

— REFERENCE ARCHITECTURE · AI FOR HUMANITY

Local Agentic AI for Education

RAG + CAG on a low-cost LLM

Trustworthy AI tutoring that runs offline, on cheap hardware — so disconnected and under-resourced classrooms aren't left behind.

Swipe through the stack →



Runs on one low-cost device

A local planner, a small quantized LLM, curriculum cached in-context (CAG) and an on-device vector store (RAG), with safeguards and a teacher — all offline. Benefits: no per-token cost, data stays on-device (FERPA), cheap hardware.



Cache the curriculum

For a fixed syllabus, preload it into the model's context and cache the KV-states (Cache-Augmented Generation). Grounded answers with no retriever — and near-instant repeat queries.



Retrieve the long tail

For the bigger library — textbooks, past papers — a small on-device vector store retrieves only what's needed. CAG for the core, RAG for the rest. Nothing leaves the device.



A small model, grounded, is enough

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



Agentic, safe, and honest — locally

A local planner routes to tutor, assessment, and feedback agents. A grounding + safety guard and an academic-integrity check run on-device — the same trust pattern, no cloud.



Private by default, teacher-led

Student data never leaves the device — FERPA/COPPA by construction. Teachers oversee via a local dashboard, and an optional offline sync refreshes the curriculum when connectivity returns.

Learner & Planner / Loop

— Learner

On-device, offline intake

EX *In a rural classroom with no internet, a student asks the offline tablet to explain photosynthesis.*

TECH Local UI on tablet / mini-PC; on-device STT; No network dependency; Local identity store

CTRL Data never leaves the device / LAN; Age-band tag from local profile; Local rate-limit

FIN Zero per-query cloud cost - inference runs on hardware already owned; only an upfront device cost.

FALL If voice capture fails, it falls back to typed input; nothing requires connectivity.

— Planner / Loop

Bounded local agent loop

EX *Splits 'explain + quiz me' into an explanation step and a check step, looping until the student gets it.*

TECH Lightweight local orchestrator; Bounded iterations (small step cap); Deterministic routing

CTRL Hard iteration + token cap (the small model is slow); Grade-scope enforcement; No tool outside the allow-list

FIN Tight step caps matter most here - every extra loop costs scarce local compute and battery.

FALL If the loop stalls, it returns the best grounded explanation so far and suggests asking the teacher.

Tutor & Assessment

Tutor

Grounded in local corpus

EX Explains photosynthesis using the offline science textbook preloaded into the model's cache.

TECH Generation over CAG + local RAG; Quantized on-device model; Citations to local sources

CTRL Cite-or-refuse against the local corpus; Read-only corpus; Reading-level matching

FIN CAG preloads the curriculum into the KV-cache so common topics need no retrieval pass at all.

FALL If a topic isn't in the local corpus, it says so rather than hallucinate - and flags it for sync later.

Assessment

Local formative checks

EX Generates three offline quiz questions on the light and dark reactions to check understanding.

TECH Local item generation; On-device scoring; Misconception detection

CTRL Standards-aligned items only; No high-stakes grading without the teacher; Local fairness checks

FIN Reuses a preloaded item bank first; only generates new items when needed, saving local compute.

FALL Uncertain scores are deferred to the teacher dashboard at the next sync.

Feedback & CAG · KV-cache

Feedback

Hint ladder, not answers

EX Nudges: 'Which part of the leaf captures light?' instead of stating chlorophyll outright.

TECH Deterministic hint ladder; Local hint-level tracking; Minimal model use

CTRL No full answers on graded work; Integrity enforced locally; Age-appropriate tone

FIN The hint ladder is mostly template-driven, sparing the slow local model.

FALL After max hints, it offers the worked example and flags for teacher follow-up.

CAG · KV-cache

Curriculum preloaded in cache

EX The whole term's science unit is preloaded into the model's cache, so it answers instantly with no lookup.

TECH Cache-augmented generation (prefilled KV); Per-unit cache builds; Versioned cache artifacts

CTRL Only approved units cached; Cache rebuilt on curriculum update; Provenance retained in the cache

FIN CAG trades a one-time cache build for near-zero per-query retrieval cost - ideal on constrained hardware.

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

Local RAG & Low-cost LLM

Local RAG

On-device retrieval

EX For a niche question outside the cached unit, it retrieves from the on-device index of extra readings.

TECH On-device vector DB (SQLite-vec / FAISS); Hybrid search + small reranker; Local embeddings

CTRL Only approved local materials indexed; Course-scoped access; Edition / version guard

FIN Runs entirely on-device - no cloud retrieval bill; the index is sized to fit local storage.

FALL No passage found -> the tutor declines and marks the gap for the next online sync.

Low-cost LLM

Quantized, on-device

EX A 4-bit quantized 7-8B model runs the whole lesson on a classroom mini-PC with no GPU.

TECH Quantized GGUF model via llama.cpp / Ollama; CPU or modest-GPU inference; Small context window managed carefully

CTRL Model + weights pinned / versioned; Output filtered by the guard layer; Deterministic settings for reproducibility

FIN The core cost story: one-time hardware, no token bills - predictable and offline-friendly.

FALL If the device is overloaded, it queues requests and serves shorter responses rather than failing.

Grounding + Safety & Academic Integrity

— Grounding + Safety

Grounded + age-appropriate

EX *Blocks an answer that wandered off the cached corpus, returning only grounded, age-appropriate content.*

TECH Local grounding verifier; On-device safety classifier; PII / toxicity filters

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

FIN Lightweight local checks add negligible compute relative to generation.

FALL Any failure -> withhold and show a safe fallback; flag for the teacher at the next dashboard sync.

— Academic Integrity

Teach, don't shortcut

EX *On a graded worksheet, it teaches the method and cites the textbook instead of filling in answers.*

TECH Local assignment-type detection; Citation enforcement; Plagiarism heuristics

CTRL No answers on graded work; Provenance preserved; Teacher-configurable

FIN Rule-based and local - it adds essentially no marginal cost.

FALL Ambiguous cases default to teaching and are flagged for the teacher.

Teacher (HITL)

— Teacher (HITL)

Human in control on sync

EX *At day's end the teacher syncs the tablet and sees which students mastered photosynthesis and which need help.*

TECH Local dashboard + deferred sync; Override + feedback into the learner model; Flag queue surfaced on connect

CTRL Human sign-off for high-stakes actions; Local audit trail; Teacher can disable any agent

FIN Batched sync avoids constant connectivity; teacher time is spent only on flagged items.

FALL Anything uncertain waits safely on-device for teacher review at the next sync.

YOUR TURN

Bringing AI to every classroom?

This is how I'd architect low-cost, offline, trustworthy education 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