FYP Plan:

Team & ownership

- **LLM/ML Engineer A (Vision + Edge)**
    - Nodule detection/segmentation & classification (2.5D UNet-lite/MobileNet-UNet), Grad-CAM.
    - Edge optimization (ONNX → TensorRT/NCNN, INT8 quant, pruning), device bring-up (Jetson / Pi 5).
- **LLM/ML Engineer B (Reporting + NLP)**
    - Structured-to-text LLM (doctor report + patient-friendly summary, multilingual).
    - Safety rails (template grounding, factuality checks), bilingual outputs.
- **Frontend Engineer**
    - Kiosk UI (touch, offline-first, Hindi + 1–2 regional languages to start).
    - Scan viewer with overlays, consent flow, tele-referral, and report export.
- **Data Engineer**
    - DICOM ingestion, de-identification, preprocessing, data versioning.
    - Training/eval pipelines, metrics dashboards, ABDM/FHIR payload builder, audit logs.

5-month roadmap (by sprint/month)

Month 1 — Foundations & baselines

Goals: get data moving, stand up baselines, choose hardware, lock metrics & success criteria.

- **Data Eng**
    - Build **DICOM → anonymized NIfTI/PNG** pipeline; store with versioning.
    - Preprocess: lung windowing, spacing normalize, slice packing (2.5D).
    - Set up experiment tracking (MLflow/W&B), secure storage, role-based access.
- **LLM/ML A**
    - Baseline **lightweight nodule segmentation** (MobileNet-UNet/2.5D) on public set.
    - Draft **explainability** (Grad-CAM) overlay pipeline.
- **LLM/ML B**
    - Define **structured schema** (JSON) the LLM will receive (counts, size, location, Lung-RADS).

    - Spin up **small LLM** baseline with strict templating for **doctor report** (EN/Hindi).
- **Frontend**
    - Low-fi UI flows (paper → Figma), pick stack (FastAPI + simple web UI).
    - Prototype viewer: load slices, step through axial stack, simple overlays.
- **All**
    - Pick **edge target** (Jetson Nano/Orin Nano preferred; Pi 5 as fallback).
    - **Success gates:** Seg Dice ≥0.75 on held-out; end-to-end latency goal draft (<30–90s/scan, device-dependent).

Month 2 — Edge, explainability, and grounded LLM

Goals: make it run on the box; ground the LLM to the vision outputs; add multilingual basics.

- **Data Eng**
    - Add **consent + audit logging** scaffolding (DPDP-compliant).
    - Build **evaluation harness** (Dice, sensitivity @ FP/scan, AUC; plus report factuality checks).
- **LLM/ML A**
    - **Quantize & compile** (ONNX → TensorRT/NCNN), INT8 with calibration.
    - Start **FP-reduction** post-proc; tune for **<45s** segmentation on device demo CT.
- **LLM/ML B**
    - **Grounded generation:** LLM only writes from provided JSON; add **factuality validator**.
    - Implement **patient summary** style guide (≤8th-grade Hindi + English).
- **Frontend**
    - **Offline-first** app shell; local queue for intermittent connectivity.
    - Multi-language switch; display **Grad-CAM + mask** overlays.
- **All**
    - Dry-run on kiosk hardware; capture latency, memory, and power numbers.
    - **Success gates:** on-device inference <60s for typical volume; factuality errors <5% on 50-case sample.

Month 3 — Clinical polish & ABDM sandbox

Goals: quality jump, ABDM/FHIR integration in sandbox, tele-referral loop.

- **Data Eng**
    - **ABDM Sandbox** onboarding; generate **FHIR bundles** (DiagnosticReport, ImagingStudy).

    - Add **tele-referral payload** + basic analytics (screened count, prevalence).
- **LLM/ML A**
    - Improve small-nodule recall; add **uncertainty flagging** (route to radiologist if low confidence).
    - Expand explainability: per-nodule saliency thumbnails.
- **LLM/ML B**
    - **Multilingual expansion** (add 1–2 regional langs, e.g., Marathi/Telugu).
    - Safety rails v2: **forbidden content filters**, numerical sanity checks (sizes, laterality).
- **Frontend**
    - **Consent UX** (ABHA/OTP flow placeholder during sandbox).
    - One-tap **"Send to specialist"** with status tracking and printable summary.
- **All**
    - **Radiologist review panel** (20–30 retrospective cases): score accuracy, completeness, clarity.
    - **Success gates:** FHIR bundle accepted in sandbox; rad panel avg ≥4/5 on clarity & correctness; small-nodule sens +5–8% vs M2.

Month 4 — Field-pilot build & hardening

Goals: productionize for pilot in rural setting; ruggedize; finalize language + power/offline behavior.

- **Data Eng**
    - **Crash-safe local store** with encryption; nightly **sync when online**.
    - Add **audit exports** and **ops dashboards** (latency, failures, case mix).
- **LLM/ML A**
    - Device-specific kernels & fusions to hit **<30–45s** end-to-end on Jetson-class box.
    - Robustness suite: motion/noise, different scanners, low-dose protocols.
- **LLM/ML B**
    - **Human-in-the-loop tools** (quick edits, regenerate sections, red-flag guidance).
    - Patient summary **speech synthesis** option (Hindi) for low literacy.
- **Frontend**
    - **Technician-friendly wizard**: 4-step flow (Identify → Load → Analyze → Share).
    - **Power/Net resilience**: resume jobs after outage; USB export fallback.

- **All**
    - Dry-run **pilot workflow** with 1 partner site; train operators (2h module).
    - **Success gates:** zero-touch run in clinic sim; outage-recovery verified; operator CSAT ≥4/5.

Month 5 — Pilot, validate, and publishable results

Goals: run a small real-world pilot; lock metrics; write the paper/report; handoff for scale.

- **Data Eng**
    - Pilot data pipeline: **prospective logs**, consent artifacts, FHIR pushes to test/lite prod.
    - Final metrics report (technical + clinical + accessibility).
- **LLM/ML A**
    - Error triage from pilot; targeted fixes (e.g., calcified nodules, apical scarring).
    - Package **v1.0 edge model** with reproducible build + checksums.
- **LLM/ML B**
    - **Readability & trust** study (10–20 patients): comprehension Qs; refine language.
    - Lock **dual-report templates**; export examples for publication appendix.
- **Frontend**
    - Polish: print layouts (A4), dark mode (sun glare), larger tap targets.
    - **In-app feedback** button; operator issue capture.
- **All**
    - **Pilot outcomes deck** for govt stakeholders (accuracy, time saved, referrals made).
    - Draft **journal/conference** manuscript (methods, edge benchmarks, ABDM integration, HCI findings).
    - **Go/No-Go** checklist for wider rollout.

Deliverables & exit criteria (v1)

- **On-device CT pipeline** (segmentation + classification + Grad-CAM) **<45s** typical scan on Jetson-class device.
- **Dual reports** (doctor + patient) in **English + Hindi** (+1 regional), **grounded** to structured findings; **factuality errors <3–5%** on expert review.
- **ABDM sandbox**: push **FHIR** DiagnosticReport/ImagingStudy with consent artifacts.
- **Kiosk UI**: offline-first, multilingual, touch-friendly; tele-referral; printable.

- **Pilot metrics**: small-nodule sensitivity uplift vs M2 baseline; operator CSAT ≥4/5; patient comprehension ≥80% on key questions.

Rural & compliance guardrails baked-in

- **Offline-first** everything; resumable jobs; low-power tolerance.
- **Privacy:** on-device processing by default; encryption at rest; explicit consent; minimal data to cloud.
- **Accessibility:** large fonts, bilingual content, optional **text-to-speech**.
- **Escalation:** uncertainty → human review; red-flag banner for urgent findings.
- **Govt alignment:** ABHA-based ID, ABDM/FHIR, audit trails for public deployments.

Cadence

- **Weekly**: team standup + demo.
- **Bi-weekly**: stakeholder/radiologist check-in.
- **Monthly**: gateway review vs success gates (ship / slip decisions).

Risks & mitigation (quick)

- **Edge latency too high:** earlier INT8 + operator-friendly ROI cropping; Orin Nano option.
- **LLM hallucinations:** strict templating + numeric validators + blocklist; human-in-loop on red-flags.
- **Scanner variability:** robust augmentation + multi-site test set; calibration guide.
- **Connectivity gaps:** store-and-forward; USB export; SMS/WhatsApp fallback for referrals (where permitted).

If you want, I can drop this into a one-pager Gantt and a RACI grid so you can hand it straight to the team.