# Kare — Code-Switching ASR Benchmark

Prepared for the Sahara CodeSwitch Africa Challenge · benchmark run 2026-09-14 · reproducible via `make benchmark` · full data in the repo's `benchmark/` directory

**Kare** is a voice-first AI health assistant for Nigeria: patients speak to it in whatever mix of English and their own language comes naturally, and it remembers them between visits. Its speech-to-text layer is the single most safety-critical piece — a mis-heard drug name or dosage flows straight into an automated interaction check. **Sahara** is the speech API this hackathon provides (built by Intron); this report benchmarks it against four alternatives to justify using it in Kare's production voice pipeline, rather than assuming it's the right choice.

**What this measures.** Nigerian patients speak in code-switched sentences — “the doctor talk say my BP dey high”: Yoruba/Pidgin grammar carrying English clinical terms. Kare's voice layer must get **both** the matrix language and the embedded English right, since the English words are usually the clinically load-bearing ones (drug names, “blood pressure”, “twice daily”). Standard ASR leaderboards are monolingual and miss this, so we built our own.

**Data.** 20 conversations (128.5 minutes) sampled from Intron's **AfriSwitchCare** (simulated doctor–patient consultations, English code-switched with Yoruba, Hausa, Igbo and Nigerian Pidgin) — 4 conversations per language, stratified across each language's code-mixing range, plus **Swahili as an out-of-family control** (a language none of Kare's users speak, so it reveals whether a model has learned West-African speech generally or merely memorised the target set). Audio is 16kHz mono; segments are cut to ≤110s (both Sahara's sync endpoint and Whisper's practical limits sit near there) and re-joined before scoring. The exact clip set is frozen by a seeded random draw so every number here is reproducible from the repo.

**Metrics, and why. WER/CER** (Word/Character Error Rate, pooled not averaged per-clip) is the standard ASR yardstick — the share of words (or characters) the model got wrong. **Med-term / number recall** is the *downstream-task* metric: Kare's agent runs triage scoring and drug-interaction checks directly off the transcript, so what matters clinically isn't overall word accuracy but whether medical terms and dosages survive — a transcript can have a mediocre WER and still be clinically usable, or a decent WER and drop the one number that mattered. **CMI** (Code-Mixing Index — a standard 0–100 score for how tightly two languages are interleaved within an utterance; higher means denser switching) **error** and **switch-point error** (same English-dictionary tagger applied to both reference and hypothesis, so its bias cancels) measure whether a model preserves the code-switching structure itself, not just individual words — a model can get most words right while still flattening a code-switched sentence into one language, which these two catch and WER alone would not.

## Headline results (5 models)

| Model | WER | CER | Med-term recall | Number recall | CMI err | RTF | $/audio-min |
|---|---|---|---|---|---|---|---|
| Sahara (Intron) STT | 53.1 | 41.0 | 71.0 | 53.7 | 4.4 | 0.15 | n/p |
| Groq · whisper-large-v3-turbo | 59.4 | 39.9 | 71.8 | 54.2 | 6.5 | 0.13 | $0.0007 |
| Groq · whisper-large-v3 | 64.1 | 43.1 | 63.1 | 59.3 | 7.6 | 0.05 | $0.0019 |
| faster-whisper tiny (local CPU) | 85.2 | 67.8 | 15.5 | 20.2 | 7.4 | 0.41 | $0.0000 |
| NCAIR/N-ATLaS (per-language)* | 99.3 | 78.3 | 8.1 | 11.6 | 23.1 | 1.08 | $0.0000 |

* NCAIR/N-ATLaS — three Whisper-small models, one fine-tuned per language, released by Nigeria's national AI institute (NITDA's NCAIR, with Awarri) — ships no Pidgin or Swahili checkpoint; scored on its own Yoruba/Hausa/Igbo subset only (41 segments), not penalised for the other two. All other models: 80/80 segments, all 5 languages. Lowest WER: **Sahara**.

## WER by language

| Model | Yoruba | Hausa | Igbo | Pidgin | Swahili | Swahili Δ |
|---|---|---|---|---|---|---|
| Sahara (Intron) STT | 66.1 | 45.5 | 61.0 | 55.7 | 34.9 | –22.2 |
| Groq · whisper-large-v3-turbo | 62.3 | 90.6 | 67.3 | 67.9 | 29.3 | –42.7 |
| Groq · whisper-large-v3 | 63.9 | 83.9 | 67.8 | 91.9 | 30.5 | –46.3 |
| faster-whisper tiny (local CPU) | 87.2 | 94.2 | 81.6 | 94.7 | 75.3 | –14.2 |
| NCAIR/N-ATLaS (per-language)* | 97.0 | 101.6 | 100.5 | — | — | — |

“Swahili Δ” = Swahili WER minus the mean of the four Nigerian languages. A large negative Δ means the model is markedly worse on the languages it's meant for than on an unrelated control — i.e. it leans on broad multilingual pretraining rather than West-African-specific coverage. Sahara has the smallest gap of the general-purpose comparison set; it is the only model built specifically for this language family.

## Downstream-task performance (what Kare's agent actually depends on)

Kare's agent calls a drug-interaction checker and a triage scorer directly off the transcript text. These two recall metrics isolate whether the words those tools need — medical terms (“insulin”, “hypertension”) and spoken numbers/dosages (“39 degrees”, “twice daily”) — actually survive transcription, independent of overall WER.

| Model | Med-term recall | Number/dosage recall | Usable for triage & drug-check? |
|---|---|---|---|
| Sahara (Intron) STT | 71.0% | 53.7% | Yes — best of the set |
| Groq · whisper-large-v3-turbo | 71.8% | 54.2% | Yes — close second |
| Groq · whisper-large-v3 | 63.1% | 59.3% | Partially |
| faster-whisper tiny (local) | 15.5% | 20.2% | No |
| NCAIR/N-ATLaS (per-language) | 8.1% | 11.6% | No (see finding below) |

**Accuracy vs. code-mixing intensity (CMI-bucketed WER, low→high):** Sahara 68.4 / 46.1 / 49.9 — Groq-turbo 75.4 / 50.1 / 58.2 — Groq-v3 90.2 / 53.2 / 57.4 — faster-whisper 91.7 / 76.9 / 89.6. Sahara is the only model that gets *better* as code-mixing intensifies — every general-purpose model degrades or stays flat.

## Qualitative findings, per model

**Sahara (Intron) STT**

**Strength:** the only model that improves as code-mixing gets denser — built for exactly this task. Best med-term/number recall alongside Groq-turbo. **Weakness:** weaker on Yoruba specifically (66.1 WER) than Hausa/Pidgin; per-minute pricing isn't published.

**Groq whisper-large-v3-turbo**

**Strength:** cheapest and fastest hosted option (RTF 0.13, $0.0007/min); best med-term recall of any model. **Weakness:** 90.6 WER on Hausa — the worst single language/model cell among the general-purpose options; degrades under heavy code-mixing.

**Groq whisper-large-v3**

**Strength:** best number/dosage recall (59.3%). **Weakness:** worst-performing Whisper variant overall; 91.9 WER on Pidgin, its worst language.

**faster-whisper tiny (local CPU)**

**Strength:** zero marginal cost, no network dependency — usable as an offline fallback. **Weakness:** catastrophic med-term recall (15.5%) makes it clinically unusable; hallucinates repeated phrases on Pidgin and misdetects the language entirely on some clips.

**NCAIR/N-ATLaS (Whisper-small, per-language)**

**Strength:** genuinely fluent monolingual Yoruba/Hausa/Igbo output when the input *is* monolingual. **Weakness:** on code-switched clinical speech it doesn't drop English words, it *transliterates* them phonetically into the target script — the reference's opening "okay, good morning" comes back respelled as Yoruba-orthography approximations of the same English sounds. Fluent output, but it erases the English half of every code-switched sentence — exactly the failure mode expected of a monolingual fine-tune on this task, and exactly what this benchmark exists to surface.

## Fairness & limitations

- AfriSwitchCare is simulated consultations read by voice artists, not real clinic audio — cleaner acoustics and less crosstalk/noise than Kare meets in the field. Absolute WERs here are a floor, not a ceiling.
- 20 conversations (80 segments) is enough to rank models, not enough for tight per-language confidence intervals; treat sub-language gaps under ~3 points as noise.
- The CMI/switch-point tagger is a dictionary lookup applied identically to every model, so cross-model comparisons are fair, but it has a systematic bias on proper nouns and loanwords.
- Two of Kare's five languages (Igbo, Hausa) have the least training data across every model tested — this benchmark makes that gap visible rather than hiding it.

**Bottom line.** Sahara wins on WER, is the only model that improves rather than degrades as code-mixing intensifies, and ties for the best downstream med-term/number recall — the three things that matter most for a clinical voice agent. That's why Kare's production voice pipeline calls Sahara, not the general-purpose alternatives benchmarked here.

Full methodology, per-clip transcripts, charts and the frozen manifest (seed 20260915) are in the repository at `benchmark/` — every number above reproduces from `make benchmark`.