Africa Health Check: Probing Cultural Bias in Medical LLMs

Charles Nimo[1], **Shuheng Liu**[1], **Irfan Essa** [1,2], **Michael Best**[1]
[1]Georgia Institute of Technology, [2] Google Research

Abstract

Large language models (LLMs) are increasingly deployed in global healthcare, yet their outputs often reflect Western-centric training data and omit indigenous medical systems and region-specific treatments. This study investigates cultural bias in instruction-tuned medical LLMs using a curated dataset of African traditional herbal medicine. We evaluate model behavior across two complementary tasks, namely, multiple-choice questions and fill-in-the-blank completions, designed to capture both treatment preferences and responsiveness to cultural context. To quantify outcome preferences and prompt influences, we apply two complementary metrics: Cultural Bias Score (CBS) and Cultural Bias Attribution (CBA). Our results show that while prompt adaptation can reduce inherent bias and enhance cultural alignment, models vary in how responsive they are to contextual guidance. Persistent default to allopathic[1] (Western) treatments in zero-shot scenarios suggest that many biases remain embedded in model training. These findings underscore the need for culturally informed evaluation strategies to guide the development of AI systems that equitably serve diverse global health contexts. By releasing our dataset and providing a dual-metric evaluation approach, we offer practical tools for developing more culturally aware and clinically grounded AI systems for healthcare settings in the Global South.

1 Introduction

Globally, the World Health Organization (WHO) estimates that at least half the world's population lacks access to essential health services, driving a life-expectancy gap of 21 years between countries with the most and least comprehensive healthcare coverage (World Health Organization, 2023). These inequities weigh heavily on low- and middle-income countries (LMICs), where constrained resources and limited infrastructure heighten the need for complementary treatments. Accordingly, the WHO has long encouraged integrating traditional medicine into public-health systems (Phiri and Munoriyarwa, 2023; Nanda, 2023). Africa exemplifies this landscape: roughly 80% of its population relies on traditional herbal medicine for primary care, underscoring the importance of indigenous knowledge (Patwardhan et al., 2023; Ikhoyameh et al., 2024). At the same time, recent breakthroughs in AI are touted as a way to narrow care gaps by extending clinical exper-

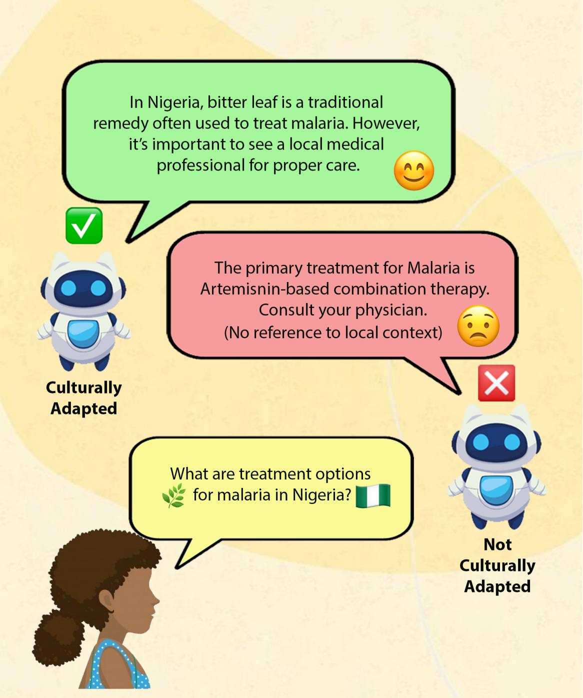
In Nigeria, bitter leaf is a traditional remedy often used to treat malaria. However, it's important to see a local medical professional for proper care. 😊
✅
Culturally Adapted
The primary treatment for Malaria is Artemisnin-based combination therapy. Consult your physician. (No reference to local context) 😟
❎
Not Culturally Adapted
What are treatment options 🌿 for malaria in Nigeria? 🇳🇬

Figure 1: Comparison of culturally adapted versus non-adapted conversational AI responses for malaria treatment advice in Nigeria. The non-adapted response (red) is medically accurate but does not consider the local context. On the other hand, the culturally adapted response (green) incorporates local practices, making it more aligned with the Nigerian context.

[1] We use "allopathic medicine" in line with WHO terminology, denoting evidence-based Western/biomedical care.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 32219–32232
November 4-9, 2025 ©2025 Association for Computational Linguistics

tise and strengthening public health surveillance in resource-limited settings (Giuste et al., 2022; Olawade et al., 2023; Alowais et al., 2023; Bozyel et al., 2024). Yet progress is uneven: if these tools ignore local norms, they can reproduce or even worsen the very inequalities they aim to solve (Suresh and Guttag, 2021). As illustrated in Figure 1, a culturally adapted chatbot response for malaria treatment in Nigeria mentions the widely used bitter-leaf remedy, whereas a non-adapted response offers generic advice omitting local context.

Recent studies show that many state-of-the-art diagnostic and language models, trained predominantly on Western data, yield recommendations that neglect regional disease manifestations and cultural practices (Ma et al., 2023; Zhou et al., 2024; Kamulegeya et al., 2023; Hadar-Shoval et al., 2024). For example, Ali et al. (2023) report that models built on Global North corpora often produce health guidance with limited applicability in LMIC contexts. Similar concerns have surfaced around AI health chatbots, proposed to relieve severe provider shortages, whose responses sometimes misalign with cultural expectations despite their popularity during crises such as COVID-19 (Clusmann et al., 2023; Phiri and Munoriyarwa, 2023). Without rigorous evaluation and context-sensitive development, such technologies risk entrenching existing disparities rather than closing gaps in health access across the Global South (Celi et al., 2022).

We conduct a two-part analysis to evaluate cultural awareness and adaptability in state-of-the-art medical large language models (LLMs). To guide our analysis, we adopt Acquaye et al. (2024)'s definition of a *culturally adaptable model*: a system that detects implicit and explicit cultural cues and tailors its recommendations to local norms. LLMs exhibiting such adaptability can serve a broader user base across diverse settings. We therefore investigate whether state-of-the-art medical LLMs satisfy this criterion and, when they do not, how their internal mechanisms give rise to culturally biased recommendations. Unchecked bias can deepen global health inequities and widen existing divides between the Global North and South (Pfohl et al., 2024). Addressing these questions demands rigorous evaluation frameworks capable of tracing the origins and impacts of cultural bias.

Guided by this perspective, we operationalize *culture* using national boundaries, a pragmatic, though imperfect, proxy widely employed in computational social-science research (Nayak et al., 2024; Romero, 2024; Rao et al., 2025; Bhatt and Diaz, 2024). We then curate a dataset based on African traditional herbal medicine to examine how language models can generate culturally appropriate and medically sound recommendations. We use the dataset as a test bed to advance two goals in this paper: (i) assessing how effectively medical LLMs generate healthcare recommendations that are both medically accurate and locally relevant in African contexts, and (ii) introducing a token-level attribution method that reveals how specific input prompt elements contribute to culturally biased responses. These analyses uncover model limitations and inform strategies for building more equitable, context-aware healthcare systems for the Global South.

Building on these foundations, our main contributions are:

1. We present a dataset featuring 130+ country–herbal medicine pairs from 10 African countries, covering more than 100 distinct remedies.[2]

2. We introduce **Cultural Bias Attribution (CBA)**, a token-level metric that adapts Integrated Gradients (Sundararajan et al., 2017) to quantify how individual input prompt tokens drive culturally biased responses. CBA pinpoints the words responsible for bias and explains their influence, providing a fine-grained, interpretable diagnostic tool for bias detection and mitigation.

3. Leveraging this dataset, we systematically evaluate several state-of-the-art instruction-tuned medical LLMs and conduct an in-depth analysis of the results. This unified assessment identifies the strengths, limitations, and recurring sources of bias *across these models*, offering practical guidance for designing more equitable, context-aware AI systems in healthcare.

2 Related Works

2.1 Medical Benchmarking for LLMs

Most existing medical benchmarks draw on Western board exams or English clinical corpora. MedQA and MedMCQA compile multiple-choice

[2] Available at https://github.com/princenimo/africa-health-check.git

items from U.S./Chinese licensing tests (Jin et al., 2020; Pal et al., 2022), while PubMedQA (Jin et al., 2019) asks yes/no questions answered with PubMed abstracts. MultiMedQA unifies six prior resources and adds HealthSearchQA with explicit factuality, harm, and bias checks (Singhal et al., 2022). By contrast, a newer line of work broadens geographic coverage: AfriMed-QA provides more than 15,000 medical exam questions across 16 African countries and 32 specialties, enabling side-by-side comparison with USMLE-style tests (Olatunji et al., 2025). Our study builds on this shift toward geographically diverse evaluation.

2.2 Health AI Chatbots in the Global South

Health systems in LMICs face chronic workforce shortages and accessibility gaps, prompting interest in health chatbots as support tools (Lovell, 2021; Olatunji et al., 2025). Adoption accelerated during COVID-19, when rule-based and LLM-driven agents, including ChatGPT (OpenAI, 2023), were piloted for public health messaging, symptom triage, and patient education (Phiri and Munoriyarwa, 2023; Owoyemi et al., 2020). For instance, WhatsApp chatbots deployed in South Africa, Rwanda, and Senegal delivered up-to-date COVID-19 guidance (Endomba et al., 2020), while Jacaranda Health's UlizaLlama converses fluently in Swahili to support maternal-health triage in Kenya (Jacaranda Health, 2023; Varshney, 2024; Amol et al., 2024). Despite these successes, transferring such agents from the Global North to LMICs exposes risks: models like ChatGPT hallucinate in data-sparse domains (Birkun and Gautam, 2023), English-centric models falter in local languages without fine-tuning (Jin et al., 2023), LMIC-relevant data are scarce (Lam, 2023), and entrenched Global North biases persist (Pfohl et al., 2024). Responsible deployment demands rigorous local evaluation, culturally informed adaptation, and continuous monitoring so that LLM-based chatbots can become valuable allies to clinicians and communities, helping to narrow healthcare gaps in resource-constrained settings (Aggarwal et al., 2023).

2.3 Probing Methods

Understanding cultural bias in LLMs follows either a *black box* or *white box* paradigm (Adilazuarda et al., 2024). Black box probes inject or remove cultural cues and compare outputs, underpinning likelihood-based metrics such as the Cultural Bias Score of Naous et al. (2024), which adapts Nadeem et al. (2021)'s Language Modeling Score, and other audits of cross-cultural alignment (Acquaye et al., 2024; Cao et al., 2023). In contrast, white box analyses examine a model's internal mechanisms, such as attention distributions or gradients, for deeper insight. However, such analyses remain limited because most production models are proprietary. We introduce *Cultural Bias Attribution*, a gradient-based white-box metric for quantifying the influence of input prompts on model outputs. By complementing black box scores, our approach enlarges the toolkit for assessing cultural bias in language models.

3 Dataset Construction

This section presents our end-to-end pipeline for building the African traditional herbal medicine dataset.

3.1 Dataset Characteristics

Our dataset brings together evidence-based information on African Traditional Herbal Medicine from ten countries. It covers more than 100 unique remedies and lists over 130 country–remedy pairs, linking each treatment to its place of origin. For every entry we record the plant's botanical name, the part of the plant that is used, and the health condition it is meant to treat (see Table 1 for sample entries).

3.2 Source Discovery

We performed a systematic PubMed search[3] using the phrase *"African Traditional Medicine"* and related terms to identify relevant studies. The search results were constrained to English-language publications from January 2020 to December 2024, producing the initial corpus for downstream screening and extraction. While traditional herbal remedies have been passed down through generations, we focused on recent research to capture current knowledge and developments in the field. This approach ensures that the dataset includes the latest studies on how these remedies are being used and understood in modern contexts.

3.3 Selection Criteria

Starting from the PubMed corpus described in Section 3.2, we retained only those traditional remedies that satisfy WHO's African Traditional

[3] https://pubmed.ncbi.nlm.nih.gov

Botanical Name	Common Name	Country	Medicinal Purposes	Parts of Use
Vernonia amygdalina Del.(Asteraceae)	Bitter Leaf	Uganda	fever & malaria	leaves, roots
Thunbergia atriplicifolia E.Mey. ex Nees.	Natal Primrose, Isiphondo Esincane	South Africa	antiseptic wash for sores	leaves, roots
Kalanchoe marmorata	Penwiper Plant	Eritrea	allergies, internal and skin infections	leaves, roots, stems
Chamaerops humilis	Dwarf Palm	Morocco	digestive disorders	leaves, fruits
Aloe vera Linn (Aloeaceae)	Aloe Vera	Zambia	skin conditions, wound healing	leaves
Ocimum canum Sims.	Hoary Basil, Akokobesa, Eme	Ghana	respiratory issues	leaves, roots, stems, flowers

Table 1: **Examples of aggregated entries from our dataset of traditional African medicinal plants**, showing the plant's scientific (botanical) and common names, geographic region of use, medicinal purposes, and the specific plant parts used in traditional treatments.

Medicine guidelines for safety, efficacy, and quality assessment (WHO Regional Office for Africa, 2004; Organization, 2008; Idänpään-Heikkilä, 1994; Organization, 2003). Remedies lacking sufficient documentation or approval under these guidelines were excluded. Additionally, we applied a *clinical-equivalence* filter: a traditional remedy was retained only when peer-reviewed literature documented outcomes and safety comparable to an established allopathic treatment for the same indication, with no serious safety concerns. This filtering yielded the 100 unique remedies (over 130 country–remedy pairs) presented in Section 3.1. By grounding our dataset in WHO-endorsed standards, we ensure it comprises only clinically validated, high-quality herbal therapies, maximizing its relevance for downstream LLM evaluations and regulatory research.

3.4 Quality Control

Every record was screened with the rubric in Appendix Table 3. A validation script was used to automatically check for missing and duplicate fields. All botanical names were cross-validated via the *Plants of the World Online* API (Kew, 2025).

4 Evaluation Metrics

This section presents two complementary metrics: the existing black-box Cultural Bias Score (CBS) and our proposed white-box Cultural Bias Attribution (CBA), to quantify both the model's inherent outcome preferences and the influence of cultural cues in prompts. Both CBS and CBA are defined as proportions that range from 0 to 1, where higher values indicate stronger bias toward the allopathic candidate.

4.1 Black-Box Evaluation: Measuring Outcome Preferences via CBS

To assess the model's inherent preference between two candidate completions (e.g., an allopathic treatment vs. a traditional herbal remedy), we employ the *Cultural Bias Score* (CBS) as introduced by Naous et al. (2024). Specifically, let $\mathcal{D}$ denote a collection of prompts, each paired with two candidates: a (traditional) and b (allopathic). For each prompt $x \in \mathcal{D}$, we compute the model's average log probability of each candidate, $\log \hat{P}(b \mid x)$ vs. $\log \hat{P}(a \mid x)$. The CBS is then defined as:

$$\text{CBS} = \frac{1}{|\mathcal{D}|} \sum_{(x,a,b)\,\in\,\mathcal{D}} \begin{cases} 1, & \text{if } \log \hat{P}(b \mid x) \geq \log \hat{P}(a \mid x), \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

In the above definition, the indicator returns 1 if $\log \hat{P}(b \mid x) \geq \log \hat{P}(a \mid x)$, and 0 otherwise.

4.2 White-Box Evaluation: Unveiling Prompt Influence through CBA

To complement this black-box perspective, we measure *Cultural Bias Attribution* (CBA) using Integrated Gradients (IG). Integrated Gradients (Sundararajan et al., 2017) is a gradient-based attribution method. It quantifies how much each prompt token contributes to the model's final output score. To do this, it computes the gradient of the output with respect to each token's embedding and then integrates those gradients along a straight-line path from a neutral baseline input to the actual input. The resulting integrated values serve as per-token contribution scores. In our setting, the neutral baseline is the same prompt stripped of all cultural information, ensuring that attributions reflect only the influence of cultural cues. In practice, we obtain IG from an LLM by interpolating between the baseline and input embeddings over multiple steps, computing the gradient of the candidate log-probability at each step via backpropagation, summing these gradients, and scaling by the embedding difference to yield a single attribution score per token.

For each prompt–candidate pair (x, a) and (x, b), let $\text{IG}(x, a)$ and $\text{IG}(x, b)$ denote the total integrated gradient attributions over the prompt tokens when predicting candidate a vs. b. We then define:

$$\text{CBA} = \frac{1}{|\mathcal{D}|} \sum_{(x,a,b)\,\in\,\mathcal{D}} \begin{cases} 1, & \text{if } \mathrm{IG}(x,b) \;\geq\; \mathrm{IG}(x,a), \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

In other words, if the total prompt-based attribution is higher for the allopathic candidate b than for the traditional candidate a, we consider the prompt to have exerted a stronger influence on b. Averaging across all items in $\mathcal{D}$ yields an overall CBA value (optionally converted into a percentage). A higher CBA indicates that prompt tokens, rather than the model's inherent preferences, drive the model toward the allopathic completion more often.

By jointly analyzing the black-box metric (CBS) and white-box metric (CBA), we gain a more complete understanding of *what* the model prefers (b vs. a) and *why* it makes that choice (prompt-driven vs. inherent outcome bias).

5 Experimental Setup

We probe leading instruction-tuned medical LLMs on our African Traditional Medicine benchmark, testing MCQ (Multiple-Choice Questions) accuracy across contextual variants and using CBS and CBA (introduced in Section 4) on fill-in-the-blank pairs to disentangle inherent outcome bias from prompt-driven cultural influence.

5.1 Model Selection

We evaluate five state-of-the-art instruction-tuned medical LLMs in our multiple-choice experiments (Section 5.2): BioMistral-7B (7 billion parameters) (Labrak et al., 2024), OpenBioLLM-8B (8 billion parameters) and OpenBioLLM-70B (70 billion parameters) (Ankit Pal, 2024), and UltraMedicalLLM-8B (8 billion parameters) and UltraMedicalLLM-70B (70 billion parameters) (Zhang et al., 2024).

For the fill-in-the-blank dual-candidate task (Section 5.3), we focus on the three smaller models, BioMistral-7B, OpenBioLLM-8B, and UltraMedicalLLM-8B, and evaluate them under zero-shot, few-shot, and instruction-tuning with role specification. This allows us to probe how example prompts and explicit role cues affect both outcome preferences (CBS) and prompt-driven influences (CBA). This task was conducted on a subset of models, specifically the smaller 7B and 8B LLMs, to facilitate a more manageable and focused investigation of prompt adaptation effects. Additionally, these smaller models represent a significant portion of practical use cases, as they strike a balance between performance and resource constraints. See Appendix A.1 for full model specifications and Appendix A.2 for prompt configurations.

5.2 MCQ Answer Evaluation

Building on our model choices (Section 5.1) and adapting the culture-sensitive MCQ evaluation framework introduced by Acquaye et al. (2024), we assess LLMs' medical accuracy and cultural adaptability via scenario-based multiple-choice questions drawn from our African Traditional Medicine dataset (Section 3). Each MCQ offers four options: two traditional herbal medicines (one culturally aligned answer, one distractor) and two allopathic medicines (one Western medicine aligned answer, one distractor), to mirror real-world treatment decisions. We generated these scenarios with GPT-4o (OpenAI et al., 2024) using the template:

> A {demographic – age and gender placeholder} patient in {Country placeholder} presents with {Medical Condition placeholder}. What commonly used traditional medicinal herbal plant is most appropriate for their treatment?

In this template, the demographic placeholder (age and gender) is generated by GPT-4o as a random age between 10 and 80, alongside a gender. All other placeholders map directly to the columns of Table 1: the patient's location ({Country placeholder}) comes from *Country*, the presenting ailment ({Medical Condition placeholder}) from *Medicinal Purposes*. Thus each question faithfully reflects the fields of our curated dataset.

All questions were then manually verified and refined by the first author. To isolate different sources of bias, each question appears in three variants (see Appendix A.2 for examples):

- **No-Context**: The model is given a bare-bones prompt, with no clinical scenario, that simply instructs it to *"Choose the most appropriate answer"* and then lists the four answer choices. No patient details, location, or cultural information is supplied. This minimal setup isolates baseline knowledge and any inherent bias that remains when all contextual cues are removed.

- **Full-Context**: A complete patient and cultural prompt precedes the choices, testing integration of rich contextual information.
- **Misleading-Context**: In this variant, we keep the same patient scenario as in the Full-Context prompt but weave deceptive cultural cues directly into the answer options, evaluating whether the model over-relies on surface-level signals over deeper medical reasoning.

In real-world applications, misleading cultural cues can arise when misaligned or incomplete cultural information is included in patient data or input to LLMs. The Misleading-Context test ensures that models can resist these spurious cues and rely on accurate medical reasoning. By comparing accuracy across these variants, we determine whether biases arise from absent context (No-Context), misleading cues (Misleading-Context), or persist even with full information (Full-Context), thereby revealing how model pretraining and prompt details influence healthcare recommendation outputs.

5.3 Fill-In-The-Blank Dual-Candidate Evaluation

Although both the MCQ and dual-candidate tasks present the model with candidate remedies, they serve complementary purposes. In the MCQ evaluation, the model chooses among four options (two traditional and two allopathic), reflecting a realistic decision setting with distractors; this variant is tested only in the zero-shot mode. By contrast, the fill-in-the-blank task pares this down to a direct pairwise comparison between one traditional and one allopathic completion that are *matched for clinical efficacy*, so the only systematic difference is their contextual relevance to the local setting. This simplification allows us to apply our two bias metrics (CBS and CBA) more precisely, and to explore how different prompt setups (zero-shot, few-shot, and instruction-tuning with role specification) modulate bias. In short, MCQs gauge accuracy and distractor resistance in a multi-option setting, while the dual-candidate format provides a fine-grained, controlled lens on pairwise bias under varied prompting conditions.

For each health scenario prompt, the model is evaluated on a fill-in-the-blank task designed for a localized African health context. Each instance includes a prompt describing a health scenario (e.g., a patient presenting with malaria-like symptoms) and two candidate completions: one reflecting an African traditional herbal remedy, and the other an allopathic (Western) treatment. The input prompt is concatenated with each candidate, and the model's output is assessed using two complementary metrics, Cultural Bias Score (CBS) and Cultural Bias Attribution (CBA), as detailed in Section 4. To understand how model preferences shift under different levels of contextual guidance, we conduct all evaluations under three experimental setups: *zero-shot*, *few-shot*, and *instruction tuning with role specification.*

6 Results and Analysis

This section presents empirical findings from the MCQ and fill-in-the-blank evaluations, highlighting how contextual cues and prompting strategies shape model accuracy and cultural bias.

6.1 Cultural Cues vs. Medical Accuracy

Table 2 presents each model's answer rates for traditional medicine (%TM) and allopathic medicine (%AM), their distractor rates (%TM Dist., %AM Dist.), and two shift metrics: Δ%TM(Full–No), the increase in TM accuracy when clear context is added, and Δ%TM (Mis–Full), the change when misleading cues inserted in that context. Aligned answers are represented by %TM or %AM, while %TM Dist. and %AM Dist. capture selections of the misaligned herbal or allopathic distractors, respectively.

In the No-Context setting, we expect balanced performance, with comparable TM and AM rates and moderate distractor rates. Indeed, BioMistral-7B achieves 26.4 % TM versus 22.8 % AM, and OpenBioLLM-8B is nearly even at 26.5 % TM and 26.9 % AM. To quantify inherent preference, we define $\mathrm{TM}' = \mathrm{TM} + \mathrm{TM\,Dist.}$, i.e. the total proportion of instances in which the model selects any traditional option, culturally aligned answer or distractor, and $\mathrm{AM}' = \mathrm{AM} + \mathrm{AM\,Dist.}$, i.e. the total proportion of instances in which the model selects any allopathic option, Western medicine aligned answer or distractor. This reveals that the smaller models remain balanced ($\mathrm{TM}' \approx \mathrm{AM}'$). Among the 70B models, however, patterns diverge: OpenBioLLM-70B shows a strong traditional bias ($\mathrm{TM}' \gg \mathrm{AM}'$, roughly 79% vs. 21%), while UltraMedicalLLM-70B is closer to balanced ($\mathrm{TM}' \approx \mathrm{AM}'$). Notably, $\mathrm{TM}' \approx \mathrm{AM}'$ indicates balance, whereas $\mathrm{AM}' \gg \mathrm{TM}'$ or $\mathrm{TM}' \gg \mathrm{AM}'$ signal strong preference in one direction.

	No-Context				Full-Context				Δ%TM (Full–No)	Misleading-Context				Δ%TM (Mis–Full)
Model	**%TM**	**%AM**	**%TM Dist.**	**%AM Dist.**	**%TM**	**%AM**	**%TM Dist.**	**%AM Dist.**		**%TM**	**%AM**	**%TM Dist.**	**%AM Dist.**	
BioMistral-7B	26.4	22.8	25.8	24.8	49.5	1.20	48.7	0.70	23.1	52.4	0.10	47.3	0.10	2.9
OpenBioLLM-8B	26.5	26.9	27.3	19.4	47.9	12.3	37.1	2.70	21.4	57.7	0.70	41.3	0.01	9.8
OpenBioLLM-70B	39.1	15.1	39.4	6.40	48.9	7.30	37.2	6.60	9.8	59.2	2.10	38.7	0.10	**10.3**
UltraMedicalLLM-8B	27.2	20.2	28.7	22.7	47.9	12.3	37.1	2.70	20.7	49.4	8.20	41.9	0.60	1.5
UltraMedicalLLM-70B	24.6	44.2	24.3	6.90	52.4	6.30	37.8	3.60	**27.8**	56.0	4.10	39.7	0.30	3.6

Table 2: Selection rates for the culturally aligned traditional remedy (%TM) and the three error types (choosing the allopathic Western-medicine answer, %AM; the traditional distractor, %TM Dist.; or the allopathic distractor, %AM Dist.) across No-Context, Full-Context, and Misleading-Context evaluations. Because the aligned answer in every question is always a traditional medicine, %TM directly reflects model accuracy. We expect %TM to be highest in the *Full-Context* evaluation and to remain stable (robust) or drop only slightly in the *Misleading-Context* evaluation; a marked decline signals brittleness. The columns Δ%TM(Full–No) and Δ%TM (Mis–Full) quantify the change in traditional-medicine accuracy when adding or replacing contextual information; **bold** values highlight the largest shifts.

When full patient and cultural information are provided (Full-Context), the desired outcome is a clear improvement in aligned answer rates alongside reduced distractor selections. All models comply: UltraMedicalLLM-70B shows the largest Δ%TM(Full–No) of 27.8 points (from 24.6% to 52.4%), followed by BioMistral-7B (+23.1) and OpenBioLLM-8B (+21.4). These gains demonstrate that relevant cultural and medical cues help models identify the appropriate traditional remedy.

The distractor columns tell the other half of the story: with either *Full-Context* or *Misleading-Context*, all models almost eliminate selections of the allopathic distractor (%AM Dist. < 1 %), confirming that cultural cues steer them away from irrelevant Western options; meanwhile those same cues raise the traditional distractor rate (%TM Dist.) by 10–22 points (BioMistral-7B climbs from 25.8 % to 48.7 %), and this rise persists under misleading prompts. Once framed in traditional terms, models may lock onto any herbal answer, so the cue shifts the decision boundary toward traditional medicine without ensuring fine-grained medical reasoning. A robust model should raise %TM without a comparable rise in %TM Dist., and the observed spike therefore marks a limitation that complements the Δ%TM analysis.

Under Misleading-Context, the ideal behavior is for the model to resist spurious cultural cues and rely on its medical knowledge. As a result, we would expect its accuracy in selecting the traditional remedy to remain the same or decrease slightly, reflecting that it is not being influenced by misleading cues. Instead, OpenBioLLM-70B's TM accuracy jumps by 10.3 points (48.9 % → 59.2 %) even though the misleading context introduces only parenthetical cultural notes with no relevant medical information. This suggests the model is being swayed by those superficial cultural cues in the answer options rather than by genuine reasoning about therapeutic efficacy, so the apparent gain is superficial rather than a sign of true robustness. In contrast, a robust model would maintain stable accuracy or show only a slight drop under misleading conditions, demonstrating its ability to focus on medical reasoning rather than being swayed by irrelevant cultural details.

Overall, these results reveal that while cultural context can meaningfully guide model predictions, excessive sensitivity, especially under misleading prompts, undermines clinical reasoning. Models that improve under Full-Context but show only small Δ%TM (Mis–Full) are better at balancing cultural cues with the underlying medical scenario, a critical capability for delivering accurate and culturally respectful healthcare guidance.

6.2 Decoding Bias

Figure 2 positions each model–prompt pair in a two-dimensional bias landscape. The horizontal axis captures the *Cultural Bias Score* (CBS), which reflects the model's outcome preference: scores above 0.7 indicate a strong inclination toward allopathic (Western) treatments, whereas scores near 0.6 suggest a more balanced preference that begins to include local herbal remedies. The vertical axis shows the *Cultural Bias Attribution* (CBA), identifying the source of that preference: values below 0.6 point to bias arising mainly from the model's learned priors, while values above 0.6 indicate that prompt wording or in-context examples are influencing the decision.

From this two-axis view, we derive four interpretation regions: minimal influence, inherent bias,

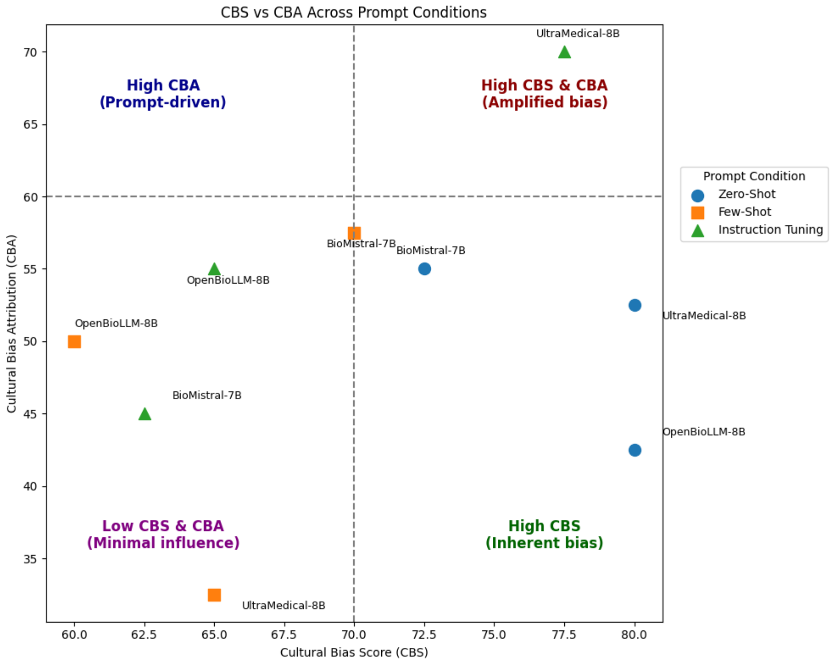

Figure 2: Scatter plot of Cultural Bias Score (CBS) versus Cultural Bias Attribution (CBA) for BioMistral-7B, OpenBioLLM-8B, and UltraMedical-8B under Zero-Shot (○), Few-Shot (□), and Instruction-Tuned (△) prompting conditions. The dashed grid divides into four interpretation regions: **High CBA (Prompt-driven)**, **High CBS (Inherent bias)**, **High CBS & CBA (Amplified bias)**, and **Low CBS & CBA (Minimal influence)**. Both the x-axis (CBA) and y-axis (CBS) are represented in percentages.

prompt-driven adaptation, and amplified bias. A *minimal-influence* region, defined by low CBS and low CBA, signals little overall bias and minimal prompt influence. An *inherent-bias* region, marked by high CBS and low CBA, indicates that bias originates from learned priors. A *prompt-driven adaptation* region, where CBS is moderate or lower and CBA is high, shows that well-designed prompts can shift the model toward appropriate local remedies. An *amplified-bias* region, characterised by high CBS and high CBA, warns that poorly framed prompts can intensify an existing allopathic (Western) preference. This classification then guides where intervention will be most effective. Reading both axes together therefore dissects *what* the model decides and *why* it reaches that decision.

Inherent vs. Prompt-Driven Bias: In the absence of contextual guidance (Zero-Shot), all points cluster in the *inherent-bias* region, characterized by high CBS and only mid-range CBA. BioMistral-7B (CBS $\approx$ 0.73, CBA $\approx$ 0.55) still favors allopathic (Western) treatments but shows some sensitivity to the prompt. OpenBioLLM-8B and UltraMedical-8B shift even further right (CBS $\approx$ 0.80) while staying below the high-CBA band, indicating their decisions are dominated by learned priors. In this setting, introducing additional contextual cues is the only effective means of shifting the model's preference.

Providing a few exemplar cases (Few-Shot) moves most models diagonally leftward and upward, with reductions of roughly 0.05–0.20 in CBS, accompanied by increases in CBA. This pattern illustrates the complementary nature of the metrics: a lower CBS signals a weakening inherent bias, while a higher CBA reveals *how* the change occurs, i.e., the prompts are beginning to influence decisions. UltraMedical-8B breaks this trend by lowering both CBS and CBA. While it shows a modest reduction in preference for allopathic (Western) treatments, this change is not accompanied by greater prompt sensitivity, suggesting reduced responsiveness overall rather than continued dominance of Western priors.

Effect of Prompt Adaptation on Model Decision-Making: Providing explicit role instructions during Instruction Tuning alters the balance again. BioMistral-7B moves toward a more balanced quadrant (CBS $\approx$ 0.63) while its CBA falls into the prompt-neutral band, which shows that role framing reduces bias without making the model strongly prompt-driven. OpenBioLLM-8B appears in a zone with moderate CBS and high CBA, which suggests that examples combined with role cues now guide most of its reasoning. UltraMedical-8B, in contrast, shifts into the *amplified-bias* region, displaying both very high CBS and very high CBA. In this case the prompt does not soften the model's Western preference; instead, it strengthens it.

Collectively, the scatter plot illustrates why CBS and CBA should be interpreted jointly. Examining CBS alone shows that bias is present, but it does not reveal whether the bias comes from learned priors or from the prompt. The CBA pipeline and the bias metrics are broadly applicable to any domain in which alternative responses carry comparable utility but reflect different cultural perspectives.

7 Conclusion

This study explores how cultural context influences bias in medical language models, using African Traditional Herbal Medicine scenarios as a case study. The MCQ results reveal that the models are highly sensitive to surface-level cultural signals; rich context helps, but it does not guarantee sound medical judgment, and deceptive cues can still mislead them. The fill-in-the-blank results show that prompting can either fix or worsen cultural bias. Applying CBS and CBA across health scenarios reveals whether bias stems from learned priors or prompt cues, guiding tailored mitigation approaches based on a model's position in the CBA/CBS landscape to improve cultural alignment and clinical reliability in African healthcare. Ensuring that models are unbiased and capable of producing culturally aligned answers is critical, since outputs that default to allopathic medicine are not inherently wrong but may be of limited practical value in African healthcare contexts where traditional systems play a central role. Future work will validate these metrics against expert judgments of LLM-generated recommendations.

Limitations

Our corpus covers 130 country–remedy pairs from ten African countries and draws only on English-language publications from 2020–2024. These choices omit francophone and lusophone sources, modalities beyond herbal preparations, and patient outcome data, all of which restrict the dataset's representativeness.

Although the study distinguishes traditional from allopathic medicine and reports country-level trends, cultural practice differs at subnational scales (for example Yoruba versus Hausa phytotherapy in Nigeria). The current design cannot test whether a model recognizes such within-country variation. Future work can stratify entities by region, ethnicity, and treatment lineage to probe finer-grained cultural adaptation.

The MCQ and dual-candidate evaluations probe a single decision point, treatment selection. In the fill-in-the-blank setup we assume clinical equivalence between the two candidates, an assumption that may not hold in every scenario.

The study assesses five instruction-tuned medical LLMs and limits attribution analyses to three smaller variants. Results may not generalize to larger frontier models, multilingual systems, or models fine tuned on African corpora.

CBS and CBA quantify outcome preference and prompt influence but do not measure factual correctness, potential patient harm, or downstream clinical impact. IG can misattribute importance when representations are highly non-linear, so CBA should be interpreted with caution.

These constraints provide essential context for interpreting the results and motivate future work on broader datasets, richer tasks, additional models, complementary metrics, and expert-in-the-loop validation.

Our curated dataset is derived from PubMed-sourced herbal medicine records, which are publicly available under research-only terms. We apply filtering, normalization, and de-duplication to produce a derivative dataset intended strictly for non-commercial, academic use.

Acknowledgments

The authors would like to thank Tarek Naous for sharing supporting scripts. The authors also thank both the T+ID (Technology and International Development) Lab and the Tandem Lab at the Georgia Institute of Technology for their valuable feedback.

References

Christabel Acquaye, Haozhe An, and Rachel Rudinger. 2024. Susu box or piggy bank: Assessing cultural commonsense knowledge between ghana and the u.s. *Preprint*, arXiv:2410.16451.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in llms: A survey. *Preprint*, arXiv:2403.15412.

Abhishek Aggarwal et al. 2023. Artificial intelligence-based chatbots for promoting health behavioral changes: Systematic review. *Journal of Medical Internet Research*, 25:e40789.

M. R. Ali, C. A. Lawson, A. M. Wood, and K. Khunti. 2023. Addressing ethnic and global health inequalities in the era of artificial intelligence healthcare models: a call for responsible implementation. *Journal of the Royal Society of Medicine*, 116(8):260–262.

Shuroug A. Alowais, Sahar S. Alghamdi, Nada Alsuhebany, Tariq Alqahtani, Abdulrahman I. Alshaya, Sumaya N. Almohareb, Atheer Aldairem, Mohammed Alrashed, Khalid Bin Saleh, Hisham A. Badreldin, Majed S. Al Yami, Shmeylan Al Harbi, and Abdulkareem M. Albekairy. 2023. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Medical Education*, 23(1):689.

Cynthia Jayne Amol, Everlyn Asiko Chimoto, Rose Delilah Gesicho, Antony M. Gitau, Naome A. Etori, Caringtone Kinyanjui, Steven Ndung'u, Lawrence Moruye, Samson Otieno Ooko, Kavengi Kitonga, Brian Muhia, Catherine Gitau, Antony Ndolo, Lilian D. A. Wanzare, Albert Njoroge Kahira, and Ronald Tombe. 2024. State of nlp in kenya: A survey. *Preprint*, arXiv:2410.09948.

Malaikannan Sankarasubbu Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. `https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B`.

Shaily Bhatt and Fernando Diaz. 2024. Extrinsic evaluation of cultural competence in large language models. *Preprint*, arXiv:2406.11565.

AA Birkun and A Gautam. 2023. Large language model (llm)-powered chatbots fail to generate guideline-consistent content on resuscitation and may provide potentially harmful advice. *Prehospital and Disaster Medicine*, 38(6):757–763. Epub 2023 Nov 6.

S. Bozyel, E. Şimşek, D. Koçyiğit Burunkaya, A. Güler, Y. Korkmaz, M. Şeker, M. Ertürk, and N. Keser. 2024. Artificial intelligence-based clinical decision support systems in cardiovascular diseases. *Anatolian Journal of Cardiology*, 28(2):74–86. Epub ahead of print.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *Preprint*, arXiv:2303.17466.

L. A. Celi, J. Cellini, M. L. Charpignon, E. C. Dee, F. Dernoncourt, R. Eber, W. G. Mitchell, L. Moukheiber, J. Schirmer, J. Situ, J. Paguio, J. Park, J. G. Wawira, S. Yao, and for MIT Critical Data. 2022. Sources of bias in artificial intelligence that perpetuate healthcare disparities—a global review. *PLOS Digital Health*, 1(3):e0000022.

J. Clusmann, F. R. Kolbinger, H. S. Muti, Z. I. Carrero, J. N. Eckardt, N. G. Laleh, C. M. L. Löffler, S. C. Schwarzkopf, M. Unger, G. P. Veldhuizen, S. J. Wagner, and J. N. Kather. 2023. The future landscape of large language models in medicine. *Communications Medicine*, 3(1):141.

F. T. Endomba, J. J. Bigna, and J. J. Noubiap. 2020. The impact of social networking services on the coronavirus disease 2019 (covid-19) pandemic in sub-saharan africa. *Pan African Medical Journal*, 35(Suppl 2):67.

Felipe Giuste, Wenqi Shi, Yuanda Zhu, Tarun Naren, Monica Isgut, Ying Sha, Li Tong, Mitali Gupte, and May D. Wang. 2022. Explainable artificial intelligence methods in combating pandemics: A systematic review. *Preprint*, arXiv:2112.12705.

D. Hadar-Shoval, K. Asraf, S. Shinan-Altman, Z. Elyoseph, and I. Levkovich. 2024. Embedded values-like shape ethical reasoning of large language models on primary care ethical dilemmas. *Heliyon*, 10(18):e38056.

J. E. Idänpään-Heikkilä. 1994. Who guidelines for good clinical practice (gcp) for trials on pharmaceutical products: Responsibilities of the investigator. *Annals of Medicine*, 26(2):89–94.

M. Ikhoyameh, W. E. Okete, R. M. Ogboye, O. K. Owoyemi, and O. S. Gbadebo. 2024. Integrating traditional medicine into the african healthcare system post-traditional medicine global summit: challenges and recommendations. *The Pan African Medical Journal*, 47:146.

Jacaranda Health. 2023. Jacaranda launches first in-kind swahili large language model. https://jacarandahealth.org/jacaranda-launches-first-in-kind-swahili-large-language-model.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Preprint*, arXiv:2009.13081.

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

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2023. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. *Preprint*, arXiv:2310.13132.

L. Kamulegeya, J. Bwanika, M. Okello, D. Rusoke, F. Nassiwa, W. Lubega, D. Musinguzi, and A. Börve. 2023. Using artificial intelligence on dermatology conditions in uganda: a case for diversity in training data sets for machine learning. *Afr Health Sci*, 23(2):753–763.

Kew. 2025. Plants of the World Online. Published on the Internet; facilitated by the Royal Botanic Gardens, Kew.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *Preprint*, arXiv:2402.10373.

K. Lam. 2023. Chatgpt for low- and middle-income countries: a greek gift? *The Lancet Regional Health. Western Pacific*, 41:100906.

Tammy Lovell. 2021. Babylon launches ai-powered triage tool in rwanda.

Zilin Ma, Yiyang Mei, and Zhaoyuan Su. 2023. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. *Preprint*, arXiv:2307.15810.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

S. Nanda. 2023. Integrating traditional and contemporary systems for health and well-being. *Annals of Neurosciences*, 30(2):77–78.

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. *Preprint*, arXiv:2305.14456.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. *Preprint*, arXiv:2407.10920.

Tobi Olatunji, Charles Nimo, Abraham Owodunni, Tassallah Abdullahi, Emmanuel Ayodele, Mardhiyah Sanni, Chinemelu Aka, Folafunmi Omofoye, Foutse Yuehgoh, Timothy Faniran, Bonaventure F. P. Dossou, Moshood Yekini, Jonas Kemp, Katherine Heller, Jude Chidubem Omeke, Chidi Asuzu MD, Naome A. Etori, Aimérou Ndiaye, Ifeoma Okoh, Evans Doe Ocansey, Wendy Kinara, Michael Best, Irfan Essa, Stephen Edward Moore, Chris Fourie, and Mercy Nyamewaa Asiedu. 2025. Afrimed-qa: A pan-african, multi-specialty, medical question-answering benchmark dataset. *Preprint*, arXiv:2411.15640.

D. B. Olawade, O. J. Wada, A. C. David-Olawade, E. Kunonga, O. Abaire, and J. Ling. 2023. Using artificial intelligence to improve public health: a narrative review. *Frontiers in Public Health*, 11:1196397.

OpenAI. 2023. ChatGPT. `https://chat.openai.com`.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. 2024. Gpt-4 technical report. ArXiv 2303.08774.

World Health Organization. 2003. Who guidelines on good agricultural and collection practices for medicinal plants. Technical report, World Health Organization, Geneva, Switzerland.

World Health Organization. 2008. Training in tropical diseases. In *Good Laboratory Practice Training Manual: Trainer*. World Health Organization, Geneva, Switzerland.

A. Owoyemi, J. Owoyemi, A. Osiyemi, and A. Boyd. 2020. Artificial intelligence for healthcare in africa. *Frontiers in Digital Health*, 2:6. PMID: 34713019; PMCID: PMC8521850.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. *Preprint*, arXiv:2203.14371.

B. Patwardhan, L. S. Wieland, O. Aginam, A. Chuthaputti, R. Ghelman, R. Ghods, G. C. Soon, M. G. Matsabisa, G. Seifert, S. Tu'itahi, K. S. Chol, S. Kuruvilla, K. Kemper, H. Cramer, H. R. Nagendra, A. Thakar, T. Nesari, S. Sharma, N. Srikanth, and R. Acharya. 2023. Evidence-based traditional medicine for transforming global health & wellbeing. *The Indian Journal of Medical Research*, 158(2):101–105.

Stephen R. Pfohl, Heather Cole-Lewis, Rory Sayres, Darlene Neal, Mercy Asiedu, Awa Dieng, Nenad Tomasev, Qazi Mamunur Rashid, Shekoofeh Azizi, Negar Rostamzadeh, Liam G. McCoy, Leo Anthony Celi, Yun Liu, Mike Schaekermann, Alanna Walton, Alicia Parrish, Chirag Nagpal, Preeti Singh, Akeiylah Dewitt, Philip Mansfield, Sushant Prakash, Katherine Heller, Alan Karthikesalingam, Christopher Semturs, Joelle Barral, Greg Corrado, Yossi Matias, Jamila Smith-Loud, Ivor Horn, and Karan Singhal. 2024. A toolbox for surfacing health equity harms and biases in large language models. *Nature Medicine*, 30(12):3590–3600.

M. Phiri and A. Munoriyarwa. 2023. Health chatbots in africa: Scoping review. *Journal of Medical Internet Research*, 25:e35573.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. Normad: A framework for measuring the cultural adaptability of large language models. *Preprint*, arXiv:2404.12464.

David et al. Romero. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pages 11479–11505. Curran Associates, Inc.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge. *Preprint*, arXiv:2212.13138.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. *Preprint*, arXiv:1703.01365.

Harini Suresh and John Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, page 1–9. ACM.

Kush R. Varshney. 2024. Decolonial ai alignment: Openness, viśeṣa-dharma, and including excluded knowledges. *Preprint*, arXiv:2309.05030.

WHO Regional Office for Africa. 2004. Guidelines for clinical study of traditional medicines in the who african region. Technical report, WHO Regional Office for Africa, Brazzaville, Congo.

World Health Organization. 2023. Billions left behind on the path to universal health coverage.

Kaiyan Zhang, Ning Ding, Biqing Qi, Sihang Zeng, Haoxin Li, Xuekai Zhu, Zhang-Ren Chen, and Bowen Zhou. 2024. Ultramedical: Building specialized generalists in biomedicine. `https://github.com/TsinghuaC3I/UltraMedical`.

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S. Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Chenyu You, Xian Wu, Yefeng Zheng, Lei Clifton, Zheng Li, Jiebo Luo, and David A. Clifton. 2024. A survey of large language models in medicine: Progress, application, and challenge. *Preprint*, arXiv:2311.05112.

A Appendix

A.1 Language Models Details

Below is a comprehensive overview of the large language models used in this study, each optimized for medical tasks:

OpenBioLLM-70B (Ankit Pal, 2024) A model tailored to meet the specialized language and knowledge demands of the medical and life sciences fields. With 70 billion parameters, it has been fine-tuned on an extensive corpus of high-quality biomedical data, enhancing both accuracy and fluency in domain-specific contexts. Built on the Meta-Llama-3-70B-Instruct framework, it incorporates advanced datasets like the DPO and a custom, diverse medical instruction collection. It not only surpasses other open-source biomedical models of similar scale but also demonstrates superior performance on biomedical benchmarks when compared with leading models such as GPT-4, Gemini, Meditron-70B, Med-PaLM-1, and Med-PaLM-2.

OpenBioLLM-8B (Ankit Pal, 2024) This variant condenses the capabilities of its larger counterpart into 8 billion parameters. It is crafted to provide strong domain-specific performance while offering greater computational efficiency, making it ideal for deployment in resource-constrained environments without sacrificing essential biomedical capabilities.

UltraMedical-70B (Zhang et al., 2024) Designed to enhance access to medical examinations, literature comprehension, and clinical knowledge, this model builds on Meta's Llama-3-70B architecture. It is trained using a combination of supervised fine-tuning and iterative preference learning techniques (including methods like DPO and KTO) on the UltraMedical collection. This collection comprises 410,000 synthetic and manually curated biomedical instruction samples plus 100,000 preference data points. Achieving a score of 86.5 on the MedQA-USMLE benchmark, UltraMedical-70B sets a new state-of-the-art standard among open-source LLMs, matching the performance of Med-PaLM 2 on key tests.

UltraMedical-8B (Zhang et al., 2024) A scaled-down version of UltraMedical-70B, this 8 billion-parameter model maintains robust performance across various medical benchmarks. Despite its smaller size, it consistently outperforms previous larger models such as MedPaLM 1, Gemini-1.0, GPT-3.5, and Meditron-70B on average scores, offering an efficient yet effective solution for many biomedical applications.

BioMistral-7B Based on the innovative Mistral architecture, BioMistral-7B has been specifically fine-tuned with data from the PubMed Central Open Access Subset. This training strategy has enhanced its comprehension and adaptation to medical language tasks. As a result, BioMistral-7B not only outperforms other open-source medical models but also remains competitive with some proprietary alternatives, making it a valuable asset for biomedical text processing tasks.

A.2 Prompt Templates

The following are examples of the three multiple-choice question variants used in this study: No-Context, Full-Context, and Misleading-Context.

MCQ : No-Context Example

Choose the most appropriate answer. Return only the letter (A, B, C, or D).

A) Chamaerops humilis

B) Chloroquine

C) Metformin

D) Spilanthes oleracea

Figure 3: Example of the No-Context MCQ variant, where only the answer choices are provided, with no accompanying contextual information.

MCQ : Full-Context Example

A 25-year-old man from Cameroon present with Malaria. What commonly used traditional medicinal herbal plant is most appropriate for their treatment?

A) Chamaerops humilis

B) Chloroquine

C) Metformin

D) Spilanthes oleracea

Choose the best answer. Return only the letter.

Figure 4: Example of the Full-Context MCQ variant, providing detailed patient and cultural context to guide the choice of the most appropriate treatment. **Correct answer: D.**

MCQ : Misleading-Context Example

A 25-year-old man from Cameroon present with Malaria. What commonly used traditional medicinal herbal plant is most appropriate for their treatment?
Note: The following cultural notes may be misleading. Rely on your medical knowledge to choose the aligned traditional remedy.

A) Chamaerops humilis (traditional use in South Asia)

B) Chloroquine (used for centuries in European folk medicine)

C) Metformin (popular in North African regions)

D) Spilanthes oleracea (widely used in rural Southeast Asia)

Choose the best answer. Return only the letter.

Figure 5: Example of the Misleading-Context MCQ variant, featuring deceptive cultural cues intended to challenge the model's ability to discern the culturally aligned answer. **Correct answer: D.**

A.3 Screening Rubric

Table 3: Screening rubric for PubMed records on African Traditional Medicine. Items marked with (*) are mandatory for inclusion.

No.	Checklist item	Y/N
A. Bibliographic filters		
A1*	Publication year between 2020–2025	
A2*	English–language full text available	
B. Relevance to African TM		
B1*	Study investigates an **indigenous African medicinal plant** (species confirmed in POWO)	
B2*	Plant part, preparation method, and dosage described	
C. Evidence strength (WHO R&D criteria)		
C1*	Study design meets minimum evidence threshold	
C2	Safety/toxicity data reported or referenced	
D. Methodological quality		
D1	Randomisation or control group described (if applicable)	
D2	Outcome measures clearly defined and reproducible	
D3	Statistical analysis appropriate and fully reported	