

PROFILE

AI Engineer with 9+ years of software engineering experience, including 3+ years shipping production LLM applications. Delivered 5 GenAI products end-to-end on a multi-tenant SaaS platform at Intellora AI — a conversational WhatsApp agent, RAG-powered retrieval, AI lead qualification, a multi-step travel-planning agent, and an AI-powered CRM — driving a +41% lift in lead-scoring precision, -38% LLM costs, and 100% sales-team adoption in 4 weeks. Deep expertise in RAG & vector search, agent orchestration, LLM evaluation & observability, and real-time streaming AI interfaces. Previously scaled systems to 50K+ concurrent users at Licious and built a telehealth platform 0→1 as founding engineer at Allo Health.

TECH STACK

AI & ML Foundations

- Machine Learning
- Deep Learning
- Natural Language Processing (NLP)
- Transformers
- Generative AI

LLMs & Foundation Models

- OpenAI GPT-4o
- Claude
- Gemini
- Llama 3
- Mistral
- Hugging Face

RAG & Vector Databases

- Pinecone
- pgvector
- Weaviate
- Qdrant
- Chroma
- Cohere Rerank
- Retrieval-Augmented Generation
- Hybrid BM25 + Dense
- Semantic Chunking
- Grounded Citations

Agents & Orchestration

- LangGraph
- LangChain
- LlamaIndex
- Vercel AI SDK
- Pydantic
- Function Calling
- MCP · Tool Use
- ReAct · Multi-Agent

Evals, MLOps & Observability

- LangSmith
- Langfuse
- Helicone
- Golden Datasets
- CI Regression Gates
- Production Monitoring
- Guardrails & PII

EXPERIENCE

AI Engineer · Intellora AI

Jan 2024 – Present

Bengaluru · Multi-tenant GenAI SaaS platform

- Shipped 5 production LLM products end-to-end on a multi-tenant SaaS platform (NestJS, Next.js, PostgreSQL + pgvector, Kafka, OAuth2/JWT); defined the 12-month AI roadmap across 4 product squads.
- WhatsApp AI Agent: conversational agent on WhatsApp Business API (GPT-4o + LangGraph tool orchestration) with context persistence, per-tenant token metering, guardrails, and human handoff — +35% qualified leads, -28% low-intent traffic.
- AI Lead Qualification: LLM scoring agent combining structured CRM features with conversation embeddings (text-embedding-3-large + pgvector), evaluated on a 5K-row golden dataset — +41% sales-accepted-lead precision over the rule-based baseline.
- AI Chat Summarisation: Claude-based hierarchical map-reduce pipeline (topics, sentiment, action items) over multi-day conversations — manager review time cut from ~12 min to <90 sec per thread.
- AI Travel Itinerary Agent: multi-step planning agent on LangGraph state machines calling flight/hotel/POI APIs as tools, with Pydantic validation and SSE streaming — <6s p95 via prompt caching and parallel tool execution.
- AI-Powered CRM: auto-drafted follow-ups, deal summaries, next-best-action, and NL-to-SQL with read-only guardrails — 100% sales-rep adoption within 4 weeks.
- RAG Pipelines: retrieval over tenant knowledge bases (semantic chunking, hybrid BM25 + dense, Cohere reranking, citations); owned the prompt-data-eval lifecycle with LangSmith/Langfuse tracing and CI regression tests gating every prompt change.
- Cut LLM spend -38% at constant eval quality via Haiku/Sonnet/Opus model routing, prompt caching, structured-output retries, and per-tenant token budgets with cost/latency observability.
- Built token-streaming chat UIs in Next.js (Server Components, SSE/WebSockets); drove team-wide adoption of Claude Code and Cursor — feature delivery time -30%.

Founding Full-Stack Engineer · Allo Health

Jun 2022 – Dec 2023

Bengaluru · Consumer telehealth, 0 → 1

- Built the entire consumer telehealth platform (Next.js, TypeScript, Node.js, PostgreSQL) — booking, payments, diagnosis flows — 0 to production in 4 months; grew engineering from 1 → 6.
- Established testing & delivery culture: Jest units, 120+ Cypress and 150+ Playwright E2E journeys in CI; led checkout A/B experiments — +22% conversion.

- Software Development Engineer II · Licious

Aug 2020 – Jun 2022

Bengaluru · Unicorn-stage D2C platform

- Engineered a distributed consumer platform serving **50K+** concurrent users (Next.js, Node.js, PostgreSQL) with real-time streaming, rate limiting, and observability; ran A/B infrastructure with 8+ concurrent experiments.
 - Scaled the team 3 → 11, defined the technical hiring bar, and authored architecture RFCs adopted across 5 product teams.
- Senior Software Engineer · Noticeboard

Jul 2018 – Aug 2020

Bengaluru

- Built a real-time chat application on WebSockets (React, Node.js, PostgreSQL), an LMS training platform, and a live geolocation attendance system; maintained **82%+** test coverage with Jenkins CI/CD.
- Software Engineer · Housejoy

Jul 2017 – May 2018

Bengaluru

- Built a consumer home-services e-commerce platform (React.js) serving 40K+ monthly users — booking, dynamic pricing, checkout; page load improved **~30%** via component-level code splitting.
- Software Development Engineer · Practo

Sep 2016 – Jun 2017

Bengaluru

- Developed doctor-facing dashboards and internal healthcare tools (React.js, Python Flask); integrated REST APIs and maintained **80%+** Jest coverage.

KEY ACHIEVEMENTS

- Shipped 5 production GenAI products in 18 months** — conversational AI, RAG, agents, evals, and streaming UIs — owning the full lifecycle from prompt design to deployment on a multi-tenant AI SaaS.
- Measurable business impact from AI:** **+41%** lead-qualification precision, **+35%** qualified leads, **+22%** checkout conversion, and **100%** sales-team adoption within 4 weeks of rollout.
- Cut LLM costs **~38%** with zero quality loss** — verified on golden-dataset evals — through model routing, prompt caching, and per-tenant token budgets; a repeatable cost-engineering playbook.
- Engineered for production speed:** **<6s p95** multi-tool agent latency and support review time cut ~12 min → **<90s** (8× faster) via parallel tool execution and hierarchical summarisation.
- 9+ years of engineering leadership at scale** — systems for 50K+ concurrent users at unicorn-stage Licious, 0-to-production in 4 months as founding engineer, teams grown 1→6 and 3→11.

Languages & Backend

Py

Python

Fa

FastAPI

TS

TypeScript

N

Node.js

Ne

NestJS

K

Kafka

Microservices

SSE · WebSockets

OAuth2/JWT · Multi-Tenant

Data, Cloud, Frontend & Tooling

Pg

PostgreSQL

R

Redis

AWS

AWS Bedrock

G

GCP

Az

Azure OpenAI

D

Docker

K8

Kubernetes

Pt

PyTorch

Re

React 18

N

Next.js

CC

Claude Code

Cu

Cursor

Cy

Cypress

Pw

Playwright

J

Jest

EDUCATION

M.Tech, Computer Science

IIIT Hyderabad

2014 – 2016

B.E., Computer Science

Medicaps Institute of Science and Technology, Indore

2009 – 2013

CORE FOCUS

RAG & Vector Search

Agent Orchestration

LLM Evals & Observability

Streaming AI Interfaces

Cost Engineering

Prompt Engineering

Fine-Tuning (LoRA/QLoRA)

Multi-Tenant SaaS

BY THE NUMBERS

5

GenAI products shipped end-to-end

9+

years of engineering experience

~38%

LLM cost at constant quality

+41%

lead-scoring precision lift

50K+

concurrent users served

<6s

p95 multi-tool agent latency