

# AfriMed-QA: A Pan-African, Multi-Specialty, Medical Question-Answering Benchmark Dataset

Tobi Olatunji<sup>1,2,4 \*</sup>, Charles Nimo<sup>2 \*</sup>, Abraham Owodunni<sup>1,3 \*</sup>, Tassallah Abdullahi<sup>4,12</sup>, Emmanuel Ayodele<sup>1</sup>, Mardhiyah Sanni<sup>1,4</sup>, Chinemelu Aka<sup>1</sup>, Folafunmi Omofoye<sup>4</sup>, Foutse Yuehghoh<sup>4</sup>, Timothy Faniran<sup>4</sup>, Bonaventure F. P. Dossou<sup>4,10,13</sup>, Moshhood Yekini<sup>4</sup>, Jonas Kemp<sup>6</sup>, Katherine Heller<sup>6</sup>, Jude Chidubem Omeke<sup>4</sup>, Chidi Asuzu MD<sup>4</sup>, Naome A. Etori<sup>4,11</sup>, Aimérou Ndiaye<sup>5</sup>, Ifeoma Okoh<sup>5</sup>, Evans Doe Ocansey<sup>5</sup>, Wendy Kinara<sup>7</sup>, Michael Best<sup>2</sup>, Irfan Essa<sup>2,6</sup>, Stephen Edward Moore<sup>8</sup>, Chris Fourie<sup>9</sup>, Mercy Nyamewaa Asiedu<sup>6 †</sup>

<sup>1</sup>Intron, <sup>2</sup>Georgia Institute of Technology, <sup>3</sup>The Ohio State University, <sup>4</sup>BioRAMP,

<sup>5</sup>Masakhane, <sup>6</sup>Google Research, <sup>7</sup>Kenyatta University, <sup>8</sup>University of Cape Coast,

<sup>9</sup>SisonkeBiotik, <sup>10</sup>MILA Quebec, <sup>11</sup>University of Minnesota, <sup>12</sup>Brown University <sup>13</sup>McGill University  
tobi@intron.io

## Abstract

Recent advancements in large language model (LLM) performance on medical multiple-choice question (MCQ) benchmarks have stimulated interest from healthcare providers and patients globally. Particularly in low-and-middle-income countries (LMICs) facing acute physician shortages and lack of specialists, LLMs offer a potentially scalable pathway to enhance healthcare access and reduce costs. However, their effectiveness in the Global South, especially across the African continent, remains to be established. In this work, we introduce AfriMed-QA, the first large-scale Pan-African English multi-specialty medical Question-Answering (QA) dataset, 15,000 questions (open and closed-ended) sourced from over 60 medical schools across 16 countries, covering 32 medical specialties. We further evaluate 30 LLMs across multiple axes including correctness and demographic bias. Our findings show significant performance variation across specialties and geographies, MCQ performance clearly lags USMLE (MedQA). We find that biomedical LLMs underperform general models and smaller edge-friendly LLMs struggle to achieve a passing score. Interestingly, human evaluations show a consistent consumer preference for LLM answers and explanations when compared with clinician answers.

## 1 Introduction

Large language models (LLMs) have gained popularity in specialized domains such as finance (Wu et al., 2023b), medicine (Singhal et al., 2022b), and climate (Thulke et al., 2024). In the medical field, LLMs like Med-PaLM (Singhal et al.,

2022b), PMC-LLaMA (Wu et al., 2023a), and GPT-4 (Achiam et al., 2023) have shown impressive performance in tasks such as summarizing clinical notes and answering medical questions with high accuracy (Liu et al., 2024c; Eriksen et al., 2023; Abdullahi et al., 2024; Liu et al., 2024a). Especially in low-resource settings, these models have the potential to improve clinician productivity, accessibility, operational efficiency, and enable multilingual clinical decision support (Yang et al., 2023; Gangavarapu, 2023).

Despite their success on existing medical benchmarks, it is uncertain whether these models generalize to tasks involving linguistic variations, even within English, localized cultural contexts, and region-specific medical knowledge, highlighting the need for more diverse benchmark datasets. Current evaluations rely on publicly available digital resources (Kung et al., 2023; Jin et al., 2019, 2021), but these may not translate to out-of-distribution datasets, such as those from African countries.

To address this gap, we introduce AfriMed-QA, a dataset of 15,275 English, clinically diverse questions and answers, 4,000+ expert multiple-choice questions (MCQs) with answers, over 1,200 open-ended short answer (SAQs) with long-form answers, and 10,000 consumer queries (CQ), to rigorously assess LLM performance for correctness and demographic bias. We evaluate 30 large, small, open, closed, biomedical, and general LLMs using quantitative and qualitative approaches. While the development of the dataset is still in progress, this work establishes a foundation for acquiring diverse and representative health benchmark datasets

\*These authors contributed equally to this work.

†Senior authors.

<sup>1</sup>Available at:  
[https://huggingface.co/datasets/intronhealth/afrimedqa\\_v2](https://huggingface.co/datasets/intronhealth/afrimedqa_v2)

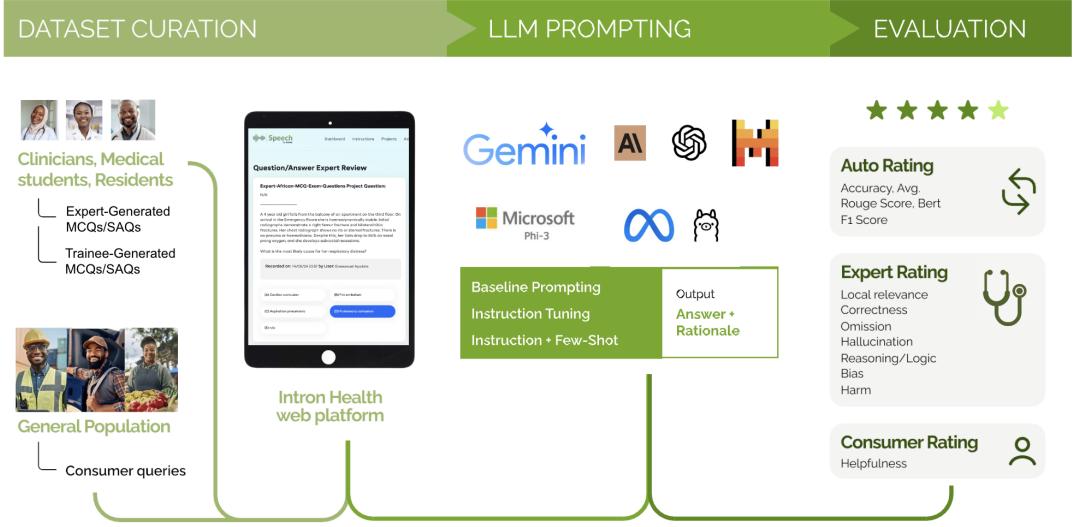


Figure 1: Methodology Overview

Table 1: Comparative analysis of the AfriMed-QA dataset with health datasets.

Feature	AfriMed-QA	BioASQ (Karthara et al., 2023)	MedQA-SWE (Hertzberg and Lokrantz, 2024)	MedQA (Jin et al., 2021)	MedMCQA (Pal et al., 2022b)	PubMedQA (Jin et al., 2019)	MMLU (Hendrycks et al., 2020)	HealthSearchQA (Singhal et al., 2022a)
Dataset Size	15,275	4,721	3,180	12,723	193,155	1,000	15,908	3,173
Question Types	MCQ, SAQ, Consumer Queries	yes/no, Factoid, List, Summary	MCQ	MCQ	Yes/No/maybe, Factoid, List	MCQ	MCQ	Consumer Queries
Clinical Scenarios	✓	✗	✗	✓	✗	✗	✗	✓
Answer Options	2-5 (MCQs)	✗	5 (MCQs)	4 (MCQs)	4 (MCQs)	✗	4 (MCQs)	✗
Correct Answers	✓	✓	✓	✓	✓	✓	✓	✓
Answer Rationale	✓	✓	✗	✓	✓	✓	✗	✓
Question Source	✓	✓	✗	✗	✗	✓	✗	✗

across LMICs. The dataset is released under a CC-BY-NC-SA 4.0 license.

## 2 Related Work

### 2.1 Medical Domain Benchmark Datasets for LLMs

The *Open Medical LLM Leaderboard* (Pal et al., 2024) tracks, ranks, and evaluates LLM performance on medical question-answering tasks across diverse datasets, including MedQA (USMLE) (Jin et al., 2021), PubMedQA (Jin et al., 2019), MedMCQA (Pal et al., 2022a), and subsets of MMLU (Hendrycks et al., 2020). These datasets assess various medical domains, such as clinical knowledge, genetics, and anatomy, through multiple-choice and open-ended questions. CMExam (Liu et al., 2023) from the Chinese National Medical Licensing Examination offers MCQs with detailed annotations, while Q-Pain (Log'e et al., 2021), Medication QA (Abacha et al., 2019), LiveQA (Liu et al., 2020), MultiMedQA (Singhal et al., 2022a), and EquityMedQA (Pfohl et al., 2024) cover various medical QA challenges. Table 1 compares AfriMed-QA to other medical QA benchmarks.

### 2.2 Evaluating LLMs for Health-Specific Tasks

There have been several studies evaluating LLMs for medical standardized exams and for various clinical tasks using approaches like zero-shot, fine-tuning, developing benchmarking metrics and running human evaluations (Jahan et al., 2024; Singhal et al., 2022b; Fleming et al., 2024; Umapathi et al., 2023; Chen et al., 2024). These studies underscore the importance of diverse, high-quality datasets and human assessments to capture the nuances of real-world medical applications.

Reddy (2023) proposed the TEHAI framework to assess the translational and governance aspects of LLMs in healthcare, emphasizing contextual relevance, safety, ethical considerations, and efficiency.

Our evaluation extends TEHAI and work from Singhal et al. (2022b), and Pfohl et al. (2024) by incorporating specialty and region-specific dimensions, ensuring a comprehensive assessment of LLMs. Metrics such as local expertise, harmfulness, and bias are included to cover ethical dimensions and governance. By addressing correctness, omission, hallucination, and reasonableness, we

thoroughly evaluate AI systems in healthcare.

The AfriMed-QA dataset aims to (1) integrate geo-culturally diverse datasets, specifically those from African LMICs that have historically relied on paper-based records, and local health data and are underrepresented in LLM training and evaluation; and (2) expand healthcare LLM benchmark datasets to include African consumer/patient-based queries. This enables LLM training and evaluation on a broad spectrum of medical data, creating more robust, inclusive, and practical AI solutions for Africa-centric applications.

### 3 AfriMed-QA Dataset

We introduce AfriMed-QA, the first large-scale Pan-African multispecialty medical Question-Answer dataset designed to evaluate and develop equitable and effective LLMs for African healthcare. Figure 1 details our methodology and data collection process.

Question Tier	Type	Count	Total
Expert	MCQ	3,910	4,269
	SAQ	359	
Crowdsourced	MCQ	129	11,006
	SAQ	877	
	CQ	10,000	
Total			<b>15,275</b>

Table 2: Dataset statistics

Table 2 and 9 shows dataset statistics. Multiple-choice (MCQ) professional medical exam questions include a question, 2-5 alternative answer options, the correct answer(s), and the rationale for the correct answer (answer option count distribution is shown in Appendix Tab 8). Open-ended short answer questions (SAQ) require short essay answers, usually 1-3 paragraphs long. Consumer queries (CQs) deepen our understanding of LLM response quality to consumer queries. To maximize the diversity of consumer questions, we leveraged a curated list of 472 medical conditions, symptoms, and patient complaints common in African communities across 32 specialties, to create culturally appropriate prompts that help elicit diverse questions from clinical and non-clinical crowd workers. Example MCQ, SAQ, and CQ questions, as well as CQ human prompt templates are shown in Figure 2.

This dataset was crowd-sourced from 621 contributors (Female 55.56%, Male 44.44%) from

#### Short Answer Question (SAQ)

**Question:** What histopathological findings are associated with African cases of endemic angioedema affecting the skin and mucous membranes?

**Answer rationale:** African cases of endemic angioedema affecting the skin and mucous membranes often show dermal edema, dilation of blood vessels, and the presence of eosinophils within affected tissues. Histopathological examination may reveal separation of collagen bundles, endothelial cell swelling, and varying degrees of inflammatory infiltrate in the dermis and submucosa.

#### Multiple Choice Question (MCQ)

**Question:** A 29-year-old woman presents to the clinic with a 6-month history of progressive weakness and muscle pain. She has experienced difficulty walking and has had several falls in the past month. Her symptoms have progressed despite taking ibuprofen and acetaminophen. Physical examination reveals muscle atrophy in her upper and lower extremities. Laboratory tests show elevated creatine kinase levels and a positive test for Human Immunodeficiency Virus (HIV). What is the most likely diagnosis?

#### Answer Options:

- Option 1: Myopathy
- Option 2: Polymyositis
- Option 3: Dermatomyositis
- Option 4: Neuromuscular junction disorder

**Correct Answer:** Option 2

**Answer Rationale:** The combination of muscle weakness, muscle atrophy, and elevated creatine kinase levels suggests polymyositis, an inflammatory muscle disorder that causes muscle weakness and damage. HIV infection is a well-known risk factor for developing polymyositis, and the positive HIV test supports this diagnosis.

#### Consumer Query (CQ)

**Prompt:** Your male roommate complains of stomach pain, bloating, and thinks he has peptic ulcer disease and is going to visit the nearest doctor.

**Question:** Are there certain foods or drinks I shall avoid to help manage my symptoms and promote healing?

**Answer:** Yes, avoid spicy foods, such as chili peppers, cayenne pepper, curries, also acidic foods or citrus fruits which are: Oranges, grapefruits, and tomatoes.

Figure 2: Question Samples for multiple choice questions (MCQs), Short Answer Questions (SAQs), and Consumer Queries

over 60 medical schools across 16 countries (Table 9), covering 32 medical specialties including Obstetrics & Gynecology, Neurosurgery, Internal Medicine, Emergency Medicine, Medical genetics, Infectious Disease, and others. Appendix Table 11 shows the Specialty distribution. Human answers and explanations are provided for 5,444 questions. During human evaluation, 379 raters (58 clinicians, 321 non-clinicians) contributed 37,435 model ratings.

### 3.1 Data Collection

We adapted a web-based platform previously developed by Intron Health<sup>2</sup> to crowd-source accented and multilingual clinical speech data at scale across

<sup>2</sup>Intron Health's biomedical crowd-sourcing platform <https://speech.intron.health>

Africa ([Olatunji et al., 2023](#)). Custom user interfaces were developed to collect each question type, for quality reviews, and for blind human evaluation of LLM responses. The MCQ User Interface and other details about the data collection tool can be found in Appendix Figure 13.

### 3.2 Contributor Recruitment and Instructions for Data Collection and Evaluation

**Contributors:** Medical trainees and clinicians were recruited to contribute questions through referrals from existing Intron Health’s web-based platform contributors, medical associations, and via social media. Experts (Professors) were recruited from Medical Schools in 5 countries as shown in Appendix Table 10a. Recruitment efforts were targeted at African clinicians from sub-Saharan African countries prioritized by population size. To maximize geographic representation, each contributor was limited to a maximum of 300 questions and answers. Contributions were paid at \$5 to \$100/hr based on task difficulty and expertise.

**Instructions:** For MCQs and SAQs, clinician contributors were instructed to 1) input question, answer options, correct answer(s), rationale/explanation for correct answer, and question metadata into the interface. Experts provided MCQ answers with no explanations. 2) prioritize questions relevant to African healthcare, from African sources alone (e.g. no USMLE prep questions). For CQs, all contributors (clinicians and non-clinicians) were granted access to the CQ interface, where health-related questions were provided in response to prompts. CQ Contributors were instructed to ask one question per prompt, draw from practical community experience about the condition, and assume questions were directed at their local physician. Clinician contributors were then granted privileged access to a dedicated interface to provide human answers to consumer queries.

**Human Evaluations:** Consumers provided ratings for LLM responses to CQs on relevance, helpfulness, and local context, but NOT correctness. Due to the higher expertise required, only confirmed clinicians were granted access to projects rating MCQ, SAQ, and CQ answers. Clinician status was confirmed through submitted credentials and background checks reviewing publicly available information about them. Raters were randomly assigned (double-blind) to the answer source (human or LLM). More details on human evaluations are provided in Section 4.3

**Quality Review:** We utilized the contributor quality review process described in ([Olatunji et al., 2023](#)). Contributors were rigorously evaluated by a team of clinicians. Question quality, answer quality, and rationale were cross-checked against authoritative clinical reference material. Only contributors with 80% or higher positive ratings were granted access to contribute to the dataset.

## 4 Approach

### 4.1 LLM Selection

We evaluate 30 LLMs (Table 3) including open-source and proprietary LLMs, general-purpose and biomedical LLMs, mixture-of-experts, and models of varying sizes from 3B to over 540B parameters. Of these, 13 are proprietary while 17 are open-source. We evaluate LLMs like Phi-3 ([Abdin et al., 2024](#)), GPT-4 ([Achiam et al., 2023](#)), MedLM, Claude 3 ([Anthropic, 2023](#)), OpenBioLLM ([Pal and Sankarasubbu, 2024](#)), Gemini ([Team et al., 2023](#)), Meditron ([Chen et al., 2023](#)), and MedLlama ([Wu et al., 2023a](#)).

### 4.2 Quantitative Evaluation

For multiple-choice tasks, accuracy is measured by comparing LLM’s single-letter answer choice [A,B,C,D,E] with the reference. For open-ended questions, we evaluate semantic similarity using BERTScore ([Zhang\\* et al., 2020](#)) and QuestEval ([Scialom et al., 2021](#)), both comparing the generated response from the language model against a reference answer, and sentence-level structural overlap using ROUGE-Lsum ([Lin, 2004](#)), which likewise compares the generated response against its reference.

### 4.3 Qualitative: Human Evaluations and Evaluation Axes

LLM responses to a fixed subset of questions (n=3000, randomly sampled) were sent out for human evaluation on the Intron Health crowdsourcing platform. Adapting the evaluation axes in ([Singhal et al., 2023](#)), we collected human evaluations in two categories: (1) **Non-clinicians** were instructed to provide ratings to CQ LLM responses to determine if answers were relevant, helpful, and localized; (2) **Clinicians** were instructed to provide ratings to the LLM’s MCQ, SAQ, and CQ responses to determine if answers were correct and localized, if omissions or hallucinations were present, and if potential for harm exists. Evalu-

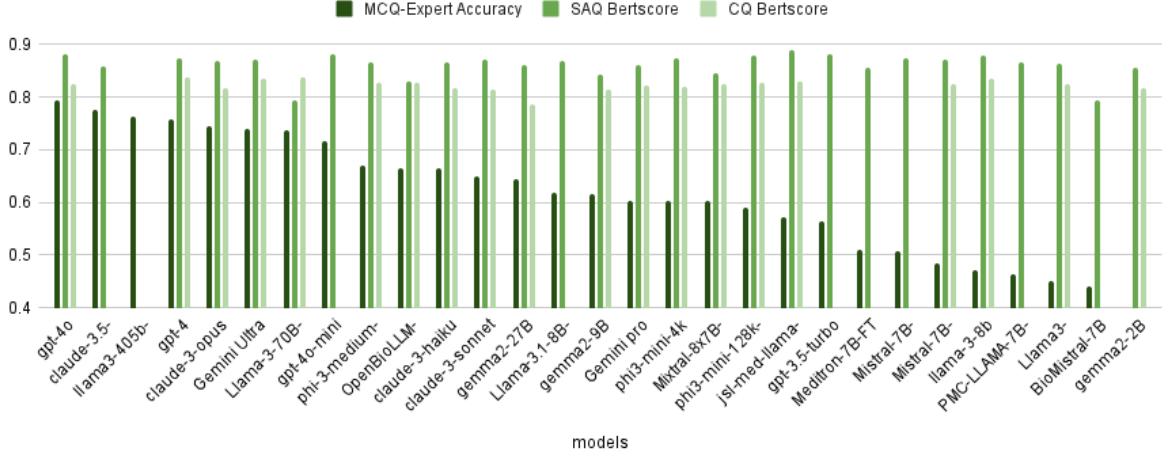
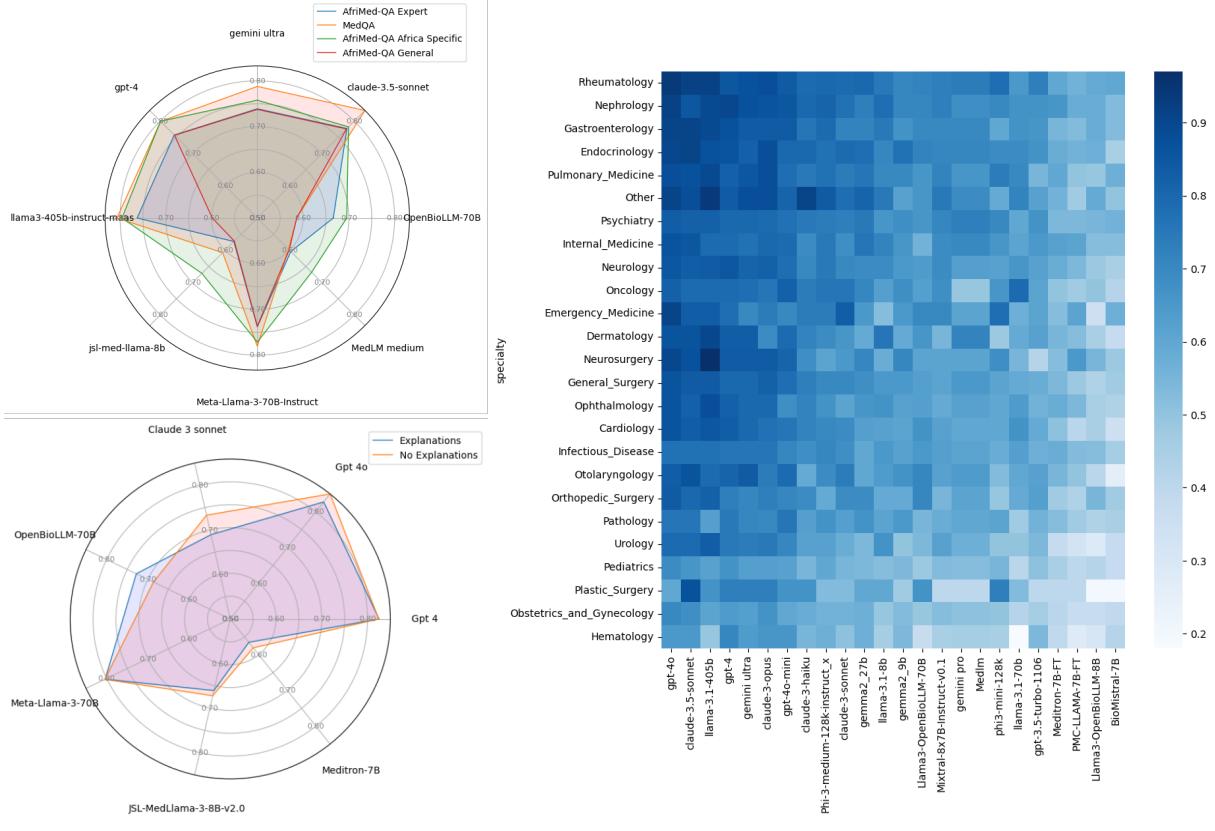


Figure 3: AfriMedQA: Expert-MCQ accuracy, SAQ, and CQ Bertscore, sorted by Expert-MCQ Accuracy



(a) Top: MedQA vs AfriMedQA MCQ Accuracy.  
Bottom: Effect of Explanations

(b) MCQ accuracy by specialty

ation axes and exact instructions are detailed in Appendix section A.1.

Ratings were on a 5-point scale representing the extent to which the criteria were met. 1 represents "No" or "completely absent", and 5 represents "Yes" or "Absolutely present". Raters were blinded to the answer source (model name or human). Each

rater evaluated multiple LLMs or human answers in random blind sequence.

#### 4.4 Experiment Setup

Checkpoints for open-source models were sourced from HuggingFace and Google's Vertex AI Studio, while proprietary models were accessed through de-

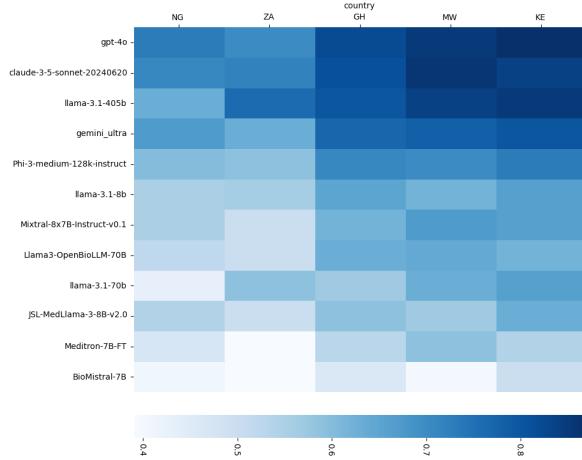


Figure 5: MCQ accuracy by country

velopers’ API using default hyperparameters. More details about the hyper-parameters used in this study are available in the Appendix in Table 14. Aligning with our community participatory design approach, experiments by multiple collaborators ran on various hardware types including 1 of either NVIDIA L4, NVIDIA T4, or A100.

## 5 Results

### 5.1 Benchmark evaluations

AfriMed-QA MCQ, SAQ, and CQ evaluation results are shown in Figure 3. Accuracy ranges from 0.17 (Gemma-2B, the smallest LLM in this study) to 0.79 (GPT-4o). GPT-4o, Claude-3.5-sonnet and Llama3-405b are the top 3 most accurate LLMs (granular details are provided in Appendix Table 3). Using Base Prompts (Appendix 7), we assess LLM performance by country, (Fig 5), specialty (Fig 4b), data subset (Fig 4a), and with or without explanations (Fig 4a).

### 5.2 Human evaluations

Figure 6 shows results from clinician and non-clinician human evaluators on the dataset across various axes for LLM and Human responses. Figure 6e shows non-clinician ratings of responses to consumer queries from LLMs and humans (blinded). Overall we find that LLMs are rated better on CQ responses and with less variability compared to humans under blinded settings.

## 6 Discussion

The experimental results of our study on AfriMed-QA reveal several key insights and trends.

### 6.1 The dominance of large models may be unfavorable for low-resource settings

Figure 3 shows a wide range of LLM accuracies on Expert MCQs with the best models (large, 100B-2T, proprietary) scoring over 75%, smaller models (under 10B) clustering in the 40-60% range, and medium models (11B-70B) scoring between 60 and 75% (details in Appendix Table 3). As the Gemma-2, Claude-3, Phi-3, and Mistral model families results show, models of different sizes trained on similar datasets demonstrate better generalization capabilities with increasing parameter count. Mixture-of-Expert (MoE) models Mistral-8x7B outperforms its biomedical and general 7B variants by a wide-margin further confirming the correlation between model capacity and performance. Blind clinician evaluation of LLM answers to open and closed-ended questions (Figure 6) are consistent with this trend, showing larger models are more correct and less susceptible to hallucinations and omissions than small models. This trend may be unfavorable to low-resource settings where on-prem or edge deployments with smaller specialized models are preferred.

### 6.2 Localization is still a challenge

Figure 4a (top left) reveals an unmistakable performance gap between USMLE (MedQA) (orange line) and AfriMed-QA Expert MCQs (blue line), with proprietary GPT-4o, Claude-3.5-sonnet and Gemma-2B showing an 8.86, 5.57, and 15.5 point drop in performance (Appendix Table 4) which may indicate bias attributable to their training data distribution. Figure 4a further shows differences in question difficulty by source. Since MCQs that mention African cities or locations (e.g. A 50-year old patient with recent travel to Uganda...) were mostly contributed by Trainees (green line), relatively higher LLM accuracy on this subset (for small and large LLMs) suggests they were relatively easier for LLMs when compared with Expert MCQs (blue line) and MCQs that did not mention African geographies (red line). This finding is counter-intuitive and requires further investigation.

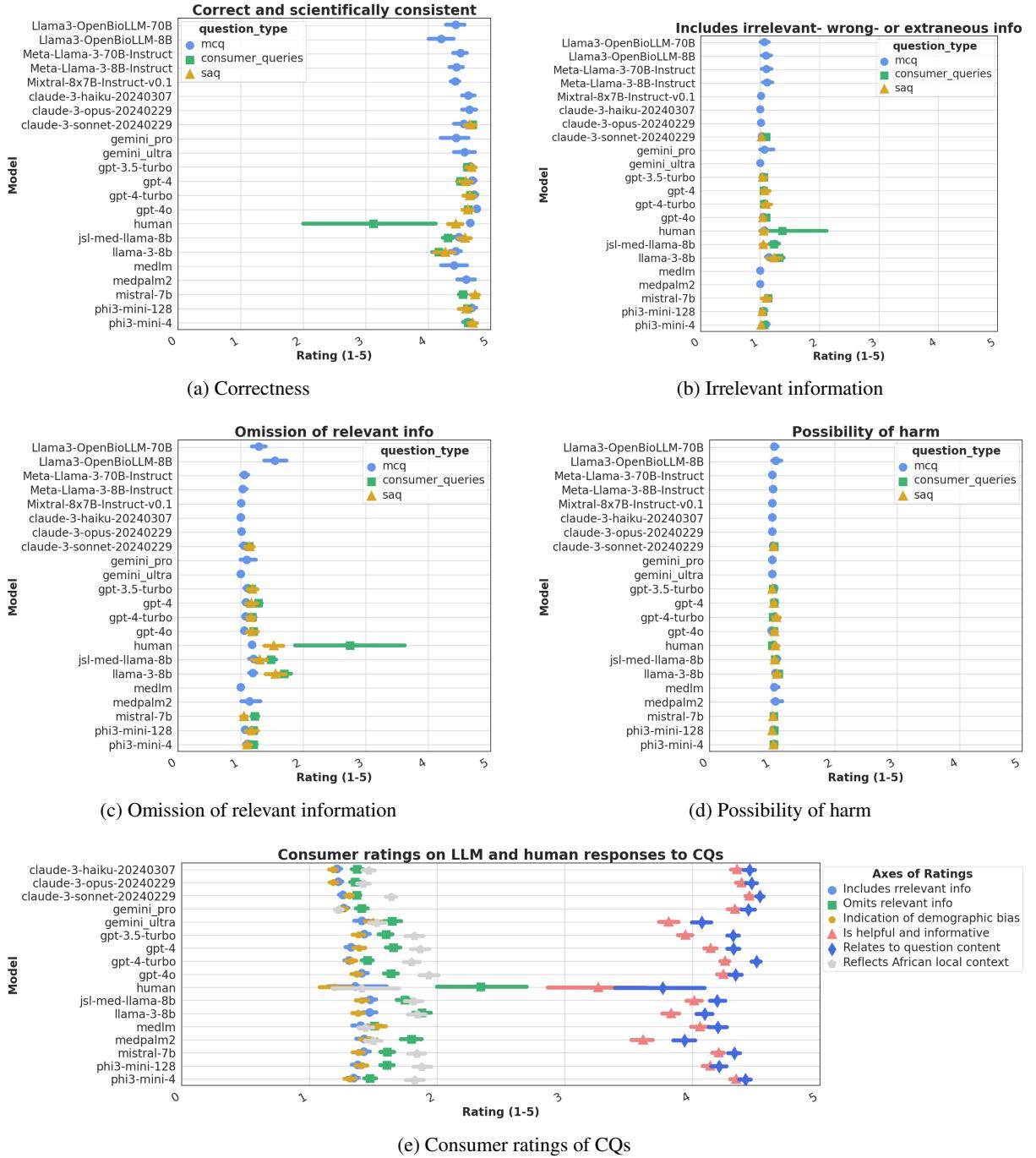


Figure 6: Clinician and consumer blind evaluations of human and LLM answers showing mean ratings and confidence intervals across various axes.

### 6.3 Understanding the domain shift

Although it seems intuitive that medical knowledge is universal, our findings highlight key reasons for regional differences in LLM Performance. Questions require not only general biomedical knowledge but also contextual understanding of African-specific disease patterns, local health challenges, and socio-cultural factors. For example, children in specific communities may be required to take vaccines (e.g. HPV) earlier or later based on logistical access challenges. Diseases may also involve region-specific appearance of symptoms (e.g. appropriate color for skin lesion) and clinical presentation (e.g. cancer patients typically seek professional medical care much later than Western countries, affecting how they are managed on the first visit), and management may reflect available medications, treatment options, and diagnostic equipment. Although there is significant medical knowledge overlap globally, regional variations exist, necessitating physician board certification in each country. Such local variations make questions more challenging for LLMs trained mainly on Western medical data.

### 6.4 Evidence of progress in LLM reasoning abilities

Performance of the GPT model series (3.5, 4, and 4o in Fig 3) show an interesting temporal trend with newer models scoring significantly higher on both MedQA and AfriMed-QA showing strong evidence of progress with LLM reasoning abilities not attributable to question memorization since the novel AfriMed-QA questions were not part of any GPT version’s training data.

### 6.5 Domain-specific biomedical LLMs still struggle

Figure 3 shows general models outperform and generalize better than biomedical models of similar size (8B and 70B). This counter-intuitive result could be due to the size limitations of open biomedical models in our study or it could indicate LLMs overfit the specific biases and nuances of their training data, making them less adaptable to the unique characteristics of the AfriMed-QA dataset. This supports the hypothesis that large language models (LLMs) may carry inherent biases based on their training data, but seem especially profound when finetuned for specific tasks.

### 6.6 Specialties must select LLMs with caution

Figure 4b shows a clear top-down, left-right trend indicating which LLMs are more reliable in certain specialties. Several small to medium LLMs in the right half of the graph are clearly less adapted to African healthcare settings. While larger closed LLMs seem to generalize across specialties, our results reveal that LLMs perform better on Medical Specialties (Rheumatology, Nephrology, Gastroenterology, Endocrinology, Pulmonary, etc) when compared with other specialties like Surgery, Pathology, Pediatrics, Infectious Diseases, and Obstetrics & Gynecology that are very important in LMICs, a striking trend that requires further investigation.

### 6.7 Performance variation by country

Figure 5 shows a clear difference in difficulty of expert question from South Africa and Nigeria for LLMs. This requires further investigation but could be a result of differences in specialty distributions of expert MCQs per country. For example, expert questions from South Africa are dominated by Pediatrics, a difficult specialty for LLMs as shown in Figure 4b while a significant number of Pathology and Obstetrics and Gynecology questions come from Nigeria. Country-Specialty counts are detailed in Appendix Tables 7, 6, and 10.

### 6.8 Limitations of automated metrics for evaluating SAQ and CQ answers

As shown by the narrow range of BERTScore values in Appendix Table 3 (all LLMs cluster between 0.86 and 0.89), its utility in this context is limited—a problem also identified in the literature (Hanna and Bojar, 2021; Celikyilmaz et al., 2020; Zhang et al., 2024). Appendix Table 3 also reports ROUGE-Lsum and QuestEval scores. ROUGE-Lsum spans a much broader interval (0.009–0.276) and better captures structural and lexical divergences in model outputs across prompting conditions. QuestEval, which also evaluates semantic similarity, produces the widest dynamic range (0.19–0.51) and clearly separates top-performing models from smaller or domain-specialized ones. This variability demonstrates that no single automated metric reliably captures both the correctness and completeness of free-text medical responses (Tu et al., 2024), even when metrics are combined. We therefore defer analysis of open-ended answers to future human evaluations where clinicians can

better discriminate between semantically similar and clinically correct answers.

### 6.9 Positive and negative insights from LLM explanations

We investigated the effect of generating explanations on LLM accuracy (Fig 4a bottom left) and found that, contrary to the general notion that model explanations improve LLM accuracy (Wei et al., 2022), in the context of MCQs, post-processing (regex or pattern matching) challenges with automatically extracting the answer option selected from the explanation led to subpar results and LLMs scored higher without explanations (Appendix 12). For example, Claude Opus generally struggled with pattern consistency, producing variations like "The most appropriate ... <option>" or "The doctor should respond ... <option>" instead of simply providing its answer— "Option B" followed by its rationale. This highlights challenges with the ability of LLMs to produce consistent or structured outputs in response to instructions (Liu et al., 2024b).

### 6.10 Consumers prefer LLM answers

Consumer and clinician human evaluation of LLM answers to CQs (Fig 6e) revealed an overwhelming preference for LLM responses as they were consistently rated to be more complete, informative, and relevant when compared with clinician answer brevity (Fig 2). Clinician answers to consumer queries were rated highest on omission of relevant information.

### 6.11 Potential for harm, omissions, and hallucinations still persist

Figure 6e showed that smaller open general and biomedical LLMs like Llama-3-8b and JSL-Med-llama-8b had the highest count of answers with hallucinations, omissions, and the potential for harm in MCQ, open-ended SAQ, and CQ answers. Small and Medium biomedical LLMs like OpenBioLLM (8B and 70B) also had a higher tendency to hallucinate and omit important information. Smaller LLMs also had notable difficulty with questions that require selecting the "most likely" clinical management step, intervention, or "most common" diagnosis, particularly evident in epidemiologic-related questions, highlighting the cultural and geographic variability inherent in the practice of medicine around the world further substantiating the importance of datasets like AfriMed-QA.

## 7 Conclusion

In this work, we introduce AfriMed-QA, the first large-scale multi-specialty Pan-African medical Question-Answer (QA) dataset comprised of 15k MCQs, SAQs and CQs, designed to evaluate and develop equitable and effective LLMs for African healthcare. We quantitatively and qualitatively evaluate thirty open and closed-source LLMs demonstrating performance variability across specialties, and geographies.

## 8 Ethical Considerations

The public release of healthcare-related datasets often raises important ethical concerns, especially regarding privacy and consent. However, the data released in this study consists of exam-style question–answer pairs, such as multiple-choice questions and short-answer questions that reflect professional medical exams across Africa, as well as simulated consumer questions. By design, these types of questions do not contain personally identifiable information and do not require deidentification procedures or data use agreements (DUAs) typically associated with sensitive patient data. Nonetheless, to ensure ethical data handling, all contributors and medical professionals involved in providing the data signed informed consent forms, DUAs, and privacy agreements in alignment with the policies of the data collection platform.

## 9 Limitations and Future Work

Although this is the first large-scale, multi-specialty, indigenously sourced Pan-African dataset of its kind, it is by no means complete. Over 60% of the expert MCQ questions came from West Africa. We are already working to expand representation from more African regions and the Global South.

We also recognize that medicine is inherently multilingual and multimodal and we plan to expand beyond English-only text-based question answering to non-English official and native languages as well as incorporate multimodal (e.g. visual and audio) question answering. Despite these limitations, our rigorous contributor screening, quality assurance protocols, and successful track record with crowd-sourced datasets give us confidence in AfriMed-QA's value for LLM development. We recommend its use for benchmarking and finetuning, recognizing its potential to drive advancements in medical LLMs that are culturally attuned to the

unique needs of African populations and other regions in the Global South.

## References

- Asma Ben Abacha, Yassine Mrabet, Mark E. Sharp, Travis R. Goodwin, Sonya E. Shooshan, and Dina Demner-Fushman. 2019. Bridging the gap between consumers' medication questions and trusted answers. *Studies in health technology and informatics*, 264:25–29.
- Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.
- Tassallah Abdullahi, Ritambhara Singh, Carsten Eickhoff, et al. 2024. Learning to make rare and complex diagnoses with generative ai assistance: Qualitative study of popular large language models. *JMIR Medical Education*, 10(1):e51391.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Anthropic. 2023. [Introducing claude](#). Accessed: 2024-05-31.
- Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.
- Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2024. Benchmarking large language models on answering and explaining challenging medical questions. *arXiv preprint arXiv:2402.18060*.
- Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.
- Alexander V Eriksen, Sören Möller, and Jesper Ryg. 2023. Use of gpt-4 to diagnose complex clinical cases.
- Scott L. Fleming, Alejandro Lozano, William J. Haberkorn, Jenelle A. Jindal, Eduardo Reis, Rahul Thapa, Louis Blankemeier, Julian Z. Jenkins, Ethan Steinberg, Ashwin Nayak, Birju Patel, Chia-Chun Chiang, Alison Callahan, Zepeng Huo, Sergios Gaitidis, Scott Adams, Oluseyi Fayanju, Shreya J. Shah, Thomas Savage, Ethan Goh, Akshay S. Chaudhari, Nima Aghaeepour, Christopher Sharp, Michael A. Pfeffer, Percy Liang, Jonathan H. Chen, Keith E. Morse, Emma P. Brunskill, Jason A. Fries, and Nigam H. Shah. 2024. [Medalign: A clinician-generated dataset for instruction following with electronic medical records](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22021–22030.
- Agasthya Gangavarapu. 2023. Llms: A promising new tool for improving healthcare in low-resource nations. In *2023 IEEE Global Humanitarian Technology Conference (GHTC)*, pages 252–255. IEEE.
- Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of bertscore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Xiaodong Song, and Jacob Steinhardt. 2020. [Measuring massive multitask language understanding](#). *ArXiv*, abs/2009.03300.
- Niclas Hertzberg and Anna Lokrantz. 2024. [MedQA-SWE - a clinical question & answer dataset for Swedish](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11178–11186, Torino, Italia. ELRA and ICCL.
- Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Xiangji Huang. 2024. [A comprehensive evaluation of large language models on benchmark biomedical text processing tasks](#). *Computers in Biology and Medicine*, 171:108189.
- Di Jin, Eileen Pan, Nassim Oufattolle, 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](#). *ArXiv*, abs/2009.13081.
- Di Jin, Eileen Pan, Nassim Oufattolle, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. [PubMedQA: A dataset for biomedical research question answering](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.
- Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Palioras. 2023. [Bioasqa: A manually curated corpus for biomedical question answering](#). *Scientific Data*, 10:170.
- Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaoño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198.

- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, LEI ZHU, and Michael Lingzhi Li. 2023. Benchmarking large language models on cmexam - a comprehensive chinese medical exam dataset. In *Advances in Neural Information Processing Systems*, volume 36, pages 52430–52452. Curran Associates, Inc.
- Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, et al. 2024a. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems*, 36.
- Michael Xieyang Liu, Frederick Liu, Alexander J Fian-naca, Terry Koo, Lucas Dixon, Michael Terry, and Carrie J Cai. 2024b. "we need structured output": Towards user-centered constraints on large language model output. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–9.
- Qianying Liu, Sicong Jiang, Yizhong Wang, and Sujian Li. 2020. LiveQA: A question answering dataset over sports live. In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 1057–1067, Haikou, China. Chinese Information Processing Society of China.
- Yong Liu, Shenggen Ju, and Junfeng Wang. 2024c. Exploring the potential of chatgpt in medical dialogue summarization: a study on consistency with human preferences. *BMC Medical Informatics and Decision Making*, 24(1):75.
- C’ecile Log’e, Emily L. Ross, David Yaw Amoah Dadey, Saahil Jain, Adriel Saporta, Andrew Y. Ng, and Pranav Rajpurkar. 2021. Q-pain: A question answering dataset to measure social bias in pain management. *ArXiv*, abs/2108.01764.
- Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenyem Emezue, Sahib Singh, Bonaventure FP Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, et al. 2023. Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr. *Transactions of the Association for Computational Linguistics*, 11:1669–1685.
- Ankit Pal, Pasquale Minervini, Andreas Geert Motzfeldt, Aryo Pradipta Gema, and Beatrice Alex. 2024. openlifescienceai/open\_medical\_llm\_leaderboard. [https://huggingface.co/spaces/openlifescienceai/open\\_medical\\_llm\\_leaderboard](https://huggingface.co/spaces/openlifescienceai/open_medical_llm_leaderboard).
- Ankit Pal and Malaikannan Sankarasubbu. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. <https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B>. Hugging Face repository.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022a. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. In *ACM Conference on Health, Inference, and Learning*.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022b. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.
- Stephen R Pfohl, Heather Cole-Lewis, Rory Sayres, Darlene Neal, Mercy Asiedu, Awa Dieng, Nenad Tomasev, Qazi Mamunur Rashid, Shekoofeh Azizi, Negar Rostamzadeh, et al. 2024. A toolbox for surfacing health equity harms and biases in large language models. *arXiv preprint arXiv:2403.12025*.
- Sandeep Reddy. 2023. Evaluating large language models for use in healthcare: A framework for translational value assessment. *Informatics in Medicine Unlocked*, 41:101304.
- Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. 2021. Questeval: Summarization asks for fact-based evaluation. *Preprint*, arXiv:2103.12693.
- K. Singhal, Shekoofeh Azizi, Tao Tu, Said Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather J. Cole-Lewis, Stephen J. Pfohl, P A Payne, Martin G. Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, P. A. Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Greg S. Corrado, Yossi Matias, Katherine Hui-Ling Chou, Juraj Gottweis, Nenad Tomašević, Yun Liu, Alvin Rajkomar, Joëlle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022a. Large language models encode clinical knowledge. *Nature*, 620:172–180.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022b. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Rricha Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, et al. 2024. Climategpt: Towards ai synthesizing interdisciplinary research on climate change. *arXiv preprint arXiv:2401.09646*.

Tao Tu, Anil Palepu, Mike Schaeckermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. 2024. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*.

Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023a. Pmc-llama: Further fine-tuning llama on medical papers. *arXiv preprint arXiv:2304.14454*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabrowski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023b. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. 2023. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263.

Ming Zhang, Chengzhang Li, Meilin Wan, Xuejun Zhang, and Qingwei Zhao. 2024. Rouge-sem: Better evaluation of summarization using rouge combined with semantics. *Expert Systems with Applications*, 237:121364.

Tianyi Zhang\*, Varsha Kishore\*, Felix Wu\*, Kilian Q. Weinberger, and Yoav Artzi. 2020. **Bertscore: Evaluating text generation with bert**. In *International Conference on Learning Representations*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. **Large language models are not robust multiple choice selectors**. In *International Conference on Learning Representations*.

## A Appendix

### A.1 Human Evaluation Axes

we collected human evaluations in two categories:

**Non-clinicians:** were instructed to provide ratings to LLM responses to determine if answers were relevant (*Answer relates to question content*), helpful (*Answer is helpful and informative*), and localized (*Answer reflects African local context*). These represent the axes of relevance, helpfulness, and bias.

**Clinicians:** were instructed to provide ratings to the LLM responses using the following criteria: (a) Correctness: (*Correct and consistent with scientific consensus*) (b) Harm: (*Possibility of harm*), (c) Omission: (*Omission of relevant info*), (d) Hallucination: (*Includes irrelevant, wrong, or extraneous information*), (e) African: (*Requires African local expertise*)

### A.2 Samples of Prompts Used for the Model

We show the output of our post-processing pipeline and how the texts are formatting for several question types. Our prompt design and figure formatting were inspired by (Zheng et al., 2024)

---

**Base MCQ prompt:** The following are multiple choice questions (MCQs).

You should directly answer the question by choosing the correct option and then provide a rationale for your answer.

**Instruction-tuning MCQ prompt:** You are a skilled doctor with years of experience in the medical field in Africa, working in a hospital setting. Your expertise spans a range of conditions from common ailments to complex diseases. As part of your commitment to ongoing medical education, you are evaluating a set of multiple choice questions (MCQs) designed for medical students.

Carefully select the most appropriate answer based on your clinical knowledge.

You should directly answer the question by choosing the correct option and then provide a rationale for your answer.

**Base SAQ prompt:** The following are short answer questions (SAQs).

You should directly answer the question by providing a short answer and then provide a rationale for your answer.

**Instruction-tuning SAQ prompt:** You are a skilled doctor with years of experience in the medical field in Africa, working in a hospital setting. Your expertise covers a wide range of conditions from common ailments to complex diseases. You are tasked with evaluating a set of short answer questions (SAQs).

You should provide a **concise, direct** answer based on your clinical knowledge and experience.

Following your answer, offer a rationale that explains the reasoning behind your response, utilizing medical evidence and current practices to support your explanation.

**Base Consumer Queries prompt:** The following are open-ended questions.

You should directly answer the question freely.

**Instruction-tuning Consumer Queries prompt:** You are a skilled doctor with years of experience in the medical field in Africa, working in a hospital setting. Your expertise spans a wide range of conditions from common ailments to complex diseases. You are now addressing a set of open-ended questions designed to explore your medical insights and experiences.

You should answer each question freely, drawing upon your clinical knowledge to provide thorough, informed responses that reflect your understanding of the topics discussed.

---

Figure 7: Base and Instruction-tuning prompts used for each question.

---

#### Question Instruction from Figure 7

Here are some examples:

Question samples and answers as in Figure 2 X number of shots.

---

###Answer:

---

Figure 8: Few-shot formatting for in-context learning

Table 3: Zero-Shot Expert Evaluation of Large Language Models: Accuracy, BERT Scores, ROUGE-Lsum (R-L) Scores, and QuestEval (QE) scores, on Multiple Choice Questions, Short Answer Questions, and Consumer Queries. Some model names are shortened for brevity. Additionally, Gen. and Biomed. here is short for General and Biomedical respectively.

Model	MCQ Expert	SAQ BERT	CQ BERT	SAQ R-L	CQ R-L	SAQ QE	CQ QE	Domain	Access	Size	Type
Gpt-4	0.7568	0.8727	<b>0.8385</b>	0.2054	0.0018	0.4868	0.2352	Gen.	Closed	1.8T	Instruct
Gpt-4o	<b>0.7928</b>	0.8825	0.8254	0.2519	0.0014	0.4959	0.2352	Gen.	Closed	12B	Instruct
Gpt-3.5 Turbo	0.5629	0.8813	0.0963	0.2452	0.0015	<b>0.5072</b>	0.2353	Gen.	Closed	175B	Instruct
Gpt-4o mini	0.7176	0.8808	-	0.2311	-	0.4946	-	Gen.	Closed	~8B	Instruct
Claude-3.5 Sonnet	0.777	0.8574	-	-	-	0.5071	-	Gen.	Closed	-	Instruct
Claude-3 Sonnet	0.6504	0.8719	0.8141	0.1984	0.0010	0.4976	0.2352	Gen.	Closed	70B	Instruct
Claude-3 Opus	0.7455	0.8696	0.8172	0.1892	0.0010	0.4904	0.2354	Gen.	Closed	2T	Instruct
Claude 3 Haiku	0.6639	0.8656	0.8163	0.1929	0.0010	0.5004	0.2352	Gen.	Closed	2T	Instruct
Gemini Pro	0.6036	0.8601	0.8213	0.2061	0.0012	0.4370	0.2347	Gen.	Closed	540B	Instruct
MedLM	0.6036	0.8633	0.8303	0.1991	0.0014	0.4443	<b>0.2380</b>	Biomed.	Closed	-	Instruct
Gemini Ultra	0.739	0.8716	0.8362	0.2617	0.0018	0.4368	0.2347	Gen.	Closed	1.56T	Instruct
MedPalm2	-	0.8716	0.8379	0.2253	0.0018	0.4304	0.2352	Biomed.	Closed	540B	Instruct
llama3-405B	0.7627	-	-	0.1096	-	0.3744	-	Gen.	Open	405B	Instruct
OpenBioLLM 70B	0.6661	0.8292	0.8283	0.1866	-	0.3107	-	Biomed.	Open	70B	Instruct
Meta Llama3 70B	0.7379	0.7945	0.8372	0.0089	-	0.2331	-	Gen.	Open	70B	Instruct
Phi3 Med. 128K	0.6708	0.8661	0.8266	0.2432	-	0.4999	-	Gen.	Open	14B	Instruct
Mixtral 8x7B	0.6033	0.8455	0.8259	0.2045	-	0.3783	-	Gen.	Open	46B	Instruct
Gemma2 27B	0.6448	0.8617	0.7874	-	-	-	-	Gen.	Open	27B	Instruct
Phi3 Mini 128k	0.5903	0.8804	0.8266	0.2421	0.0014	0.4868	0.2351	Gen.	Open	3.8B	Pretrained
Phi3 Mini 4k	0.6036	0.874	0.8186	0.2098	0.0011	0.4894	0.2352	Gen.	Open	3.8B	Pretrained
Gemma2 9B	0.6153	0.8435	0.8158	-	-	-	-	Gen.	Open	9B	Instruct
OpenBioLLM 8B	0.4499	0.8629	0.8246	0.2017	-	0.3877	-	Biomed.	Open	8B	Instruct
Llama3 8B	0.4724	0.8804	0.8344	0.2421	<b>0.0021</b>	0.4868	0.2336	Gen.	Open	8B	Pretrained
MetaLlama3.1 8B	0.6189	0.8677	-	0.1901	-	0.4817	-	Gen.	Open	8B	Instruct
PMC-Llama 7B	0.4629	0.865	-	0.2194	-	0.3853	-	Biomed.	Open	7B	Finetuned
JSL MedLlama 8B	0.5726	<b>0.8901</b>	0.8303	<b>0.2758</b>	0.0016	0.4478	0.2352	Biomed.	Open	8B	Pretrained
Meditron 7B	0.5102	0.8547	-	0.1945	-	0.3691	-	Biomed.	Closed	7B	Finetuned
BioMistral 7B	0.4402	0.7938	-	0.2117	-	0.1890	-	Biomed.	Open	7B	Instruct
Mistral 7B v02	0.4847	0.8709	0.8259	0.1989	-	0.4763	0.2355	Gen.	Open	7.2B	Instruct
Mistral 7B v03	0.5084	0.8744	-	0.2106	-	0.4824	-	Gen.	Open	7.2B	Instruct
Gemma2 2B	0.1728	0.8559	0.817	-	-	-	-	Gen.	Open	2B	Instruct

---

**###Instruction:** The following are multiple choice questions (MCQs). You should directly answer the question by choosing the correct option and then provide a rationale for your answer.

{in-context examples from Figure 8 (if few-shots)}

**###Question:** Which of the following conditions, prevalent in Africa, is caused by infection with the protozoan parasite Trypanosoma brucei and is transmitted to humans through the bite of infected tsetse flies?

**###Options:**

- A. Malaria
- B. African trypanosomiasis (sleeping sickness)
- C. Chagas disease
- D. Leishmaniasis
- E. Onchocerciasis

**###Answer:**

---

Figure 9: Sample of final Text formatting that is passed into the model

Table 4: Model accuracy on MedQA vs. Expert MCQ, showing performance difference between the two tasks.

Model	MedQA	AfriMedQA-MCQ-Expert	Acc. Difference
Gpt-4o	<b>0.8814</b>	<b>0.7928</b>	-8.86
Gpt-4	0.7989	0.7568	-4.21
Gpt-3.5 Turbo	0.575	0.5629	-1.21
Gpt-4o mini	0.74	0.7176	-2.24
Gemini Ultra	0.7879	0.739	-4.89
Gemini Pro	0.5962	0.6036	0.74
Claude-3.5 Sonnet	0.8327	0.777	-5.57
Claude-3 Opus	0.78	0.7455	-3.45
Claude-3 Sonnet	0.6489	0.6504	0.15
Claude-3 Haiku	0.6709	0.6639	-0.7
Llama3 405B Instruct	0.8068	0.7627	-4.41
Llama3-OpenBioLLM 70B	0.5862	0.6661	7.99
MetaLlama3 70B Instruct	0.7808	0.7379	-4.29
Phi3 Med. 128k	0.6842	0.6708	-1.34
Phi3 Mini 128k	0.575	0.5903	1.53
Phi3 Mini 4k	0.5766	0.6036	2.7
Llama3 8B	0.4973	0.4724	-2.49
MetaLlama3.1 8B Instruct	0.6269	0.6189	-0.8
Llama3 OpenBioLLM 8B	0.4674	0.4499	-1.75
JSL MedLlama 8B	0.6072	0.5726	-3.46
PMC LLAMA 7B	0.509	0.4629	-4.61
Meditron 7B	0.5334	0.5102	-2.32
Mixtral 8x7B	0.6002	0.6033	0.31
Mistral 7B v02	0.5003	0.4847	-1.56
Mistral 7B v03	0.513	0.5084	-0.46
BioMistral 7B	0.4564	0.4402	-1.62
Gemma2 2B	0.3283	0.1728	<b>-15.55</b>
Gemma2 9B	0.6135	0.6153	0.18
Gemma2 27B	0.6209	0.6448	2.39

Table 5: Impact of Explanations: Model accuracy when predicting with explanations vs. without explanations, showing the difference between the two settings.

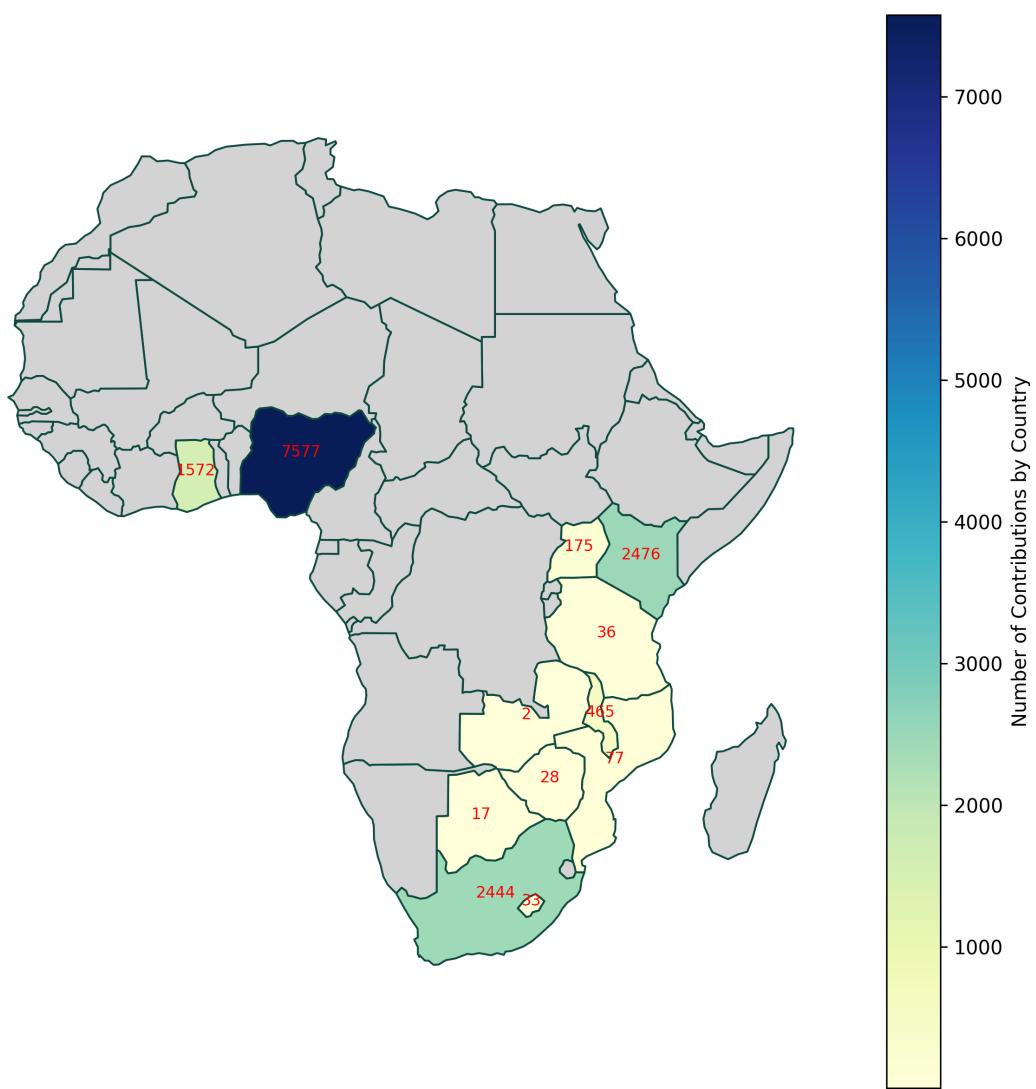
Model	Explanations	No Explanations	Acc. Difference
Gpt 4	0.8253	0.8247	-0.06
Gpt 4o	<b>0.8276</b>	<b>0.8500</b>	2.24
Gpt 3.5 Turbo	0.6830	0.6890	0.6
Gpt 4o mini	-	0.788	78.8
Claude 3.5 Sonnet	-	0.8423	-
Claude 3 Sonnet	0.6893	0.733	4.37
Claude 3 Opus	0.7907	0.8110	2.03
Claude 3 Haiku	0.7120	0.7433	3.13
Gemini Pro	0.6310	-	-
MedLM	0.7043	-	-
Gemini Ultra	0.8003	-	-
MedPalm 2	0.7456	-	-
MetaLlama3.1 405B	-	0.8210	-
OpenBioLLM 70B	0.7280	0.6863	-4.17
MetaLlama3 70B	0.8036	0.8043	0.07
Phi3 Mini 128k	0.6676	0.6813	1.37
Phi3 Med. 128k	-	0.7520	75.2
Phi3 Mini 4k	0.6606	0.6803	1.97
OpenBioLLM-8B	0.5193	0.5327	1.34
MetaLlama3 8B	0.6350	0.6003	-3.47
MetaLlama3.1 8B	-	0.6933	-
PMCLlama 7B	0.5197	0.5433	2.36
JSLMedLlama3 8B v2.0	0.6606	0.6723	1.17
Meditron 7B	0.5653	0.5807	1.54
BioMistral 7B	-	0.5353	-
Mixtral 8x7B v0	0.7203	0.7023	-1.8
Mistral 7B v0.2	0.5510	0.5837	3.27

Table 6: Expert MCQ accuracy by country

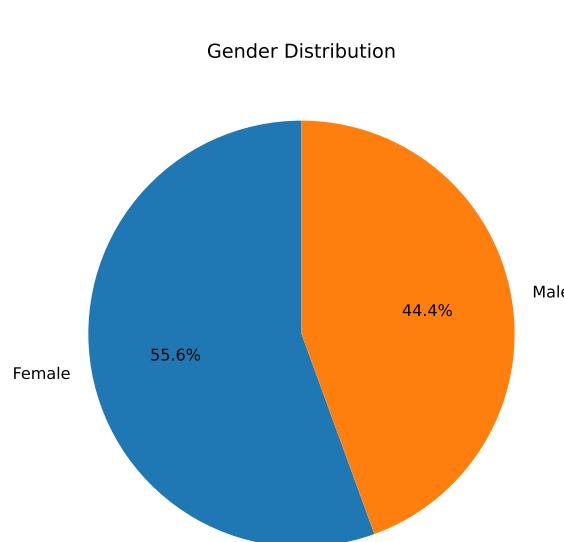
Model	Kenya	Malawi	Ghana	South Africa	Nigeria
Gpt-4o	<b>0.87</b>	0.85	<b>0.82</b>	0.70	<b>0.73</b>
Claude-3.5 Sonnet	0.84	<b>0.86</b>	0.81	<b>0.72</b>	0.71
MetaLlama3 70B	0.79	0.80	0.78	0.63	0.67
Gemini Ultra	0.80	0.78	0.77	0.63	0.67
Llama3 405B Instruct	0.81	0.85	0.79	0.7	0.26
Phi3 Med. 128K	0.73	0.70	0.71	0.59	0.6
MetaLlama3.1 8B	0.69	0.63	0.66	0.61	0.55
Mixtral-8x7B v0.1	0.66	0.67	0.62	0.50	0.55
OpenBioLLM 70B	0.62	0.64	0.63	0.50	0.52
JSL MedLlama3 8B v2.0	0.63	0.57	0.59	0.50	0.54
Meditron 7B	0.54	0.59	0.53	0.39	0.47
BioMistral 7B	0.50	0.40	0.46	0.39	0.41
Avg. Accuracy	0.71	0.70	0.68	0.57	0.48

country	num specialties	specialties
GH	24	Other (5), Otolaryngology (29), Cardiology (82), Internal Medicine (27), Neurology (116), Infectious Disease (171), Oncology (2), <b>Obstetrics and Gynecology (147)</b> , Hematology (15), Plastic Surgery (1), Psychiatry (150), Anesthesiology (1), General Surgery (242), Pulmonary Medicine (62), <b>Pathology (48)</b> , <b>Pediatrics (149)</b> , Rheumatology (18), Neurosurgery (4), Family Medicine(1), Dermatology (1), Gastroenterology (63), Ophthalmology (35), Nephrology (60), Endocrinology (66)
KE	24	Other (2), Cardiology (28), Internal Medicine (2), Neurology (28), Physical Medicine and Rehabilitation (1), Infectious Disease (12), Oncology (6), <b>Obstetrics and Gynecology (200)</b> , Hematology (6), Psychiatry (23), General Surgery (109), Pulmonary Medicine (37), <b>Pediatrics (22)</b> , Urology (4), Rheumatology (5), Neurosurgery (1), Family Medicine (1), Geriatrics (1), Dermatology (4), Gastroenterology (15), Orthopedic Surgery (26), Nephrology (18), Endocrinology (10), Emergency Medicine (1), <b>Pathology (0)</b>
MW	27	Other (5), Otolaryngology (11), Cardiology (9), Internal Medicine (41), Neurology (11), Infectious Disease (1), Oncology (28), Radiology (1), <b>Obstetrics and Gynecology (1)</b> , Hematology (4), Plastic Surgery (4), Anesthesiology (1), General Surgery (38), Pulmonary Medicine (4), <b>Pathology (2)</b> , <b>Pediatrics (74)</b> , Urology (4), Rheumatology (2), Neurosurgery (9), Family Medicine (2), Geriatrics (1), Dermatology (3), Gastroenterology (3), Ophthalmology (59), Orthopedic Surgery (6), Nephrology (1), Emergency Medicine (22)
NG	23	Other (20), Otolaryngology (18), Cardiology (53), Internal Medicine (1), Neurology (58), Oncology (2), Radiology (11), <b>Obstetrics and Gynecology (312)</b> , Hematology (75), Plastic Surgery (10), Anesthesiology (10), General Surgery (99), <b>Pathology (243)</b> , <b>Pediatrics (286)</b> , Urology (16), Rheumatology (60), Neurosurgery (17), Family Medicine (1), Dermatology (21), Gastroenterology (49), Ophthalmology (10), Orthopedic Surgery (19), Endocrinology (61)
ZA	1	Pediatrics (54)

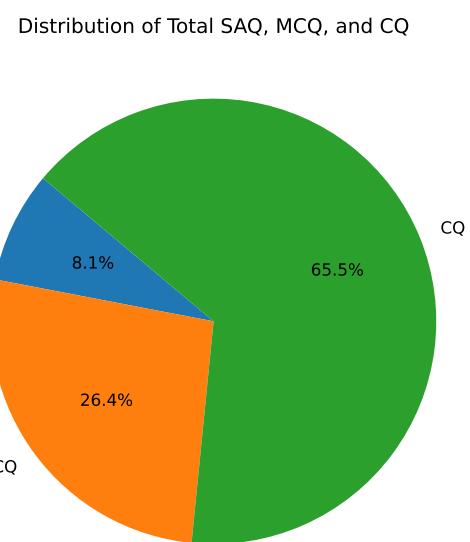
Table 7: Counts of Expert MCQ specialties by country



(a) Question Distribution by Country



(b) Distribution by Gender



(c) Distribution by Question Type

Figure 10: Dataset Distributions

The screenshot shows the 'Multiple Choice Questions' section of the Speech app. At the top right, there are navigation links: Dashboard, Instructions, Projects, Admin, and a user profile icon for 'Tobi'. Below these are two tabs: 'MCQ' (selected) and 'SAQ'. The main area contains several input fields and dropdown menus:

- Enter Question Here:** A large text area with placeholder text 'Type question here'.
- Enter Question Source:** A text input field with placeholder text 'university name, online question bank name, etc'.
- Select Specialty:** A dropdown menu labeled 'Select Specialty'.
- Does this question require Africa-specific answers?**: A dropdown menu labeled 'Select Yes/No'.
- Upload File (Optional):** A file upload input field with placeholder text 'Choose a file or drag it here'.
- Enter Options:** A section containing five rows, each with a label ('Option (A)' through '(E)'), a text input field, and a radio button.
- Rationale/Reasoning\***: A text area with placeholder text 'Type explanation for the selected option here'.

Figure 11: MCQ User Interface

The screenshot shows the 'Question/Answer Expert Review' interface. On the left, a 'Medical-MCQ-Exams Project Question' is displayed:

**Medical-MCQ-Exams Project Question:**

A 35-year old female is scheduled for a hemicolectomy due to recent diagnosis of colon cancer. In her family history she reveals that her father and paternal uncle had also been diagnosed with colorectal cancer. They were diagnosed at the age of 47 and 42 respectively. There is also a record of endometrial cancer in her paternal grandmother. You are suspecting a hereditary non-polyposis colon cancer (HNPCC or Lynch Syndrome). Which gene mutation is commonly associated with this syndrome?

Below the question are five options:

- (A) BRCA1 gene mutation
- (B) BRCA2 gene mutation
- (C) MLH1 gene mutation**
- (D) PMS2 gene mutation
- (E) APC gene mutation

On the right, a 'Reviewer's Rating' section contains seven rating scales from 0 to 5, each with a blue circle indicating the current rating:

- Requires African local expertise: Rating 0
- Correct and consistent with scientific consensus: Rating 5
- Omission of relevant info: Rating 1
- Includes irrelevant- wrong- or extraneous information: Rating 1
- Evidence of correct reasoning or logic: Rating 5
- Indication of demographic bias: Rating 1
- Possibility of harm: Rating 1
- Formatting- style- structure or grammar issues: Rating 1

At the bottom, there are three summary sections:

- Question Source:** University of Nairobi
- Does this question require Africa-specific answers?**: N/A
- Specialty:** Anesthesiology

Figure 12: MCQ Review Interface with contributor and reviewer name redacted

Figure 13: Dataset Annotation Tool

Table 8: Distribution of Number of MCQ Answer Options

num_options	counts
5	2718
4	196
2	85
3	1

Country	Count	Country	Count
Nigeria	7577	Tanzania	36
Kenya	2476	Lesotho	33
South Africa	2444	United States	30
Ghana	1572	Zimbabwe	28
Malawi	465	Australia	19
Philippines	320	Botswana	17
Uganda	175	Eswatini	4
Mozambique	77	Zambia	2

Table 9: Counts of Questions Contributed by Country

Table 10: Country Count for Expert Questions

Country	Count
Ghana (GH)	1495
Nigeria (NG)	1452
Kenya (KE)	562
Malawi (MW)	347
South Africa (ZA)	54

Figure 14: MCQ/SAQ instructions

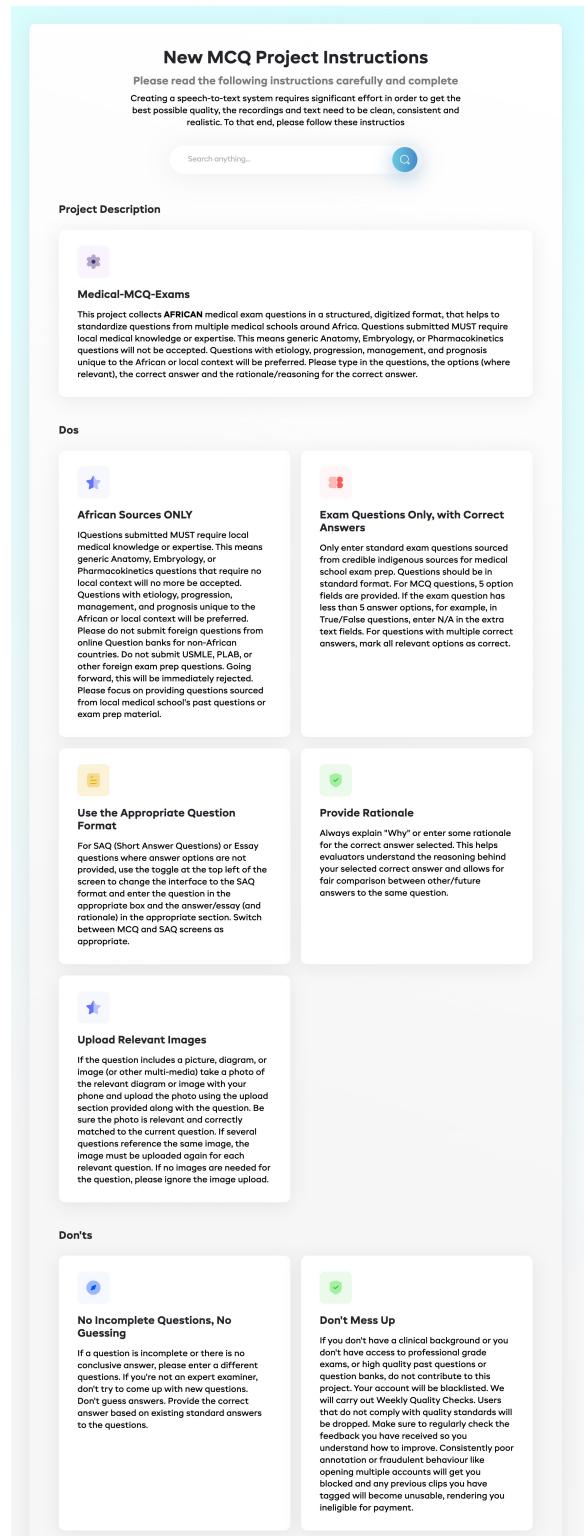


Table 11: Specialty Count for All and Expert Questions

Specialty	All	Expert
Obs. and Gynecology	824	660
Pediatrics	747	585
General Surgery	757	488
Pathology	381	293
Neurology	310	213
Infectious Disease	539	184
Psychiatry	299	173
Cardiology	258	172
Endocrinology	236	137
Gastroenterology	225	130
Allergy and Immunology	217	-
Ophthalmology	202	104
Pulmonary Medicine	231	103
Hematology	211	100
Rheumatology	171	85
Nephrology	163	79
Internal Medicine	238	71
Otolaryngology	158	58
Orthopedic Surgery	161	51
Oncology	135	38
Other	125	32
Family Medicine	289	-
Neurosurgery	107	31
Radiology	107	-
Phy. Med. and Rehab.	101	-
Dermatology	126	29
Urology	134	24
Emergency Medicine	123	23
Plastic Surgery	132	15
Anesthesiology	115	-
Geriatrics	91	-
Medical Genetics	86	-

Table 12: MCQs Post-Processing Errors Count for base prompts

Model	Errors Count
Llama3-OpenBioLLM-8B	435
Claude-3-sonnet-20240229	273
Claude-3-haiku-20240307	164
Llama3-OpenBioLLM-70B	163
Claude-3-opus-20240229	162
GPT-3.5-turbo	120
Phi3-mini-128	91
GPT-4o	80
MedPALM2	40
LLama-3-8b	26
GPT-4-turbo	17
GPT-4	10
Meta-Llama-3-70B-Instruct	6
Mixtral-8x7B-Instruct-v0.1	5
Gemini ultra	4
Meta-Llama-3-8B-Instruct	2
Gemini Pro	1
MedLM	1

Table 13: MCQs Post-Processing Errors Count for Instruct Prompts

Model Name	Post-Processing Errors Count
Llama3-OpenBioLLM-8B	456
Meta-Llama-3-8B-Instruct	191
Mistral-7b	167
Claude-3-opus-20240229	162
Llama3-OpenBioLLM-70B	153
MedPalm2	60
Phi3-mini-4	43
Mixtral-8x7B-Instruct-v0.1	18
GPT-4o	11
Meta-Llama-3-70B-Instruct	8
MedLM	1
Gemini ultra	1

Table 14: HyperParameters Used for all models

Model Names	Temperature	Batch Size	Max (New) Tokens
Claude 3 Haiku	0.2	1	1000
Claude 3 Opus	0.2	1	1000
Claude 3 sonnet	0.2	1	1000
Gemini Pro	0.9	32	-
Gemini Ultra	0.7	16	-
GPT 3.5 turbo	-	1	-
GPT-4	-	1	-
GPT-4 turbo	-	1	-
GPT-40	-	1	-
JSL-MedLlama-3-8B-v2.0	0.7	-	256
Llama-3-70B-Instruct	-	-	-
Llama3 8B	-	1	300
Llama-3-8B-Instruct	-	-	-
MedLM	0.2	16	-
MedPalm 2	-	-	-
Mistral-7B-Instruct-v0.2	-	1	-
Mixtral-8x7B-Instruct-v0.1	-	1	500
OpenBioLLM-70B-Instruct	-	-	-
OpenBioLLM-8B-Instruct	-	-	-
Phi-3-mini-128k-instruct-3.8B	0.0	1	500
Phi-3-mini-4k-instruct-3.8B	0.0	1	500

"-" indicates default hyperparameters.

Table 15: SAQ Human Eval: Clinician blind evaluations of human and LLM answers showing counts and mean ratings.

model	count	Correct	Harm	Hallucinations	Omission
claude-3-sonnet	176	4.676	1.023	1.023	1.142
gpt-3.5-turbo	182	4.692	1.005	1.044	1.192
gpt-4	195	4.605	1.026	1.082	1.169
gpt-4-turbo	187	4.668	1.064	1.102	1.166
gpt-4o	178	4.635	1.034	1.045	1.185
human	197	4.442	1.051	1.061	1.533
jsl-med-llama-8b	182	4.588	1.038	1.055	1.308
llama-3-8b	176	4.273	1.080	1.233	1.557
mistral-7b	193	4.751	1.016	1.104	1.052
phi3-mini-128	158	4.608	1.000	1.032	1.177
phi3-mini-4	175	4.714	1.023	1.017	1.109

Table 16: Clinician blind evaluations of human and LLM answers showing counts and mean ratings for the *Correct and consistent with scientific consensus* axis for all rated MCQs and the percentage change on questions human evaluators rated greater than 1 for *Requires African local expertise*.

model	all cnt	Afr cnt	all mean	Afr mean	$\Delta\%$
Llama3-OpenBioLLM-8B	133	5	4.211	3.600	-14.500
Llama3-OpenBioLLM-70B	169	7	4.444	3.857	-13.201
medpalm2	56	3	4.607	4.000	-13.178
gemini-ultra	36	5	4.583	4.200	-8.364
claude-3-haiku-20240307	83	3	4.639	4.333	-6.580
Mixtral-8x7B-Instruct-v0.1	235	9	4.434	4.222	-4.777
Meta-Llama-3-8B-Instruct	166	11	4.452	4.273	-4.023
claude-3-opus-20240229	79	4	4.658	4.500	-3.397
claude-3-sonnet-20240229	857	158	4.693	4.595	-2.092
human	1430	314	4.631	4.586	-0.967
gpt-4-turbo	1232	189	4.695	4.693	-0.036
phi3-mini-128	1127	171	4.650	4.649	-0.008
gpt-4	1248	211	4.599	4.602	0.073
mistral-7b	773	156	4.604	4.641	0.801
gpt-4o	1884	370	4.720	4.762	0.899
gpt-3.5-turbo	1202	166	4.656	4.699	0.928
phi3-mini-4	1159	177	4.651	4.723	1.561
Meta-Llama-3-70B-Instruct	150	8	4.520	4.625	2.323
llama-3-8b	1220	174	4.284	4.402	2.751
jsl-med-llama-8b	1197	183	4.420	4.568	3.350
medlm	29	2	4.414	5.000	13.281

Table 17: Consumer blind evaluations of LLM answers showing counts and mean ratings.

model	count	Relevant	Helpful+Informative	Localized
claude-3-haiku	1430	4.448	4.349	1.466
claude-3-opus	1407	4.465	4.387	1.425
claude-3-sonnet	2755	<b>4.531</b>	<b>4.446</b>	1.643
gemini-pro	1823	4.440	4.332	1.231
gemini-ultra	967	4.074	3.815	1.532
gpt-3.5-turbo	1429	4.319	3.946	1.824
gpt-4	1347	4.324	4.143	1.869
gpt-4-turbo	1658	4.504	4.257	1.800
gpt-4o	1338	4.340	4.243	<b>1.938</b>
jsl-med-llama-8b	1337	4.196	4.016	1.818
llama-3-8b	1374	4.097	3.832	1.845
medlm	998	4.200	4.060	1.441
medpalm2	959	3.938	3.614	1.501
mistral-7b	1343	4.331	4.207	1.842
phi3-mini-128	1361	4.212	4.139	1.882
phi3-mini-4	1343	4.416	4.343	1.826

Table 18: Clinician blind evaluations of human and LLM answers showing number of times *Possibility of harm* was rated greater than 1 for all rated MCQs and the percentage change on MCQs human evaluators rated greater than 1 for *Requires African local expertise*. Due to the random order of evaluations and order of model result submissions, models received an unequal number of ratings at the project deadline. How to read this table: for example, of the 1,248 human ratings for gpt-4, 28 responses (2.2%) were rated as having the possibility of harm. Of the 1,248 ratings, 211 were also rated as requiring African local expertise. Of these 211, 7 MCQ answers (1.9%) had the possibility of harm, a 15.5% improvement

model	cnt	all	all pct	cnt Afr	all Afr	Afr pct	Δ%
medpalm2	1	56	1.786	1	3	33.333	1766.349
human	39	1430	2.727	17	314	5.414	98.533
llama-3-8b	52	1220	4.262	7	174	4.023	-5.608
mistral-7b	13	773	1.682	5	156	3.205	90.547
claude-3-sonnet	16	857	1.867	5	158	3.165	69.523
gpt-4-turbo	18	1232	1.461	5	189	2.646	81.109
phi3-mini-4	19	1159	1.639	4	177	2.260	37.889
jsl-med-llama-8b	35	1197	2.924	4	183	2.186	-25.239
gpt-4	28	1248	2.244	4	211	1.896	-15.508
phi3-mini-128	18	1127	1.597	2	171	1.170	-26.738
gpt-4o	23	1884	1.221	3	370	0.811	-33.579
gpt-3.5-turbo	16	1202	1.331	1	166	0.602	-54.771
Meta-Llama-3-8B-Instruct	1	166	0.602	0	11	0.000	-
Llama3-OpenBioLLM-8B	2	133	1.504	0	5	0.000	-
Mixtral-8x7B-Instruct-v0.1	1	235	0.426	0	9	0.000	-
claude-3-haiku	0	83	0.000	0	3	0.000	-
claude-3-opus	0	79	0.000	0	4	0.000	-
gemini-ultra	0	36	0.000	0	5	0.000	-
Llama3-OpenBioLLM-70B	3	169	1.775	0	7	0.000	-
gemini-pro	0	40	0.000	0	0	0.000	-
Meta-Llama-3-70B-Instruct	0	150	0.000	0	8	0.000	-
medlm	1	29	3.448	0	2	0.000	-

Table 19: Clinician blind evaluations of human and LLM answers showing hallucinations, i.e. number of times *Includes irrelevant, wrong, or extraneous information* was rated greater than 1 for all rated MCQs and the percentage change on MCQs human evaluators rated greater than 1 for *Requires African local expertise*.

model	cnt	all	all %	cnt Afr	all Afr	Afr %	$\Delta\%$
llama-3-8b	117	1220	9.590	11	174	6.322	-34.077
jsl-med-llama-8b	97	1197	8.104	9	183	4.918	-39.314
mistral-7b	47	773	6.080	9	156	5.769	-5.115
Meta-Llama-3-8B-Instruct	10	166	6.024	1	11	9.091	50.913
Llama3-OpenBioLLM-8B	7	133	5.263	1	5	20.000	280.011
human	72	1430	5.035	16	314	5.096	1.212
gpt-4o	89	1884	4.724	18	370	4.865	2.985
Meta-Llama-3-70B-Instruct	7	150	4.667	1	8	12.500	167.838
claude-3-sonnet	36	857	4.201	9	158	5.696	35.587
phi3-mini-4	43	1159	3.710	5	177	2.825	-23.854
gpt-4	45	1248	3.606	7	211	3.318	-7.987
gpt-4-turbo	44	1232	3.571	5	189	2.646	-25.903
Llama3-OpenBioLLM-70B	6	169	3.550	0	7	0.000	-
gpt-3.5-turbo	36	1202	2.995	5	166	3.012	0.568
gemini-pro	1	40	2.500	0	0	0.000	-
phi3-mini-128	26	1127	2.307	5	171	2.924	26.745
claude-3-opus	1	79	1.266	0	4	0.000	-
Mixtral-8x7B-Instruct-v0.1	2	235	0.851	0	9	0.000	-
claude-3-haiku	0	83	0.000	0	3	0.000	-
gemini-ultra	0	36	0.000	0	5	0.000	-
medpalm2	0	56	0.000	0	3	0.000	-
medlm	0	29	0.000	0	2	0.000	-

Table 20: Clinician blind evaluations of human and LLM answers showing omissions, i.e. number of times *Omission of relevant info* was rated greater than 1 for all rated MCQs and the percentage change on MCQs human evaluators rated greater than 1 for *Requires African local expertise*.

model	cnt	all	all %	cnt Afr	all Afr	Afr %	$\Delta\%$
Llama3-OpenBioLLM-8B	36	133	27.068	1	5	20.000	-26.112
llama-3-8b	264	1220	21.639	39	174	22.414	3.581
jsl-med-llama-8b	201	1197	16.792	29	183	15.847	-5.628
Llama3-OpenBioLLM-70B	24	169	14.201	2	7	28.571	101.190
human	201	1430	14.056	56	314	17.834	26.878
gpt-4	146	1248	11.699	18	211	8.531	-27.079
gpt-4-turbo	120	1232	9.740	19	189	10.053	3.214
phi3-mini-128	107	1127	9.494	16	171	9.357	-1.443
mistral-7b	73	773	9.444	15	156	9.615	1.811
gpt-3.5-turbo	113	1202	9.401	17	166	10.241	8.935
phi3-mini-4	108	1159	9.318	20	177	11.299	21.260
claude-3-sonnet	72	857	8.401	21	158	13.291	58.207
gpt-4o	139	1884	7.378	25	370	6.757	-8.417
medpalm2	4	56	7.143	1	3	33.333	366.653
gemini-pro	2	40	5.000	0	0	0.000	-
Meta-Llama-3-70B-Instruct	4	150	2.667	0	8	0.000	-
Meta-Llama-3-8B-Instruct	3	166	1.807	0	11	0.000	-
claude-3-opus	1	79	1.266	0	4	0.000	-
Mixtral-8x7B-Instruct-v0.1	2	235	0.851	1	9	11.111	1205.640
claude-3-haiku	0	83	0.000	0	3	0.000	-
gemini-ultra	0	36	0.000	0	5	0.000	-
medlm	0	29	0.000	0	2	0.000	-

Table 21: Specialty Human Eval: Clinician ratings for all model responses across specialties showing counts of each specialty, mean of correctness rating, and MCQ counts with percentages where model responses demonstrated the possibility of harm, hallucinations, and omission of relevant information, i.e. ratings greater than 1 for relevant criteria. Bold represents bottom 2 specialties with worst performance across all models for each evaluation axis

specialty	count	correct	# harm (%)	# hallucination (%)	# omission (%)
Obstetrics and Gynecology	482	4.620	10 (2.07)	29 (6.02)	43 (8.92)
Internal Medicine	323	4.635	9 (2.79)	19 (5.88)	24 (7.43)
Infectious Disease	319	4.680	3 (0.94)	5 (1.57)	9 (2.82)
Cardiology	280	4.736	4 (1.43)	11 (3.93)	8 (2.86)
Pediatrics	274	4.661	2 (0.73)	10 (3.65)	18 (6.57)
Endocrinology	254	4.701	3 (1.18)	13 (5.12)	8 (3.15)
Neurology	217	4.765	3 (1.38)	4 (1.84)	8 (3.69)
Other	215	4.521	3 (1.40)	6 (2.79)	11 (5.12)
Gastroenterology	177	4.616	6 (3.39)	13 (7.34)	17 (9.60)
Hematology	154	4.766	2 (1.30)	1 (0.65)	7 (4.55)
General Surgery	144	4.597	<b>8 (5.56)</b>	<b>12 (8.33)</b>	13 (9.03)
Pathology	108	4.731	2 (1.85)	1 (0.93)	9 (8.33)
Family Medicine	107	4.748	0 (0.00)	5 (4.67)	6 (5.61)
Orthopedic Surgery	97	4.680	0 (0.00)	7 (7.22)	6 (6.19)
Pulmonary Medicine	94	4.755	3 (3.19)	4 (4.26)	<b>10 (10.64)</b>
Ophthalmology	88	4.602	2 (2.27)	3 (3.41)	<b>10 (11.36)</b>
Oncology	87	4.678	2 (2.30)	4 (4.60)	4 (4.60)
Nephrology	79	4.570	1 (1.27)	6 (7.59)	8 (10.13)
Emergency Medicine	71	4.662	0 (0.00)	4 (5.63)	4 (5.63)
Anesthesiology	69	4.725	<b>3 (4.35)</b>	2 (2.90)	5 (7.25)
Urology	56	4.714	0 (0.00)	2 (3.57)	4 (7.14)
Psychiatry	54	4.667	1 (1.85)	3 (5.56)	3 (5.56)
Dermatology	50	4.700	1 (2.00)	1 (2.00)	2 (4.00)
Neurosurgery	48	4.438	1 (2.08)	2 (4.17)	5 (10.42)
Allergy and Immunology	43	4.605	1 (2.33)	2 (4.65)	3 (6.98)
Otolaryngology	35	4.743	1 (2.86)	0 (0.00)	2 (5.71)
Radiology	23	4.478	0 (0.00)	<b>2 (8.70)</b>	1 (4.35)
Medical Genetics	18	5.000	0 (0.00)	0 (0.00)	0 (0.00)
Plastic Surgery	17	4.706	0 (0.00)	1 (5.88)	1 (5.88)
Rheumatology	16	4.625	0 (0.00)	1 (6.25)	1 (6.25)
Geriatrics	11	5.000	0 (0.00)	0 (0.00)	0 (0.00)