Neel Raval

pnraval2007@gmail.com • (+91) 8217897162 • Bengaluru, India
linkedin.com/in/neel-raval • neelluu.com • github.com/neelraval13

Summary

Full-stack software engineer with 3+ years of experience architecting AI-powered products across healthcare, enterprise, and consumer domains. Currently leading engineering at Odyssey Therapeia, where I drive the development of a patented computer vision pipeline for cervical cancer screening alongside the company's broader product surface. Deep expertise in agentic LLM systems, multimodal AI (Claude, Gemini, Gemma), structured-output contracts, RAG, function calling, and streaming inference, with production backends on FastAPI, Next.js, and AWS. Co-inventor on 1 pending patent and co-author of peer-reviewed research published in the International Journal of Gynecological Cancer. Focused on shipping AI systems that are honest about their limits, validated end-to-end, and measurable in clinical and business impact.

Education

Bachelor of Technology (B.Tech) in Electronics and Computer Engineering, *Vellore Institute of Technology* 07/2019 – 05/2023

Work Experience

Engineering Lead, Software & Product **08/2024 – Present**
Odyssey Therapeia *Bengaluru, India*

- Spearheading development of an AI-powered diagnostic pipeline for a proprietary cervical cancer screening device; architecting computer vision and machine learning models to analyze live video feeds from a camera-equipped speculum and deliver real-time tissue analysis and clinical parameters to assist gynecologists.
- Architected an extensible, scalable tender discovery portal for parent company Regency using Next.js and TypeScript; engineered a plug-and-play ingestion architecture automating data aggregation from 9 government platforms, eliminating manual search overhead and accelerating the enterprise bidding pipeline.
- Engineered an end-to-end billing and case management workflow, replacing a manual 4-invoice generation process with a streamlined module and establishing a robust data foundation for real-time case analytics.
- Designed a comprehensive, real-time inventory management system on Convex to track the complete lifecycle of chemotherapy machines and accessories; implemented reactive stock alerts, warranty tracking, and a centralized hospital and physician database.
- Architected and deployed a fully mobile-optimized client web platform for ReACH with advanced SEO and Answer Engine Optimization (AEO) strategies; engineered a seamless appointment booking flow that increased scheduled consultations by ~35% and improved organic search visibility within 3 months of launch.
- Reduced AWS infrastructure costs by 20% through cloud architecture review, resource right-sizing, and CI/CD optimization.
- Built a full 3D Avatar production pipeline by adapting and enhancing open-source tooling, giving the company complete control over the production process and removing dependency on paid external tools.

Software Developer I **07/2023 – 08/2024**
Ankr Health *Bengaluru, India*

- Developed the frontend of a physician answering service using HTML, CSS, JavaScript, and Bootstrap, integrating Auth0 for secure authentication and role-based access control to enhance user interaction and satisfaction.
- Built a cross-platform mobile application for iOS and Android using React Native, expanding the company's reach and enabling patient-facing features across both app stores.
- Customized, white-labeled, and deployed an open-source application for internal use with Next.js, Prisma, PostgreSQL, and Docker, improving operational efficiency and reducing third-party tooling costs.
- Designed and shipped responsive marketing landing pages with optimized Core Web Vitals and on-page SEO, boosting online engagement and lead generation.

MERN Stack Intern **07/2022 – 09/2022**
Yellow Sapphire Technologies (Bizco) *Remote*

- Built an e-commerce Kirana marketplace using the MERN stack (MongoDB, Express, React, Node.js) to digitize neighborhood retail and enhance the online shopping experience.
- Integrated the Paystack payment gateway with offline-first payment handling, reducing payment errors and increasing successful transaction rates.
- Implemented IBM Cloudant with PouchDB for offline data synchronization of store purchases, ensuring reliable data handling under intermittent connectivity.
- Deployed the REST API on Heroku and frontend on Netlify with CI/CD pipelines, ensuring seamless integration, faster load times, and higher user engagement.
- Leveraged Redis, TailwindCSS, PouchDB, TypeScript, ChakraUI, and Zustand to deliver a performant, maintainable product within tight timelines.

Neel Raval

pnraval2007@gmail.com • (+91) 8217897162 • Bengaluru, India
linkedin.com/in/neel-raval • neelluu.com • github.com/neelraval13

Projects

Jogen GitHub

AI Resume-Tailoring Advisor - React 19, FastAPI, Anthropic Claude, SSE, Pydantic

- Architected a streaming AI advisor that ingests resumes across 5 file formats (PDF, DOCX, TXT, MD, LaTeX) and job descriptions, returning structured tailoring analysis - 0-100 fit scoring, verbatim line edits, and skill-gap roadmaps - via a typed three-stage pipeline (parse → fetch-jd → analyze) with ∼25s warm latency.
- Engineered a custom Server-Sent Events transport with a UTF-8-safe stream parser and a Pydantic contract gate that schema-validates every LLM response before it reaches the client, surfacing parse failures explicitly instead of leaking raw model output downstream; hardened with a 37-test suite (sub-second runtime) and a ∼96 MB Docker image.

Flexius GitHub

AI Gym Coach - Next.js 16, React 19, Turso, Drizzle, PWA, Web Push

- Architected a provider-neutral LLM abstraction with adapters for 3 providers (Claude, Gemini, OpenAI) behind a single interface, enabling vendor-agnostic tool-calling and auto-detection of user-supplied API keys via prefix matching swapping models becomes a one-line change at the call site.
- Engineered a two-phase agentic loop over 13 typed tools (workout logging, plan mutation, history queries) on an 11-table Drizzle/Turso schema with 42 ownership-scoped query functions, plus web-search grounding as a fallback when no tool fires cleanly separating "query private data" from "query the public web" under one chat interface.
- Built an offline-first PWA end-to-end (∼22k LOC across 200 files, 116 React components) with a custom service worker, 3 IndexedDB stores for write-queueing, and auto-replay on reconnect users log sets and view plans with zero connectivity and sync atomically when back online, with PR detection guarded against multi-device race conditions.

Reel Analyzer API GitHub

Multi-tenant Vision-LLM SaaS - FastAPI, Redis, Gemini, Qwen2.5-VL via Ollama, iOS Shortcuts

- Architected a provider-neutral vision-LLM backend with a `BaseAnalyzer` abstraction over Gemini (cloud) and Qwen2.5-VL (local via Ollama), using Pydantic schemas as decode-time response constraints the model emits guaranteed-valid structured JSON instead of free text parsed post-hoc.
- Engineered the multi-tenant control plane on Redis: Stripe-style API keys hashed with SHA256 at rest, two-window sliding rate limits (10 req/min burst + 100 req/day cost guardrail) returning RFC-compliant 429s with `Retry-After` headers, schema-versioned analysis cache that invalidates on provider/model/schema change without migrations, and permanent usage counters separated from ephemeral enforcement state.
- Hardened for production with an SSRF-defending URL allow-list, an error taxonomy carrying orthogonal `http_status` and `retryable` flags wired into Tenacity retry predicates, request-scoped `structlog` JSON logging, and a cookie-authenticated admin dashboard that returns 404 (not 401) when disabled to hide its existence from probes.

Sanjeevani GitHub

Offline Clinical Decision Support - Gemma 4 (multimodal, edge), Ollama, FastAPI, SSE

- Architected a 6-stage agentic clinical reasoning chain (feature extraction → differentials → urgency triage → investigations → referral → action plan) with strict per-stage JSON schemas, running fully offline on Gemma 4 E4B via Ollama 0 bytes leave the device, addressing connectivity and data-sovereignty constraints of rural Indian PHCs.
- Engineered the multimodal feature-extraction stage with 5 explicit anti-hallucination guardrails, forcing the model to describe visual findings independently of the clinician's text input rather than confabulating findings to match the narrative the failure mode that breaks trust in clinician-facing imaging AI.
- Built a streaming FastAPI backend with Server-Sent Events emitting per-stage progress across 5 event types (`step_start`, `step_complete`, `summary`, `complete`, `error`), plus an optional 7th stage that produces a 150-200 word patient-friendly summary in 7 Indian languages (Kannada, Hindi, Telugu, Tamil, Malayalam, Bengali, Marathi).

SOC-Analyst OpenEnv GitHub

Agentic RL Benchmark - OpenEnv (Meta), FastAPI, Pydantic, HuggingFace Spaces

- Built a stateful RL-style environment for training and evaluating LLM agents on Security Operations Center workflows 18 alerts across 3 difficulty tiers, exposed over 7 REST endpoints conforming to Meta's OpenEnv interface (`reset`/`step`/`state`) so any compatible agent or training loop can be plugged in against a frozen benchmark.
- Designed a 7-action agent contract (classify, query_context, escalate, contain, dismiss, correlate, submit_report) culminating in a 10-alert multi-stage campaign correlation problem at 50% signal-to-noise, where 5 kill-chain alerts (recon → initial access → C2 → lateral movement → exfiltration) are buried in 5 decoys and must be linked via shared IPs and temporal ordering.
- Engineered a two-channel reward system - dense per-step shaping with 6× severity weighting and asymmetric penalties (missed true-positives cost 1.67× more than false alarms) plus task-specific terminal graders combining accuracy, F1 on chain reconstruction, and process-quality signals like "did the agent investigate before classifying."

Neel Raval

pnraval2007@gmail.com • (+91) 8217897162 • Bengaluru, India
linkedin.com/in/neel-raval • neelluu.com • github.com/neelraval13

Skills

AI & Machine Learning
LLMs • Agentic LLM Systems • Multimodal AI • Computer Vision • Tool-Use & Function Calling • Structured Outputs (Pydantic / JSON Schema) • Retrieval-Augmented Generation (RAG) • Embeddings • Prompt Engineering • LLM Evaluation & Reward Shaping • Streaming Inference • Anthropic • Anthropic Claude • Google Gemini • OpenAI • Local LLM Deployment (Gemma, Qwen2.5-VL via Ollama)

Engineering & Frameworks
Python • TypeScript • JavaScript • FastAPI • Pydantic • Next.js • React • React Native • Node.js • Express • Django • REST APIs • GraphQL • tRPC • gRPC • WebSockets • Server-Sent Events (SSE) • Streaming • Rate Limiting • Authentication (Auth0, OAuth) • Drizzle • Prisma • TailwindCSS • Zustand • Progressive Web Apps (PWA) • Testing (Vitest, Jest, Mocha, Chai) • Structured Logging (structlog)

Database, Cloud & DevOps
AWS • Nginx • Docker • PostgreSQL • Redis • MongoDB • Turso • Convex • CI/CD • GitHub Actions • Git • Vercel • Render • HuggingFace Spaces

Leadership & Product Strategy
System Architecture • Technical Roadmapping • Engineering Leadership • Product Strategy • Code Review • Mentorship • Agile / Scrum

Publications & Patents

Patent: Cervical Imaging Using AI-Powered Speculum with a Deployable Camera
Indian Patent Office • Filed, Pending Grant *01/2026*

- Co-inventor of a novel computer vision-assisted medical screening methodology and device leveraging live video feeds for real-time cervical tissue analysis.

Visispec: An AI-Enabled Vaginal Speculum Integrating Coaxial Illumination and Intra-Blade Imaging for Cervical Screening
International Journal of Gynecological Cancer • Read Paper *02/2026*

- Co-authored peer-reviewed abstract detailing advanced AI screening methodologies; delivered an oral presentation of the research findings at the AICOGS 2026 conference, Delhi.

Design of Obstacle Assault Game and Turtle Sidestep Game Using Non-Fungible Tokens (NFT)
4th National Conference on Communication Systems (NCOS '22) *2022*

- Co-authored research on integrating NFT-based ownership mechanics into interactive gaming experiences on the Ethereum blockchain.

Activities

Hackathon Juror & Mentor **01/2026 - Present**
WitchHunt *Remote*

- Serving as a juror evaluating hackathon submissions across technical execution, product thinking, and engineering rigor.
- Mentoring two participating teams through architecture reviews, technical problem-solving, and shipping guidance under hackathon timelines.

Founder & Vice President **07/2020 - 07/2021**
Hack Club *VIT Chennai, India*

- Founded and led a 50+ member technical coding club; organized hackathons, workshops, and educational initiatives that drove community engagement and hands-on developer upskilling.