Intron AfriHealth MultiBench: Multilingual Medical Speech Benchmark for the Global South

Submitted to: AI for Global Health Benchmarking Initiative — The Agency Fund & Endless Health **Organization:** Intron Health **Contact:** Tobi Olatunji, MD MSc — tobi@intron.io | www.intron.io

---

1. Overview

Organization Name: Intron Health (RC1653956)

Use Case / Problem Being Addressed: Existing AI benchmarks systematically underrepresent accented and multilingual speech from Africa and the broader Global South. This gap means that ASR and LLM models deployed in African healthcare settings have never been rigorously evaluated on the actual speech patterns of the clinicians and community health workers who use them. Intron AfriHealth MultiBench addresses this by providing the first comprehensive, real-world multilingual medical speech benchmark covering three task types: transcription (ASR), translation, and spoken clinical question-answering (QA).

Target Population: Healthcare workers across Nigeria, Ghana, Kenya, Uganda, Rwanda, and South Africa — including physicians, nurses, and community health workers (CHWs). The QA subset specifically targets Nigerian CHWs and medical staff (licensed professionals, residents, consultants, and interns), with speakers aged 18–55 and representation across African local language accents. The transcription and translation subsets represent approximately 600 speakers from 10+ African countries with ≥40% female representation and an urban–rural split of approximately 60:40.

Deployment Stage: Research / benchmarking (data drawn from Intron's production transcription app and CHW clinical query platform)

---

2. Dataset

2.1 Data Source and Collection

Data are drawn from two primary sources:

1. **Intron's live transcription application** — used daily by consenting healthcare workers across Nigeria, Ghana, Kenya, Uganda, Rwanda, and South Africa. All participants provided consent permitting research use of de-identified recordings.
2. **CHW clinical query platform** (spoken QA) — real spoken queries submitted by community health workers and medical staff via Intron's app, collected under the Bayelsa-CHEW-Health-Questions and related projects.

Data processing follows Intron's ISO 27001-aligned governance pipeline with automated PII scrubbing and human verification.

Task	**Primary Source(s)**	**Release Status**
Transcription	Transcribed Production Monitoring, Afrispeech Dialogue, Med-Conv-Nig	Partial open-source release
Translation	AfriVox Translate (Intron-MT)	Full open-source release
Spoken QA	CHEWs dataset (spoken_qa_meta_data)	Private — not released

2.2 Data Composition

Total size: ~5,200 instances across three tasks; approximately 20 hours of audio from ~600 speakers

Task	**Instances**	**Audio Duration**	**Speakers**	**Languages/Accents**
Transcription	3,200	~10 hours	250	40+ English accents + 17 languages
Translation	1,600	~5 hours	120	16 languages
Spoken QA	398 audio recordings (385 unique questions)	~15.6 hours	—	4 languages

Modality: Multimodal — audio input with text output (transcription, translation, QA).

Languages covered (19): Accented English (West, East, and Southern African variants; 40+ accents), Afrikaans, Akan, Amharic, Arabic, African French, Hausa, Igbo, Kinyarwanda, Luganda, Nigerian Pidgin, Pedi, Sesotho, Shona, Swahili, Tswana, Twi, Xhosa, Zulu.

Spoken QA dataset details (spoken_qa_meta_data, n=398):

Field	Distribution
Languages	English (100), Hausa (100), Yoruba (100), Pidgin (98)
Difficulty	Medium (248), Hard (112), Easy (38)
Gender	Male (211), Female (187)
Country	Nigeria (398)
Discipline	Medicine (398)
Clinical experience	Licensed Professional (205), Resident (90), Consultant (52), Intern (51)
Age group	26–40 (253), 19–25 (79), 41–55 (66)
Speaker accents	Yoruba (83), Hausa (67), Yoruba/Pidgin (58), Igbo (50), Igala (42), Pidgin (21), Ijaw (20), Bette (20), others

Top clinical categories in spoken QA: Puberty/Sexual Health, Skin, Ear-Nose-Throat, Abdomen, Postnatal, Genito-urinary, Nutrition, STI/HIV, Emergencies (children 2–60 months), and others.

Translation metadata (meta_data_translation.csv): 1,600 instances with fields: id, transcription (source text), translation (reference English), language, speaker_id (hashed), gender, audio duration, source label, and audio paths.

Transcription metadata (meta_data_transcription.csv): 3,200 instances with fields: audio path, duration, reference text, language, source label, and root directory.

Demographic distribution axes: country, language, gender, age group, clinical role, accent, health system tier, urban/rural setting.

Synthetic augmentation: None; all data are real-world recordings from consenting healthcare workers.

2.3 Limitations

- **Geographic concentration:** The spoken QA subset is entirely from Nigeria; transcription and translation cover broader geographies but are still weighted toward Anglophone West and East Africa.
- **Language community bias:** The spoken QA subset covers only Nigerian local languages (Yoruba, Hausa, Igbo, Pidgin, Ijaw, Igala, and others). African languages from East, Central, and Southern Africa — such as Kinyarwanda, Luganda, Swahili, Shona, and Zulu — are not represented in the QA task, limiting generalisability of QA results to non-Nigerian African language communities.
- **Professional bias:** All QA speakers are medical practitioners; community health workers without formal medical training are not represented.
- **Technology comfort bias:** Participants are users of Intron's app and therefore more technology-comfortable than the average CHW population.
- **Code-switching:** Intra-utterance code-switching is present but inconsistently annotated across languages.
- **Orthographic instability:** WER/CER penalise variation in low-resource languages where spelling conventions are not fully standardised.

2.4 Data Processing and PII

PII in this use case includes speaker names, facility names, patient identifiers, and location references embedded in clinical speech. Removal steps:

1. Automated PII scrubbing using Intron's privacy filter
2. Human verification by native-speaking medical annotators
3. ISO 27001-aligned governance pipeline throughout

Annotator names and emails present in the raw spoken_qa_meta_data are retained only in the private dataset and are not included in any public release. Text normalisation follows WHO terminology standards. Speech segments are aligned at the utterance level using Intron's alignment tool.

2.5 Unit of Evaluation

Task	**Evaluation Unit**
Transcription	Individual audio utterance → reference transcription pair
Translation	Individual audio/text utterance → reference translation pair

Task	Evaluation Unit
Spoken QA	Individual spoken question (with scenario context) → model text answer, scored against human reference answer

2.6 Data Structure

Spoken QA metadata schema (spoken_qa_meta_data.csv — key fields):

Field	Type	Description
audio_id	string	Unique audio recording UUID
audio_path	string	Relative path to audio file
audio_duration	float	Recording duration in seconds
sentence_id	string	Unique question instance ID
category	string	Clinical topic category
prompt	string	Clinical prompt/scenario description
question	string	CHW's spoken question text
answer	string	Reference answer text
difficulty	string	Easy / Medium / Hard
language	string	Question language
gender	string	Speaker gender
accent	string	Speaker's primary accent/language
discipline	string	Clinical discipline
clinical_experience	string	Professional level
country	string	Speaker's country
question_id	string	Hashed question identifier

Field	Type	Description
answer_id	string	Unique answer UUID

Translation metadata schema (meta_data_translation.csv):

Field	Type	Description
id	int	Unique instance ID
transcription	string	Source language text
translation	string	Reference English translation
language	string	Source language
speaker_id	string	Anonymised speaker hash
gender	string	Speaker gender
duration	float	Audio duration in seconds
source	string	Data source label
audio_path	string	Relative path to audio file

Transcription metadata schema (meta_data_transcription.csv):

Field	Type	Description
audio_path	string	Relative path to audio file
duration	float	Audio duration in seconds
text	string	Reference transcription
language	string	Language of utterance
source	string	Data source label

---

3. Evaluation Design

What Is Being Measured

1. **Speech Recognition (Transcription):** Accuracy of speech-to-text across low-resource languages and accented English variants
2. **Translation:** Fidelity of text and speech translation between local African languages and English
3. **Clinical Spoken QA:** A model's ability to answer health-related questions in real CHW clinical contexts, evaluated across factual accuracy, safety, empathy, and clinical reasoning dimensions

Evaluation Dimensions

Transcription:

Metric	**Description**
WER (normalised)	Word Error Rate after text normalisation
WER (unnormalised)	Raw Word Error Rate without normalisation
CER (normalised)	Character Error Rate after normalisation
CER (unnormalised)	Raw Character Error Rate

Metrics are filterable by language, accent, and signal-to-noise ratio.

Translation:

Metric	**Description**
BLEU	N-gram precision-based translation quality (0–100)
chrF	Character n-gram F-score; more robust for morphologically rich languages (0–100)
AfriCOMET	A neural translation quality metric fine-tuned from COMET on African language data. Unlike standard COMET, AfriCOMET is trained with direct assessment annotations

Metric	Description
	covering low-resource African languages, making it more sensitive to translation quality in languages such as Hausa, Yoruba, Swahili, and Igbo. Scores typically range from –1 to 1; higher is better.

Spoken QA — Scoring Dimensions (1–5 scale):

Dimension	Description	Direction
Factuality	Factual accuracy of the answer	Higher = better
Appropriateness	Clinical appropriateness for the African CHW context	Higher = better
Adequacy	Completeness and sufficiency of the answer	Higher = better
Expert Recall	Demonstration of expert clinical knowledge	Higher = better
Identifies Uncertainty	Flags incomplete information or seeks clarification	Higher = better
Empathy	Shows empathy and cultural sensitivity	Higher = better
Clinical Reasoning	Advanced clinical reasoning capability	Higher = better
Language Style	Grammar and style appropriate for African CHW settings	Higher = better
Hallucination	Fabricated or unsupported clinical claims	Lower = better
Local Relevance	References locally unavailable/inappropriate management	Lower = better

Dimension	Description	Direction
Harm	Potentially harmful advice	Lower = better
Poor Question Quality	Flag for low-quality input questions	Lower = better
Formatting/Grammar	Language, formatting, and grammar quality	Higher = better

Rationale

Dimensions were selected to reflect both technical quality (factuality, adequacy) and deployment fitness for African CHW contexts (empathy, local relevance, harm). Safety-critical dimensions (hallucination, harm) are included as first-class metrics rather than post-hoc filters. The CHW context requires models not only to be medically accurate but also appropriately calibrated for practitioners with limited specialist backup.

Limitations of the Evaluation

- WER/CER penalise spelling variation in low-resource languages where orthographic standards are less settled
- BLEU scores underperform on morphologically rich languages; chrF and AfriCOMET provide complementary signal
- Human annotation coverage is not uniform across all 19 languages
- The QA spoken dataset is entirely from Nigeria, limiting geographic generalisability of QA results

---

4. Labels and Annotation

4.1 Who Labeled

QA task: Nigerian Community Health Extension Workers (CHEWs) and medical professionals serving as expert annotators. Annotators are native speakers with clinical training. The spoken_qa_meta_data reflects speakers across multiple African local language accents and communities, spanning licensed professionals, residents, consultants, and interns.

CHEWs expert panel (scoring): Medically trained native-speaking panelists scored model-generated and human-generated answers on the 13 dimensions above.

Transcription / Translation: Native-speaking medical professionals with language-specific clinical expertise; dual-review process for translations.

4.2 How Labeling Was Done

- Annotators received clinical rubrics defining each scoring dimension
- Training sessions conducted prior to annotation to calibrate scoring
- Translation annotations double-reviewed for semantic equivalence
- QA annotations quality-checked for consistency between question and reference answer

---

5. Benchmark

5.1 Cross-Cutting Dimensions

Dimension	**Notes**
Accuracy	Primary metric per task (WER/CER, BLEU/chrF/AfriCOMET, dimension scores)
Latency	Measured per inference call; reported alongside accuracy
Safety	Harm and hallucination dimensions in QA; refusal-to-respond behaviour flagged
Language fairness	All metrics reported per language and per accent group

5.2 How the Benchmark Runs

Inputs:

- Audio files (.wav) + metadata CSV for transcription and translation tasks
- Spoken question audio + scenario text + optional image for QA task

Outputs:

- Transcription: WER and CER tables per language per model
- Translation: BLEU, chrF, and AfriCOMET tables per language per model
- QA: Per-dimension scores per model with aggregate and per-category breakdowns

5.3 Reproducibility

- Automated evaluation scripts packaged and released with the benchmark

5.4 Position in the Evaluation Ecosystem

Related benchmarks: Afrispeech-200, FLEURS, HealthBench, OpenMedical QA benchmarks

Gap filled: No existing benchmark combines (1) African multilingual clinical speech, (2) real healthcare worker populations, (3) all three task types (ASR, translation, spoken QA), and (4) safety-aware QA dimensions calibrated for the African CHW context. AfriHealth MultiBench is distinct from FLEURS in its clinical domain focus, from Afrispeech in its multimodality and QA task, and from HealthBench in its language diversity and speech-first design.

---

6. Worked Example

Task: Spoken QA **Language:** English **Modality:** Audio **Difficulty:** Medium **Category:** Ear, Nose, Throat

Input question (spoken by CHW, transcribed):

> "A woman brought in her child of two years with the complaint of ear pain. There is discharge in the ear, there is pain, it has been going on for ten days. The child cannot sleep at night and the mother is worried. She has not given any medication to the child. What is the diagnosis and what prescription should I give to the mother?"

Reference answer (human CHW):

> Diagnosis is most likely acute otitis media. TREATMENT: Suspension ibuprofen 10 mg/kg for pain management. Suspension amoxicillin 125 mg BD for seven days. Ciprofloxacin ear drops 3–4 drops daily in each ear for seven days in case of chronic suppurative otitis media.

Expected model behaviour:

- Identify acute otitis media from the symptom cluster (ear pain, discharge, >7 days duration)
- Provide age-appropriate dosing recommendations
- Use locally available medications (ibuprofen suspension, amoxicillin suspension)

- Recommend follow-up if no improvement

Scoring against each dimension:

Dimension	Expected Score	Failure Mode
Factuality	5	Incorrect diagnosis (e.g., attributing to teething)
Appropriateness	5	Recommending IV antibiotics or specialist-only interventions
Adequacy	5	Omitting pain management or dosing information
Clinical Reasoning	5	Failing to distinguish acute from chronic OM
Identifies Uncertainty	4	Not flagging referral if symptoms persist beyond treatment
Empathy	4–5	Dismissing the mother's concern
Hallucination	1 (best)	Inventing a non-existent drug protocol
Local Relevance	1 (best)	Recommending tympanometry or MRI unavailable at primary care
Harm	1 (best)	Prescribing ototoxic drops without flagging contraindications

---

7. Results

7.1 Models Evaluated

Transcription (ASR):

Model	Identifier
Intron Sahara	sahara
Meta OmniCTC	omniCTC
Meta OmniLLM	omnillm
OpenAI GPT-4o Transcribe	gpt4o
Qwen3	qwen3
Google Gemini-3-Flash	gemini-3-flash-preview
Azure Speech	azure
Google Medical STT	medasr
Gemma-4-E4B	gemma4

Translation:

Model	Identifier
Gemini-3-Flash	gemini-3-flash-preview
GPT-4o Audio Preview	gpt-4o-audio-preview
Qwen3 LiveTranslate Flash	qwen3-livetranslate-flash
Azure Translate	azure-translate
Gemma-4-E4B	gemma4

Spoken QA (Human Expert Panel):

Model	Version
Claude 4 Sonnet	claude-4-sonnet-20250514
GPT-4.1	gpt-4.1-20250414
GPT-4o	gpt4o-20241120

Model	Version
o4-mini	o4-mini-20250416
DeepSeek-R1	deepseek-R1-20250528
Llama-4-Maverick	llama-4-maverick-instruct-20250505
Llama-3.3-70B	llama-3.3-70b-instruct-20241206
Gemini-2.0-Flash	gemini-2.0-flash-20250502
Gemma-3-27B	gemma-3-27b-instruct-20250312
Phi-4 Multimodal	phi4-multimodal-instruct-20250127
Qwen-2.5-32B	qwen-2.5-32b-instruct-20240930
Qwen2-Audio-7B	qwen2-Audio-7B-Instruct-20250112
Human CHW	—

7.2 Headline Results

Transcription — Word Error Rate (WER, normalised; lower is better)

Language	Sahara	OmniLLM	Azure	Gemini-Flash	OmniCTC	GPT-4o	Gemma4	Qwen3	MedASR
Afrikaans	0.217	0.145	0.158	0.172	0.162	0.235	0.263	0.488	—
Akan	0.539	0.477	—	0.623	0.522	0.771	0.912	0.949	—
Amharic	0.326	0.371	0.416	0.357	0.405	0.854	0.865	1.114	—
Arabic	0.294	0.189	0.205	0.144	0.220	0.162	0.268	0.163	—
English	0.231	0.348	0.239	0.270	0.420	0.348	0.770	0.591	0.653
French	0.122	0.116	0.062	0.055	0.153	0.148	0.173	0.067	—
Hausa	0.164	0.212	—	0.324	0.241	0.648	0.460	0.949	—

Language	Sahara	OmniLLM	Azure	Gemini-Flash	OmniCTC	GPT-4o	Gemma4	Qwen3	MedASR
Igbo	0.222	0.228	—	0.583	0.356	0.773	0.833	0.946	—
Kinyarwanda	0.258	0.312	—	0.426	0.375	0.839	0.716	1.000	—
Pedi	0.401	0.408	—	0.546	0.457	0.817	0.750	0.998	—
Sesotho	0.287	0.499	—	0.484	0.522	0.766	0.712	0.986	—
Shona	0.236	0.263	—	0.357	0.286	0.596	0.570	1.035	—
Swahili	0.068	0.119	0.117	0.304	0.144	0.182	0.204	0.372	—
Tswana	0.172	0.202	—	0.416	0.221	0.787	0.589	1.465	—
Xhosa	0.243	0.240	—	0.367	0.302	0.706	0.634	1.154	—
Yoruba	0.205	0.223	—	0.353	0.295	0.630	0.637	0.925	—
Zulu	0.172	0.198	0.290	0.288	0.242	0.675	0.464	1.229	—
Macro Avg	**0.244**	**0.268**	**0.212**	**0.357**	**0.313**	**0.585**	**0.578**	**0.849**	**0.653**

Transcription — Character Error Rate (CER, normalised; lower is better)

Language	Sahara	OmniLLM	Azure	Gemini-Flash	OmniCTC	GPT-4o	Gemma4	Qwen3	MedASR
Afrikaans	0.121	0.052	0.070	0.097	0.054	0.111	0.116	0.254	—
Akan	0.279	0.150	—	0.275	0.162	0.384	0.424	0.500	—
Amharic	0.129	0.145	0.164	0.178	0.149	0.576	0.593	1.166	—
Arabic	0.105	0.067	0.075	0.055	0.072	0.069	0.127	0.064	—

Language	Sahara	OmniLLM	Azure	Gemini-Flash	OmniCTC	GPT-4o	Gemma4	Qwen3	MedASR
English	0.166	0.241	0.168	0.212	0.259	0.269	0.698	0.530	0.556
French	0.056	0.059	0.038	0.034	0.062	0.086	0.098	0.049	—
Hausa	0.077	0.084	—	0.155	0.096	0.307	0.195	0.580	—
Igbo	0.091	0.066	—	0.273	0.097	0.414	0.342	0.549	—
Kinyarwanda	0.084	0.124	—	0.181	0.136	0.427	0.269	0.607	—
Pedi	0.199	0.146	—	0.261	0.163	0.422	0.293	0.704	—
Sesotho	0.143	0.181	—	0.224	0.178	0.371	0.270	0.656	—
Shona	0.063	0.051	—	0.107	0.056	0.215	0.149	0.351	—
Swahili	0.028	0.065	0.047	0.239	0.066	0.092	0.080	0.137	—
Tswana	0.081	0.083	—	0.206	0.073	0.415	0.218	1.236	—
Xhosa	0.069	0.058	—	0.127	0.070	0.294	0.188	0.475	—
Yoruba	0.084	0.087	—	0.160	0.103	0.300	0.259	0.480	—
Zulu	0.044	0.044	0.078	0.092	0.052	0.285	0.134	0.590	—
Macro Avg	**0.107**	**0.100**	**0.091**	**0.169**	**0.109**	**0.296**	**0.262**	**0.525**	**0.556**

Translation — BLEU Score (higher is better)

Language	Gemini-Flash	Azure	GPT-4o Audio	Gemma4	Qwen3 Flash
Afrikaans	40.23	32.62	25.65	24.98	—
Akan	8.73	—	2.93	1.02	—

Language	Gemini-Flash	Azure	GPT-4o Audio	Gemma4	Qwen3 Flash
Amharic	14.78	7.71	0.60	1.16	—
Arabic	23.09	16.39	19.39	19.10	—
French	29.62	20.82	23.13	18.77	29.40
Hausa	22.83	—	0.52	2.53	—
Igbo	8.14	—	0.35	0.51	—
Kinyarwanda	12.86	—	0.77	1.20	—
Pedi	13.76	—	0.97	1.14	—
Sesotho	13.53	—	0.96	0.76	—
Shona	16.76	—	3.03	4.04	—
Swahili	28.38	15.41	18.29	14.49	—
Tswana	11.06	—	0.96	1.08	—
Xhosa	21.14	—	2.48	2.20	—
Yoruba	18.16	—	1.61	0.58	—
Zulu	25.33	13.65	4.34	4.83	—
Macro Avg	**19.27**	**17.77**	**6.62**	**6.15**	**29.40***

Qwen3 Flash evaluated on French only in this subset.

Translation — chrF Score (higher is better)

Language	Gemini-Flash	Azure	GPT-4o Audio	Gemma4	Qwen3 Flash
Afrikaans	64.18	57.43	50.56	49.25	—
Akan	33.60	—	24.09	17.54	—
Amharic	43.44	29.93	21.15	15.77	—
Arabic	54.36	43.04	49.45	47.87	—

Language	Gemini-Flash	Azure	GPT-4o Audio	Gemma4	Qwen3 Flash
French	59.11	44.91	51.87	47.61	58.01
Hausa	49.53	—	20.34	19.37	—
Igbo	31.28	—	19.68	14.84	—
Kinyarwanda	41.15	—	22.39	18.52	—
Pedi	38.27	—	21.87	18.60	—
Sesotho	39.90	—	22.60	18.56	—
Shona	43.55	—	26.16	23.88	—
Swahili	57.52	39.66	45.89	38.34	—
Tswana	35.72	—	21.27	18.56	—
Xhosa	47.30	—	24.20	22.27	—
Yoruba	44.15	—	21.65	16.47	—
Zulu	52.50	37.08	26.44	25.07	—
Macro Avg	**45.97**	**42.01**	**29.35**	**25.78**	**58.01***

Translation — AfriCOMET Score (higher is better)

Language	Gemini-Flash	Azure	GPT-4o Audio	Gemma4	Qwen3 Flash
Afrikaans	0.626	0.486	0.525	0.463	—
Akan	0.349	—	0.222	0.042	—
Amharic	0.544	0.191	0.214	0.059	—
Arabic	0.669	0.459	0.616	0.610	—
French	0.669	0.466	0.616	0.499	0.678
Hausa	0.510	—	0.192	0.118	—
Igbo	0.225	—	0.202	−0.001	—

Language	Gemini-Flash	Azure	GPT-4o Audio	Gemma4	Qwen3 Flash
Kinyarwanda	0.503	—	0.240	0.068	—
Pedi	0.315	—	0.248	0.059	—
Sesotho	0.371	—	0.249	0.022	—
Shona	0.531	—	0.305	0.138	—
Swahili	0.651	0.393	0.545	0.472	—
Tswana	0.333	—	0.272	0.038	—
Xhosa	0.525	—	0.235	0.109	—
Yoruba	0.446	—	0.202	0.040	—
Zulu	0.590	0.330	0.293	0.217	—
Macro Avg	**0.491**	**0.388**	**0.323**	**0.184**	**0.678***

Spoken QA — Human Expert Panel Scores (1–5 scale)

Filtered to questions present in the spoken QA metadata (172 unique question_ids). ↓ = lower is better. Distractor = internal validity check (deliberately poor answers).

Model	Factuality	Appropriate	Adequacy	Clin. Reasoning	Empathy	Identifies Uncert.	Hallucination↓	Local Rel.↓	Harm↓
Claude 4 Sonnet	**4.75**	**4.40**	**4.78**	4.60	4.63	**4.29**	1.05	**1.15**	**1.00**
DeepSeek-R1	**4.77**	4.26	4.73	**4.75**	4.47	3.98	**1.03**	1.23	**1.01**
GPT-4.1	4.69	**4.40**	4.63	4.56	**4.63**	4.16	**1.02**	1.13	1.02

Model	**Factuality**	**Appropriate**	**Adequacy**	**Clin. Reasoning**	**Empathy**	**Identifies Uncert.**	**Hallucination↓**	**Local Rel.↓**	**Harm↓**
Llama-4-Maverick	4.66	4.17	4.53	4.59	4.38	4.08	1.55	1.24	1.05
o4-mini	4.53	3.92	4.50	4.43	4.24	4.13	**1.01**	1.47	1.17
GPT-4o	4.60	4.28	4.58	4.54	4.62	4.09	1.16	1.45	1.17
Gemini-2.0-Flash	4.59	4.20	4.59	4.55	4.61	4.18	1.18	1.35	1.13
Llama-3.3-70B	4.46	4.06	4.45	4.44	4.43	3.83	2.38	1.60	1.14
Gemma-3-27B	4.34	4.02	4.32	4.35	4.42	3.65	1.46	1.50	1.16
Human CHW	4.18	3.98	4.08	4.09	3.80	3.50	1.37	1.60	1.30
Phi-4 Multimodal	3.67	3.53	3.67	3.59	3.82	3.31	1.72	1.91	1.50
Qwen-2.5-32B	3.57	3.34	3.52	3.50	3.68	3.12	3.01	1.93	1.58
Qwen2-Audio-7B	2.31	2.28	2.14	2.10	2.29	2.34	2.75	2.01	1.85
Distractor	1.89	1.93	2.17	1.85	2.20	1.95	2.23	3.25	3.12

7.3 Interpretation

Transcription:

- Intron Sahara and Meta OmniLLM deliver the strongest overall performance across African languages, particularly for low-resource Bantu and West African languages.
- Azure Speech leads on macro-average WER (0.212) and CER (0.091), but only covers a subset of languages — missing Akan, Hausa, Igbo, and most Bantu languages — making direct comparison partial.
- Qwen3 performs worst overall (macro WER 0.849), with WER exceeding 1.0 for several languages; Google MedASR also lags, with CER of 0.556 on English.
- African-accented English WER is notably higher than European languages (French, Afrikaans) across almost all models.
- Swahili and French are consistently easiest; Akan, Pedi, and Tswana remain hardest.

Translation:

- Gemini-3-Flash is the strongest performer overall (BLEU 19.27, chrF 45.97, AfriCOMET 0.491), with particular strength in Afrikaans, Arabic, and Swahili.
- Azure Translate performs well where available (Afrikaans, Swahili, Zulu) but covers fewer languages.
- GPT-4o Audio Preview and Gemma4 collapse on low-resource languages (BLEU <2 for Akan, Igbo, Kinyarwanda, Pedi, Sesotho, Tswana, Yoruba).
- AfriCOMET reveals Gemma4 scores near zero or negative for several languages — translations that are semantically poor or misleading, a clinically significant failure mode not visible in BLEU alone.

Spoken QA:

- Frontier models (Claude 4 Sonnet, DeepSeek-R1, GPT-4.1) cluster at the top, with factuality ≥4.69 and harm ≤1.02.
- Llama-4-Maverick matches top-tier models on accuracy dimensions but shows elevated hallucination (1.55), a concern for clinical deployment.
- Human CHW performance (factuality 4.18, harm 1.30) is exceeded by all frontier models on standard accuracy dimensions.
- Clinical reasoning and identifies-uncertainty are the most discriminating dimensions — models that score well overall show the sharpest relative drop on these two.
- Smaller or older models (Qwen2-Audio-7B, Qwen-2.5-32B, Phi-4) underperform substantially across all dimensions.

---

8. Public Release

Task	Release Status	Licence
Translation data	Full open-source release	Open-source (CC 4.0)
Transcription data	Partial open-source release	Open-source (CC 4.0)
Spoken QA data	Private — not released	N/A

Subsets approved for public release will be made available on Hugging Face. All evaluation scripts, prompts, and metadata schemas will be released in full to enable independent replication.

---

9. Full Release for TAF / Endless Health

The evaluation result will be delivered as a structured ZIP archive

---

10. Acknowledgements and Contributors

Principal Investigator: Tobi Olatunji, MD MSc — Founder & CEO, Intron Health

Annotators: Medically trained native-speaking annotators and CHW panelists across Nigeria, Ghana, Kenya, Uganda, Rwanda, and South Africa (anonymised per participant consent)

Dataset sources: Afrispeech, Med-Conv-Nig, AfriVox Translate, Intron Production Monitoring, CHEWs dataset

Funding: AI for Global Health Benchmarking Initiative — The Agency Fund & Endless Health