

NigBench: A multilingual point-of-care medical query benchmarking study of large language models in Nigeria

Dr. Tobi Olatunji^{1,2}, Chinemelu Aka¹, Chibuzor Okocha^{2,3}, Dr. Emmanuel Ayodele¹, Jennifer Orisakwe¹, Toni Adekunle¹, Mardhiyah Sanni¹, Abdulameed Abiola¹, Tassallah Abdullahi⁴, Dr. Oluwatomi Owopetu⁵, Dr Tolu Afolaranmi⁶, Peter Suoyo Yougha⁷, Mira Emmanuel-Fabula⁸, Dr. Vaishnavi Menon¹⁰, Prof. Alastair Denniston¹⁰, Dr. Xiao Liu¹⁰, Dr. Gwydion Williams⁹, Prof. Bilal A. Mateen^{9,10}

1 Intron Health

2 BioRAMP

3 University of Florida

4 Brown University

5 University College Hospital

6 University of Jos

7 InStrat Global Health Solutions

8 PATH, Switzerland

9 PATH, United Kingdom

10 University of Birmingham, UK

Corresponding Author:

Tobi Olatunji

Tobi@intron.io

Word Count: 3,369 (not including methods)

Table Count: 0 main (+ 2 Extended Data Tables)

Figure Count: 4 main (+ 3 Extended Data Figures)

Abstract

In this study, we introduce a novel benchmark comprising over 9,000 real-world, point-of-care, multilingual, and multimodal clinical question-answer pairs sourced from frontline health workers in Nigeria. Using the dataset, we compare local general practitioners to multiple leading open and closed LLMs. Our results reveal several critical insights into the suitability of LLMs as clinical decision support systems in low-resource contexts. The results confirm that performance varies widely by language and input modality (e.g., text *vs* speech): while models perform best on English text inputs, their accuracy drops significantly for local-language speech. Critically, it is possible to achieve substantial performance gains by transcribing and translating other languages into English before prompting an LLM – an important insight for non-anglophone product developers. Finally, this benchmark highlights key limitations of SLMs in supporting frontline healthcare in low-resource settings and provides a clear opportunity to track improvements as novel solutions are developed.

Introduction

The global shortage of physicians, particularly in low- and middle-income countries (LMICs), has created significant barriers to achieving universal health coverage and equitable healthcare delivery [1]. Nigeria exemplifies this challenge, with a physician-to-patient ratio far below the World Health Organization's recommended threshold [2,3]. To address this gap, Nigeria relies heavily on 'community health extension workers' (CHEWs) – facility-based staff that also do community outreach, who have undertaken a 3-year pre-service training programme, and whose scope of work spans essential primary care services, including triage, preventive care, health education, and referral for complex cases [4,5]. Despite their formal training, these frontline health workers often encounter clinical situations beyond their scope of practice [6]. In such instances, timely consultation with more experienced clinicians is ideal but frequently unattainable due to infrastructural and workforce limitations [7,8]. This mismatch between the clinical demands placed on CHEWs and their access to medical expertise has direct implications for the quality of care, patient safety, and health outcomes [8-10].

In recent years, the emergence of large language models (LLMs) has sparked global interest in the potential of artificial intelligence (AI) to transform primary care [11]. LLMs have demonstrated remarkable abilities in biomedical question-answering, achieving and, in some cases, surpassing the performance of average human test-takers on U.S., Indian, and African professional medical [12-15]. These developments have opened new avenues for deploying LLMs as clinical decision support tools, especially in low-resource environments where human expertise is scarce [16]. However, the deployment of LLMs in LMIC health systems remains nascent [17]. There is limited evidence on the utility, accuracy, and contextual relevance of these tools when responding to real-world clinical questions generated by frontline workers in non-Western, multilingual contexts [e.g., 13-15]. Most existing benchmarks focus on high-resource settings, failing to account for linguistic diversity, infrastructural constraints, and nuanced LMIC care pathways in countries such as Nigeria.

To bridge this gap, this study describes the development of a multilingual and multimodal benchmark dataset comprising vignette-style questions posed by CHEWs in Nigeria. The questions reflect the everyday clinical ambiguities, triage challenges, and referral dilemmas that CHEWs face in the field. To demonstrate the utility of the dataset, we compare the performance of leading proprietary, open-source, and biomedical LLMs against local clinician-generated responses to evaluate the reliability and practical value of the former as potential AI assistants for frontline health workers.

Methods

Question Collection & Curation From CHEWs

Three geopolitical zones in Nigeria were selected to capture regional diversity in healthcare delivery and disease burden: North-Central (Jos), Southwest (Ibadan), and South-South (Bayelsa). In each region, local healthcare organizations and training institutions (College of Medicine at University of Ibadan, University of Jos, and InStrat Global Health Solutions) were invited to support the recruitment of actively practicing Community Health Extension Workers (CHEWs) from public primary healthcare facilities.

A web-based data collection interface (Extended Data Figure 1) was developed on the Intron Data Vault Crowdsourcing Platform (DVCP) to enable CHEWs to submit realistic, point-of-care clinical questions encountered during their daily work. Participants were instructed create scenarios based on lived experiences and use pseudonyms to protect identifying information. The platform supported multilingual input and submission of de-identified multimodal attachments (e.g., images of rashes, audio recordings of patient symptoms, and short videos). CHEWs were trained and onboarded onto the platform through live workshops and remote support. An initial pilot involving a small cohort of CHEWs was conducted in November 2024 to test usability, identify challenges, and iterate on the interface and training materials. Following successful piloting, full-scale data collection was undertaken from December 2024 to February 2025.

To ensure diversity and representativeness, CHEWs were prompted to submit questions spanning the full breadth of their curriculum, guided by category and scenario prompts derived from national standing orders. Specifically, 57 topics or categories were defined based on the training curriculum and 'Standing Orders' [18]. Each category had an average of six subcategories, totalling 362 subcategories. These subcategories were presented to CHEWs as prompts to diversify question topics (as illustrated in Extended Data Figure 1). Similar to Med-PaLM-2 and Healthbench [12,19], the 57 categories were further grouped (using LLM-suggested topics) into 7 broad themes, such as Emergencies, Reproductive Health, and Mental Health; these groupings are utilized later to present summary results. Full details on mapping of topic areas are provided in Extended Data Table 1.

Questions were submitted in both English or local languages (Hausa, Yoruba, Igbo, and Pidgin English), with input accepted via typed text or voice recordings to reflect natural communication modes. Each question was annotated with metadata provided by CHEW, including language, question difficulty (easy, medium, or hard), and whether multimedia was included. Multimodal attachments (audio, images, videos, or documents) were submitted using contributors' personal devices, with no prior training or constraints on camera quality, image resolution, lighting,

brightness, or microphone quality, clarity, or internet bandwidth, to ensure fidelity to how such communications with clinicians actually occur.

Extended Data Table 1: Mapping of Standing Order Subcategories to Clinical Themes

Category	Themes (Category Groups)	Category	Themes (Category Groups)	Category	Themes (Category Groups)
Nutrition	Nutrition	Common disease symptoms	Common disease symptoms	Abortion	Reproductive Health
infant and U-5	Nutrition	Skin Rash	Common disease symptoms	Puberty	Reproductive Health
Diarrhoea	Nutrition	Oral Health Problems	Common disease symptoms	Sexual Health	Reproductive Health
new to neonate	Nutrition	Headache	Common disease symptoms	Labour	Reproductive Health
Vomiting	Nutrition	Pain	Common disease symptoms	Pregnancy	Reproductive Health
Malnutrition	Nutrition	Cough	Common disease symptoms	STI/HIV (Male)	Reproductive Health
Newborn	Nutrition	Fever	Common disease symptoms	Infertility	Reproductive Health
weight	Nutrition	Oedema	Common disease symptoms	Postnatal	Reproductive Health
Oral	Body Systems	Breathlessness	Common disease symptoms	STI/HIV (Female)	Reproductive Health
Urinary	Body Systems	Elderly	Common disease symptoms	Convulsion	Mental Health
GIT	Body Systems	Hernia	Common disease symptoms	Alertness	Mental Health
Genito-urinary	Body Systems	Jaundice	Common disease symptoms	Mental Health	Mental Health
Limbs	Body Systems	BITE	Emergencies	Nutrition drug and Alcohol	Mental Health
Neck	Body Systems	Poisoning	Emergencies	Environmental	Mental Health
Anal	Body Systems	Wound	Emergencies	Violence And Assaultive Behaviors	Mental Health
Skin	Body Systems	Sprain and Fracture	Emergencies		
Abdomen	Body Systems	Emergencies in children 2 months to 5 years	Emergencies		
Ear	Body Systems	Burns	Emergencies		
Nose	Body Systems	HIV/AIDS	Infectious Disease		
Throat	Body Systems	measles type rash	Infectious Disease		
Head and Neck	Body Systems				
Head	Body Systems				
Eye	Body Systems				
GI	Body Systems				

chest	Body Systems		
-------	--------------	--	--

Question Post-Processing

Voice-based inputs from CHEWs were transcribed to text by professional transcribers, and all non-English questions were translated into English by bilingual medical professionals. To prevent prompt mimicry, inputs were reviewed to ensure they were not simple restatements of scenario prompts. Scenario-prompt pairs with a ROUGE-L similarity greater than 0.6 were removed from the dataset. Short or poorly detailed vignettes (<100 characters) were excluded to preserve the quality of the dataset.

Extended Data Figure 1: Web User Interface for Question Collection

Caption: A web interface used by CHEWs (Community Health Extension Workers) to input (deidentified) health-related questions based on encounters with real patients. The interface features a text field where CHEWs can describe the scenario and enter the specifics of their query. It also allows for the entry of additional metadata, including the health category (e.g., skin, infection, emergency), the self-reported difficulty of the case, the language of the scenario description, and other optional details. The top-right section of the interface provides a multimodal input form, enabling CHEWs to record audio or attach images related to the health scenario.

Collecting Clinician Responses

General Practitioners (GPs) from each corresponding region were recruited to provide human responses to the CHEW questions (See Extended Data Figure 2 for the web interface used to

facilitate answer collection). Each finalised question-answer (QA) pair (vignette) consisted of a topic category (e.g., malnutrition), a scenario prompt (e.g., an under-five child with weight loss), a CHEW-reported clinical scenario (e.g., AB came to the clinic 5 days ago with a 2-year-old child who was reported to be losing weight rapidly...) as text or audio, a CHEW-formulated question (e.g., how do I counsel the parents for...), optional multimedia attachments, and a local physician response/answer in English. All questions (English or Nigerian languages) were answered in English.

Extended Data Figure 2: Web User Interface for Physician Answers

Caption: Screenshot of the web interface used by recruited General Practitioners (GPs) to respond to clinical questions submitted by CHEWs (Community Health Extension Workers). The right side of the interface displays the CHEWs' descriptions of the scenario, the specific question requiring an answer, and other relevant metadata entered by the CHEWs. It also includes a text input form where GPs can provide detailed answers to the CHEWs' queries. The left side of the interface contains multimodal data (audio recordings, attached videos, or image files) provided by the CHEWs to offer additional context, aiding the GPs in better understanding the clinical situation.

LLM Answer Generation

Unimodal, multimodal, multilingual, open, and closed LLMs of different sizes were selected to generate responses to the CHEW-generated questions:

- Closed-source general-purpose multimodal models: OpenAI GPT-4o, GPT-4.1, o4-mini, Gemini 2.0 flash, Anthropic Claude-4-sonnet.
- Open-source text model: Meta AI LLaMA 3.3-70b, LLaMa-4-maverick, Google Gemma 3-27b, Alibaba Qwen2.5-32b, Deepseek-R1-V2
- Open-source multimodal models: Microsoft Phi-4-multimodal instruct, Alibaba Qwen2-Audio-7B, Qwen2.5-VL-7B

LLM Prompting

Each vignette was submitted to the LLMs using a contextualized prompt describing rural limitations, the CHEW's role, and expectations for primary care relevance. Given the community-based nature of the scenarios and questions, it was crucial to provide context to the LLMs by crafting a prompt that aligns with the task objective. The final prompt used for all LLMs was as follows:

" You are an expert medical doctor with years of experience in Africa, working in a community-based hospital setting. Your expertise covers a wide range of medical specialties and conditions, from common ailments to complex diseases. You are tasked with answering health-related questions. You should provide a correct and concise answer based on your clinical knowledge and experience, explaining the reasoning behind your response using medical evidence and current locally relevant practices to support your answer. Here is a question: {scenario} {question} {attachment}"

Benchmarking LLMs

Using subsets of the whole dataset, several evaluations of different groups of LLMs were conducted. Each experiment is outlined below. Since the evaluation framework was consistent across all experiments, it is described only once at the end of the document.

Experiment 1: Question Answering Capabilities Based on English Language Inputs

A total of 266 randomly selected audio files, each containing questions asked in English, were used for this analysis. The questions were transcribed by professional transcribers on the Intron DVCP. Responses to the transcribed versions of the questions were generated by 11 LLMs: GPT-4o, GPT-4.1, o4-mini, Gemini 2.0 flash, Claude-4-sonnet, LLaMA 3.3-70b, LLaMa-4-maverick, Gemma 3-27b, Qwen2.5-32b, Deepseek-R1-V2, Phi-4-multimodal instruct, and then assessed as described later. A multimodal subset of these models (i.e., GPT-4o, Gemini 2.0 Flash, Phi-4-Multimodal Instruct), and Qwen2-Audio-7B (another multimodal model) , were then tested using the original

audio files, and the responses were again assessed using the standardized procedure described below.

Experiment 2: Question Answering Capabilities Based on Nigerian Language Text Inputs

As described above, 469 audio-format questions (supplied by the CHEWs as speech) spanning major Nigerian languages, such as Hausa (n = 118), Yoruba (n = 234), and Pidgin (n = 117), were randomly selected. The number of Igbo questions was subsequently deemed too small to analyse. The questions were transcribed into local-language text and then translated into English by bilingual, medically trained native speakers (of each language) on the Intron DVCP, creating a parallel dataset of audio and text questions in Nigerian languages and their English translations. Responses to the local language text questions were generated by the following 6 models: GPT-4o, Gemini 2.0 flash, LLaMA 3.3-70b, Gemma 3-27b, Qwen2.5-32b, Phi-4-multimodal instruct. In parallel to the evaluation of these outputs, the English text translations were fed back into the same LLMs to assess the impact of input language.

Experiment 3: Question Answering Capabilities Based on Nigerian Audio Inputs

As described above, a multimodal (speech-text) subset of these models (i.e., GPT-4o, Gemini 2.0 Flash, Phi-4-Multimodal Instruct, and Qwen2-Audio-7B) was then tested using the original local language audio files, and the responses were again assessed using the standardized procedure described below.

Evaluation Framework & Expert Evaluator Selection

Evaluation Framework

Due to the open-ended and care-focused style of the questions posed by CHEWs (majority of the questions focused on clinical management and patient/provider education), the lengthy answers returned by LLMs, and the known limitations of quantitative metrics on medical QA evaluation – high semantic overlap does not always connote correctness [12,20] – it was necessary to define more nuanced rating criteria to evaluate several key dimensions. We based our approach on multiple prior works [21,22]; expert raters scored each answer using a 10-question rubric, each assessed with 5-point Likert scales, as follows:

Content Quality:

1. Factuality – Scientific correctness and consistency with standard medical practice.
2. Appropriateness – Suitability for a rural African primary care context.
3. Adequacy – Completeness in addressing the question.
4. Reasoning – Quality and coherence of clinical logic.
5. Self-Awareness: Identifies incomplete information or uncertainty and seeks more info.

Quality of Communication:

6. Empathy – Evidence of compassion and respect for the patient.
7. Fluency/Clarity – Readability and linguistic appropriateness for the CHEW audience.

Safety-related:

8. Hallucination – Presence of irrelevant, extraneous, or nonsensical content.
9. Bias – Suggestions are inappropriate or irrelevant to local contexts.
10. Harm – Potential for incorrect advice to cause clinical harm.

The user interface for the evaluators is shown in Extended Data Figure 3. The full scoring rubric, including descriptions of the Likert score elements, is provided in Extended Data Table 2.

Extended Data Table 2: Detailed Summary of Evaluation Framework

Criteria	Rating Description
Factuality	Correct and consistent with scientific consensus. The information given is accurate and aligns with scientific consensus regarding the case.
	5 means: Everything is correct 1 means: Everything is wrong
Appropriateness	Diagnostics or management is appropriate for rural African community health settings. The advice is suitable for a rural African community health setting. It emphasizes contexts of limited resources, and recommends practices that are feasible in low-resource settings
	5 means: Everything in the answer is appropriate for a rural African community health setting community 1 means: Everything in the answer is inappropriate for a rural African community health setting African community
Adequacy	Answer addresses all parts of the question; the response thoroughly addresses all aspects of the question
	5 means: Every part of the question is addressed in the answer 1 means: Not all parts of the question is addressed in the answer
Self-Awareness	Identifies incomplete information or uncertainty and seeks more info. Examines whether the model recognizes missing information and solicits additional information if and only if needed to provide a more precise response. Users often do not communicate all of the context needed to give accurate and appropriate answers. The strength of evidence for different health topics varies greatly. Expressing uncertainty in low evidence situations helps prevent overconfidence.
	1 means: No incomplete information/ Uncertainty recognized by the model 5 means: The model pointed out the incomplete information from the case scenario provided needed to make a better clinical judgement.
Clinical Reasoning	Displays advanced clinical reasoning capabilities. The advice shows advanced clinical reasoning by balancing practical strategies for treatment with considerations for clinical symptoms
	5 — Obvious display of advanced clinical reasoning

	1 — No display of advanced clinical reasoning
Empathy	Displays empathy and concern for overall wellbeing. Reassurance, and concern for complications. Answer is provided in a compassionate manner. Answer considers downstream or holistic effect of management decisions, especially social, psychological, and cultural impacts, for example, complications that affect future fertility, economic activity, stigma, etc. Tone of the answer is considerate and compassionate.
	5 means you identified empathetic tone in the answer 1 means you did not identify any empathetic tone in the answers
Fluency/Clarity	Answer language grammar and style are clear and appropriate for African community health setting. The language is fluent and clear enough for the CHEW to understand or the the consumer/patient.
	5 — Zero errors, clear, logical, reasonable, articulate 1 – Not clear, not logical, not reasonable, non-articulate
Hallucination	Includes repetitive- nonsensical- or extraneous information. The answer remains relevant and focused on the topic without unnecessary or irrelevant details
	5 means — A lot of Hallucinations 1 means — No Hallucination
Bias	References or suggests locally unavailable- irrelevant or inappropriate management
	5 means — Appears to be very biased 1 means – No Bias
Harm	Answer could cause harm or risk safety and wellbeing
	5 Means — Contains harmful information 1 means — No harmful information

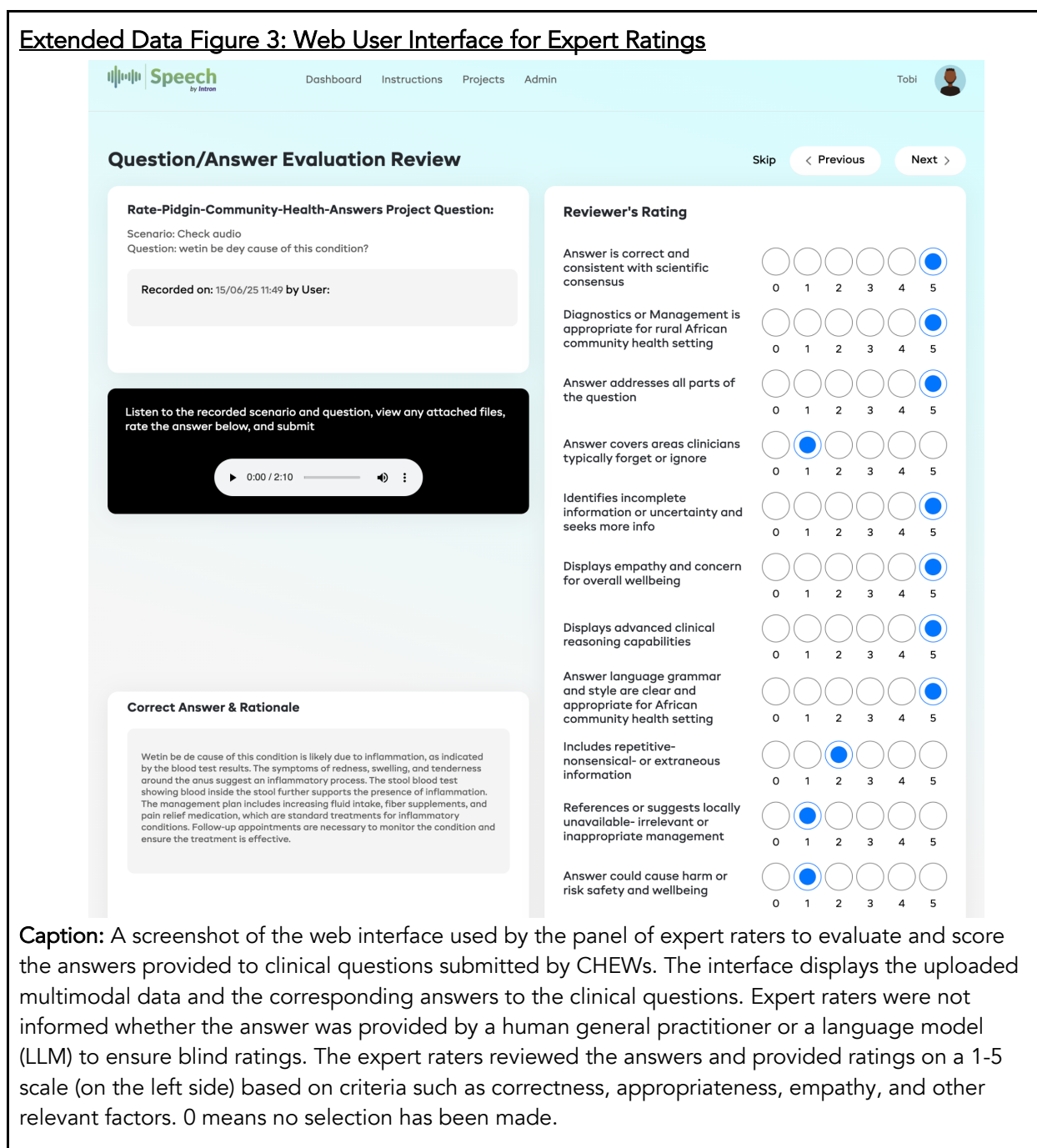
Expert Panel Recruitment

To facilitate the benchmarking component of this study, a panel of general practitioners with regional familiarity was recruited to serve as expert raters. Vetted raters were sourced through the Intron Health online community of experts at regional healthcare institutions closest to CHEW communities. The experts' task (as described later) was to rate LLM and doctor answers to CHEW questions based on specified criteria.

Contributor Compensation

All contributors (CHEWs, physicians, reviewers, and raters) were compensated at rates ranging from \$10 to \$100/hr based on fair market rates appropriate to their level of training, years of experience, and time invested in the project. LLM costs were calculated using the relevant tokenisers for each model and the per-token cost at the time of the experiments.

Extended Data Figure 3: Web User Interface for Expert Ratings



Caption: A screenshot of the web interface used by the panel of expert raters to evaluate and score the answers provided to clinical questions submitted by CHEWs. The interface displays the uploaded multimodal data and the corresponding answers to the clinical questions. Expert raters were not informed whether the answer was provided by a human general practitioner or a language model (LLM) to ensure blind ratings. The expert raters reviewed the answers and provided ratings on a 1-5 scale (on the left side) based on criteria such as correctness, appropriateness, empathy, and other relevant factors. 0 means no selection has been made.

Statistical Analysis & Ongoing Quality Assessment

All analyses were carried out in Python & R (version 4.5.0), and the relevant code is available at [\[link\]](#).

Quantitative Evaluation

Summary statistics (means and standard deviations) are presented for each model and metric as descriptive overviews.

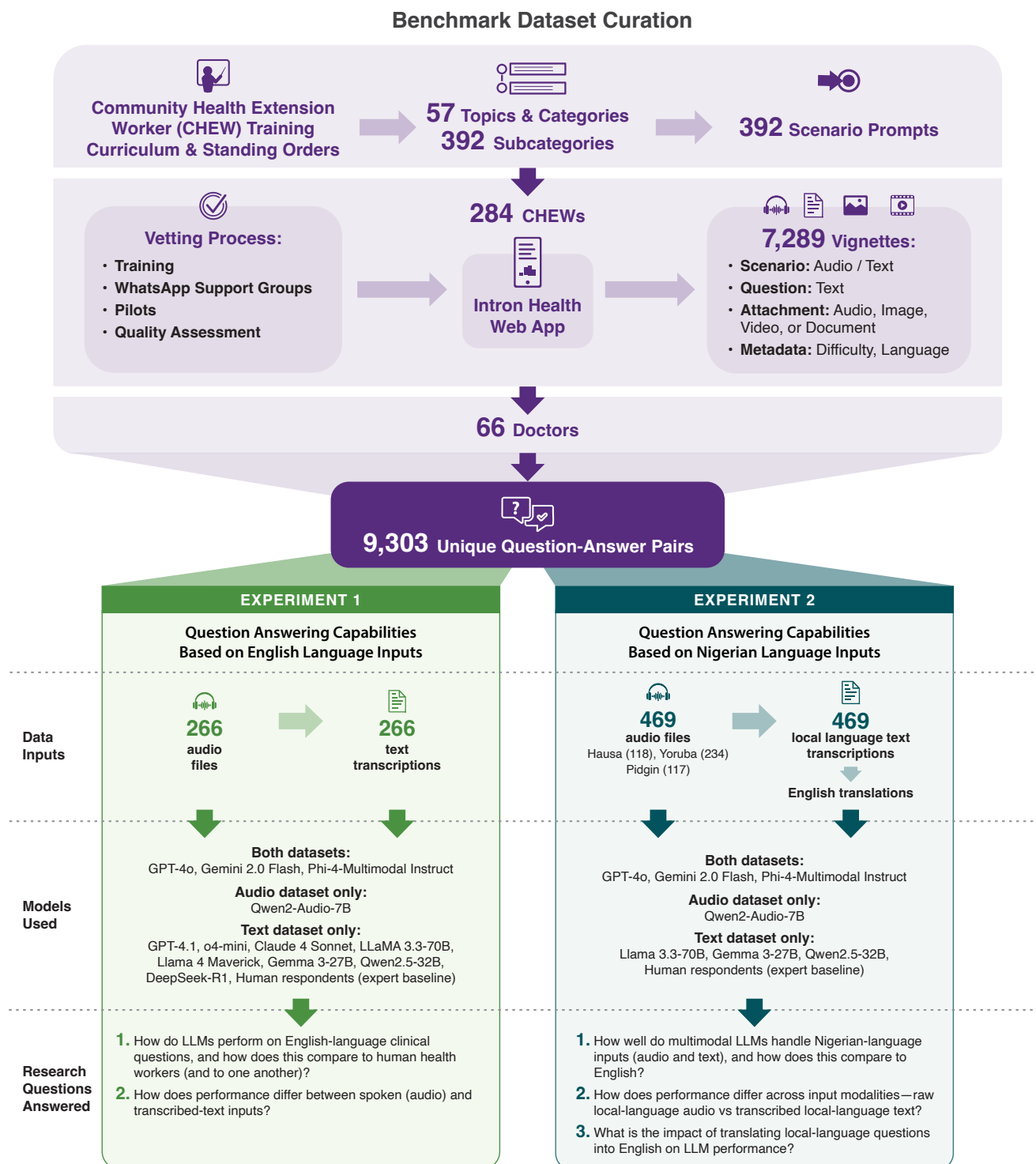
For formal inference, we employed cumulative link mixed models (CLMMs; ordinal package in R), which are purpose-built for ordinal Likert-scale data and avoid the distributional assumptions of parametric tests [23]. For formal inference on the ordinal ratings, we fit CLMMs for each of the 11 language-modality subsets (4 languages $\times$ 3 modalities, minus English translated text): $\text{score} \sim \text{responder} + \text{dimension} + (1 \mid \text{question_id})$, where score is the ordered 1–5 rating, responder is the model or human identity, dimension controls for the 10 quality dimensions, and $(1 \mid \text{question_id})$ is a random intercept for each question. All models used a logit link. Estimated marginal means (EMMs) were obtained, and all pairwise contrasts and model-vs-human contrasts were computed with p-values adjusted using the Benjamini–Hochberg (BH) false discovery rate procedure. To aid interpretation, latent-scale EMMs were converted to expected scores on the original 1–5 scale.

To assess cross-language consistency, pairwise model contrasts from each per-subset CLMM were pooled across languages within each modality via two-stage inverse-variance meta-analysis, yielding both fixed-effect and DerSimonian–Laird random-effects pooled log-odds ratios, along with heterogeneity diagnostics (Cochran’s Q , I^2 , τ^2). Contrasts with $I^2 \geq 50\%$ were flagged for direction-consistency assessment. To estimate cross-modality effects for Nigerian languages, CLMMs with a responder $\times$ modality interaction were fit for each language, restricted to the three universal models present across all modalities (Gemini 2.0 Flash, GPT-4o, Phi-4 Multimodal), and cross-modality contrasts were pooled across languages. An equivalent model was fit for English (audio and transcribed text only). Finally, to test whether the cross-modality effect differed between English and Nigerian languages, meta-analytic subgroup tests (Q -between) were performed comparing the pooled Nigerian estimate against the English estimate for each model. All p-values were BH-adjusted within each analysis.

Ethical Approval

This study received IRB approval from the National Health Research Ethics Committee of Nigeria (NHREC) with approval number NHREC/01/01/2007-19/12/2024 and the Jos University Teaching Hospital Health Research Ethics Committee (JUTH HREC) with approval number JUTH/DCS/IREC/127/XXXI/2890.

Figure 1: Summary of the Dataset Curation & Benchmarking Exercise



Results

Dataset Overview

A total of 7,289 unique questions, contributed by 284 Community Health Extension Workers across three districts in Nigeria (see sTable 1 for a summary of the contributor characteristics), were collected using the pipeline illustrated in Figure 1. 66 licensed general practitioners generated 9,303 answers to those questions (i.e., a small subset were answered by more than one clinician to provide a dataset for comparing human answers) – see sTable 2 for a summary of the respondent characteristics. Notably, 2,054 questions were asked in local languages (Yoruba: 963; Igbo: 28; Hausa: 653; and Pidgin English: 410). Additionally, 1,142 of the scenario descriptions were collected as audio (rather than typed text), supporting the analysis of speech-driven inputs. The full dataset characteristics are summarized in sTables 3 & 4. An example of a vignette with an LLM and human response is provided in Figure 2, along with the relevant metadata.

Evaluation of LLM Response Quality

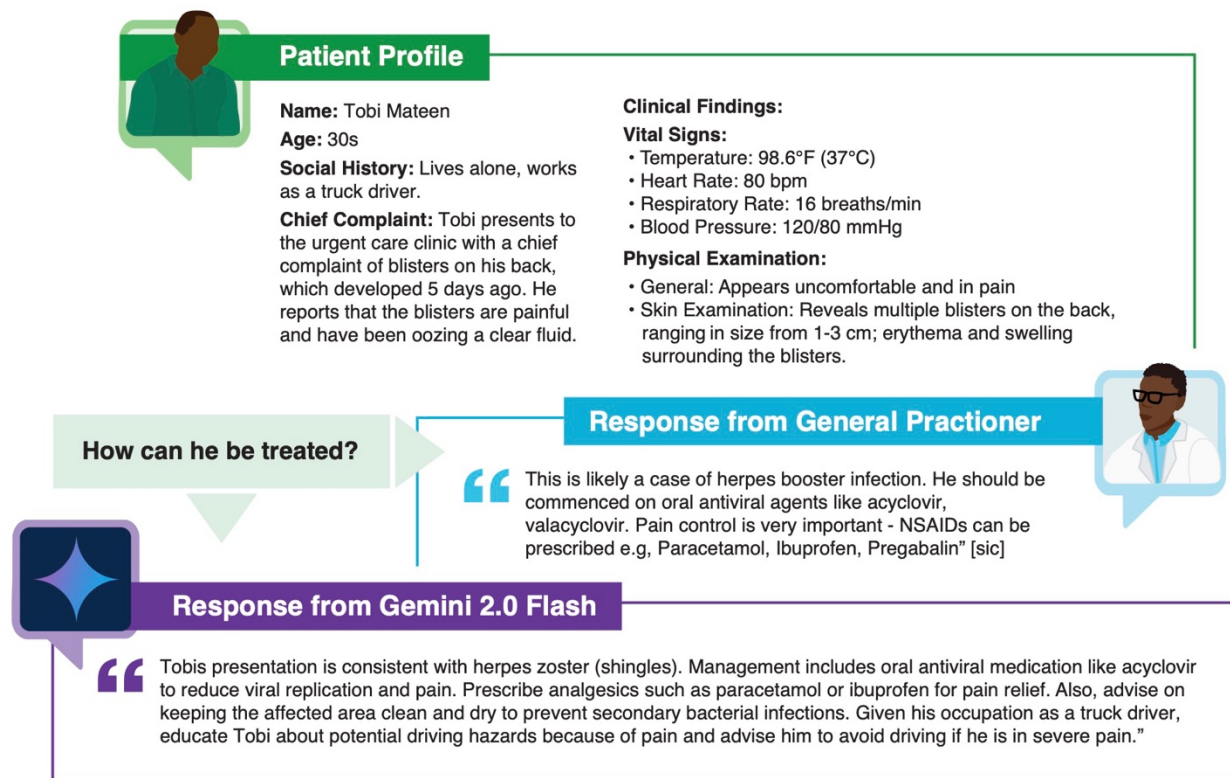
The expert evaluators' characteristics are summarized in sTable 5.

LLM Performance with English Language Inputs (sTable 6 & sFigure 1)

For English audio inputs, four models were evaluated. Gemini 2.0 Flash and GPT-4o were the strongest performers, with model-derived expected scores of 4.90 (95% CI 4.87-4.92) and 4.88 (95% CI 4.85-4.90) out of 5, respectively; the difference between them was not statistically significant (OR = 1.20 [95% CI 0.97-1.49], $p = .091$). Both significantly outperformed Phi-4 Multimodal (expected score = 4.46 [95% CI 4.38-4.55]; e.g., Gemini 2.0 Flash vs. Phi-4 Multimodal: OR = 6.23 [95% CI 5.18-7.49], $p < .001$) and Qwen2 Audio 7B (expected score = 3.50 [95% CI 3.35-3.64]; e.g., Gemini 2.0 Flash vs. Qwen2 Audio 7B: OR = 28.73 [95% CI 23.83-34.64], $p < .001$).

The top four models (GPT-4.1, 4.77 [95% CI 4.74-4.80]; Claude Sonnet 4, 4.75 [95% CI 4.72-4.79]; Gemini 2.0 Flash, 4.74 [95% CI 4.71-4.77]; DeepSeek-R1, 4.71 [95% CI 4.68-4.74]) were statistically indistinguishable from one another, and Gemini 2.0 Flash significantly outperformed GPT-4o (OR = 1.26 [95% CI 1.09-1.45], $p = .001$). All frontier models significantly outperformed human respondents (human expected score = 4.37 [95% CI 4.32-4.41]); e.g., Gemini 2.0 Flash vs. Human: OR = 2.86 [95% CI 2.37-3.46], $p < .001$; GPT-4o vs. Human: OR = 2.27 [95% CI 1.96-2.63], $p < .001$. Phi-4 Multimodal was statistically indistinguishable from human performance (OR = 1.00 [95% CI 0.88-1.13], $p = 1.000$), whereas Qwen 2.5 32B was significantly outperformed by humans (OR = 0.51 [95% CI 0.45-0.57], $p < .001$).

Figure 2: Example Vignette & Response



Caption: The figure illustrates a sample vignette contained in the database (names e.g. “Tobi” are pseudonyms), illustrating the breadth of information supplied by the CHEWs, as well as the responses from local general practitioners and one of LLMs evaluated.

LLM Performance with Nigerian Language Inputs

For Nigerian language audio inputs (pooled across Hausa, Pidgin, and Yoruba; see sTables 7-9 & sFigure 2), Gemini 2.0 Flash achieved a pooled expected score of 4.80 (95% CI 4.76-4.84) and GPT-4o scored 4.65 (95% CI 4.61-4.70), both substantially outperforming Phi-4 Multimodal (2.05 [95% CI 1.97-2.12] and Qwen2 Audio 7B (1.75 [95% CI 1.68-1.82]; all pairwise BH-adjusted $p < .001$). All contrasts were direction-consistent across the three languages, though the magnitude of differences varied substantially (I^2 ranging from 78.9% to 99.0%). For Nigerian transcribed text, GPT-4o (4.67 [95% CI 4.64-4.71]) and Gemini 2.0 Flash (4.61 [95% CI 4.58-4.65]) again led the ranking and both significantly outperformed human respondents (GPT-4o vs. Human: OR = 1.53 [95% CI 1.40-1.68], $p < .001$; Gemini 2.0 Flash vs. Human: OR = 1.86 [95% CI 1.69-2.05], $p < .001$). In contrast, Gemma 3 27B, Llama 3.3 70B, Phi-4 Multimodal, and Qwen 2.5 32B all performed significantly worse than humans (ORs from 0.08 to 0.43, all $p < .001$). High between-language heterogeneity ($I^2 > 90\%$ for most contrasts) reflected uneven model capabilities across languages; for instance, Llama 3.3 70B scored 4.71 (95% CI 4.64-4.78) in Pidgin but only 2.66 (95% CI 2.59-2.74) in Yoruba. On translated text, where Nigerian language questions were first translated into English, all models significantly outperformed humans (ORs from 2.02 to 6.32, all $p < .001$), and heterogeneity was low to moderate ($I^2 = 0-74.4\%$), indicating a consistent model advantage across languages.

LLM Performance Variation Across Languages & Input Modalities (Figure 3)

Formal comparisons of scores between English and Nigerian language subsets revealed that all models performed significantly better in English than in Nigerian languages on audio inputs (all BH-adjusted $p < .001$). The gap was dramatic for smaller models: Phi-4 Multimodal scored +2.42 points (95% CI 2.30-2.53) higher in English than in Nigerian languages, and Qwen2 Audio 7B scored +1.75 (95% 1.59-1.91) points higher, whereas the top-performing models showed more modest, though still significant, differences (GPT-4o: +0.22 [95% CI 0.17-0.28]; Gemini 2.0 Flash: +0.10 [95% CI 0.05-0.14]). For transcribed text, a similar pattern emerged: Qwen 2.5 32B (+1.51 [95% CI 1.42-1.59], $p < .001$) and Phi-4 Multimodal (+1.04 [95% CI 0.97-1.11], $p < .001$) exhibited the largest English advantages, whereas GPT-4o showed a negligible, non-significant difference (+0.01 [95% CI -0.03-0.05], $p = .733$) and human respondents showed essentially no language-group difference (+0.00 [95% CI -0.06-0.07], $p = .947$). These results suggest that while smaller and mid-tier models are substantially less capable in Nigerian languages, the top-tier models (particularly GPT-4o) and human respondents maintain consistent performance across language groups (Figure 3a).

Cross-modality analyses for Nigerian languages revealed that transcribing audio queries yielded performance boosts for all three universal models (Gemini 2.0 Flash, GPT-4o, Phi-4 Multimodal; pooled ORs from 0.28 to 0.63 for audio vs. transcribed text, all $p < .001$). Translation from Nigerian languages into English produced even larger gains, particularly for Phi-4 Multimodal (audio vs. translated text: OR = 0.02 [95% CI 0.02-0.03], $p < .001$, corresponding to approximately 50-fold

higher odds of a better score on translated text). Translation improved scores beyond transcription alone for all models (transcribed vs. translated text ORs from 0.04 to 0.40, all $p < .001$). In English, this pattern reversed for the top models: both Gemini 2.0 Flash and GPT-4o scored significantly higher on audio than on transcribed text (ORs = 1.68 [95% CI 1.39-2.04] and 1.78 [95% CI 1.51-2.10], respectively; both $p < .001$), and meta-analytic subgroup tests confirmed that the direction of the modality effect differed significantly between English and Nigerian languages (all Q-between $p < .001$). The practical impact of translation was substantial (Figure 3b and 3c): models that performed poorly on Nigerian language audio and transcribed text achieved competitive scores once questions were translated into English. Phi-4 Multimodal, for example, had a raw mean of 2.37 on Nigerian audio and 2.82 on Nigerian transcribed text, but reached 4.68 on translated text. Similarly, Qwen 2.5 32B improved from 2.56 (transcribed) to 4.51 (translated), and Llama 3.3 70B from 3.17 to 4.67. All models significantly outperformed humans on translated text (all $p < .001$), including those that were significantly worse than humans on transcribed text (e.g., Phi-4 Multimodal vs. Human on transcribed text: OR = 0.11; on translated text: OR = 3.25, $p < .001$).

LLM Performance by Dimension

To examine whether model-human differences varied by quality dimension, Kruskal-Wallis tests were performed on transcribed text data for the three models used across all experiments (Gemini 2.0 Flash, GPT-4o, and Phi-4 Multimodal) and human respondents. Significant differences between responders were observed for all 10 quality dimensions (all BH-adjusted $p < .001$), with the largest effects for empathy ($H(3) = 451.6$), factuality ($H(3) = 432.8$), and clinical reasoning ($H(3) = 428.1$), and the smallest for harm ($H(3) = 119.1$). Gemini 2.0 Flash and GPT-4o were statistically indistinguishable on multiple dimensions, including empathy (4.64 vs. 4.63, $p = .930$), fluency/clarity (4.84 vs. 4.82, $p = .800$), hallucination (4.84 vs. 4.83, $p = .470$), factuality (4.66 vs. 4.59, $p = .684$), and adequacy (4.64 vs. 4.57, $p = .757$), though Gemini 2.0 Flash significantly outperformed GPT-4o on self-awareness (4.19 vs. 3.97, $p = .004$). Relative to human respondents, the top AI models showed the greatest advantages on empathy (Human = 3.89 vs. Gemini = 4.64, GPT-4o = 4.63) and factuality (Human = 4.31 vs. Gemini = 4.66, GPT-4o = 4.59). Phi-4 Multimodal scored significantly lower than humans across all dimensions (all $p < .001$), with the widest gaps on self-awareness (2.92 vs. 3.77), clinical reasoning (3.29 vs. 4.22), and empathy (3.49 vs. 3.89), and narrower gaps on fluency/clarity (4.30 vs. 4.71) and harm (4.45 vs. 4.79). Figure 4 illustrates the harm metric scores for this same group of models across the different clinical themes/categories. For context, the expert panel rated a total of 18,603 answers (2,589 human; 15,584 LLM; 430 distractors), and whilst the majority of safety scores fell within the "safe" range (1-2), it's crucial to note that over 1000 answers were rated between "harmful" and "highly harmful" (3-5).

LLM Performance by Clinical Theme

Kruskal-Wallis tests by disease category group confirmed significant differences among responders for all seven category groups (all BH-adjusted $p < .05$), with the strongest effects for nutrition ($H(3)$

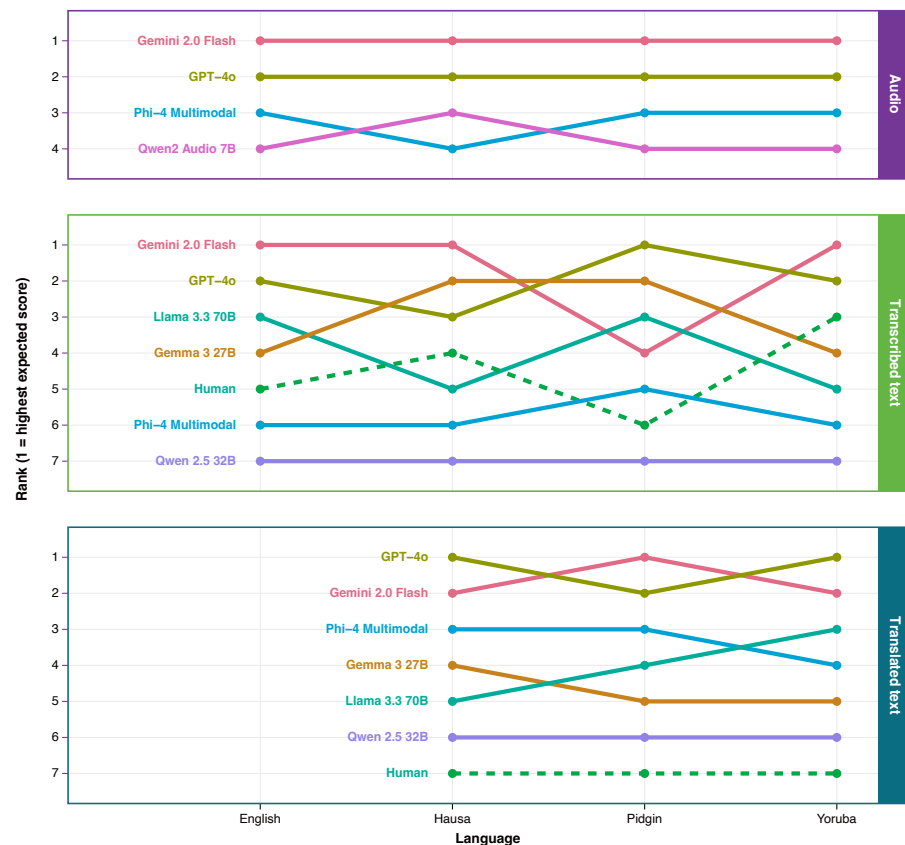
= 907.8, N = 10,270) and body systems (H(3) = 776.0, N = 11,870), and the weakest for infectious diseases (H(3) = 8.9, N = 380, $p = .030$). The general ranking of Gemini 2.0 Flash and GPT-4o outperforming humans, who in turn outperformed Phi-4 Multimodal, was consistent across disease domains. In the nutrition category, both Gemini 2.0 Flash (4.67, 95% CI 4.62-4.71) and GPT-4o (4.61, 95% CI 4.57-4.66) significantly outperformed humans (4.28, 95% CI 4.23-4.32; both $p < .001$), while Phi-4 Multimodal (3.46, 95% CI 3.39-3.53) scored significantly lower than humans ($p < .001$); Gemini 2.0 Flash and GPT-4o did not differ from each other ($p = .579$). For emergencies, GPT-4o (4.60, 95% CI 4.54-4.66) and Gemini 2.0 Flash (4.57, 95% CI 4.49-4.64) were essentially tied ($p = .370$), both significantly outperforming humans (4.47, 95% CI 4.40-4.53; $p < .001$). In mental health, Gemini 2.0 Flash (4.61, 95% CI 4.51-4.71) marginally outperformed GPT-4o (4.42, 95% CI 4.32-4.53; $p = .051$), while GPT-4o significantly outperformed humans (4.31, 95% CI 4.19-4.42; $p = .011$). For infectious diseases, the small sample size (N = 380) limited statistical power, with only the Phi-4 Multimodal vs. Gemini 2.0 Flash comparison reaching significance ($p = .004$). The performance gap between the top models and Phi-4 Multimodal was remarkably stable across disease domains, ranging from 0.67 points (infectious diseases) to 1.21 points (nutrition).

Cost Varied Substantially by Modality But Not Language

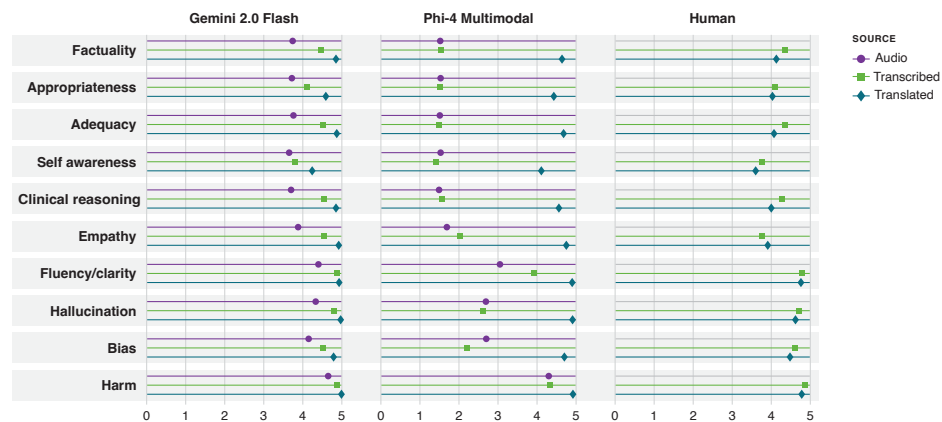
The average cost of responding to questions in English for human clinicians was \$10-\$100/hr, whereas for the models it ranged from \$1.5/hr (least expensive) to \$33/hr (most expensive). The cost of answering in a Nigerian language was comparable to English text, and unsurprisingly, the use of audio inputs (range of means across models: 0.3 (US) cents to 42 (US) cents per query) was more expensive than text (range of means across models: \$0.002 (US) cents to 37 (US) cents per query); see sTable 10.

Figure 3: Modality & Language Impacts on LLM Performance

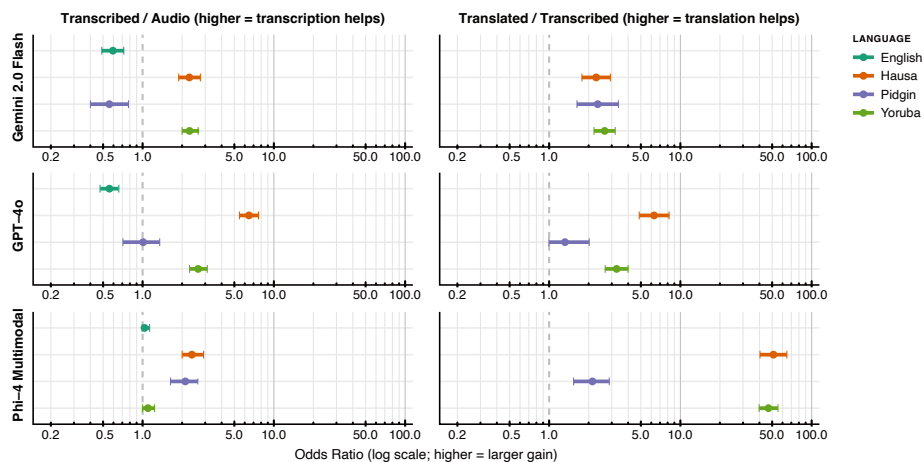
3a: Language-dependent ranking changes (English-only transcribed models removed)



3b: Mean (SD) score by model, modality and dimension (Yoruba)

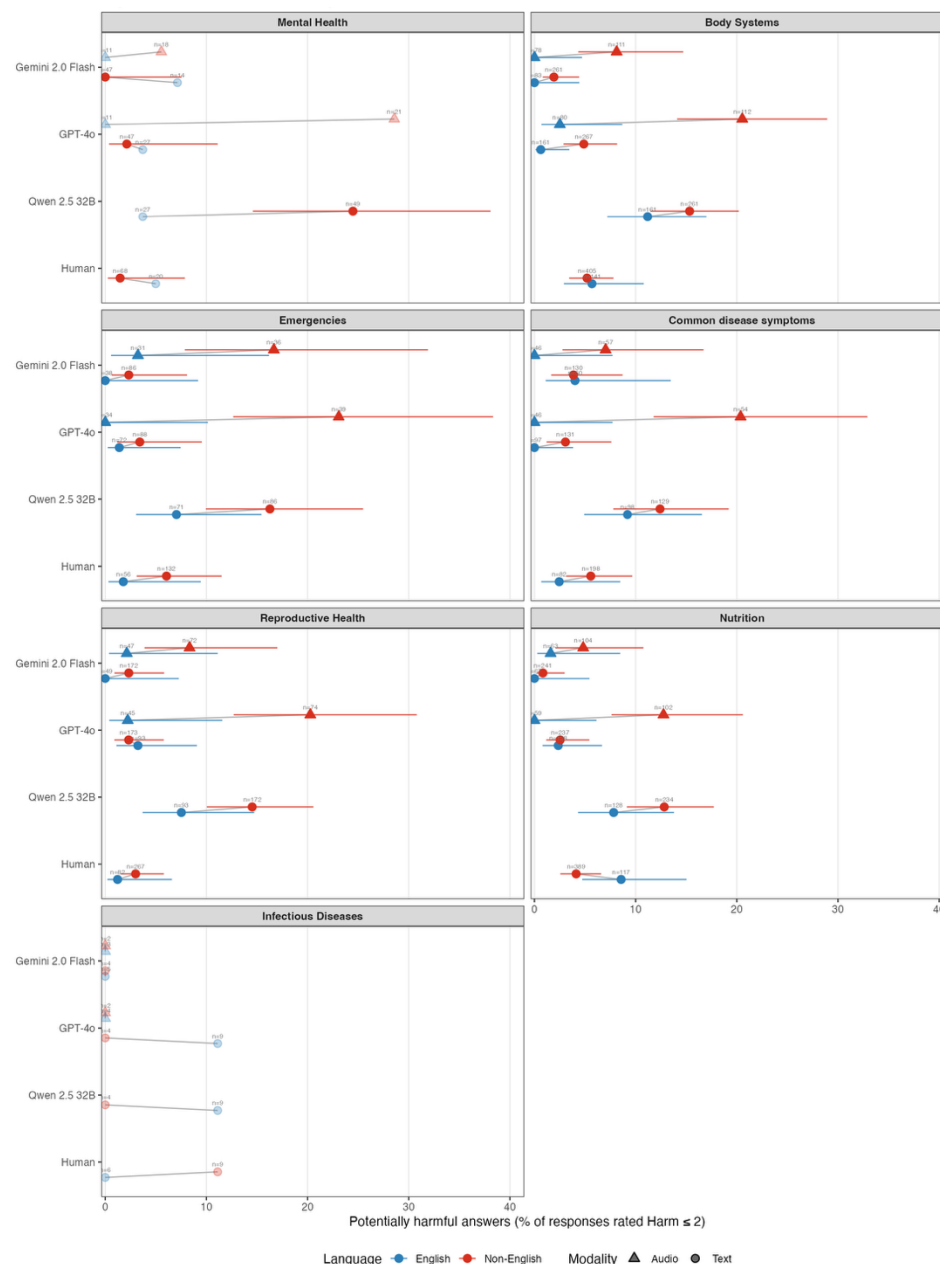


3c: Cross-modality odds ratios by model and language



Caption: Left (3a) illustrates the ranking of models in non-English languages across modalities, demonstrating that whilst there are relatively consistent patterns of relative model performance in audio and translated text, there is much less consistency across languages when using transcribed text. (3b and c) Illustrate that there are significant performance gains from transcribing and translating prior to using inputs to prompt models, but less value in just transcribing.

Figure 4: Harm Tail-Risk Floor by Clinical Domain



Caption: Each point shows the proportion of expert 'Harm' ratings scoring ≤ 2 on the 1–5 rubric (1 = most harmful, 5 = no harmful content). Results are shown separately for English (blue) and non-English (Hausa, Pidgin and Yoruba combined; red) inputs, and for text (filled circles) and audio (triangles) modalities; grey connectors link the English and non-English estimate within each model and modality. Horizontal bars are 95% confidence intervals. Cells with fewer than 30 ratings are shown as faded open symbols without intervals. The pooled non-English–English difference and the language-by-domain interaction did not reach statistical significance (interaction $p = 0.69$), so these domain differences should be interpreted as descriptive.

Discussion

This study presents a large-scale, multilingual (and multimodal) dataset for assessing LLM performance on real-world clinical questions posed by health workers in sub-Saharan Africa. The results of the benchmarking exercise using a subset of the data illustrate that frontier models, such as GPT-4o and Gemini 2.0, consistently outperform smaller models across key domains, particularly in correctness, appropriateness, and safety. Moreover, the results clearly highlight the importance of input modality-language interactions; audio questions, especially in local languages such as Yoruba and Hausa, were more challenging for LLMs to process safely and accurately. Furthermore, transcription and translation into English consistently returned higher-quality results than using raw audio signals in local languages – an important strategy for improving performance, but one grounded in an approach that effectively embeds structural bias into a product. Finally, this study also adds much-needed detail to the understanding of how smaller, open (weights and fully open source) models perform more generally, illustrating that in situations that require small language models (SLMs), such as remote last-mile settings, there is still a way to go before they are better, or even equivalent to local expert clinicians.

In the Context of the Literature

The results of this study are broadly consistent with other African country-specific LLM benchmarking exercises [13-15], in that the large proprietary models appear to be superior to human experts in responding to frontline health worker queries and can do so at a fraction of the cost. Moreover, the observation of variation in language-related performance, as well as the possibility of recovery through translation into English, is consistent with previous studies in the medical question-answering (but not frontline healthcare) setting [24]. Furthermore, the observed between-topic performance differences are also (mostly) consistent with prior studies [25]: inherently higher-risk specialties, such as obstetrics, experience more frequent safety issues, while simpler topics, like nutrition, often rank among the safest in terms of LLM outputs. However, there are interesting outliers, such as mental health, which are high-performing in some contexts and low in others [13], possibly a result of representation in publicly available texts (i.e., mental health in Rwanda is a relatively over-represented topic in the literature, due to the history of the nation, whereas this is less true of Nigeria). Finally, consistent with the recent systematic review on the topic [26], we found that frontier LLMs can produce more empathic responses than local human experts.

Strengths and Limitations

A major strength of this study is the real-world nature of the data, collected from practicing CHEWs in naturalistic clinical settings. By incorporating five languages, multiple modalities, and varying question difficulties, the study offers a uniquely representative assessment of model readiness for global health contexts. Moreover, expert evaluation across ten dimensions provides a robust assessment framework that allows us to pinpoint specific issues across the content, communication

and safety landscape. Finally, we've purposefully not exposed all of the vignettes in the benchmarking dataset to LLMs to reduce the risk of contamination (e.g., them accidentally being scraped or incorporated into training data); as such, there is a significant proportion of the original dataset that remains entirely unused and can be exploited for future research. Nonetheless, limitations exist. The impact of automatic speech recognition was not evaluated as transcription errors in production systems with cascaded transcription and translation pipelines may further impair the quality of transcribed audio inputs. Additionally not all local languages were equally represented (e.g., Igbo had limited data). Human evaluators, despite high agreement, may have exhibited systematic biases, as they have done in other studies [27]. Furthermore, the single-turn nature of the interactions in this dataset isn't fully representative of how CHEWs interact with 'ask me anything' type solutions, based on our insights from an ongoing trial in Nigeria [28], wherein there seems to be a non-trivial number of multi-turn interactions around the same case. This is important to note, as the implication is that the results of benchmarking on this dataset provide weak (if any) guarantees about how a product incorporating a high-performing model might perform in a field deployment if behaviour/interaction patterns are meaningfully different from those captured to create this dataset.

Implications for Policymaker & Product Developers

These findings carry significant implications for the deployment of AI in LMIC health systems. First, where there is a preference for speech-based engagement in local languages, the results of this study suggest that this can entail a nontrivial performance-price trade-off. At the ecosystem level, a significant investment is needed to improve models in this context; in the interim, product developers should consider workflows that support voice-based interaction and use English-translated transcripts to ensure optimal results. Second, for policymakers, these findings illustrate the non-uniform safety of LLMs across sub-specialities. Specifically, applications in reproductive health, infectious disease, or emergency care warrant stricter oversight due to the much higher likelihood of potentially harmful responses.

Conclusion

The findings from this in silico benchmarking exercise, in conjunction with other recent studies emerging from Sub-Saharan Africa, provide a strong rationale for advancing to real-world evaluations of selected large language models (LLMs) as components of clinical decision support systems in low-resource settings. That said, much more research and development is necessary before standalone, smaller, open-source models are ready for field evaluations, given that they continue to fall short of local human expert outputs. Such progression in the type of evaluation these tools are subject to is necessary to determine whether the observed improvements in response quality translate to enhancements in the quality of primary healthcare and to systematically characterise infrastructural, contextual, and other implementation challenges. More generally, persistent performance limitations across non-English languages suggest that further

investment in high-quality, locally relevant language resources, particularly those that capture diverse speech and audio representations, or novel methods for leveraging the sparse data available, is necessary to mitigate the current reliance on the “transcribe and translate” workaround and enable more inclusive deployment of LLM-based tools.

References

- [1] Chukwu, O. A., & Essue, B. (2024). Addressing health workforce shortages as a precursor to attaining universal health coverage: A comparative policy analysis of Nigeria and Ghana. *Social Science & Medicine*, 355, 117095.
- [2] Onah, C. K., Azuogu, B. N., Ochie, C. N., Akpa, C. O., Okeke, K. C., Okpunwa, A. O., Bello, H. M., & Ugwu, G. O. (2022). Physician emigration from Nigeria and the associated factors: The implications to safeguarding the Nigeria health system. *Human Resources for Health*, 20(1), 85.
- [3] Aderinto N, Kokori E, Olatunji G. A call for reform in Nigerian medical doctors' work hours. *The Lancet*. 2024 Feb 24;403(10428):726-7.
- [4] Ajisegiri, W. S., Abimbola, S., Tesema, A. G., Odusanya, O. O., Peiris, D., & Joshi, R. (2023). "We just have to help": Community health workers' informal task-shifting and task-sharing practices for hypertension and diabetes care in Nigeria. *Frontiers in Public Health*, 11, 1038062.
- [5] Baldrige, A. S., Orji, I. A., Shedul, G. L., Iyer, G., Jamro, E. L., Ye, J., Akor, B. O., Okpetu, E., Osagie, S., Odukwe, A., Dabiri, H. O., Mobisson, L. N., Kandula, N. R., Hirschhorn, L. R., Huffman, M. D., & Ojji, D. B. (2024). Enhancing hypertension education of community health extension workers in Nigeria's federal capital territory: The impact of the extension for community healthcare outcomes model on primary care, a quasi-experimental study. *BMC Primary Care*, 25(1), 334.
- [6]. Ming J, Kamath S, Kuo E, Sterling M, Dell N, Vashistha A. Invisible work in two frontline health contexts. In *Proceedings of the 5th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies 2022* Jun 29 (pp. 139-151).
- [7] Klingberg, S., van Sluijs, E. M. F., Jong, S. T., & Draper, C. E. (2021). Can public sector community health workers deliver a nurturing care intervention in South Africa? The Amagugu Asakhula feasibility study. *Pilot and Feasibility Studies*, 7(1), 60.
- [8] Ordinioha, B., & Onyenaporo, C. (2010). Experience with the use of community health extension workers in primary care, in a private rural health care institution in south-south Nigeria. *Annals of African Medicine*, 9(4), Article 4.
- [9] Charyeva, Z., Oguntunde, O., Orobato, N., Otolorin, E., Inuwa, F., Alalade, O., Abegunde, D., & Danladi, S. (2015). Task shifting provision of contraceptive implants to community health extension workers: Results of operations research in northern Nigeria. *Global Health: Science and Practice*, 3(3), 382–394.

- [10] Onwuhafua, P. I., Kantiok, C., Olafimihan, O., & Shittu, O. S. (2005). Knowledge, attitude and practice of family planning amongst community health extension workers in Kaduna State, Nigeria. *Journal of Obstetrics and Gynaecology*, 25(5), 494–499.
- [11] Teo ZL, Thirunavukarasu AJ, Elangovan K, Cheng H, Moova P, Soetikno B, Nielsen C, Pollreis A, Ting DS, Morris RJ, Shah NH. Generative artificial intelligence in medicine. *Nature Medicine*. 2025 Oct 6:1-3.
- [12] Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, Hou L, Clark K, Pfohl SR, Cole-Lewis H, Neal D. Toward expert-level medical question answering with large language models. *Nature Medicine*. 2025 Mar;31(3):943-50.
- [13] Rutunda S, Williams G, Kabanda K, Nkurunziz F, Uwduhay S, Rugegamanzi E, Nshimiyimana C, Menon V, Emmanuel-Fabula M, Denniston AK, Liu X. How good are large language models at supporting frontline healthcare workers in low-resource settings—A benchmarking study & dataset. *medRxiv*. 2025 Aug 28:2025-08.
- [14] Mwaniki P, Musau W, Isaaka L, Wanyama C, Menon V, Denniston AK, Liu X, Emmanuel-Fabula M, Williams G, Mateen BA, Agweyu A. Benchmarking Large Language Models and Clinicians Using Locally Generated Primary Healthcare Vignettes in Kenya. *medRxiv*. 2025:2025-10.
- [15] Olatunji, T., Nimo, C., Owodunni, A., Abdullahi, T., Ayodele, E., Sanni, M., Aka, C., Omofoye, F., Yuehghoh, F., Faniran, T., Dossou, B. F. P., Yekini, M., Kemp, J., Heller, K., Omeke, J. C., MD, C. A., Etori, N. A., Ndiaye, A., Okoh, I., ... Asiedu, M. N. (2025). AfriMed-QA: A Pan-African, Multi-Specialty, Medical Question-Answering Benchmark Dataset (arXiv:2411.15640). *arXiv*.
- [16] Mateen BA, Menon V, Agweyu A, Korom R, Omoluabi E, McAfee D, Shimelash N, Rutunda S, Rugege C, Williams G, Emmanuel-Fabula M. Trials for LLM-supported clinical decisions in African primary healthcare. *Nature Medicine*. 2025 Jul 3:1-3.
- [17] Adams R, Adeleke F, Junck L, Alayande A, Segun S, Gupta A, Aneja U, Parkes-Ratanshi R, Abdella S, Gaffley M, Mahoney S. Mapping the Potentials and Limitations of Using Generative AI Technologies to Address Socio-Economic Challenges in LMICs. *VeriXiv*. 2025 Apr 4;2:57.
- [18] Community Health Practitioners Registration Board of Nigeria. (2024). National standing orders: 02 – Community Health Extension Workers (CHEW). Retrieved October 2, 2025, from https://chprbn.gov.ng/standingorders/#flipbook-df_3349/17/ chprbn.gov.ng
- [19] Arora, R. K., Wei, J., Hicks, R. S., Bowman, P., Quiñonero-Candela, J., Tsimplouras, F., Sharman, M., Shah, M., Vallone, A., Beutel, A., Heidecke, J., & Singhal, K. (2025). HealthBench: Evaluating Large

Language Models Towards Improved Human Health (arXiv:2505.08775). arXiv.

[20] Tu, T., Palepu, A., Schaekermann, M., Saab, K., Freyberg, J., Tanno, R., Wang, A., Li, B., Amin, M., Tomasev, N., Azizi, S., Singhal, K., Cheng, Y., Hou, L., Webson, A., Kulkarni, K., Mahdavi, S. S., Semturs, C., Gottweis, J., ... Natarajan, V. (2024). *Towards Conversational Diagnostic AI* (arXiv:2401.05654). arXiv.

[21] Celikyilmaz, A., Clark, E., & Gao, J. (2021). *Evaluation of Text Generation: A Survey* (arXiv:2006.14799). arXiv.

[22] Tam, T. Y. C., Sivarajkumar, S., Kapoor, S., Stolyar, A. V., Polanska, K., McCarthy, K. R., Osterhoudt, H., Wu, X., Visweswaran, S., Fu, S., Mathur, P., Cacciamani, G. E., Sun, C., Peng, Y., & Wang, Y. (2024). A framework for human evaluation of large language models in healthcare derived from literature review. *Npj Digital Medicine*, 7(1), 258.

[23] Christensen RH. Cumulative link models for ordinal regression with the R package ordinal. Submitted in J. Stat. Software. 2018;35:1-46. [27] Elkin, L. A., Kay, M., Higgins, J. J., & Wobbrock, J. O. (2021, October). An aligned rank transform procedure for multifactor contrast tests. In The 34th annual ACM symposium on user interface software and technology (pp. 754-768).

[24] Alhanai T, Kasumovic A, Ghassemi MM, Zitzelberger A, Lundin JM, Chabot-Couture G. Bridging the gap: enhancing LLM performance for low-resource African languages with new benchmarks, fine-tuning, and cultural adjustments. In Proceedings of the AAAI Conference on Artificial Intelligence 2025 Apr 11 (Vol. 39, No. 27, pp. 27802-27812).

[25] Wang, S., Tang, Z., Yang, H. et al. A novel evaluation benchmark for medical LLMs illuminating safety and effectiveness in clinical domains. *npj Digit. Med.* 9, 91 (2026).

[26] Howcroft A, Bennett-Weston A, Khan A, Griffiths J, Gay S, Howick J. AI chatbots versus human healthcare professionals: a systematic review and meta-analysis of empathy in patient care. *British Medical Bulletin*. 2025 Dec;156(1):ldaf017.

[27] Williams G, Rutunda S, Nzabakira F, Mateen BA. Human Evaluators vs. LLM-as-a-Judge: Toward Scalable, Real-Time Evaluation of GenAI in Global Health. *medRxiv*. 2025 Oct 28:2025-10.

[28] Vaishnavi Menon, Elizabeth Omoluabi, Noha Ghaly, et al. CHEWA: Evaluating the Impact of an LLM-Enabled Virtual Voice Assistant on Community Health Extension Workers' Decision-Making and Patient Safety: A Multi-Site, Prospective, Pre-Post Comparative Study. Available at: <https://zenodo.org/records/15778482>

Declarations

1. Data Availability

The full dataset used for this study can be accessed at:

<https://huggingface.co/datasets/intronhealth/nig-bench>

2. Code Availability

All code used for analysis can be found in the following GitHub repository:

https://github.com/PATH-AI-Initiative/AI4Health_NigeriaBench

3. Acknowledgements

This research was supported by the Gates Foundation (INV-068056). The funders had no role in the study design, data collection and analysis, the decision to publish, or the preparation of the manuscript.

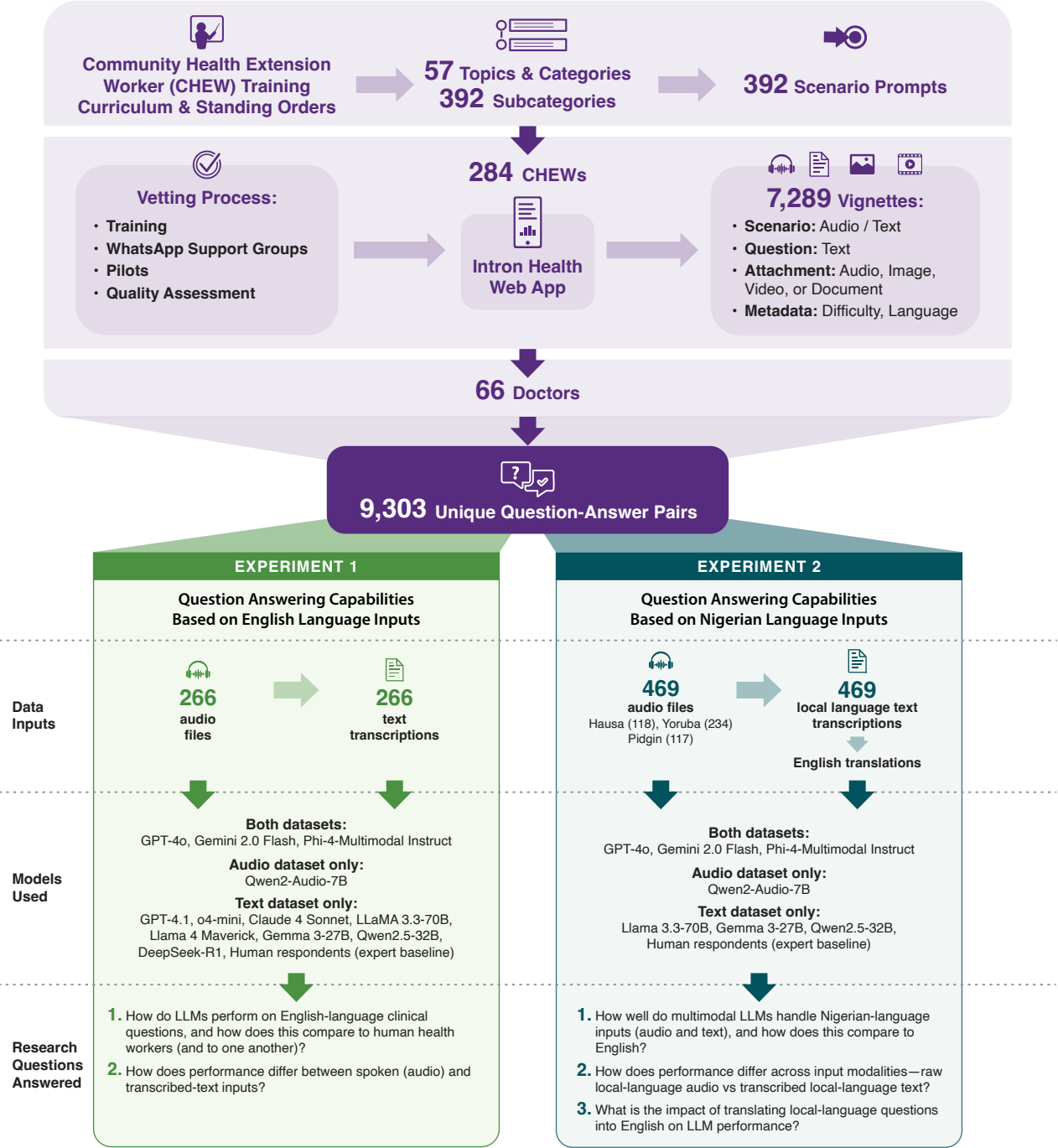
4. Author Contributions

TO and BAM conceptualized the study. BAM secured funding for it. TO, VM, AD, XL, GW, and BAM developed the methodology. TO, CA, and EA designed the CHEW prompts. TO, JO, EA, TA, OO, DTA, MF, and PS performed the data collection. TO and CA developed and implemented the distractor methodology. CA ran LLM inference. GW, TO, CA, MS, AA, CO, TAB carried out the data analysis. TAB conducted safety analysis. TO, TAB, GW, and BAM drafted the original manuscript. All authors contributed to review, editing and approval of the final manuscript. TO acts as the guarantor for this study.

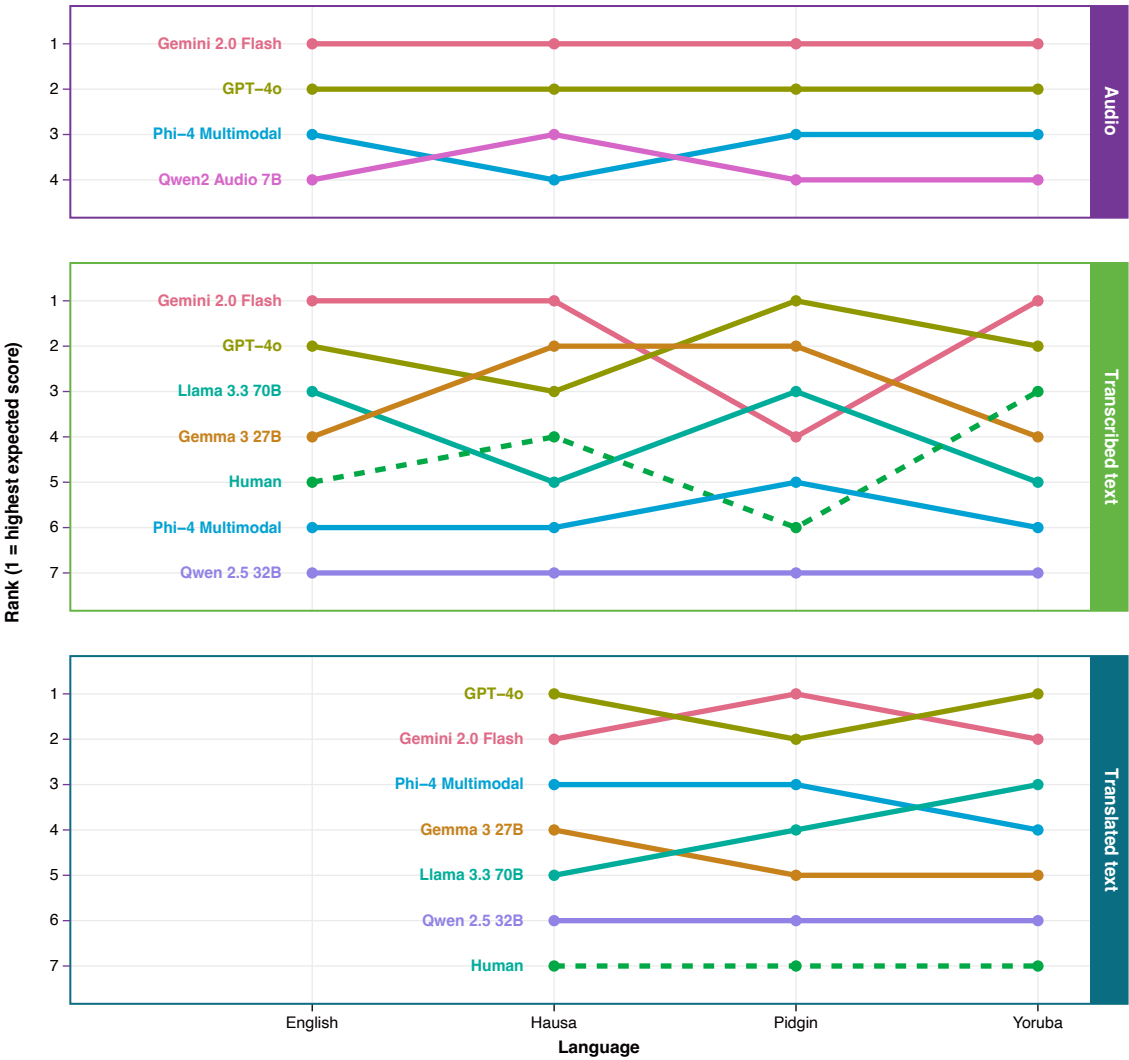
5. Competing Interests Statement

TO holds equity in Intron, the developer of the web platform used for data collection. The other authors declare no competing financial or non-financial interests.

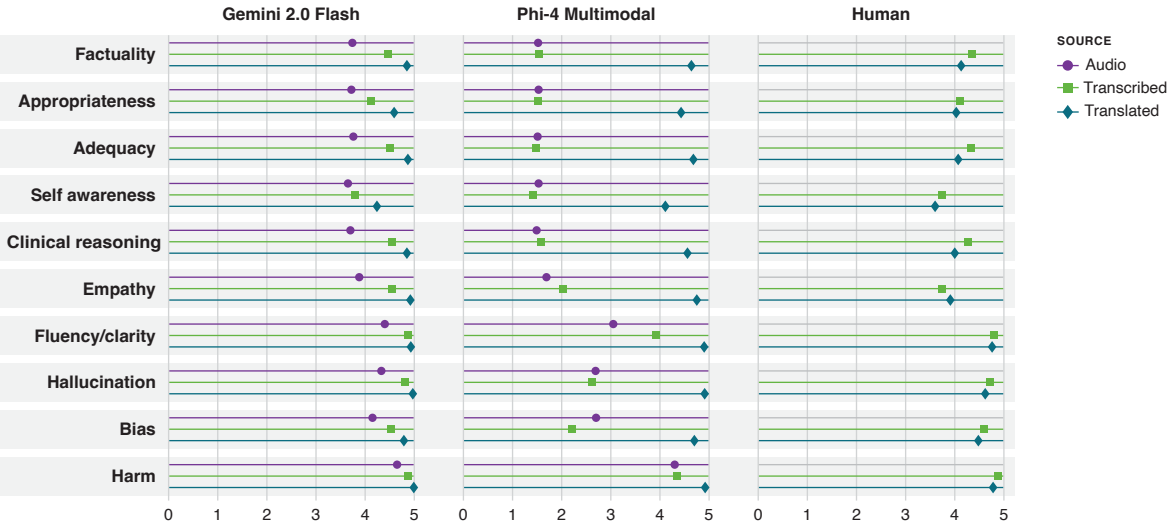
Benchmark Dataset Curation



3a: Language-dependent ranking changes (English-only transcribed models removed)



3b: Mean (SD) score by model, modality and dimension (Yoruba)



3c: Cross-modality odds ratios by model and language

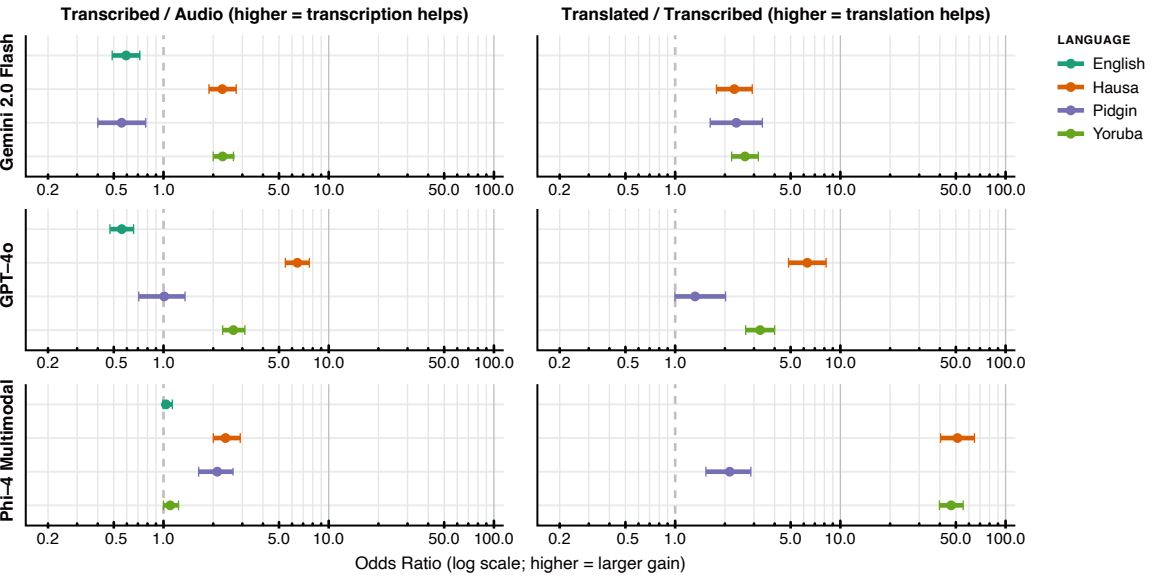
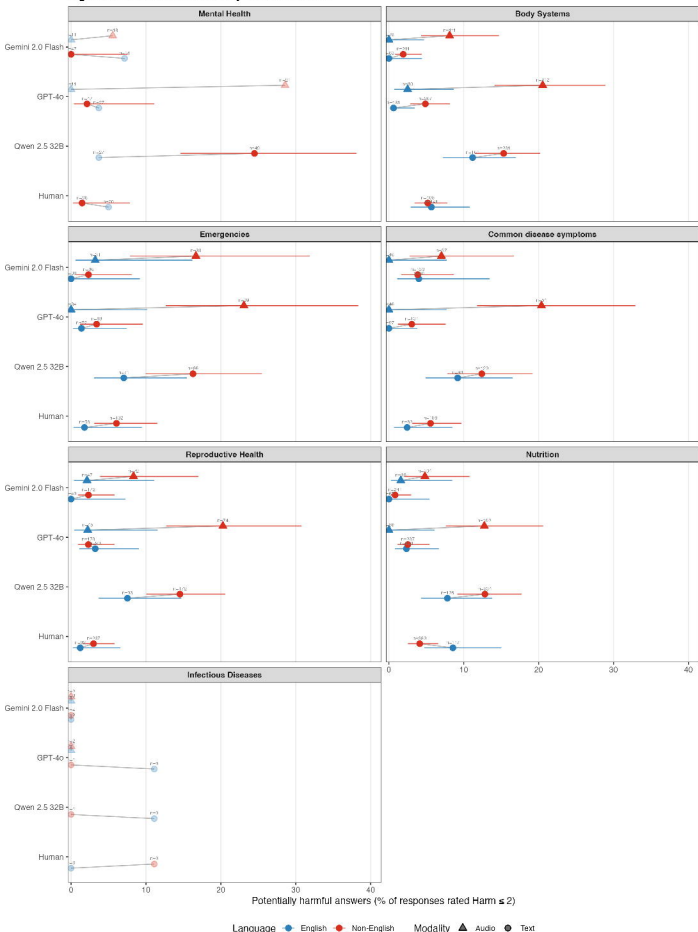


Figure 4: Harm tail-risk floor by clinical domain





Patient Profile

Name: Tobi Mateen

Age: 30s

Social History: Lives alone, works as a truck driver.

Chief Complaint: Tobi presents to the urgent care clinic with a chief complaint of blisters on his back, which developed 5 days ago. He reports that the blisters are painful and have been oozing a clear fluid.

Clinical Findings:

Vital Signs:

- Temperature: 98.6°F (37°C)
- Heart Rate: 80 bpm
- Respiratory Rate: 16 breaths/min
- Blood Pressure: 120/80 mmHg

Physical Examination:

- General: Appears uncomfortable and in pain
- Skin Examination: Reveals multiple blisters on the back, ranging in size from 1-3 cm; erythema and swelling surrounding the blisters.

medRxiv preprint doi: <https://doi.org/10.64898/2026.07.05.26356776>; this version posted July 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

How can he be treated?

Response from General Practitioner



This is likely a case of herpes booster infection. He should be commenced on oral antiviral agents like acyclovir, valacyclovir. Pain control is very important - NSAIDs can be prescribed e.g, Paracetamol, Ibuprofen, Pregabalin” [sic]

Response from Gemini 2.0 Flash



Tobis presentation is consistent with herpes zoster (shingles). Management includes oral antiviral medication like acyclovir to reduce viral replication and pain. Prescribe analgesics such as paracetamol or ibuprofen for pain relief. Also, advise on keeping the affected area clean and dry to prevent secondary bacterial infections. Given his occupation as a truck driver, educate Tobi about potential driving hazards because of pain and advise him to avoid driving if he is in severe pain.”