) and Adolescent SRH are less frequently represented, indicating potential gaps in focus or reporting. This distribution reflects CHWs' priorities and the diversity of health challenges they address.

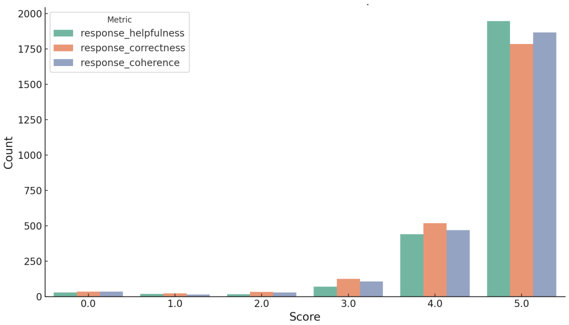

Figure 6: Distribution of Scores Across Response Metrics. The chart compares the ratings for response helpfulness, response correctness, and response coherence. Most ratings are concentrated at the maximum score of 5, indicating strong performance across all three metrics. Slight variability is observed in correctness, reflecting occasional inaccuracies in system responses. This suggests overall high user satisfaction with the AI's response quality.

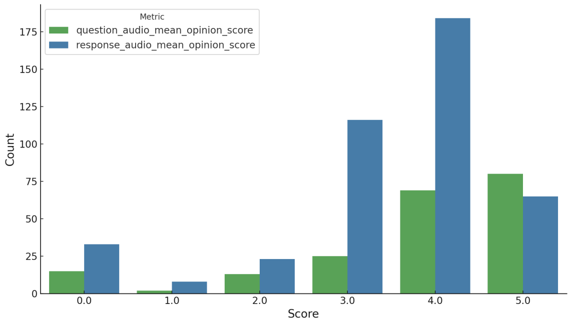

Figure 7: Audio Mean Opinion Scores: Question vs. Response. The chart compares the audio quality ratings for user questions and AI responses. Question audio scores are generally higher, with many ratings concentrated around 4 and 5, indicating good input clarity. Response audio scores show more variability, with a notable number of mid-range ratings (2–3), suggesting room for improvement in the synthesized speech output. This emphasizes the need for enhancements in the text-to-speech pipeline to ensure consistent audio quality.

- Other significant categories included Supply Chain of Medical Commodities and Nutrition, reflecting logistical and health-related challenges in the field.

4. Performance ratings:
   - Across response helpfulness, correctness, and coherence, the AI system performed exceptionally well, with most ratings concentrated at the maximum score of 5.
   - Translation fluency, both for questions and responses, exhibited more variability, with midrange scores (2–3) indicating room for improvement in handling multilingual interactions.

Future work

Continuing this work, the focus will be on two critical areas. First, the development of a localized English and

Kinyarwanda benchmark dataset tailored for CHWs. The dataset will encompass real-world scenarios, dialogues, and terminologies relevant to the Rwanda/African community health context, enabling the fine-tuning and evaluation of LLMs for more accurate and culturally appropriate responses.

Second future work, to explore and improve the assistive capabilities of LLMs to support CHWs in providing preliminary diagnoses. By integrating medical knowledge with context-aware reasoning, LLMs could become valuable tools for CHWs, assisting them in symptom assessment, identifying potential health risks, and guiding patients on the next steps for care, ultimately improving healthcare delivery in under-resourced settings.

Acknowledgment

We would like to thank the Center for the Fourth Industrial Revolution (C4IR) for their immense support in the implementation and management of the project, for liasing us with the Government institutions including the Rwanda Biomedical Center, The Rwandan ministry of Health and the Rwandan Ministry of ICT and Innovation. Extend our thanks to The Gates Foundation for funding this project through the Grand Challenge program[2]. , and the Rwanda Biomedical Center, who were partners in this project for providing us access with the necessary technical, field information and access to Community health workers.

References

Al Ghadban, Y.; Lu, H. Y.; Adavi, U.; Sharma, A.; Gara, S.; Das, N.; Kumar, B.; John, R.; Devarsetty, P.; and Hirst, J. E. 2023. Transforming healthcare education: Harnessing large language models for frontline health worker capacity building using retrieval-augmented generation. *medRxiv*, 2023–12.

Board, R. G. 2017. Rwanda Community Health Workers Programme: 1995–2015. 20 Years of Building Healthier Communities.

Chen, Z.; Cano, A. H.; Romanou, A.; Bonnet, A.; Matoba, K.; Salvi, F.; Pagliardini, M.; Fan, S.; Köpf, A.; Mohtashami, A.; et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Elamin, M.; Chanie, Y.; Ewuzie, P.; and Rutunda, S. 2023. Multilingual Automatic Speech Recognition for Kinyarwanda, Swahili, and Luganda: Advancing ASR in Select East African Languages. In *4th Workshop on African Natural Language Processing*.

Okolo, C. T.; Agarwal, D.; Dell, N.; and Vashistha, A. 2024. " If it is easy to understand then it will have value": Examining Perceptions of Explainable AI with Community Health Workers in Rural India. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–28.

Pal, J.; Dasika, A.; Hasan, A.; Wolf, J.; Reid, N.; Kameswaran, V.; Yardi, P.; Mackay, A.; Wagner, A.; Mukherjee, B.; et al. 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*, 1–12.

Ramjee, P.; Chhokar, M.; Sachdeva, B.; Meena, M.; Abdullah, H.; Vashistha, A.; Nagar, R.; and Jain, M. 2024. ASHABot: An LLM-Powered Chatbot to Support the Informational Needs of Community Health Workers. *arXiv preprint arXiv:2409.10913*.

Reimers, N.; and Gurevych, I. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rutunda, S.; Kabanda, K.; and Stan, A. 2023. Kinyarwanda TTS: Using a multi-speaker dataset to build a Kinyarwanda TTS model. In *4th Workshop on African Natural Language Processing*.

Rwanda, M. 2013. National community health stratetegic plan July 2013–June 2018. *Kigali: MoH Rwanda*.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

Whidden, C.; Kayentao, K.; Liu, J. X.; Lee, S.; Keita, Y.; Diakité, D.; Keita, A.; Diarra, S.; Edwards, J.; Yembrick, A.; et al. 2018. Improving Community Health Worker performance by using a personalised feedback dashboard for supervision: a randomised controlled trial. *Journal of global health*, 8(2).

Wu, C.; Qiu, P.; Liu, J.; Gu, H.; Li, N.; Zhang, Y.; Wang, Y.; and Xie, W. 2024. Towards evaluating and building versatile large language models for medicine. *arXiv preprint arXiv:2408.12547*.

[2] https://gcgh.grandchallenges.org/grant/unleashing-benefits-large-language-models-llms-low-resource-languages