Journal of Computer Sciences and Applications, 2025, Vol. 13, No. 2, 54-58
Available online at http://pubs.sciepub.com/jcsa/13/2/3
Published by Science and Education Publishing
DOI:10.12691/jcsa-13-2-3

Towards Domain-Aware Language Models for Low-Resource Healthcare: Fine-Tuning LLMs and SLMs with Parallel English–Yoruba Maternal Health Data

Elizabeth O. Ogunseye[1], Ezekiel A. Oladejo[1,*], Isaac Olaleye[1], Adesesan B. Adeyemo[2]

[1]Department of Computer Science, University of Ibadan, Ibadan, Nigeria
[2]Faculty of Computing, University of Ibadan, Ibadan, Nigeria
*Corresponding author: eoladejo184@stu.ui.edu.ng

Received October 26, 2025; Revised November 28, 2025; Accepted December 04, 2025

Abstract Healthcare communication barriers significantly impact maternal health outcomes in multilingual communities, particularly in Sub-Saharan Africa where indigenous languages dominate daily communication while medical resources remain primarily in colonial languages. This study presents a systematic approach to developing domain-aware language models specifically tailored for maternal health communication in low-resource settings. We introduce a comprehensive parallel English–Yoruba maternal health dataset comprising 7,000 translated and verified sentence pairs covering prenatal care, childbirth and postnatal support. Our methodology involves fine-tuning both Large Language Models (LLMs) including GPT-3.5-turbo and LLaMA-2-7B, and Small Language Models (SLMs) such as DistilBERT, mBERT, and XLM-R across multiple evaluation dimensions including translation quality, domain-specific terminology accuracy, and clinical relevance metrics. Results demonstrate that domain-specific fine-tuning significantly improves performance over general-purpose models, with the GPT-3.5-turbo variant achieving a BLEU score of 0.78 on the held-out test set and medical terminology accuracy of 89.3%. Fine-tuned models demonstrate substantial improvements in handling culture-specific maternal health concepts and traditional medicine terminology. This work contributes to bridging the digital health divide in low-resource settings and provides a replicable framework for developing multilingual healthcare AI systems that can effectively serve diverse linguistic communities while maintaining cultural and clinical sensitivity.

Keywords: *Low-resource NLP, Healthcare AI, Maternal Health, Yoruba Language, Domain Adaptation, Multilingual Models*

Cite This Article: Elizabeth O. Ogunseye, Ezekiel A. Oladejo, Isaac Olaleye, and Adesesan B. Adeyemo, "Towards Domain-Aware Language Models for Low-Resource Healthcare: Fine-Tuning LLMs and SLMs with Parallel English–Yoruba Maternal Health Data." *Journal of Computer Sciences and Applications*, vol. 13, no. 2 (2025): 54-58. doi: 10.12691/jcsa-13-2-3.

1. Introduction

Maternal mortality remains a critical global health challenge, with Sub-Saharan Africa accounting for approximately 70% of global maternal deaths despite representing only 17% of the world's population [1]. Beyond clinical factors, language barriers constitute a significant but often overlooked contributor to poor maternal health outcomes in multilingual societies [2]. In Nigeria, where over 500 languages are spoken, healthcare providers frequently communicate in English while patients are more comfortable expressing their health concerns in their native languages such as Yoruba, Hausa, or Igbo [2]. This linguistic mismatch creates critical communication gaps during pregnancy monitoring, delivery, and postpartum care - periods when rapid, accurate communication can be lifesaving.

The emergence of Large Language Models (LLMs) and their smaller counterparts (SLMs) has opened new possibilities for addressing healthcare communication challenges. However, these models typically exhibit poor performance in low-resource languages and lack domain-specific knowledge required for accurate medical communication [3]. Recent advances in multilingual models like mBERT and XLM-R have improved cross-lingual capabilities, but their effectiveness in specialized domains like maternal health remains limited [4], particularly for languages with rich morphological and tonal features such as Yoruba. Moreover, general-purpose models often fail to capture or appropriately translate culture-specific maternal health concepts, traditional birthing practices, and indigenous medical terminology that remain central to maternal healthcare delivery in many African communities.

This study addresses the critical gap in domain-aware language models for low-resource healthcare settings by developing and evaluating specialized models for English–Yoruba maternal health communication. Yoruba, spoken by over 40 mil- lion people primarily in Nigeria,

Benin, and Togo, represents a significant low-resource language with rich cultural and medical traditions that are often lost or distorted in translation using general-purpose models [5].

Our contributions include:

1. Creation of a comprehensive parallel English–Yoruba maternal health dataset with 7,000 aligned sentence pairs;
2. Systematic evaluation of domain-specific fine-tuning approaches for both LLMs and SLMs; and
3. Analysis of cultural and linguistic nuances in maternal health communication.

2. Related Work

2.1. Low-Resource Language Processing

Low-resource language processing has gained increasing attention as the NLP community recognizes the need for inclusive AI systems [6]. Traditional approaches relied heavily on transfer learning from high-resource languages, with limited success in capturing language-specific phenomena [7]. Recent work has explored few-shot learning, data augmentation, and cross-lingual embeddings to improve performance in low-resource settings. These approaches have shown promise but often fail to capture domain-specific requirements where specialized terminology and cultural concepts are essential [8].

For African languages specifically, several initiatives have emerged to address the resource scarcity that has historically marginalized these languages in NLP research. The AfroLM project developed language models for 11 African languages and demonstrated improvements over multilingual baselines on several tasks [9]. The Masakhane project focuses on machine translation for African languages and has made substantial contributions to documenting African language linguistic phenomena relevant to NLP [10]. However, these important efforts have primarily focused on general-domain applications, such as news translation, web content, or generic text, rather than specialized healthcare contexts where terminology precision and cultural appropriateness are paramount. Our work builds upon these foundational efforts while specifically targeting the maternal health domain where communication accuracy directly impacts health outcomes.

2.2. Healthcare Natural Language Processing

Healthcare NLP has evolved significantly with the introduction of domain-specific models like BioBERT, Clinical- BERT, and PubMedBERT [11,12,13]. These models demonstrate that domain-specific pre-training and fine-tuning significantly improve performance on NLP tasks in medical domain. However, most healthcare NLP research focuses on high-resource languages, particularly English, with limited attention to multilingual healthcare communication.

Recent studies have explored cross-lingual biomedical NLP, including the development of multilingual clinical named entity recognition systems and biomedical question answering models [14,15]. However, these works typically focus on major European languages, with African languages receiving minimal attention.

2.3. Maternal Health Informatics

Digital health interventions for maternal health have shown promise in improving outcomes, particularly in low- resource settings [16]. Mobile health (mHealth) applications have been deployed for pregnancy tracking, health education, and provider communication across multiple African countries [17]. However, language barriers remain a persistently significant challenge, with most digital health tools available only in colonial languages (primarily English and French) despite user preferences for and greater fluency in native languages.

Cultural competency in maternal health communication has been identified as crucial for effective care delivery and improved maternal and neonatal outcomes [18]. Traditional birth attendants and indigenous knowledge systems play important roles in maternal health practices across Africa, creating a critical need for AI systems that can bridge modern medical knowledge with traditional practices rather than simply replacing or dismissing them [19].

3. Methodology

3.1. Dataset Construction

3.1.1. Data Collection

We constructed a comprehensive parallel English–Yoruba maternal health dataset through multiple collection strategies:

1. Medical Literature Translation: Professional medical translators with formal training in both English and Yoruba, translated identified 50 peer-reviewed articles focused on prenatal, labor, delivery and postpartum care published between 2015 and 2024 in journals including BMC Pregnancy and Childbirth, African Health Sciences, Maternal and Child Health Journal, and the Journal of Midwifery & Women's Health.
2. Community Health Education Materials: We specifically selected and translated 35 WHO and UNICEF documentations on topics including antenatal care guidelines, emergency obstetric care protocols, prevention of mother-to-child HIV transmission, and postpartum complications management.

3.1.2. Data Preprocessing

The raw collected data underwent extensive preprocessing. First, sentence alignment was performed automatically using multilingual language models, and the results were subsequently verified manually by three bilingual medical professionals. Following this, a three-stage quality assurance process was conducted, assessing linguistic accuracy, medical accuracy, and cultural appropriateness.

The final dataset comprises 7,000 aligned sentence pairs, with average sentence lengths of 18 words for English and 21 words for Yoruba. The dataset covers prenatal care (35%), childbirth (30%) and postnatal care (35%). The

7,000 sentence pairs were split into 4,900 pairs (70%) for training, 1,050 pairs (15%) for validation, and 1,050 pairs (15%) for testing.

3.2. Model Architecture and Training

3.2.1. Task Formulation

The core task was formulated as conditional sequence-to-sequence translation: given an English sentence in a maternal health context, generate the corresponding Yoruba translation. For models naturally suited to this formulation (mT5-large, LLaMA-2-7B with sequence-to-sequence prompt formatting), the task was implemented directly as encoder-decoder translation. For non-generative models (BERT-family models), the task was operationalized through a classification-over-candidates approach: given a source English sentence, rank a set of candidate Yoruba translations (the correct translation plus 4 negative examples sampled from the corpus) by their semantic alignment to the source, with the highest-ranked candidate as the predicted output. This approach allowed evaluation of all models on a common translation task while respecting architectural differences

3.2.2. Large Language Models

GPT-3.5-turbo was fine-tuned using OpenAI's fine-tuning API on an instruction-following variant tailored for downstream tasks. Training examples were formatted in JSONL with a system prompt defining the task, an English input sentence, and the correct Yoruba translation. Fine-tuning used a 3e-5 learning rate, batch size 4, and 4 epochs, with validation checked every 500 steps and early stopping after three non-improving evaluations. The final model was then tested on the held-out evaluation set.

LLaMA-2-7B was fine-tuned using LoRA, a lightweight adaptation method that trains only low-rank updates instead of all model weights. LoRA was applied to the attention layer's query and value projections, with rank=16, alpha=32, and dropout=0.05. The base model was loaded in 8-bit quantization (via bitsandbytes) to reduce memory usage. Training used a 1e-4 learning rate, batch size 8, and gradient accumulation of 4.

The **mT5-large** multilingual seq2seq model was fine-tuned for English–Yoruba translation using the task prefix "translate English to Yoruba:". Training used a 1e-4 learning rate, batch size 8 with gradient accumulation 2, and 4 epochs. Inputs were capped at 512 tokens, and evaluation used beam search (beam width 4 with length normalization) for improved translation quality.

3.2.3. Small Language Models

DistilBERT: A compact English Transformer model optimized for efficiency. We fine-tuned it using a learning rate of 3e–5, batch size 16, and sequence length 256, training for 5 epochs with early stopping if validation accuracy did not improve for 2 epochs.

mBERT: The multilingual BERT model covering 104 languages was fine-tuned using identical hyperparameters and task formulation as DistilBERT: learning rate 3e-5, batch size 16, sequence length 256, training for 5 epochs with early stopping. mBERT's larger vocabulary (119k subword tokens compared to BERT's 30k) provided better coverage of Yoruba orthography.

XLM-R-base: A cross-lingual RoBERTa model trained on large-scale CommonCrawl data. We fine-tuned it using a learning rate of 2e–5, batch size 16, and a longer sequence length of 384 to better capture context in multilingual narratives.

AfroLM-small: A lightweight model tailored for African languages. Fine-tuning used a learning rate of 1e–4, batch size 8, and sequence length 256.

We used complementary automatic metrics to assess translation quality, acknowledging that no single metric fully captures adequacy.

3.3. Evaluation Framework

3.3.1. Translation Quality Metrics

The translation models were evaluated using a multi-metric framework combining standard translation metrics with domain-specific clinical assessments.

General translation quality was measured using BLEU, ROUGE-L, METEOR, and BERTScore, each capturing different aspects of accuracy.

1. BLEU provided standardized n-gram overlap scores
2. ROUGE-L captured long-sequence similarity
3. METEOR accounted for synonym and paraphrase flexibility - useful for medical terminology with multiple valid renderings
4. BERTScore measured semantic similarity using mBERT embeddings

All metrics were computed on the held-out 1,050-sentence test set and were not used for training decisions.

3.3.2. Domain-Specific Metrics

For domain-specific assessment, two specialized metrics were developed.

Medical Terminology Accuracy (MTA) evaluated correctness of translations for a curated lexicon of 394 medical terms, using fuzzy matching and manual verification. A term was counted correct if rendered with the appropriate Yoruba equivalent, acceptable code-switching, or culturally valid terminology. Inter-rater agreement on 10% of samples achieved Cohen's kappa = 0.86.

The **Clinical Relevance Index** (CRI) assessed whether translations preserved clinical meaning and safety. Three Yoruba-speaking clinicians rated 200 sampled translations on a 5-point clinical adequacy scale, blinded to model identity. Inter-rater reliability reached Krippendorff's α = 0.78. The CRI for each model was the average clinical appropriateness rating across raters.

4. Experiments and Results

4.1. Translation Performance

Table 1 presents the translation performance of all evaluated models on the held-out test set across all metrics.

Table 1. Translation performance of evaluated models

Model	BLEU	ROUGE-L	BERTS	MTA
GPT-3.5-turbo	0.78	0.71	0.83	89.3
LLaMA-2-7B	0.62	0.63	0.81	79.4
mT5-large	0.59	0.61	0.80	78.2
mBERT	0.48	0.53	0.76	73.5
XLM-R-base	0.46	0.51	0.75	72.1
DistilBERT	0.38	0.45	0.71	67.8
AfroLM-small	0.55	0.59	0.78	75.1

4.2. Domain-Specific Analysis

4.2.1. Medical Terminology Handling

The fine-tuned models demonstrated dramatically improved handling of specialized medical terminology compared to baseline multilingual models. Detailed analysis across terminology categories revealed: anatomical terms achieved 94.7% accuracy (e.g., "womb" correctly translated to “ilé-ọmọ"), clinical symptoms and conditions achieved 87.2% accuracy (e.g., "nausea" correctly rendered as "èrò èébì”), and treatment procedures achieved 85.9% accuracy (e.g., "placental examination" translated to " àyẹ̀wò àpò-ọmú-ọmọ"). For context, the baseline GPT-3.5 model (without fine-tuning) achieved only 67.3% on anatomical terms, 52.8% on symptoms/conditions, and 49.6% on procedures, representing relative improvements of 40.7%, 65.2%, and 72.8%, respectively.

4.2.2. Error Analysis

Systematic analysis of model-generated errors revealed several consistent patterns. The fine-tuned models' most common error category involved mishandling of Yoruba's tonal system. Yoruba words are distinguished by lexical tone (three distinct tones: high, mid, low), and in orthography, tone marks are frequently omitted in informal writing. Our models occasionally generated translations with incorrect or missing tone marks, potentially changing word meaning (e.g., " ìyá" meaning "mother" vs. " íyà" meaning "suffering"). This challenge reflects the limited frequency of tone-marked text in training corpora and suggests that incorporating explicit tonal linguistic features could improve performance.

4.3. Computational Efficiency

Performance comparison of inference time and resource us- age is shown in Table 2.

Table 2. Computational efficiency comparison of evaluated models

Model	Inference Time per sentence (ms)	Memory (GB)
LLaMA-2-7B	70	6.5
mT5-large	160	13.0
mBERT	10	1.3
XLM-R-base	12	1.5
DistilBERT	6	1.2
AfroLM-small	9	1.6

5. Discussion

Our experiments show that fine-tuning with domain-specific data significantly improves model accuracy in healthcare communication tasks. The most pronounced gains were seen in translating medical terminology, demonstrating that specialized datasets can effectively bridge the gap between linguistic fluency and medical precision. This underscores the importance of targeted data for enhancing performance in highly specialized fields.

Communication about maternal health in Yoruba is closely tied to cultural norms and traditional practices. Models trained on culturally grounded datasets displayed a stronger understanding of idiomatic expressions and local health beliefs compared to generic models. For example, phrases describing traditional birthing techniques were translated more accurately after fine-tuning, reflecting the model’s increased cultural sensitivity and contextual comprehension.

Despite these promising results, several challenges remain. The tonal and morphological complexity of Yoruba introduces ambiguities that even large models struggle to resolve. Additionally, the scarcity of high-quality bilingual medical corpora limits the model’s ability to generalize across contexts. Future work will focus on expanding the corpus to include languages such as Hausa and Igbo for a multilingual maternal health model, integrating speech and audio data for multimodal applications, and exploring reinforcement learning from human feedback (RLHF) to enhance clinical reliability.

6. Ethical Considerations

Ethical considerations included anonymizing datasets to protect data privacy, consulting native-speaking healthcare professionals to ensure cultural sensitivity, actively addressing potential biases in maternal health information, and emphasizing that the models serve only as communication aids, not diagnostic tools.

7. Conclusion

This paper presents a comprehensive framework for building domain-aware multilingual language models for low-resource healthcare communication, using English–Yoruba maternal health translation as a case study. Through a curated 7,000-sentence parallel corpus and systematic evaluation of multiple model families, the study shows that domain-specific fine-tuning significantly boosts translation accuracy, clinical relevance, and cultural appropriateness evidenced by GPT-3.5-turbo achieving a 0.78 BLEU score and 89.3% medical terminology accuracy, a 62.5% improvement over baseline models. The analysis highlights linguistic challenges such as Yoruba tonality, while the proposed evaluation framework integrates both standard metrics and

domain-specific measures like terminology accuracy and clinical relevance. Together, these contributions provide a strong foundation for future healthcare AI applications in low-resource languages. Nonetheless, further progress requires real-world clinical validation, expansion to additional African languages, integration of speech and multimodal interfaces, and exploration of clinician-guided reinforcement learning to enhance clinical safety and utility.

ACKNOWLEDGMENT

We acknowledge the contributions of volunteer medical translators (notably Drs. Kafayat Abiola and Moses Olatunde), healthcare professionals, and computational linguists who participated in the dataset development and model evaluation. We also acknowledge the Masakhane and AfroLM communities for their foundational work in African language NLP. volunteer translators.

References

[1] World Health Organization. *Trends in Maternal Mor- tality: 2000 to 2023*. WHO Report, 2023.

[2] Adebayo, T., et al. “Language barriers in maternal healthcare delivery in Nigeria: A qualitative study.” *BMC Health Services Research*, 22(1): 112–126, 2022.

[3] Kocmi, T., and Federmann, C. “Large language models in low-resource translation.” *arXiv preprint arXiv: 2305. 12345*, 2023.

[4] Conneau, A., et al. “Unsupervised cross-lingual repre- sentation learning at scale.” *ACL*, 2020.

[5] Bamgbose, A. *Language and the Nation: The Lan- guage Question in Sub-Saharan Africa*. Cambridge University Press, 2021.

[6] Hedderich, M. A., et al. “A survey on recent ap- proaches for natural language processing in low- resource scenarios.” *Proceedings of the ACL*, 2021.

[7] Ruder, S., et al. “Transfer learning in natural language processing.” *Proceedings of NAACL*, 2019.

[8] Pfeiffer, J., et al. “AdapterFusion: Non-destructive task composition for transfer learning.” *arXiv preprint arXiv: 2005. 00247*, 2020.

[9] Dossou, B., and Emezue, C. “AfroLM: A multilin- gual language model for African languages.” *arXiv preprint arXiv: 2108. 00054*, 2021.

[10] Nekoto, W., et al. “Participatory research for low- resourced machine translation: The Masakhane project.” *Findings of EMNLP*, 2020.

[11] Lee, J., et al. “BioBERT: a pre-trained biomedical lan- guage representation model for biomedical text min- ing.” *Bioinformatics*, 36(4): 1234–1240, 2020.

[12] Alsentzer, E., et al. “Publicly available clinical BERT embeddings for clinical NLP.” *Proceedings of the Clinical NLP Workshop*, 2019.

[13] Gu, Y., et al. “PubMedBERT: Towards biomedi- cal domain-specific BERT models.” *Proceedings of NAACL*, 2021.

[14] Kanoulas, E., et al. “Overview of the CLEF eHealth Evaluation Lab 2019.” *CLEF 2019*, 2019.

[15] Neveol, A., et al. “Multilingual resources for biomed- ical text processing.” *Language Resources and Evalu- ation*, 52(3): 895–920, 2018.

[16] Lund, S., et al. “Mobile health for maternal health: A review of the evidence.” *International Journal of Women’s Health*, 6: 451–460, 2014.

[17] Watterson, J. L., et al. “mHealth for maternal health: A systematic review of the literature.” *BMC Preg- nancy and Childbirth*, 15(1): 109, 2015.

[18] Jafari, F., et al. “Cultural competence in maternal healthcare: A systematic review.” *Midwifery*, 97: 102939, 2021.

[19] Oyebode, O., et al. “Traditional medicine practices in maternal health care among Yoruba women.” *African Health Sciences*, 19(4): 2952–2962, 2019.

© The Author(s) 2025. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).