

MOHIT KUMAR SINGH

Senior Software Engineer | Generative AI | Agentic AI | iOS

Noida, Uttar Pradesh, India • +91 94121 43791 • mohit.singh.coder@gmail.com • [linkedin.com/in/mohit-singh-ios](https://www.linkedin.com/in/mohit-singh-ios) • avioratech.com/Mohit • **Notice Period: Immediate**

PROFESSIONAL SUMMARY

Senior Software Engineer with 6+ years of production engineering, now building LLM-powered agents and RAG systems in Python. Designed a LangGraph booking agent that calls tools through an MCP server, pauses for human approval before any irreversible action, and survives restarts with a PostgreSQL-backed checkpointer. Built and deployed a grounded RAG service on AWS, measured with a hand-labelled evaluation set. Brings 12+ shipped iOS apps, on-device AI in a live App Store product, and experience leading engineering teams.

TECHNICAL SKILLS

Agentic AI:	LangGraph (state graphs, conditional routing, interrupts, checkpointing), Human-in-the-Loop approval, LLM Tool / Function Calling, Model Context Protocol (MCP), AI Agents
Generative AI & RAG:	RAG, Embeddings, Vector Search (Qdrant), Prompt Engineering, Grounded Generation with Citations, Hallucination Control, OpenAI API, Gemini, Groq, Apple Foundation Models
LLM Evaluation & Data:	Evaluation Dataset Design, Rubric-Based Labelling and Grading, Agent Trajectory Evaluation, Error Analysis, A/B Comparison
Backend & Python:	Python, FastAPI, asyncio, Pydantic, SQLAlchemy (async), Alembic, PostgreSQL, Server-Sent Events (SSE), REST APIs, Webhooks, JWT / Firebase Auth
Cloud, DevOps & Testing:	Docker, Docker Compose, AWS (ECS Fargate, ECR, SSM Parameter Store), Cloudflare Tunnel, Firebase, pytest, Integration Testing, Structured Logging, GitHub Actions
iOS:	Swift, SwiftUI, UIKit, Combine, Swift Concurrency, SwiftData, Core Data, MVVM, Fastlane, Bitrise, XCTest, App Store Connect

PROFESSIONAL EXPERIENCE

Independent AI Engineering — *Generative AI & Agentic AI Engineer* Noida, India | Aug 2026 – Present
Full-time building, testing and deploying AI systems in Python. Two projects, both described below.

ATLAS — *Agentic Flight Booking Assistant (LangGraph, MCP, FastAPI)* Sep 2026 – Present

- Built a LangGraph agent that searches live airline inventory (Duffel API) through an MCP tool server, holds seats and takes payments (Cashfree), using OpenAI tool calling.
- Gated every irreversible action behind human approval with LangGraph interrupts. The application, not the model, authorizes: approvals are idempotent, expire with the fare, and seats the airline has refused can never be offered again.
- Streamed agent progress to a web UI over FastAPI and SSE (202 + event stream), with ordered event ids, a full-state snapshot on reconnect, and one turn at a time per conversation.
- Made payments safe to retry: HMAC-verified webhooks, deduplication on a unique provider event id, and a payment state machine in which paid is final.
- Hardened for deployment: Firebase JWT verification (RS256), conversations owned per user, a PostgreSQL-backed LangGraph checkpointer so a paused agent survives restarts, and JSON logs carrying request, conversation and turn ids with token and latency per LLM call.
- Wrote 62 integration tests against a real server and PostgreSQL, with fakes only for the LLM, airline and payment providers. Shipped as a 246 MB multi-stage, non-root Docker image.

Grounded Answer Engine — *RAG Service over Indian Income-Tax Law, Deployed on AWS* Aug 2026 – Sep 2026

- Built a retrieval-augmented generation service over 3,286 indexed legal provisions, where every claim must cite a retrieved provision or the service declines to answer.
- Created a 40-question gold evaluation set with expected sections, refusal labels and a hand-graded correctness rubric, plus agent trajectory checks (expected tool order, redundant calls).
- Raised grounded answers from 74% to 100% and correct refusals from 72% to 98% (medians of three runs); raised retrieval hit rate from 83% to 93% with named-section lookup and duplicate suppression.
- Ran the pipeline as a LangGraph graph with a tool-using agent branch, pluggable embedder and vector-store interfaces, and streamed progress over SSE.
- Deployed two containers to AWS ECS Fargate behind a Cloudflare Tunnel with no inbound ports, keeping the API key in SSM Parameter Store.

Adventure India Journey Pvt Ltd — *Senior iOS Engineer*

Delhi, India | Aug 2024 – Jul 2026

- Led technical delivery and architecture for a 5-member iOS team on a travel booking platform.
- Reduced app cold-launch time by ~70% by deferring non-critical API calls and serving first paint from local cache.
- Designed an offline-first architecture with SwiftData, keeping itineraries available with no connectivity.
- From Oct 2025, studied Generative AI and Agentic AI in parallel with delivery (Python, LLM APIs, RAG, agent design), which led to the AI projects above.

Nickelfox Technologies Pvt Ltd — *Senior Software Engineer (iOS)*

Noida, India | Jun 2021 – Aug 2024

- Technical lead for a 4-developer iOS team, delivering 7+ client apps across healthcare, banking, AI and wellness.
- Integrated LLM-based persona chat into a consumer app (Ele): prompt design, streaming responses and failure handling, through to App Store release.
- Introduced MVVM with modular UI and networking layers, cutting onboarding time for new client projects by 30%; automated releases with Fastlane and Bitrise.

Earlier roles — *Software Engineer (iOS)*

Gurugram, India | Mar 2020 – Jun 2021

- Uniqubic Solutions: iOS companion app for *Sea of Thieves* with offline dataset caching. SD Global Services: two-sided on-demand marketplace app (Soothe) with real-time booking and push notifications.

PRODUCT

EvenSteven — *On-Device AI Expense App*

Live on App Store & Google Play | Jul 2026 – Present

apps.apple.com/app/id6787313313 · avioratech.com/EvenSteven

- Built an on-device receipt understanding pipeline with VisionKit OCR and Apple Foundation Models that extracts line items from receipts.
- Implemented provider-agnostic LLM routing (Gemini + Groq) with per-feature quotas and fallback; sole engineer from design to store release.

PRIOR EXPERIENCE & EDUCATION

- **Founder & Managing Director** | MOSRON Technologies Pvt Ltd (2015 – 2019) — ran an LED lighting manufacturer (up to 14 staff, 30 dealers) before moving into software.
- **B.Tech, Electronics & Communication Engineering** (71%) — IMS Engineering College, Ghaziabad (2010 – 2014)