VOICES OF CARE

A Hausa Clinical Speech Benchmark for Community Health AI

Final Submission Document

Health AI Benchmarking Initiative — The Agency Fund × Endless Health

This document compiles the complete project record: dataset, documentation, labels, benchmarks, results, model and deployment metadata, contextual considerations, and limitations.

Grantee	EHA Clinics Limited / eHealth Africa — Kano, Nigeria
Technical partner	Data Science Nigeria (DSN) / EqualyzAI Limited — Lagos, Nigeria
Program	The Agency Fund × Endless Health — Health AI Benchmarking Initiative
Benchmark of record	Final benchmark, released 9 August 2026
Code repository	github.com/eHealthAfrica/voices-of-care (Apache-2.0)
Dataset	huggingface.co/datasets/eHealthAfrica/voices-of-care-hausa (CC BY 4.0)
Ethics clearance	SHREC/2026/7706 · Kano State MoH HREC · 13 May 2026 · NHREC/17/03//2018
Corresponding lead	Tahir Buhari, Senior Research Lead, EHA Group · tahir.buhari@ehealthnigeria.org
Document date	5 November 2026 · Version 1.2 (corrects the per-speaker recording count in §2.2; previous revision 3 September 2026)

1. Executive Summary

Voices of Care is the first publicly documented, clinically validated Hausa speech benchmark for community health AI. The setting: across northern Nigeria, many patients reach the health system only through community health workers, in Hausa — a language of more than 70 million speakers that is almost absent from speech-technology benchmarks. Under this grant, EHA Clinics / eHealth Africa, with Data Science Nigeria / EqualyzAI as technical partner, produced a 30,000-recording (~50-hour) Hausa clinical speech corpus spanning ten priority health domains, defined two benchmark tasks (automatic speech recognition and 11-class micro-intent classification), and ran a complete, fully reproducible evaluation against six models.

The benchmark is evaluated on a held-out test set of 5,375 recordings, built so that no speaker and no sentence from the training data appears in the test data — a model cannot score well simply by recognising a familiar voice or sentence. Each model was trained five times from different random starting points, results carry confidence ranges, and performance is reported separately by gender, dialect, and age band. The two leading fine-tuned ASR models are statistically tied: Whisper Large-v3 with LoRA fine-tuning at 15.14% ± 0.20pp WER and XLSR-53 at 15.59% ± 0.12pp WER ($p = 0.16$), both with character error rates below 4%. The champion intent classifier, Afro-XLMR-large, reaches 87.34% ± 1.11pp accuracy. An end-to-end voice pipeline built from the two champions retains 79.43% intent accuracy against an 86.00% perfect-transcript ceiling — transcription costs 6.57 accuracy points, not a collapse, supporting the viability of Hausa voice front-ends for high-frequency clinical intents.

Every published number can be regenerated from the code, configurations, and frozen data splits committed to the code repository. An earlier evaluation round reported lower error rates, but its data split allowed the same speakers to appear in both training and test data; the final protocol removed that overlap, which is why the final figures are slightly higher and more trustworthy.

All required deliverables are complete: the finalized evaluation dataset, complete documentation, benchmark results from six models, and public release of evaluation results are all delivered. In line with the program's guidance that datasets be shared privately with the program team, and specifically with Endless Health's request to withhold public dataset publication pending frontier-model comparison, dataset repository access has been restricted (see Section 12) while the code, results, and documentation remain public.

2. The Dataset

2.1 Data source and provenance

The corpus derives from the clinical communication context of EHA Clinics' active primary-care and community-health operations in Kano, Nigeria. 15,000 sentence templates were authored to reflect authentic community health worker (CHW)–patient conversations — symptom reports, information-seeking questions, counselling, referral, and treatment discussions — through a combination of clinician authoring and AI-assisted generation, followed by duplication removal, medical-appropriateness validation by clinicians across the ten domains, and linguistic validation by native-Hausa reviewers with a minimum of five years of community-health experience.

The audio is real human speech: approximately 50 consented native Hausa speakers recorded the templates in prompted sessions on a purpose-built recording platform (EqualyzCrowd) under field-simulated conditions. Each sentence was recorded by two different speakers of different genders — a gender-paired, cross-speaker design (14,928 of 14,966 unique sentences carry exactly this pairing; zero same-speaker pairs). We state plainly what this is and is not: the utterances are scripted readings of clinically authored templates, not recordings of live patient encounters. This design was chosen deliberately — it is what makes the corpus PII-free by construction and releasable under an open licence in a domain where real encounter audio could not be — and it is disclosed as the benchmark's primary limitation (§12): scores are an upper bound on spontaneous-dialogue performance.

2.2 Scale

Voiced instances	30,000 recordings (~50 hours of audio)
Unique sentences	15,000 clinically authored templates (14,966 unique IDs in manifest)
Speakers	~50 native Hausa speakers (25 male / 25 female), ~600 recordings (~1 hour) each
Gender balance	15,000 male / 15,000 female recordings — exactly balanced
Health domains	10 (maternal health, malaria, immunization, NCDs, mental health & substance abuse, pediatrics & child nutrition, infectious diseases, reproductive & sexual health, health & wellness, environmental health)
Audio format	48 kHz capture (Ogg containers, mostly Vorbis with a minority of FLAC); resampled to 16 kHz mono for ASR

The corpus contains 30,000 recordings from about 50 speakers. Fifty speakers is a modest contributor base, offset by depth per speaker (about an hour of audio each), an exact gender balance, and coverage of four dialect regions. The held-out test set behind all reported results contains nine of those speakers (six female, three male), so subgroup results describe those nine voices rather than the wider population (see Section 11).

2.3 Modality

Voice with verbatim transcription and English gloss: each instance pairs an audio recording with its ground-truth Hausa transcript (in Boko orthography, preserving hooked and glottalized characters ƙ, ɓ, ɗ, ƴ), a reference English translation, an 11-class micro-intent label, an emotion-tone label, and a speaker-type label. The dataset is monolingual Hausa.

2.4 Privacy, consent, and rights

Sentence content is PII-free by construction: no names, dates, locations, facility names, or record numbers. All speaker-contributors gave informed consent in Hausa with third-party verification of comprehension — a plain-language Hausa information sheet, a yes/no comprehension-check consent form, an audio-recorded consent discussion stored de-linked from voice samples, and a written signature or witnessed thumbprint. Contributors were explicitly informed their de-identified recordings would be released under CC BY 4.0 for health-AI research, could withdraw at any time, and were fairly compensated.

The study holds formal ethical clearance from the Kano State Ministry of Health State Health Research Ethics Committee, granted 13 May 2026 (reference SHREC/2026/7706; committee NHREC registration NHREC/17/03//2018), with Tahir Buhari as Principal Investigator. EHA Clinics is Data Controller and DSN/EqualyzAI Data Processor, with roles fixed in the sub-award agreement and data-processing addendum. Rights to the dataset are held by EHA Clinics under the sub-award's work-product and licensing clauses and Nigerian law on commissioned works.

2.5 Sampling and representativeness

Speakers were recruited across northern Nigeria and stratified across the four primary Hausa dialect clusters — Kano (Kananci, 20 speakers), Katsina (Katsinanci, 10), Zaria (Zazzaganci, 15), and Sokoto (Sakkwatanci, 5) — with balanced gender, age bands spanning 15–29, 30–45, and 45+, and a mix of secondary and tertiary education levels. The population represented is adult native Hausa speakers in northern Nigeria communicating about community health topics.

This benchmark is not demographically representative and never claims to be: the corpus over-represents the Kano cluster relative to Sokoto, and the evaluated test partition contains nine speakers. Results for each subgroup are published with speaker counts, and any group with too little data shows a range rather than a single number, precisely so readers can see how much evidence sits behind each figure.

3. Health-Specific Privacy Considerations

Sensitive health topics: the corpus deliberately includes content on reproductive and sexual health, mental health and substance abuse, HIV, and TB — domains where Hausa-language voice tools are most needed and where template-based (rather than real-encounter) collection is the appropriate safeguard. Because all utterances are scripted, no real individual's health status, history, or condition is disclosed anywhere in the corpus. No minors participated: contributors were aged approximately 18–60.

Voice is inherently biometric. Under Nigerian law voice recordings are sensitive personal data (NDPR Art. 1.3(xiii)); processing was conducted on an explicit-consent basis with heightened safeguards: encrypted (≥256-bit AES) access-controlled storage, multi-factor authentication, monthly-audited access logs, NITDA 72-hour breach notification procedures, and NDPR Art. 2.11 safeguards for cross-border transfers (including to the funder in the US). After the project, raw voice files are securely deleted; only the released version — with speaker identities removed and recordings screened for any identifying audio — is retained.

We are open about remaining risks. Speakers appear only under neutral codes (speaker_001 to speaker_050), but because a partner copy of the source data included speaker identifiers, we do not claim full anonymity — and we say so plainly. The dataset licence and the consent terms both forbid any attempt to re-identify speakers. Data was collected and processed in Nigeria, reviewed by the funder in the US, and computed on cloud servers in Europe, with all transfers following Nigerian data-protection safeguards.

4. Interaction Structure and Unit of Analysis

The unit of analysis is the single utterance: one voiced rendition of one clinically authored sentence, analogous to a single turn in a CHW–patient exchange. Each released row carries a stable per-recording identifier (audio_id), the sentence template identifier (sentence_id) linking the two gender-paired renditions of a sentence, a deterministic speaker pseudonym (speaker_id), and demographic strata (gender, dialect, age_band). Labels and benchmarks apply at utterance level for ASR and at unique-sentence level for intent classification; both scopes are documented in the split design (§7.2). Speaker-type labels (patient statement, patient question, caregiver statement, caregiver question) preserve the conversational role each utterance plays.

5. Labels

The corpus carries two primary label families and two secondary ones, each documented below.

5.1 Ground-truth transcript (ASR reference)

Label name	hausa_transcription — the verbatim Hausa transcript (a cleaned version is also provided)
What it measures	The verbatim Hausa text of the utterance in Boko orthography — the reference against which model transcriptions are scored.
Why it matters	Transcription accuracy is the gatekeeper for every downstream use: a CHW tool that mishears "zazzaɓin cizon sauro" (malaria fever) cannot triage it. Hausa's hooked consonants and diacritics make character-level fidelity clinically meaningful.
Who labelled	Source templates verified against each recording by trained native-Hausa reviewers (10 nominated from the EHA Group) via the EqualyzCrowd review workflow; automated QA flagged items with WER > 0.40 for native-speaker adjudication.
Level	Query-response pair (single utterance).
High/low example	High-quality output: "tari na ya ƙaru tun da hayaki ya fara a kewayen gidana" (exact match, diacritics preserved). Low-quality: "tari na ya karu tunda hayaki ya fara kewaye gidana" (diacritics dropped, word boundaries merged — 4 word errors).

5.2 Micro-intent (11-class classification target)

Label name	micro_intent / intent_label
What it measures	The communicative purpose of the utterance — what the patient or caregiver is trying to do (report a symptom, seek information, disclose risk history, raise a medication concern, etc.).
Why it matters	Intent is the routing signal for CHW decision support: it determines whether an utterance needs triage, counselling content, escalation, or follow-up scheduling.
Who labelled	Two-stage process: primary annotation from the written transcript by trained Hausa-speaking annotators with community-health backgrounds; independent second-annotator re-labelling of a random sample with accept/reject adjudication at review. Labelling was from text, not audio, so labels reflect linguistic content rather than delivery. No LLM was used for labelling.
Level	Unique sentence (both renditions of a sentence share the label).
High/low example	High: "ina jin ƙarancin numfashi..." → Symptom reporting (correct). Low: a treatment-plan discussion labelled Prevention & counselling — the most frequent confusion pair, 10 of 25 true Treatment items in the baseline confusion analysis.

The 11 classes and their distribution (class imbalance is a reported property of the benchmark, not an accident):

Micro-intent	Corpus	Intent train	Intent test
Information seeking	7,313	3,312	163
Clinical assessment	6,978	3,151	162
Prevention and counselling	6,435	2,860	188
Symptom reporting	4,492	2,002	117
Risk / history / context disclosure	2,046	904	55
Treatment	914	412	25
Follow-up / status update	540	246	11
Vaccine recommendation / administration	462	209	11
Referral	364	158	12
Diagnosis / results communication	356	168	4
Medication concern	100	47	1

5.3 Secondary labels

emotion_tone (Neutral, Concerned, Calm, Reassuring) and speaker_type (patient/caregiver × statement/question) were annotated in the same two-stage pass. They are not part of the core benchmark scoring and are held back from the public dataset release, which carries the transcripts, intent label, sentence identifier and speaker demographics.

6. Health-Specific Evaluation Dimensions: Scope Declaration

In scope for this benchmark: (a) accuracy of transcription for the target user population, including whether Hausa's special characters and spellings are captured correctly (measured through the character error rate, CER); (b) correct recognition of communicative intent, including the clinically consequential rare intents (referral, medication concern) whose per-class reliability is separately reported; (c) equity of performance across gender, dialect, and age, measured through mandated disaggregated reporting with statistical floors; and (d) awareness of system scope, addressed through the explicit scope declaration in §8 of this document and on the dataset card.

Out of scope: medical appropriateness of generated advice, harm avoidance in generation, and uncertainty-driven escalation behaviour. These are generation-model properties; Voices of Care evaluates the recognition and understanding layers (ASR and intent) that sit in front of any generation or decision layer. The benchmark's error-propagation analysis (§7.5) quantifies exactly how recognition errors degrade downstream understanding — the measured foundation any generation-layer evaluation in this context will need.

7. Benchmarks and Results

7.1 Benchmark definitions

Benchmark 1 — ASR. Word Error Rate (primary) and Character Error Rate (secondary), computed with jiwer 4.0.0 through a version-pinned text normalizer (v1.3.0, recorded per row; both raw and normalized transcripts ship so third parties can re-score under their own normalizer). CER is a reliable measure for Hausa because a single character — a diacritic, a hooked consonant such as ƙ or ɗ, or a doubled consonant — can change a word's meaning, and differing word-boundary conventions can make WER overstate errors: a low CER alongside a moderate WER means the model heard the words correctly and slipped mainly on spelling or word boundaries. We recognise that WER and CER do not capture semantic equivalence; we therefore publish both raw and cleaned transcripts so that semantic or LLM-based scoring can be layered on top, and we treat validated LLM-judge evaluation for Hausa as future work.

Benchmark 2 — Intent classification. Overall accuracy and macro-F1 over the 11 classes, with per-class precision/recall/F1 and confusion analysis. Macro-F1 is reported alongside accuracy because the label set is heavily imbalanced; the gap between them is itself a reported quantity.

Both benchmarks are repeatable by construction: the data splits are frozen, every configuration is saved, all results files are committed to the code repository, and each model is trained several times from different random starting points so results carry an explicit uncertainty range.

7.2 Split protocol

The main data split keeps speakers and sentences fully separate between training and test: no speaker, and no sentence, appears on both sides. This is the harder, more honest protocol — a model cannot score well by recognising a voice or a sentence it has already seen. The separation has a cost: of the 30,000 recordings, 140 lack speaker information and could not be placed, and a further 5,913 were set aside to keep the separation strict. The test set was unaffected, and all counts are published.

Configuration	Train	Validation	Test	Protocol
Speech recognition (final)	16,717	1,855	5,375	Speakers and sentences fully separated
Earlier protocol (superseded)	24,003	2,999	2,998	Sentences separated; speakers could overlap
Intent classification	13,469	748	749	Split by unique sentence

7.3 Speech recognition (ASR) results

Held-out test set, 5,375 recordings, nine speakers. Results are the average across five training runs, shown with their spread; MMS-1B-fl102 is a single-run reference.

Model	Params	Method	Seeds	Test WER	Test CER
Whisper Large-v3 (champion)	1.55B	LoRA fine-tune, bf16	5	15.14% ± 0.20pp	3.66% ± 0.05pp
XLSR-53	315M	Full fine-tune, bf16	5	15.59% ± 0.12pp	3.60% ± 0.04pp
MMS-1B-all	1B	Hausa adapter, bf16	5	22.81% ± 0.95pp	5.43% ± 0.33pp
MMS-1B-fl102 (reference)	1B	Hausa adapter, bf16	1	30.74%	7.12%

The two leaders are statistically indistinguishable: the speaker-clustered paired bootstrap between their runs of record puts the delta at +0.51pp with a 95% CI of [−0.19, +1.26], p = 0.16 — the ranking should not be over-read. Both are decisively better than the adapter models (7.5–8.0pp, p = 0.001). CER below 4% for the leaders means over 96% of characters are reproduced correctly; errors concentrate at word boundaries and morphology rather than phonetic misrecognition — a strong result for low-resource clinical Hausa.

7.4 Intent classification results

Model	Role	Seeds	Accuracy	Macro-F1
Afro-XLMR-large (champion)	Headline	4 of 5*	87.34% ± 1.11pp	68.98% ± 1.83pp
mBERT	Reference	5 of 5	81.38% ± 1.17pp	60.71% ± 1.22pp

*One of the five Afro-XLMR training runs failed to learn (it predicted a single category throughout) and is excluded; figures are the average of the four successful runs. The gap between accuracy and macro-F1 is the key intent finding: categories with fewer than about 250 training examples are recognised much less reliably than the larger ones. More data for the rare categories — not a different model — is the fix.

7.5 Disaggregated results and error propagation

Fairness disaggregation (Whisper Large-v3, WER, 95% speaker-clustered CIs): Female 13.3% (12.4–14.2, 6 speakers) vs Male 18.6% (14.3–24.7, 3 speakers); Kananci 16.9%, Katsinanci

14.8%, with Sakkwatanci and Zazzaganci below the reporting floor (interval-only); age 15–29 at 14.1% vs 45+ at 17.8%. Where a subgroup has too little data for a reliable figure (fewer than 30 recordings or fewer than 3 speakers), we show only a range and exclude it from comparisons. All subgroup results describe the nine test-set speakers, not the wider population.

Error propagation (ASR → intent): on the test set (4,837 unique sentences), the champion classifier reaches 86.00% accuracy on perfect transcripts. Feeding ASR hypotheses costs: XLSR-53 −5.71pp, Whisper −6.57pp, MMS-1B-all −7.09pp, MMS-1B-fl102 −16.06pp. The relationship is non-linear — transcription error becomes disproportionately expensive downstream as it grows. The champion end-to-end pipeline (Whisper → Afro-XLMR) retains 79.43% intent accuracy.

7.6 Relationship to partner baselines

Our partner DSN/EqualyzAI produced earlier results under a different evaluation setup (a different test set, software configuration, and text-cleaning approach). Those figures are kept as reference points but are not directly comparable to the final results and are never treated as a scoreboard.

8. Clinical vs Non-Clinical Scope Declaration

Intended scope: support for frontline health workers — specifically, evaluation of the speech-recognition and intent-understanding layers that would sit inside CHW-facing voice tools (symptom capture, triage navigation assistance, structured documentation) for Hausa-speaking communities in northern Nigeria.

This benchmark and its underlying dataset are explicitly NOT: a clinical decision support system; a validated clinical device; a basis for autonomous clinical decision-making; or an instrument for diagnosis. AI outputs evaluated against this benchmark are intended to supplement and precede human judgment — a CHW or clinician remains the decision-maker. The dataset card carries matching use restrictions (no re-identification, no voice cloning, no biometric or surveillance use). Benchmark results must be interpreted strictly within this declared scope, and the per-class intent results make the boundary concrete: rare but clinically consequential intents (medication concern, treatment, follow-up) are not yet reliably recognized and must route to human review in any deployment.

9. Model and Deployment Metadata

Item	Detail
Models evaluated	ASR: openai/whisper-large-v3 (LoRA r=32, α=64), facebook/wav2vec2-large-xlsr-53 (full fine-tune), facebook/mms-1b-all and facebook/mms-1b-fl102 (Hausa adapters). Intent: Davlan/afro-xlmr-large, bert-base-multilingual-cased.
Architecture	Fine-tuning evaluation harness (no RAG, no prompting layer): seq2seq encoder–decoder vs CTC architectures compared under identical splits, metrics, and seeds.
Infrastructure	AWS cloud GPUs (NVIDIA A10G), with every training and evaluation run logged for full traceability.
Evaluation dates	Evaluation ran July–August 2026; final results released 9 August 2026.
Deployment medium	Benchmark infrastructure, not a deployed product. Target deployment context for evaluated models: CHW mobile and IVR channels; single-stream inference latency is published per model (Whisper 1.257s median/utterance vs XLSR-53 0.017s on A10G) as a deployment-relevant differentiator, with the caveat that deployment-class device latency is not yet measured.
Versions	All dataset, code, and metric versions are recorded in the repository for exact reproduction.

The full evaluation used about $390 of cloud GPU time (an earlier round used about $220) — well within the program's compute budget, with costs logged per model and per stage. Training cost inverts parameter-count intuition: XLSR-53's full fine-tune ($210.68) cost roughly twice Whisper's LoRA run ($97.39) for statistically indistinguishable accuracy, while at inference the compact CTC model is ~70× faster — the benchmark's central deployment trade-off finding.

10. Regional and Contextual Considerations

Domain	Health (community health / primary care)
Target user	Community health workers and CHW-facing digital health tools
Target beneficiary	Hausa-speaking patients and caregivers in northern Nigeria
Problem statement	AI voice tools proposed for CHW deployment are almost never evaluated on the language, dialects, and clinical register they would serve; Hausa (70M+ speakers) had no clinical speech benchmark before this work.
Country / region	Nigeria — northern states (Kano, Katsina, Sokoto, Kaduna/Zaria)
Geographic focus	Mixed urban and rural; field-simulated recording conditions
Languages	Hausa (ha/hau), monolingual, with English reference glosses
Standard applied	Nigerian Federal Ministry of Health and WHO guidance for clinical content validation; native-Hausa clinician and linguist consensus for linguistic appropriateness

Northern Nigeria carries a disproportionate share of the national disease burden (maternal mortality 993 per 100,000 live births; ~67M annual malaria cases; DPT3 coverage ≈57%), and many patients reach the health system only through CHWs, in Hausa. Communication is a determinant of health in this setting, which is why the benchmark evaluates the communication layer.

11. Biases, Limitations, and Failure Modes

The benchmark's technical report enumerates eight limitations; the dataset card carries the same list. In summary:

- Read speech, not spontaneous dialogue — scores are an upper bound on live-conversation performance.
- Nine evaluated speakers — every headline figure is measured on nine test voices; no representativeness claim is made, and three subgroups had too little data for a reliable single figure.
- Dialect and age labels come from our partner's records rather than EHA's own collection, so they carry the partner's annotation choices.
- Field-simulated recording conditions, not genuine field recordings over CHW devices and networks.
- Class imbalance — several intent classes have single-digit test support (medication concern: 1 test item) and cannot yet be evaluated reliably.
- One automated stability check proved stricter than intended; we kept its original outcome rather than adjusting it after the fact. It does not affect the winning models and will be re-tuned in the next release.
- Two minor reproducibility caveats are documented openly: some software versions are not yet fully locked, and one repeat evaluation differed by 0.04 percentage points. Neither affects any conclusion.

Measured failure modes: word-boundary and morphology errors dominate ASR mistakes (diacritics, gemination, segmentation); intent misclassification concentrates in semantically adjacent pairs (treatment ↔ prevention/counselling; risk disclosure ↔ symptom reporting); rare-intent unreliability is driven by training-set size; and transcription error degrades downstream intent accuracy non-linearly — a weak recognizer (fl102) costs 2.3× the intent accuracy of a mid recognizer despite only 1.35× the WER. Error rates are higher for male speakers and for speakers aged 45 and above than for female and younger speakers; these gaps are reported openly, with confidence ranges. Every correction made during the project is dated and preserved in the technical report rather than silently edited.

12. Final Deliverables and Access

Deliverable	Where
Finalized evaluation dataset	huggingface.co/datasets/eHealthAfrica/voices-of-care-hausa (audio, transcripts, labels, demographic fields, and a contamination canary)
Complete documentation (data, labels, benchmarks, metadata)	Dataset card, technical report, and datasheet in the repositories, plus this document
Benchmark run on ≥1 model	4 ASR + 2 intent models, 5 seeds each, two campaigns
Public evaluation results on Hugging Face	Reference results published on the dataset card; full mirrors in the code repository
Public dataset release (where feasible)	See below
Private dataset share with program team	Reviewer access grantable on HF within 24 hours

12.1 Data sharing status and the Endless Health requirement

Public release of the dataset was the project's plan from the outset, with 31 July communicated as the target date across our correspondence. The request to hold the dataset back for a fair frontier-model comparison reached us on 7 August — after the partner copy had been public since mid-July, and while our own release was in its final upload step (it went live on 9 August). The request and the release crossed in transit rather than the request being disregarded. Once the requirement was clear, access was restricted: the eHealth Africa dataset repository is now private, and the partner copy is being restricted in coordination with DSN/EqualyzAI. Reviewer access is available to the program team on request, and public access will be restored once the comparison window closes. Two built-in protections also address the contamination concern: every release carries a unique canary marker, so any future model trained on this data can be detected, and download counts during the public window were minimal.

13. Next Steps and Contact

Planned next: evaluation of additional state-of-the-art and frontier models, tighter software version locking, and a release that supports per-domain breakdowns.

Contact: Tahir Buhari · Senior Research Lead, EHA Group · tahir.buhari@ehealthnigeria.org · +234 817 667 5896

EHA Clinics Limited / eHealth Africa · 4–6 Independence Road, Kano, Nigeria