Demo: Building Maternal Health LLMs for Low-Resource Settings

Lyvia Lusiji[1], Stanslaus Mwongela[1], Francesco Piccinno[2], Jay Patel[1], Stephen Obonyo[1,*], Ellen Sebastian[2], Annalisa Pawlosky[2], Mfoniso Ukwak[2,†] Kelvin Ndambuki[1], Sylvia Mbugua[1], Sathy Rajasekharan[1], Dennis Troper[2]
[1]Jacaranda Health, [2]Google
{llusiji, smwongela, jpatel, sobonyo,}@jacarandahealth.org
{kndambuki, smbugua, srajan}@jacarandahealth.org
{ellensebastian, piccinno, troper, apawlosky}@google.com
mjumoh@gmail.com

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities across various domains, including healthcare applications. However, developing and deploying these models typically requires substantial computational resources and large training datasets, creating significant barriers for low-resource languages. This paper presents a tailored pipeline for designing and serving low-resource LLMs in maternal health. We introduce two key contributions: a model design and adaptation method optimized for healthcare applications in low-resource settings, and, a model deployment and serving pipeline, featuring an automated auditor framework for continuous quality assessment of model responses in production. Our approach is validated through UlizaMama, a deployed LLaMA3-based LLM serving over 12,000 daily maternal health queries in Kenya.

1 Introduction

Maternal Healthcare (MH) is a critical global health priority, particularly in Sub-Saharan African (SSA) countries, where limited access to clinically accurate information contributes to high maternal mortality rates [1]. For instance, in Kenya, the Kenya National Bureau of Statistics estimates the maternal mortality rate is 342 deaths per 100,000 live births [2], which is 18 times higher compared to a first-world country such as the United States. Although [3] reports that more than 90% of these deaths are preventable with accurate and timely health information, extrapolating the Kenyan statistics to other African countries, some SSA might not meet the Sustainable Development Goal (SDG) 3.1 of reducing the global maternal mortality rate to less than 70 per 100,000 live births by 2030.

Large Language Models (LLMs) have demonstrated remarkable capabilities in solving complex Natural Language Processing (NLP) tasks, from text summarization to question answering and text generation. Their application is expanding rapidly across various domains, including healthcare, where they show promise in tasks such as clinical decision support [4], medical diagnosis [5, 6], patient care [7], and medical research [8]. In this paper, we present the application of LLMs in answering maternal health questions in a multi-lingual and low-resource setting; where training data is scarce. The framework is inspired by a deployed autoregressive Llama 3-based LLM—UlizaMama—currently serving more than 12,000 maternal queries daily in Kenya in Swahili, English, and code-mixed. Our contributions are as follows:

[*]Corresponding author: sobonyo@jacarandahealth.org
[†]Work done while at Google.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

- **LLM training for healthcare applications in low-resource.** We present a method that can be used to train low-resource language LLMs in maternal health. It includes unique pretraining and fine-tuning strategies that leverage available data effectively.
- **Maternal health LLM deployment and serving.** In critical settings such as healthcare, LLM deployment requires multiple checks to ensure accurate responses. We therefore present a method that enables the triaging of user questions, allowing the LLM to respond to less critical questions while directing those that indicate danger signs, such as severe bleeding, to human agents for timely, expert response. Deployment and serving also includes an automated auditor framework for monitoring the medical accuracy, cultural sensitivity, and clarity of the model's responses in production.

2 Related Work

Several studies have proposed the application of generative LLMs in general healthcare as well as maternal health domains. [9] created a specialized LLM for the medical domain by fine-tuning LLaMA on real-world patient-doctor conversations [10]. Compared to the ChatGPT, the ChatDoctor model was able to provide more fine-grained and accurate responses to medical questions. [11] demonstrated how OpenAI's GPT 4.0 with Retrieval Based Augmentation (RAG) can be used to improve healthcare education in low and middle-resource countries. The authors' model allowed frontline medical workers to provide timely and accurate maternal information to women in rural India. In addition the existing related work, several other studies have proposed LLMs for the healthcare domain including (i) fine-tuning LLaMA and Mistral models with RAG for medical chatbot applications [12]; (ii) evaluating LLMs such as GPT-3.5 and GPT-4 in specialized medical contexts [13]; (iii) employing RAG for question-answering on emerging health issues using social media data [14]; (iv) integrating LLMs and knowledge graphs for research in traditional Chinese medicine [15]; (v) supporting genetic counseling with LLM-based systems [16]; (vi) leveraging knowledge graphs and LLMs for diagnosis prediction [17]; (vii) assessing medical fitness through retrieval-augmented approaches [18]; (viii) optimizing the interpretation of clinical guidelines [19]; (ix) automating document writing in clinical trials [20]; and (x) advancing diabetic care [21]. For in-depth reviews of LLMs in healthcare, we refer the reader to [22].

3 Training, Deployment and Serving

3.1 Framework Overview

We propose a comprehensive approach for designing and serving low-resource LLMs in healthcare, as shown in Figure 1. The framework is composed of unique LLM training, deployment and serving methods that are tailored for low-resource languages and the healthcare domain. We discuss the framework components in detail in the following sections.

3.2 Training Data Design

The UlizaMama data design pipeline systematically prepares high-quality data for training low-resource LLMs in the MNH domain. This multi-stage process begins with data collection where real-world SMS interactions between mothers and helpdesk teams are gathered, ensuring domain relevance. The data is then filtered to remove irrelevant content, assess quality by addressing errors and incomprehensible text, eliminate duplicates, and safeguard privacy by removing personal information.

Following the filtering step, the data labelling phase structures the data into question-answer pairs, tagging each for language (English or Swahili) and classifying intent to support both training and deployment safety mechanisms. To address data scarcity, data augmentation techniques are employed, including cross-lingual translation, conversational flow enhancements (adding greetings, empathy, and follow-up questions), cultural and contextual adaptation of medical advice, and the addition of safety and escalation messaging. Crucially, safety datasets from sources like Beaver Tails [23] are also augmented to ensure the model generates safe and appropriate responses. The pipeline concludes with data validation, encompassing medical accuracy checks by specialists and linguistic validation by native speakers to ensure data integrity and cultural appropriateness before model training.

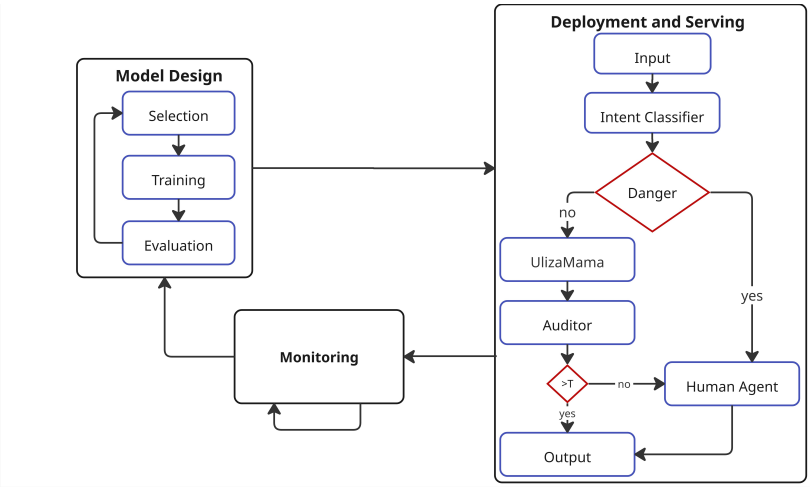

Figure 1: The framework for training, deployment, and serving low-resource LLMs in healthcare.

3.3 Model Selection and Training

The first step is to select a model based on permissive open-source licenses, robust performance with computational efficiency, and strong multilingual support (for low-resource). LLaMA models (LLaMA 3) were chosen for their alignment with these criteria. To adapt the model for our use case, we leveraged Parameter-Efficient Fine-Tuning (PEFT) techniques to balance performance with computational constraints[24]. This involves Continual Pre-Training of the LLaMA base model on a diverse open source (with English-Swahili translation) corpus (see data details in Appendix - Table 4), strategically blended with general, language-specific, and Maternal and Newborn Health (MNH) domain data using batch-level interleaving to enhance both Swahili proficiency and specialized medical knowledge. This pre-training utilizes LoRA[25] for efficiency. After the continual pre-training, the model is fine-tuned on a curated dataset of 400K MNH questions and answers samples, which includes both English and Swahili queries. This dataset is created by human experts at Jacaranda Health, a Kenyan health NGO, and is designed to cover a wide range of maternal health topics. UlizaMama specifically implements a Multi-Adapter LoRA Architecture with six specialized adapters[26] for distinct maternal health subdomains (i.e, Pregnancy, Baby, Postpartum Family Planning, Diet and Nutrition, Medication, and Other MNH), each trained on a specific data subset. This modular approach allows for efficient, specialized responses by dynamically combining the adapters using Task-wise Interference Elimination and Sampling (TIES) [27] to obtain a unified model.

3.4 Model Evaluation

- **Training Evaluation.** This phase of model design involves continuous evaluation during training to ensure stability and prevent overfitting. This is accomplished by monitoring several key metrics, including: (1) cross-entropy loss, (2) perplexity, (3) token accuracy, and (4) gradient norm. These metrics are crucial for assessing the model's learning abilities during training and informing hyperparameter optimization to enhance model performance.
- **Quantitative Post-training Evaluation.** We conducted a quantitative evaluation using a domain-specific benchmark of 2,831 out-of-sample MNH questions curated from our validation set. The benchmark was stratified by language and intent, covering English (1,416), Swahili (1,003), and code-mixed (418) queries, with human-generated helpdesk responses serving as the ground truth. To assess response quality, we employed a suite of standard NLP metrics, including ROUGE[28], METEOR[29]and BERTScore [30]. We benchmarked UlizaMama-V2 (based on Llama-2-7B) and UlizaMama-V3 (based on Llama-3-8B) against several instruction-tuned models of a similar scale: Gemma-1.1-7B-it [31], Mistral-7B-Instruct [32], and their respective base models, Llama-2-7B-hf [33] and Llama-3-8B. For a broader comparison, we also included results from OpenAI's GPT-4 model.

Model	Swahili			Code-Mixed		
	BertScore	**METEOR**	**ROUGE-L**	**BertScore**	**METEOR**	**ROUGE-L**
Gemma-7b-it	0.8227	0.0617	0.0684	0.8128	0.0643	0.0652
Mistral-7B-it	0.7524	0.0532	0.0424	0.7468	0.0649	0.0495
Llama2-7b-hf	0.8047	0.0872	0.0757	0.7971	0.0804	0.0676
Llama-3-8B	0.7787	0.0955	0.0596	0.7723	0.1054	0.0636
GPT-4	0.8560	0.1905	0.1820	0.8395	0.1745	0.1558
UlizaMama-V2	0.8913	0.4235	0.4211	0.8664	0.3461	0.3344
UlizaMama-V3	**0.9025**	**0.4902**	**0.4843**	**0.8798**	**0.4319**	**0.4199**

Table 1: Average Performance Metrics: MNH-Swahili and MNH-Code-Mixed Queries

To further assess performance, we evaluated UlizaMama on established medical and general benchmarks. This dual evaluation aimed to (i) benchmark its specialized capabilities against general-purpose models and (ii) measure the degree of catastrophic forgetting resulting from domain-specific fine-tuning [34]. The results in Appendix - Table2 show that while UlizaMama excels on domain-specific MNH benchmarks, its performance on broader medical and general knowledge tasks (MedMCQA, MedQA) is slightly diminished compared to general-purpose models. This suggests that deep specialization on a narrow domain can erode a model's general reasoning capabilities, a significant challenge when real-world user queries are often unpredictable and may fall outside the core training domain.

- **Qualitative Post-training Evaluation.** The fine-tuning data was translated to Hausa then same framework used to build UlizaMama Hausa. To evaluate the quality of its responses, we sampled Hausa questions from our validation set and generated the corresponding responses with UlizaMama (Hausa version), GPT o3-mini, and GPT-4. The 200 responses were evaluated by four human experts based on seven core aspects: grammar, spelling, punctuation, fluency (coherence), empathy and greeting, bias and cultural sensitivity, and medical accuracy. Each aspect was independently scored on a 1–5 scale, enabling a detailed analysis of the model's performance across both linguistic and clinical dimensions. A score of 1 corresponds to a very poor response about the qualitative aspect, 2: poor, 3: acceptable or fair, 4: good, and 5: very good or excellent. In addition to the scoring criteria, the evaluators were also allowed to provide any general comments. Results in Figure 2 show that UlizaMama (Hausa version) consistently outperforms GPT o3-mini and GPT-4 across all the critical aspects.

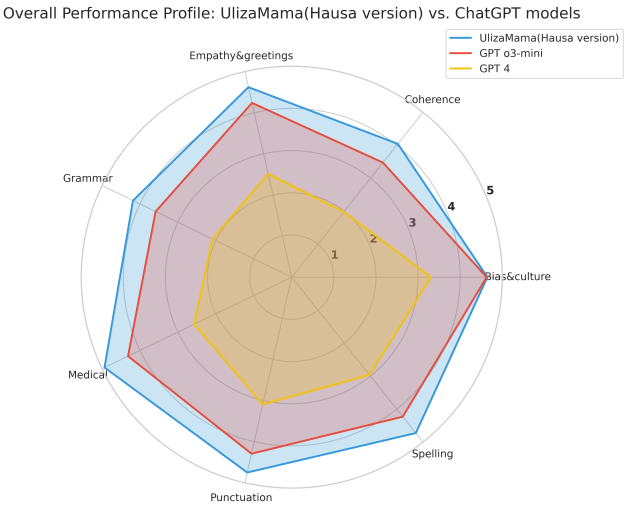

Figure 2: Comparison of qualitative evaluation results for UlizaMama (Hausa version)

3.5 Deployment and Serving

- **Intent Classification Model.** The UlizaMama deployment pipeline begins with a hierarchical intent classifier to triage incoming user queries. This two-level model first identifies a broad category (e.g., diet-nutrition) and then a more granular intent (e.g., fruits). All the questions are then channeled to UlizaMama for a response generation. However, queries flagged as danger-signs, such as severe bleeding, are immediately escalated to human agents for intervention; the generated response serves as a suggestion and is followed by a support call. All other queries and their generated responses are passed to the auditor for quality assurance before delivery. The intent classification currently recognizes 220 distinct intents across categories, including 38 intents under baby care, danger signs (31), diet (33), family planning (18), pain (7), headache (2), pregnancy (36) and root(55). The classifier is based on a fine-tuned XLM-RoBERTa model [35], trained on over 162,000 manually annotated questions. For more details on the specific intent names under each category, see Appendix A.7.
- **Model Serving and Inference.** UlizaMama is served with vLLM [36] to improve memory utilization during inference through the PagedAttention algorithm. This memory management has enabled the deployment of 16-bit precision UlizaMama, allowing for more efficient use of GPU memory. The deployment is configured with 2 L4 NVIDIA GPUs with 24GB of memory each on a Google Cloud's Vertex AI server with monthly costs of approximately $1,460.
- **Auditor.** To automate the human evaluation process, we use an LLM-as-a-judge, awarding scores for grammar, agent behavior, and medical accuracy. Agent behavior is evaluated based on the empathetic tone, cultural sensitivity, and clarity of the response. As discussed by [37], we use rule-augmented prompting to provide a defined scoring rubric to OpenAI's o3-mini, which the LLM uses to evaluate and score each response generated by UlizaMama in these three focus areas. For explainability, the auditor also generates narrative justifications for each score. See a sample of an auditor's prompt in Appendix A.3

4 Discussion

Our key contribution is demonstrating that a domain-specific adaptation strategy can produce an LLM that significantly outperforms larger, general-purpose LLMs, especially in handling the linguistic nuances of low-resource environments. On our MNH benchmark, UlizaMama-V3 substantially surpasses its base model (Llama 3-8B) and even GPT-4, a success directly attributable to our data design and adaptation strategies.Upon human evaluation, the Hausa version of UlizaMama demonstrates superior performance in critical areas for health information dissemination within a specific cultural context, achieving the highest scores across all the evaluated aspects compared to GPT o3-mini and GPT-4. However, evaluation on general medical benchmarks reveals the inherent trade-offs of specialization, illustrating the challenge of catastrophic forgetting, where deep fine-tuning on a narrow domain can erode general knowledge. Despite a slight performance degradation on broad question answering tasks, UlizaMama-V3's strong performance on PubMedQA suggests our continual pretraining successfully retained relevant reasoning capabilities. Beyond quantitative metrics, the framework's impact stands out in its deployment architecture, which prioritizes safety and accessibility. The integration of an intent classifier and an automated auditor serves as a critical responsible AI mechanism, creating a hybrid human-AI system that provides a practical and scalable template for deploying generative AI in healthcare and other sensitive domains. This system balances automation's efficiency with the indispensable oversight of human expertise. Finally, by sending out LLM-generated responses via SMS, the framework directly addresses equitable access, overcoming digital barriers.

Acknowledgments and Disclosure of Funding

We would like to extend our sincere gratitude to Google for funding, coaching, and support throughout this project. We also appreciate AWS for compute credits, Meta for valuable advice, and the Patrick J. McGovern Foundation for financial support, all of which were instrumental in advancing our research.

References

[1] Noluthando Ndlovu, Andrew Gray, Bonga Mkhabela, Nqobile Myende, and Candy Day. Health and related indicators 2022. *South African Health Review*, 2022(1):1–121, 2022.

[2] Duncan N Shikuku, Irene Nyaoke, Onesmus Maina, Martin Eyinda, Sylvia Gichuru, Lucy Nyaga, Fatuma Iman, Edna Tallam, Ibrahim Wako, Issak Bashir, et al. The determinants of staff retention after emergency obstetrics and newborn care training in kenya: a cross-sectional study. *BMC Health Services Research*, 22(1):872, 2022.

[3] WHO et al. Trends in maternal mortality estimates 2000 to 2023: estimates by who, unicef, 2025.

[4] Stefan Lukac, Davut Dayan, Visnja Fink, Elena Leinert, Andreas Hartkopf, Kristina Veselinovic, Wolfgang Janni, Brigitte Rack, Kerstin Pfister, Benedikt Heitmeir, et al. Evaluating chatgpt as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. *Archives of Gynecology and Obstetrics*, 308(6):1831–1844, 2023.

[5] Zekai Chen, Mariann Micsinai Balan, and Kevin Brown. Boosting transformers and language models for clinical prediction in immunotherapy. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 332–340, 2023.

[6] Vivian Weiwen Xue, Pinggui Lei, and William C Cho. The potential impact of chatgpt in clinical and translational medicine. *Clinical and Translational Medicine*, 13(3):e1216, 2023.

[7] Mohd Javaid, Abid Haleema, and Ravi Pratap Singh. Chatgpt for healthcare services: An emerging stage for an innovative perspective. *BenchCouncil Transactions on Benchmarks, Standards & Evaluations*, 3(1), 2023.

[8] Malik Sallam. The utility of chatgpt as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations. *MedRxiv*, pages 2023–02, 2023.

[9] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6), 2023.

[10] Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 9241–9250, 2020.

[11] Yasmina Al Ghadban, Huiqi Lu, Uday Adavi, Ankita Sharma, Sridevi Gara, Neelanjana Das, Bhaskar Kumar, Renu John, Praveen Devarsetty, and Jane E Hirst. Transforming healthcare education: Harnessing large language models for frontline health worker capacity building using retrieval-augmented generation. *medRxiv*, pages 2023–12, 2023.

[12] Arunabh Bora and Heriberto Cuayáhuitl. Systematic analysis of retrieval-augmented generation-based llms for medical chatbot applications. *Machine Learning and Knowledge Extraction*, 6 (4):2355–2374, 2024.

[13] Xi Chen, Li Wang, MingKe You, WeiZhi Liu, Yu Fu, Jie Xu, Shaoting Zhang, Gang Chen, Kang Li, and Jian Li. Evaluating and enhancing large language models' performance in domain-specific medicine: Development and usability study with docoa. *Journal of Medical Internet Research*, 26:e58158, 2024.

[14] Sudeshna Das, Yao Ge, Yuting Guo, Swati Rajwal, JaMor Hairston, Jeanne Powell, Drew Walker, Snigdha Peddireddy, Sahithi Lakamana, Selen Bozkurt, et al. Two-layer retrieval-augmented generation framework for low-resource medical question answering using reddit data: Proof-of-concept study. *Journal of Medical Internet Research*, 27:e66220, 2025.

[15] Yuchen Duan, Qingqing Zhou, Yu Li, Chi Qin, Ziyang Wang, Hongxing Kan, and Jili Hu. Research on a traditional chinese medicine case-based question-answering system integrating large language models and knowledge graphs. *Frontiers in Medicine*, 11:1512329, 2025.

[16] Takuya Fukushima, Masae Manabe, Shuntaro Yada, Shoko Wakamiya, Akiko Yoshida, Yusaku Urakawa, Akiko Maeda, Shigeyuki Kan, Masayo Takahashi, and Eiji Aramaki. Evaluating and enhancing japanese large language models for genetic counseling support: Comparative study of domain adaptation and the development of an expert-evaluated dataset. *JMIR Medical Informatics*, 13:e65047, 2025.

[17] Yanjun Gao, Ruizhe Li, Emma Croxford, John Caskey, Brian W Patterson, Matthew Churpek, Timothy Miller, Dmitriy Dligach, and Majid Afshar. Leveraging medical knowledge graphs into large language models for diagnosis prediction: design and application study. *Jmir Ai*, 4: e58670, 2025.

[18] Yu He Ke, Liyuan Jin, Kabilan Elangovan, Hairil Rizal Abdullah, Nan Liu, Alex Tiong Heng Sia, Chai Rick Soh, Joshua Yi Min Tung, Jasmine Chiat Ling Ong, Chang-Fu Kuo, et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *npj Digital Medicine*, 8(1):187, 2025.

[19] Simone Kresevic, Mauro Giuffrè, Milos Ajcevic, Agostino Accardo, Lory S Crocè, and Dennis L Shung. Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. *NPJ digital medicine*, 7(1):102, 2024.

[20] Nigel Markey, Ilyass El-Mansouri, Gaetan Rensonnet, Casper van Langen, and Christoph Meier. From rags to riches: Using large language models to write documents for clinical trials. *arXiv preprint arXiv:2402.16406*, 2024.

[21] Shayan Mashatian, David G Armstrong, Aaron Ritter, Jeffery Robbins, Shereen Aziz, Ilia Alenabi, Michelle Huo, Taneeka Anand, and Kouhyar Tavakolian. Building trustworthy generative artificial intelligence for diabetes care and limb preservation: a medical knowledge extraction case. *Journal of Diabetes Science and Technology*, page 19322968241253568, 2024.

[22] Lameck Mbangula Amugongo, Pietro Mascheroni, Steven Brooks, Stefan Doering, and Jan Seidel. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*, 4(6):e0000877, 2025.

[23] Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36: 24678–24704, 2023.

[24] Vladislav Lialin, Vijeta Deshpande, and Anna Rumshisky. Scaling down to scale up: A guide to parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.15647*, 2023.

[25] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1 (2):3, 2022.

[26] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[27] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115, 2023.

[28] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[29] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[30] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[31] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[32] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL `https://arxiv.org/abs/2310.06825`.

[33] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL `https://arxiv.org/abs/2307.09288`.

[34] Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023.

[35] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[36] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[37] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*, 2024.

[38] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.

[39] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR, 2022.

[40] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

[41] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[42] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL `https://arxiv.org/abs/2005.14165`.

[43] Derek Xu, Tong Xie, Botao Xia, Haoyu Li, Yunsheng Bai, Yizhou Sun, and Wei Wang. Does few-shot learning help llm performance in code synthesis? *arXiv preprint arXiv:2412.02906*, 2024.

A Appendix

A.1 Data and Resources

The underlying dataset for training and evaluating the model is proprietary and cannot be shared publicly because we do not have permission from the mothers (users) who provided the data. A live demo of the trained model is hosted at `https://5vjpjyirfk.us-east-2.awsapprunner.com/predict_ui`. Later, we will publish the training code used for training and evaluating UlizaMama V3.

A.2 Qualitative Evaluation on Medical and General Benchmarks

Another area of evaluation involved assessing UlizaMama's performance on established medical benchmarks. This served two purposes: (i) to compare UlizaMama against general-purpose LLMs in the MNH domain, and (ii) to evaluate the extent of catastrophic forgetting that may have occurred during domain-specific fine-tuning [34]. For this, we selected three widely used medical benchmarks: PubMedQA [38], MedMCQA [39] and MedQA [40]. While MedMCQA and MedQA offered a multiple choice questions that primarily test factual recall, PubmedQA requires yes, no or maybe responses and is grounded in biomedical abstracts, thus evaluating both reasoning and comprehension. For comparison, we also evaluated the LLaMA-2-7B-hf, Mistral-7B-Instruct, and Gemma-1.1-7B-IT. To assess the rate of catastrophic forgetting, we used MMLU(Massive Multitask Language Understanding) [41] general benchmark to compare UlizaMama's performance to that of the base model, LLaMA-2-7B-hf. As a metric, we focused on accuracy given that the benchmarks had clearly defined ground-truth answers. In the setup we also employed 5-shot learning as used by [41], which provides the model with a small number of example input-output pairs before evaluating its performance. This strategy mimics the process of evaluating a human on similar tasks and has been shown to enhance model adaptability and generalization to new tasks [42], thus improving their performance [43].

Model	PubMedQA	MedMCQA	MedQA (4 options)	MMLU General
Gemma 7b-it	73	39	41	38
Mistral-7B-it	74	42	44	38
Llama2-7b-hf	73	38	36	29
UlizaMama-V2	73	35	28	32
UlizaMama-V3	76	37	32	36

Table 2: Performance Metrics: Different Medical and General Benchmarks

A.3 Auditor Prompt Engineering Design

The auditor prompt is designed using a rule-augmented approach to guide the LLM-as-a-judge in evaluating responses from UlizaMama. The prompt provides a detailed scoring rubric, breaking down the evaluation into key aspects such as Agent Behavior and Medical Accuracy. For each aspect, specific criteria and point allocations are defined to ensure consistent and objective assessment. This structured format enables the auditor to not only assign a score but also provide specific, actionable feedback. Table 3 presents a sample of the instructions provided to the auditor LLM.

A.4 Continual PreTraining Details

Continual pre-training was performed to adapt the base Llama model, enhancing its Swahili language capabilities and maternal health domain knowledge. This was achieved by further training the model on a diverse corpus of over 30 million samples from various sources. A key innovation was the use of batch-level interleaving, which alternated between data sources to ensure balanced exposure and prevent overfitting. The process was conducted efficiently using LoRA on a 4×A100 40GB GPU infrastructure. See Table 4 for detailed pre-training parameters.

Audit Aspect	Audit Instructions
Agent Behavior (5 points)	**Greeting:** Check if the mother was greeted. Highlight examples noted in quotation points to indicate where greetings are missing. Terms to use:'Habari mama', 'Hi mum', 'Hey mum', 'Hello mum'. 0.5 points **Empathy:** Evaluate if the response empathized with the mother. This is shown by words like sorry for this, good question, thank you for your question, pole kwa hilo, pole kwa maumivu, swali nzuri, samahani, asante kwa swali lako. 1 mark **Clarity:** Assess if the response uses clear and simple terms. Highlight complex terms with quotation marks and include them in the output. Terms such as 'ama' and 'tafadhali' if used appropriately, do not need a translation. 1 point **Repetitions:** Identify any unnecessary repetitions. Use quotation marks to highlight repeated words or sentences, and clearly outline this in the output section. 0.5 points **Response Length:** Determine if the response is too long (over 640 characters). Use the character limit to highlight this in the output section. 0.5 points **Language Match:** Confirm if the language in the mother's question matches the language used in the response. Use brackets to indicate words in a different language. If there are words that are in a different language and have not been put in brackets, highlight these with quotation marks in the output section. 1 point **Query Completion:** Check if the response asked the mother to confirm that her question was answered. The use of terms such as 'Have we answered your question?' or 'Has your question been answered?' or 'Je tumejibu swali lako?' is recommended. Use "quotation marks" for missing confirmation. 0.5 points
Medical Accuracy (5 points)	**Normal/Abnormal:** Check if the response clarifies the normality or the abnormality of the issue. 1 mark **Explain:** Provide medical information or an explanation for the query. State possible causes, conservative management in brief, and danger signs to be watched out for. Use single quotation marks for medical explanations. 3 marks **Plan:** Give instructions on when to seek medical attention or how to get further information on the topic raised. Use single quotation marks for instructions on seeking medical attention. This also covers any plan given about the next steps for the reported item. A plan can also be an instruction to talk to her healthcare provider for more information on the issue, however brief it may be. It may also just be an instruction to seek care should the mum experience the danger signs stated. 1 mark

Table 3: Sample Auditor Prompt Instructions. All the other aspects are audited similarly.

A.5 UlizaMama Finetuning Details

The Llama-3 base model was fine-tuned using a multi-adapter LoRA strategy. This approach involved training six specialized adapters, each focused on a distinct maternal health domain to create modular expertise. The training was carried out on a $4\times$A100 GPU infrastructure, optimized with DeepSpeed ZeRO-2 and gradient checkpoint. For deployment, the specialized adapters were merged into a single, unified model using the TIES-merging technique. See Table 5 for detailed fine-tuning parameters.

Table 4: Continual Pre-training Details

Parameter	Value
Data Corpus	
General Diverse Datasets	21M samples
Swahili Instructions	4.1M samples
English Instructions	3.2M samples
Jacaranda Domain-Specific	1.3M samples
Swahili Wikipedia	171K samples
Training Hyperparameters	
PEFT Technique	LoRA
Batch Size	2 per GPU
GPUs	4
Gradient Accumulation	16 steps
Effective Batch Size	128
Learning Rate	1×10^{-4}
Warm-up Ratio	0.05 (Linear)
LR Scheduler	Cosine Decay
Max Sequence Length	1024 tokens
Gradient Clipping	1.0

Table 5: Fine-tuning Details for UlizaMama

Parameter	Value
PEFT Configuration	
PEFT Technique	Multi-adapter LoRA
LoRA Rank (r)	64
LoRA Alpha	128
Target Modules	attention & feed-forward
Multi-Adapter Architecture	
Number of Adapters	6
Training Hyperparameters	
GPUs	4
Optimization	DeepSpeed ZeRO-2
Epochs	2 (with early stopping)
Memory Optimization	Gradient Checkpointing
Adapter Merging	
Technique	TIES, TIES-SVD
Merging Density	0.8

A.6 Multi-adapter LoRA Architecture

UlizaMama employs a multi-adapter architecture with six specialized LoRA adapters, each fine-tuned on a distinct subdomain of maternal health to provide comprehensive and nuanced responses. The adapters include Pregnancy, Baby, Family Planning, Diet, Medication, and Others MNH. Data distribution of these adapters is shown in Figure 3.

A.7 Intent Classifier Taxonomy

The UlizaMama deployment pipeline incorporates a hierarchical intent classifier to effectively triage incoming user queries. This two-level model is designed to first identify a broad, primary-level category, termed a "root intent," followed by a more granular, secondary-level intent within that initial classification. This structured approach facilitates precise query routing, ensuring that user needs are accurately understood and addressed.

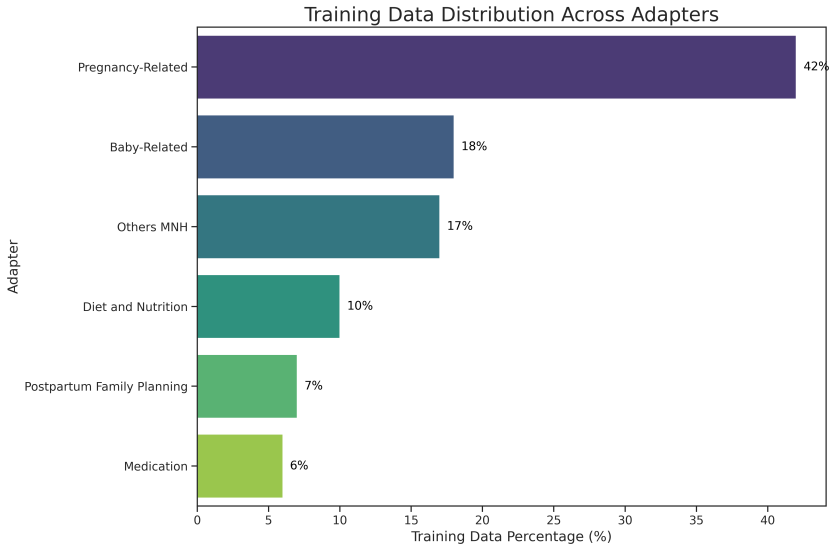

Figure 3: Multi-Adapter Architecture for UlizaMama. Each adapter is trained on a specific subset of the training data to ensure focused expertise in its respective domain.

The comprehensive taxonomy, comprising 220 distinct intents, was developed through extensive manual annotation of questions by human agents. This meticulous process ensures a fine-grained understanding of user inquiries, enabling the delivery of contextually relevant, clinically accurate, and culturally sensitive responses. The system supports queries in Swahili, English, and Sheng (a Swahili-English code-mix). Table 6 provides a comprehensive list of all 220 recognized intents, systematically grouped under their corresponding primary categories.

Table 6: Comprehensive List of All Recognized Intents by Category

Category	Intents
root	ambulance, anc_visit, baby_basking_sun, baby_carrying, baby_cold_body, baby_cold_congestion, baby_deworming, baby_diaper, baby_drooling_foamy_saliva, baby_eye_issues, baby_fainting_or_convulsions_fits, baby_fever_root_intent, baby_fontanelle, baby_growth_development, baby_head_shaking, baby_head_tilt, baby_hiccups, baby_immunity, baby_jaundice, baby_massage, baby_milestones, baby_neck_support, baby_noises, baby_sleeping_position, baby_skin_rash, baby_smell, baby_spitting_up, baby_stool, baby_sweating, baby_swelling, baby_teeth, baby_umbilical_cord_care, baby_vomiting, baby_weight_loss_gain, baby_yellow_eyes, breastfeeding_bleeding, breastfeeding_cracked_nipples, breastfeeding_engorgement, breastfeeding_flat_nipples, breastfeeding_leaking_milk, breastfeeding_low_milk_supply, breastfeeding_mastitis, breastfeeding_pain, breastfeeding_pumping, breastfeeding_sore_nipples, breastfeeding_weaning, bleeding, chest_pain, contraception_general, cough, fever_general, fistula, headaches, hiv_aids, malaria, malnutrition, nausea, ok_thanks, pregnancy_black_line, pregnancy_clothing, pregnancy_weight_loss_gain, sleeping_position, spitting, stretchmarks, survey_response, sweating, swelling, tea, timing, tiredness
baby_root_intent	baby_acne_pimples, baby_bath, baby_bcg_injection_scar, baby_breathing, baby_colic, baby_complementary_feeding_6months, baby_constipation, baby_convulsions_epilepsy, baby_cord, baby_coughing_blood, baby_danger_signs, baby_dehydration_or_not_peeing, baby_diarrhoea, baby_difficulty_fast_breathing, baby_ear_infection, baby_eye_issues, baby_feeding, baby_fontanel_bulging_sunken, baby_fever_danger_sign, baby_head_injury, baby_hearing, baby_hiccups_excessive, baby_immunization, baby_movement, baby_neck_stiff, baby_newborn_jaundice, baby_rash, baby_seizures, baby_skin_condition, baby_sleep, baby_swallowing_problems, baby_teething, baby_thrush, baby_umbilical_cord_not_fallen_or_healed, baby_vomiting_severe, baby_weight_gain, baby_whooping_cough, baby_yellow_skin
pregnancy_root_intent	abortion, acne_pimples, baby_bump, pregnancy_general_v2, pregnancy_stool, pregnancy_working, back_pain, body_odour, cravings, discharge, dizzy, early_pregnancy, edema, fatigue, frequent_urination, heartburn, itching, morning_sickness, nosebleeds, palpitations, pelvic_pain, pregnancy_acne, pregnancy_anemia, pregnancy_cramps, pregnancy_diabetes, pregnancy_exercise, pregnancy_fever, pregnancy_flu_cold, pregnancy_headaches, pregnancy_high_blood_pressure, pregnancy_insomnia, pregnancy_nausea, pregnancy_swelling, pregnancy_weight_gain_loss, pregnancy_yeast_infection, quickening, vaginal_discharge, vomiting
diet_nutrition_root_intent	appetite, avocado, banana, beans, bread, calories, carbs, cereals, chicken, coffee, cooking_oil, diet_during_breastfeeding, diet_general, diet_for_pregnancy, eggs, fats, fish, fluids, food_allergies, fruits, garlic, ginger, grains, healthy_eating, herbs, honey, junk_food, meat, milk, mushrooms, nuts, porridge, salt, snacks, spices, sugar, sugarcane, vegetables, water
danger_sign	baby_cord_not_fallen_or_healed, baby_dehydration_or_not_peeing, baby_diarrhoea, baby_difficulty_fast_breathing, baby_fainting_or_convulsions_fits, baby_fever_danger_sign, baby_movement_decreased, baby_no_crying, baby_not_feeding, baby_not_urinating, baby_rash_severe, baby_seizures_convulsions, baby_severe_bleeding, baby_severe_vomiting, baby_skin_colour_changes, baby_sleepy_unresponsive, baby_swelling_severe, baby_temperature_low, baby_difficulty_breathing, bleeding_vaginal_heavy, blurred_vision, convulsions_fits, fever_high, foul_smelling_discharge, headache_severe, high_blood_pressure_danger, loss_of_consciousness, persistent_vomiting, severe_abdominal_pain, severe_bleeding, severe_headache, swelling_face_hands_feet, blurred_vision_danger, chest_pain_danger, difficulty_breathing_danger, fever_danger, severe_pain
family_planning_root_intents	birth_control_implants, condom, contraception_side_effects, family_planning_general, family_planning_methods, family_planning_myths, iud, male_contraceptives, menstrual_cycle, natural_family_planning, not_fp_related, periods, resuming_sex, safety, sexually_transmitted_infections, sterilization, vasectomy, withdrawal_method
pain_root_intent	pain_back_butt, pain_breast, pain_leg, pain_lower_abdomen, pain_muscle_joint, pain_other, pain_stomach
pain_head_root_intent	pain_head, pain_head_general