Project name

Nku — Offline Medical AI for Pan-African Triage

Your team

W. Elorm Yevudza Jnr, MD/MS — Solo developer. Born and raised in Ghana. Incoming surgery resident, NewYork-Presbyterian Queens. MD/MS Columbia VP&S (2025); BA Neuroscience, Amherst College (2019). Maintains clinical connections with health professionals across Ghana for pilot coordination and field validation.

Problem statement

In Sub-Saharan Africa, fewer than 2.3 physicians serve every 10,000 people — far below the WHO's recommended 44.5 health workers [1,2]. Over 450 million people lack accessible primary care [3]. Community Health Workers (CHWs), the frontline of care, frequently lack reliable diagnostic tools due to equipment deficiencies and supply stock-outs [4] — yet nearly all carry Android smartphones [5].

Powerful clinical AI models exist, but require reliable cloud connectivity. In rural Sub-Saharan Africa, 25% of the population lacks mobile broadband entirely [6]. Cloud-based AI is impractical where it is needed most.

Target user: A CHW in a rural region with a $60+ TECNO or Infinix phone (3GB+ RAM) and no stable internet [7]. She needs immediate triage guidance — offline, on her existing device — to determine which patients require urgent referral. Transsion brands (TECNO, Infinix, itel) hold >50% of the African smartphone market [8].

Impact & Deployment logistics: Distributing a 2.3GB LLM to a CHW in a rural region is a primary logistical hurdle. We address this via a multi-tiered infrastructure strategy:

1. **Pilot Sideloading:** Supervisors provision phones centrally via MDM or side-load the model directly via MicroSD card, requiring zero village internet bandwidth.
2. **Play Asset Delivery (PAD):** The 50MB core app is installed via the Play Store. It automatically downloads the 2.3GB model as an `install-time` asset when the CHW intercepts 4G/LTE cellular connectivity in larger towns.
3. **Peer-to-Peer Viral Sharing:** African smartphone culture involves local peer-to-peer file transfer. Only one CHW per clinic needs to download the model via 4G; they can then use Android's native *Nearby Share* or *Xender* to beam the 2.3GB `.gguf` file to other CHWs' offline phones at ~30MB/s over Bluetooth or offline peer-to-peer protocols.
4. **Zero-Rated Data:** For scaled Ministry of Health rollout, the Google Play download URL is "zero-rated" through partnerships with major Mobile Network Operators (MNOs) (e.g., MTN, AirtelTigo), ensuring the massive download does not deduct from the CHW's personal cellular data balance.

For reviewers installing the APK directly: The app automatically detects if MedGemma is missing on first triage and downloads the 2.3GB Q4_K_M model from HuggingFace, validates its SHA-256 checksum, and proceeds — no manual setup required. Simply install the APK, connect to Wi-Fi, and run a triage.

Overall solution

Every line of clinical reasoning executes on the phone itself.

Nku (Ewe: "eye") runs MedGemma entirely on $60+ Android smartphones. It is a proof-of-concept prototype; field validation with CHWs is the critical next step.

MedGemma 4B is irreplaceable in this system. It is the sole clinical reasoning engine, performing the interpretation that transforms structured sensor data and symptoms into triage assessments — a capability

no smaller model possesses. Cloud inference fails completely in low-connectivity zones. Only MedGemma, quantized to Q4_K_M and deployed via llama.cpp JNI on ARM64, enables the offline + accurate combination Nku requires.

The Nku Cycle is a multi-stage orchestration pipeline where MedGemma serves as the clinical reasoning engine within a self-adapting workflow:

Stage	Component	Size	Function
1. Sense	Nku Sentinel (5 detectors)	0 MB	Camera + microphone → structured vital signs
2. Translate	Android ML Kit (On-Device)	~30MB/lang	Translates 59 supported local languages to English (offline). Unsupported indigenous languages are translated via cloud fallback; if no internet, they are not allowed.
3. Reason	MedGemma 4B (Q4_K_M)	2.3GB	Clinical reasoning (in English) on symptoms + sensor data
4. Translate	Android ML Kit	~30MB/lang	English → supported local language output (offline) or via cloud fallback
5. Speak	Android System TTS	0 MB	Spoken result in local language
Fallback	World Health Organization / Integrated Management of Childhood Illness (WHO/IMCI) rules	0 MB	Deterministic triage if MedGemma unavailable (e.g., insufficient available RAM)

Crucially, because optical sensors historically exhibit diagnostic bias against darker skin tones (classified as Types V and VI on the Fitzpatrick skin typing scale), every Nku camera modality in the "Sense" stage is engineered to be "Fitzpatrick-aware" to ensure equitable accuracy. Sensor confidence must exceed 75% for inclusion in MedGemma's prompt; below-threshold readings trigger a localized ⚠ warning.

Adaptive Memory Management & Fallback: Android `mmap` pages the 2.3GB MedGemma model dynamically to fit within 3GB+ RAM devices without thrashing. An explicit runtime check ensures ~1.2 GB of available RAM before load. If unavailable, or if the OS OOM killer intervenes, Nku gracefully falls back to deterministic WHO/IMCI rule-based triage. Every stage operates independently for maximum stability.

Structured Prompting & Compression: MedGemma requires structured data, not raw arrays. Nku explicitly compresses verbose multimodal sensor arrays (e.g., 478 MediaPipe facial landmarks) into concise biomarkers (e.g., `Edema index: 0.52`). **Crucially, this halves token consumption, bypassing the strict 2048 token KV-Cache limit on 3GB RAM devices.** This unlocks enough token space for MedGemma to utilize full **Chain-of-Thought (CoT)** reasoning before outputting its JSON response, delivering a median 53% accuracy improvement [9] and +20pp triage gain (detailed in Appendix G).

Technical details

Edge AI — Quantization: We achieve 69% model size reduction (8GB → 2.3GB) via Q4_K_M quantization, retaining 81% of MedQA accuracy. IQ2_XS (1.3GB) with medical imatrix calibration outperforms Q2_K but suffers a critical 12.6pp accuracy drop to 43.8%, losing its ability to reliably parse multi-morbidity in high-stakes triage. Thus, Q4_K_M is deployed (full benchmark table in Appendix B).

HeAR Respiratory Screening: Sub-Saharan Africa bears a massive burden of TB, COPD, and childhood pneumonia [1, 30, 31]. Nku activates Google's HeAR Event Detector (1.1MB TFLite, FP32 fallback) to screen 2–5 seconds of cough/breathing audio in ~50ms. Unlike traditional dense audio embeddings, the 1.1MB Event Detector passes a *structured summary* (`Cough Probability: 0.82` , `Risk Score: High`) directly into the prompt. MedGemma excels at reasoning over these explicit probabilities alongside other symptoms (detailed in Appendix D).

When MedGemma is unavailable, the app displays a transparency banner identifying the triage as guideline-based (WHO/IMCI) with actionable recovery steps — all in the CHW's selected language.

Safety: 8-layer `PromptSanitizer` at every model boundary (zero-width stripping, homoglyph normalization, whitespace normalization, base64 detection, regex patterns, character allowlist, delimiter escaping, length capping). Auto-pause at 42°C. Always-on "Consult a healthcare professional" disclaimer. *Note on CVE-2025-69872 (`diskcache`): Since the HuggingFace Hub client pins `diskcache` , our optional inference API implements strict local directory restrictions and cache partitioning to mitigate unauthorized cache injection.*

46 Pan-African languages (14 clinically verified): ML Kit on-device for supported national languages (100% offline). Unsupported indigenous languages trigger a UI connectivity alert and use Cloud Translate fallback mechanics; all final reasoning occurs entirely on-device in English.

Prize Track: Main + Edge AI — Q4_K_M compression (8GB→2.3GB), mmap loading on $60+ phones (3GB+ RAM), llama.cpp JNI (NDK 29, ARM64 NEON), systematic 4-level quantization benchmark (IQ2_XS with medical imatrix calibration), 100% on-device inference with MedGemma bundled via Play Asset Delivery (2.3GB, install-time), and CHW-initiated respiratory screening via HeAR on-device: Event Detector (MobileNetV3-Small, 1.1MB TFLite with FP32 fallback) classifies 8 health sound events in ~50ms, with risk scores and event classes fed to MedGemma for TB/COPD/pneumonia triage.

Open source: Nku is fully open source under the Apache License 2.0. Source code, scripts, and calibration data on GitHub. Quantized model weights on HuggingFace (subject to Google Gemma Terms of Use).

See `kaggle_submission_appendices.md` for full references [1–32], language list (46), calibration scenarios, MedGemma reasoning examples, sensor pipeline details, and safety architecture.

Development tooling: *Google Antigravity (Gemini 3 Flash/Pro, Gemini 3.1 Pro, Claude Opus 4.5/4.6); OpenAI Codex IDE (GPT 5.3 Codex).*