REFERENCE ARCHITECTURE · TOP-1% · AI FOR HUMANITY

Local Agentic AI for Healthcare

RAG + CAG on a low-cost LLM

Trustworthy clinical AI that runs offline, on cheap hardware — for rural clinics, field medicine, and privacy-critical settings where the cloud isn't an option.

Swipe: the stack, then advanced capabilities →

Atiqul Islam

atiqulislam.dev

THE WHOLE STACK

Runs on one low-cost device

A local planner, a small quantized clinical LLM, guidelines cached in-context (CAG) and an on-device patient record + references (RAG), with safeguards and a clinician — all offline. PHI never leaves the device (HIPAA); no per-token cost; cheap hardware.

CAG

Cache the guidelines

Clinical protocols and guidelines are preloaded into the model's context and KV-cache. Grounded answers with no retriever — and near-instant repeat queries for a stable knowledge base.

LOCAL RAG

Retrieve the record

The patient's record — a local FHIR store — and references are retrieved on-device, only when needed. CAG for the guidelines, RAG for the record. Nothing leaves the device.

THE MODEL

A small model, grounded, is enough

A quantized open model (3-8B, GGUF) runs on a laptop or mini-PC. Grounding in the guidelines and record compensates for size — the trust layer matters more than raw model power.

AGENTIC + SAFE

Agentic, safe — locally

A local planner routes to triage, documentation, and med-check agents. A grounding + safety guard (cite-or-refuse) and a clinical critic run on-device — the same trust pattern, no cloud.

PRIVATE & LED

Private by default, clinician-led

PHI never leaves the device — HIPAA by construction. Clinicians stay in control via a local dashboard, and an optional offline sync refreshes guidelines when connectivity returns.

Engineered to be top-1%

Production-grade capabilities layered on the core stack — across performance, access, grounding, safety, and privacy.

Runs on a ~$300 mini-PC | 100% offline-capable | 0 PHI egress | $0 per-token cost | Hot-swap LoRA

Illustrative design targets — not benchmarks.

- Performance on cheap hardware
- Access & multimodal
- Grounding & accuracy
- Trust, safety & verification
- Privacy, security & lifecycle

Performance on cheap hardware

Quantized + speculative decoding
4-bit GGUF with draft-model speculative decoding for fast CPU tokens.

KV + semantic response cache
Near-instant answers for repeated questions.

Adaptive compute budget
Cheaper path for easy queries; heavy reasoning only when needed.

Access & multimodal

On-device voice (ASR + TTS)
Hands-free; usable by low-literacy patients.

On-device vision triage
Reads wounds, skin lesions, X-rays at the point of care.

Multilingual / low-resource
Local and under-served languages, fully offline.

Grounding & accuracy

GraphRAG over clinical ontologies
SNOMED CT, ICD-10, RxNorm, LOINC knowledge graph.

Hybrid retrieval + reranker
CAG + RAG fused and ranked by a cross-encoder.

Structured FHIR extraction
Free-text notes become validated FHIR, on-device.

Trust, safety & verification

Uncertainty-aware abstention
Low-confidence cases escalate instead of guessing.

Continuous eval + red-team
Golden-set evals and adversarial probes run locally.

Clinical hard-stops & rules
Deterministic allergy, dosing, interaction red-lines.

Tamper-evident audit
Hash-chained, verifiable local decision log.

Privacy, security & lifecycle

Encrypted-at-rest + secure enclave
Attested boot; the data boundary is the device.

Federated learning
Improve a shared model with no PHI shared.

Signed guideline packs + CRDT sync
Fresh knowledge, conflict-free offline.

LoRA adapter hot-swap
Specialty/region adapters without re-downloading.

Agentic actions with approval
FHIR write-back only after clinician sign-off.

Patient / Encounter & Planner / Loop

Patient / Encounter

On-device, offline intake

EX *At a remote clinic with no reliable internet, a nurse enters a patient's symptoms: fever, cough, 3 days.*

TECH Local clinic UI; on-device capture; No network dependency; Local patient identity

CTRL PHI never leaves the device / LAN; Consent state attached at intake; Local access control

FIN No per-encounter cloud cost; it runs on owned clinic hardware.

FALL If capture is incomplete, it prompts for the missing vitals before proceeding.

Planner / Loop

Bounded clinical loop

EX *Splits the case into triage, documentation, and medication-safety steps, looping until each is grounded.*

TECH Lightweight local orchestrator; Bounded iterations + token cap; Deterministic routing

CTRL Hard step cap; clinician-scope enforcement; Tool allow-list per agent; No autonomous clinical action

FIN Tight loop caps conserve scarce local compute and keep latency clinically acceptable.

FALL If the loop stalls, it returns the grounded summary so far and routes to the clinician.

Triage & Clinical Docs

Triage

Urgency, never autonomous

EX *Flags the fever + cough + age combination as needing same-day clinician review, not self-care advice.*

TECH Guideline-grounded triage over CAG; Vitals + red-flag rules; Quantized on-device model

CTRL Cite-or-refuse to guidelines; Red-flag -> mandatory clinician; Conservative bias (escalate when unsure)

FIN CAG-cached guidelines mean triage rarely needs retrieval, saving local compute.

FALL Any uncertainty or red flag escalates to the clinician rather than reassuring the patient.

Clinical Docs

SOAP draft for review

EX *Drafts a structured encounter note (SOAP) from the captured symptoms for the clinician to review.*

TECH Note generation (SOAP) on-device; FHIR-shaped output via local RAG; Citations to guideline sources

CTRL Draft only - the clinician signs; Cite-or-refuse; PHI stays local

FIN Reuses templates + the local model; no cloud transcription or documentation fees.

FALL Low-confidence sections are marked '[needs clinician]' rather than asserted.

Med Check & CAG · Guidelines

Med Check

Interactions + allergies

EX *Catches that the suggested antibiotic interacts with the patient's existing prescription and flags it.*

TECH Local drug-interaction DB + RAG; Dose / allergy rule checks; Quantized model for reasoning

CTRL Never prescribes - advises only; Deterministic interaction rules; Allergy list honored strictly

FIN Rule + local-DB checks dominate; the model is used sparingly, keeping compute low.

FALL If the drug reference lacks data, it flags 'unverified - clinician check' instead of clearing the med.

CAG · Guidelines

Guidelines preloaded in cache

EX *National treatment guidelines for common conditions are preloaded into the cache for instant grounding.*

TECH Cache-augmented generation; Per-guideline cache builds; Versioned cache artifacts

CTRL Only approved guidelines cached; Rebuilt on guideline update; Provenance retained

FIN A one-time cache build replaces per-query retrieval - critical on constrained clinic hardware.

FALL On a cache miss, it falls through to local RAG rather than failing.

Local RAG · FHIR & Low-cost LLM

Local RAG · FHIR

FHIR + refs, on-device

EX *Retrieves the patient's local FHIR history and a drug monograph the cache didn't hold.*

TECH On-device vector DB; local FHIR store (HAPI); Hybrid search + reranker; Local embeddings

CTRL PHI access scoped + audited; Only approved references indexed; Version / edition guard

FIN Fully on-device retrieval - no cloud bill; the index is sized to clinic storage.

FALL No grounded source -> the agent declines and flags the gap for the next sync.

Low-cost LLM

Quantized, on-device

EX *A 4-bit quantized model runs triage, notes, and med checks on a clinic mini-PC with no connectivity.*

TECH Quantized GGUF via llama.cpp / Ollama; CPU / modest-GPU inference; Careful context management

CTRL Model pinned / versioned; Output filtered by guard + critic; Deterministic settings

FIN One-time hardware, no token bills - predictable cost where connectivity and budgets are thin.

FALL Under load, it queues and shortens responses rather than failing the encounter.

Grounding + Safety & Clinical Critic

Grounding + Safety

Grounded + scope-limited

EX *Blocks a draft that gave a definitive diagnosis, downgrading it to a grounded differential for clinician review.*

TECH Local grounding verifier; Safety + hallucination checks; PHI-leak filters

CTRL Unsafe output blocked, not shown; Citations required; Mandatory clinician escalation on flags

FIN Deterministic checks cost little versus the risk of an unsafe clinical answer.

FALL Any failure -> withhold and route to the clinician with the reason logged.

Clinical Critic

Adversarial second pass

EX *A second model pass challenges the note: 'differential omits pneumonia given these vitals' - and it's added.*

TECH Critic / judge prompt as a second pass; Checks against guidelines + vitals; Confidence scoring

CTRL Must pass the critic before the clinician sees it; Disagreements escalate, not auto-resolve; Decision recorded for audit

FIN A short critic pass is far cheaper than a missed condition; reserved for clinical outputs only.

FALL If critic and agent conflict, the case goes to the clinician unresolved rather than guessing.

Clinician (HITL)

Clinician (HITL)

Human signs everything

EX *The clinician reviews the AI's triage, note, and med flags, edits, and signs - the AI never acted alone.*

TECH Local dashboard + review queue; Override + feedback loop; Deferred sync of audit on connect

CTRL Human sign-off mandatory; Full local audit trail; Clinician can disable any agent

FIN Risk-tiered review focuses clinician time on flagged + material cases.

FALL Everything uncertain waits safely for the clinician - the default is always a human.

YOUR TURN

Bringing AI to every clinic?

This is how I'd architect low-cost, offline, trustworthy clinical AI — local RAG + CAG on a small model. Building it in the open as AI for Humanity.

DM me “assessment”

Atiqul Islam
Fractional Cloud & AI Architect for SMBs · atiqulislam.dev