VOICES OF CARE

A Hausa Clinical Speech Benchmark for Community Health AI

Final Submission Document

Health AI Benchmarking Initiative — The Agency Fund × Endless Health

Prepared in the structure of the Grantee Technical Roadmap & Submission Template (v1.0). 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	voc-2026-08 (v1.1.1, released 9 August 2026; v1.0 preserved as v10-legacy)
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	10 August 2026 · Version 1.1

1. Executive Summary

Voices of Care is the first publicly documented, clinically validated Hausa speech benchmark for community health AI. 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 two complete, fully reproducible evaluation campaigns against six models.

The benchmark of record is the v1.1 campaign (voc-2026-08), evaluated on a strictly speaker-disjoint and text-disjoint held-out test partition of 5,375 recordings, five training seeds per headline model, with speaker-clustered bootstrap confidence intervals and disaggregated reporting 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 regenerates from committed artifacts: pinned configurations, MLflow-tracked runs, frozen splits (seed 42), a version-pinned metric stack (jiwer 4.0.0, text normalizer 1.3.0), and results mirrors committed to the public code repository. An earlier v1.0 campaign (voc-2026-07, July 2026) is preserved byte-reproducibly as the v10-legacy configuration; its headline WERs (13.31% / 13.41%) are lower because its split protocol permitted speaker overlap between train and test — the v1.1 protocol removed that leakage, which is why the v1.1 figures are higher and more honest.

Against the initiative's final deliverables (template §13): 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 is being restricted (see §13) while the code, results, and documentation remain public.

2. The Dataset

2.1 Data source and provenance (template §3.1)

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 (template §3.2)

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), ~300 sentences (~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/Opus, MP3); resampled to 16 kHz mono for ASR

The corpus exceeds the initiative's minimum of 1,000 input/output pairs by 30×. The contributor base (~50 speakers) is below the indicative 100-user minimum; per the template's note on smaller datasets, this was flagged to the technical team and is mitigated by depth per speaker, exact gender balance, and four-way dialect stratification. The evaluated test roster is nine speakers (six female, three male) under the speaker-disjoint protocol — every disaggregated claim in this document is explicitly bounded by that roster (§12).

2.3 Modality (template §3.3)

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 — a non-English dataset, as the initiative prefers.

2.4 Privacy, consent, and rights (template §3.4)

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 (template §3.5)

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. Disaggregated results are published with speaker counts and below-floor suppression flags (cells with fewer than 30 utterances or fewer than 3 speakers publish confidence intervals only, and are excluded from ranking claims) precisely so readers can see how thin each cell is.

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). Post-project, raw voice files are securely deleted; only the de-identified release is retained.

Residual risk is disclosed rather than claimed away. The release is pseudonymous, not anonymous: speaker identifiers ship as deterministic pseudonyms (speaker_001…speaker_050), but the upstream partner release publishes speaker identifiers alongside the audio at the pinned revision this release joins against, so re-identification through that public join is possible and is documented as the primary residual mechanism on the dataset card. A secondary residual comes from demographic columns leaving small cells (the v1.1 test partition holds exactly one Zazzaganci speaker). Both routes are governed by the licence and consent terms' no-re-identification obligation, stated on the card. Cross-jurisdiction processing (Nigeria collection/processing; US funder review; EU-region cloud compute in AWS eu-west-1) is acknowledged per template §4.

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). The corpus does not contain multi-turn sessions or longitudinal conversation histories; speaker_type labels (patient statement, patient question, caregiver statement, caregiver question) preserve the discourse role each utterance plays.

5. Labels

The corpus carries two primary label families and two secondary ones. Documentation below follows template §6.1.

5.1 Ground-truth transcript (ASR reference)

Label name	hausa_transcription / transcript_raw (+ transcript_normalized, normalizer v1.3.0)
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 hayaƙi 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 provided for research use and are not part of the core benchmark scoring.

6. Health-Specific Evaluation Dimensions: Scope Declaration

In scope for this benchmark: (a) clarity and fidelity of transcription for the target user population, including Hausa-specific orthographic fidelity measured through 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 matters for Hausa: a single character — a diacritic, a hooked consonant, gemination — can change meaning, and the benchmark measured that normalization choices alone can move reported WER by up to 12.57pp on this corpus, which is why the normalizer is pinned and versioned.

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: frozen splits (seed 42), pinned configurations whose hash is logged on every MLflow run, committed results mirrors, and statistical machinery (speaker-clustered bootstrap percentile intervals; five training seeds per headline model reported as mean ± sample standard deviation) that makes uncertainty explicit rather than implied.

7.2 Split protocol

The v1.1 primary ASR split is speaker-disjoint AND text-disjoint: no speaker and no sentence appears on both sides of any boundary, stratified by dialect and gender at speaker level. This is the harder, more honest protocol — a model cannot score by recognising a voice or a sentence it trained on. Enforcing double disjointness costs data (5,913 recordings dropped by construction; 140 with no speaker attribution excluded), which is published, not absorbed.

Configuration	Train	Validation	Test	Protocol
asr (v1.1, primary)	16,717	1,855	5,375	Speaker- and text-disjoint, seed 42

Configuration	Train	Validation	Test	Protocol
v10-legacy (frozen v1.0)	24,003	2,999	2,998	Sentence-disjoint only (speakers overlap)
intent (v1.0 = v1.1)	13,469	748	749	Unique-sentence split, seed 42

7.3 ASR results — campaign of record (voc-2026-08)

Held-out test partition, 5,375 recordings, nine speakers. Mean ± sample standard deviation across five training seeds (0/1/42/1234/2024); MMS-1B-fl102 is a single-seed reference. Every cell resolves to an MLflow run via committed mirrors.

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 Afro-XLMR seed collapsed to a single predicted class and is excluded as degenerate under a published arithmetic rule; the basis is stated, not elided. The ~18-point accuracy-to-macro-F1 gap is the benchmark's central intent finding: classes with fewer than ~250 training examples score below 0.75 F1 while all larger classes score ≥ 0.88 — per-class reliability is governed by training-set size (with class separability as the instructive exception: Referral reaches 0.96 F1 on just 158 examples). The path to closing the gap is targeted data collection, not architectural change.

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%. Cells below the floor (fewer than 30 utterances or 3 speakers) are suppressed from point estimates and rankings. These are statements about nine evaluated speakers, not population estimates — the disaggregation machinery, floors included, is itself a deliverable of v1.1.

Error propagation (ASR → intent): on the v1.1 test partition (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 the v1.0 campaign and partner baselines

The v1.0 campaign (voc-2026-07, shipped 21 July 2026) reported Whisper 13.31% / XLSR-53 13.41% WER on a 2,998-recording test split that was sentence-disjoint but allowed speaker overlap between train and test. v1.1's leakage audit motivated the stricter protocol; the v1.0 partition is preserved byte-reproducibly as the v10-legacy config so those numbers remain checkable. The apparent regression from 13.3% to 15.1% is the price of removing speaker leakage — it is the more truthful number.

DSN/EqualyzAI's earlier fp16-era results (Whisper 23.84%, XLSR-53 17.80%, MMS-1B-fl102 64.72% WER; intent 88.8% accuracy) are carried in the benchmark as calibration anchors, transcribed in a committed mirror. They were produced by a different harness on a different (1,346-example) test basis with undocumented normalization, and are treated as harness-comparability diagnostics, never as a scoreboard: re-running the fl102 anchor configuration inside the EHA harness reproduced 32.61pp, a −32.11pp difference that measures the distance between measurement setups, not model quality.

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 EC2 G5 (NVIDIA A10G 24GB), Databricks Runtime 16.4 LTS ML, PyTorch 2.6.0 / CUDA 12.4, bf16 throughout; MLflow experiment tracking with config hash and dataset digest on every run.
Evaluation dates	v1.0 campaign: July 2026 (voc-2026-07). v1.1 campaign of record: July–August 2026 (voc-2026-08), released 9 August 2026 as v1.1.1.
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	Dataset v1.1.1 · code repository tagged v1.1.1 · results provenance pinned to revision e684c322 · normalizer 1.3.0 · jiwer 4.0.0.

Campaign compute cost is published per model and stage from MLflow-stamped GPU hours: the v1.1 campaign totals $390.67 (EC2 on-demand basis; dollar rates flagged as assumed, GPU-hours are the durable measured quantity), following the v1.0 campaign's $219.63 — both well inside the program's GPU envelope. 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 subgroup cells sit below the reporting floor.
- Derived dialect labels — per-recording dialect/age labels come from the partner's public release at a pinned revision, joined offline with a frozen coverage gate; they inherit the upstream annotation decisions.
- 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.
- Deployment-class latency unmeasured — published latency is server-class single-stream instrumentation.
- A pre-registered seed-stability gate (R-02) was later measured to be calibrated against a spread smaller than run-to-run noise; the disposition stands as fired (re-adjudicating post hoc would fit the gate to the outcome), the champion is unaffected, and re-calibration is a pre-registered v1.2 item.

- Reproducibility is bounded and the bounds are published: the dependency stack is not fully pinned, and one same-artifact evaluation discrepancy of +0.04pp was observed; both remain open and disclosed in v1.1.

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. Male-speaker and 45+ WER run higher than female and younger cells in the current partition, disclosed with intervals rather than smoothed over. Transparency notes: every correction made during the campaign is dated and preserved in the technical report rather than silently edited, including the withdrawal of an earlier incorrect reading of the partner's 64.72% figure as a scoring artifact.

12. Final Deliverables and Access

Template deliverable	Status	Where
Finalized evaluation dataset	Delivered	huggingface.co/datasets/eHealthAfrica/voices-of-care-hausa (v1.1.1; 3 configs: asr, intent, v10-legacy; 12 columns incl. dual transcripts + contamination canary)
Complete documentation (data, labels, benchmarks, metadata)	Delivered	Dataset card (Data Statements v2 + Croissant-RAI metadata) · docs/TECHNICAL_REPORT.md · docs/DATASHEET.md (Gebru et al. framework) · this document
Benchmark run on ≥1 model	Delivered — 6 models	4 ASR + 2 intent models, 5 seeds each, two campaigns
Public evaluation results on Hugging Face	Delivered	Reference results published on the dataset card; full mirrors in the code repository
Public dataset release (where feasible)	Prepared — access restricted at program request	See below
Private dataset share with program team	Available on request	Reviewer access grantable on HF within 24 hours

12.1 Data sharing status and the Endless Health requirement

The template records that datasets are expected to be shared privately with the program for quality review, with public/private/hybrid release at the organization's discretion; the Agency Fund has since relayed Endless Health's requirement to withhold public dataset publication pending a fair frontier-model comparison. EHA's v1.1.1 dataset release on Hugging Face is accordingly being placed under restricted access (private/gated), with reviewer access available to the program team immediately; the partner-hosted copy of the source data is being restricted in coordination with DSN/EqualyzAI. Two protections relevant to the contamination concern are built into the release itself: a contamination canary (a unique GUID embedded per a published rule, so any future model trained on this data is detectable) and dual raw/normalized transcript columns that keep third-party re-scoring possible without republication. The evaluation code, configurations, results, and documentation remain fully public — the benchmark is auditable without the audio being crawlable.

13. Submission Record and Version History

This document compiles the full conversation record with the program team, as requested.

Date (2026)	Event
30 June – 1 July	Final submission to The Agency Fund: dataset (30,000 recordings), final project documentation report, intent classification results (fp16-era baseline, 88.8% accuracy).
2–5 July	ASR addendum (2 of 3 models) and updated dataset/report delivered; program requested zipped audio, filename-keyed CSV, and reproduction codebase.
7–9 July	Technical partner unable to release evaluation scripts (proprietary infrastructure); EHA secured commitment to written methodology documentation and launched a fully independent, open evaluation pipeline under Group CEO oversight.
17 July	Partner delivered updated reports with reproducibility sections and corrected figures (five training-pipeline bugs resolved; fl102 64.72% → 20.13% on their harness). EHA pipeline at 4 of 5 phases, 634 tests green.
21–22 July	v1.0 milestone shipped (voc-2026-07): 6-model leaderboard, 3 seeds, $219.63 GPU spend; results and code published at github.com/eHealthAfrica/voices-of-care.
23 July – 9 Aug	v1.1 "Rigor & Disaggregation": speaker+text-disjoint re-split with leakage audit, 5-seed extension, speaker-clustered bootstrap CIs, gender/dialect/age disaggregation with reporting floors, latency instrumentation, error-propagation re-run, contamination canary, 1,693+ tests green. v1.1.1 dataset released to Hugging Face 9 August.
10 August	This submission document; dataset access restricted per Endless Health requirement; v1.2 (SOTA roster & reproducibility hardening) roadmapped.

Planned next iteration (v1.2, roadmapped): evaluation of additional state-of-the-art and frontier models, full dependency pinning to close the two published reproducibility limits, re-calibrated stability gates, and a domain-tagged release enabling per-domain disaggregation.

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