YOKESH KS

GenAI Developer | Agentic AI & RAG Systems | Lead Full Stack Engineer

Chennai +91 8098444187 ksyokesh98@gmail.com https://www.yokesh.in/

yokesh-ks
yokesh-ks

Summary

AI Engineer with 4+ years of enterprise software engineering experience and deep specialization in Generative AI, Retrieval Augmented Generation (RAG), and Agentic AI architectures. Designing and deploying production-grade AI systems using Python, LangChain, LangGraph, and LlamaIndex with a strong focus on hands-on development, system architecture, and production reliability. Experienced in building multi-agent systems, implementing Model Context Protocol (MCP) style context management, and deploying scalable AI platforms on Azure and Google Cloud Vertex AI with expertise in hybrid cloud architecture and cross-platform service integration. Strong background in enterprise platforms, healthcare compliance (FHIR R4), and MLOps, bridging traditional software engineering with modern applied AI system design.

Experience

Lead Full Stack Engineer | Incresco **January 2024 - Present**

Leadership & Ownership

- Lead a 5-person engineering team across multiple product lines, driving architectural reviews, technical roadmap planning, and engineering best practices while serving as primary technical point of contact for cross-functional stakeholders.
- Mentor engineers through weekly code reviews and Architectural Decision Records (ADRs), improving code quality by 35% and reducing feature to production cycle time by 30%.
- Translate business requirements into scalable system designs, coordinating with product, sales, and operations teams to deliver high impact technical solutions.

Architecture & Platform Engineering

- Architected and standardized a micro-frontend architecture using Next.js, enabling independent deployment of CRM, Support, and Admin modules; reduced build times from 12 minutes to 3 minutes per module.
- Engineered a zero-downtime ETL migration pipeline migrating 1M+ customer records and 120GB of data from vTigerCRM and Zoho with 100% data integrity across 10+ B2B client tenants through automated validation, checkpoint/resume capability, and batch processing.
- Led system design and implementation of a custom quote-to-booking workflow with Hotel Reservation System integration, reducing sales cycle time by 40% and eliminating manual handoffs between sales and operations.

Platform & Strategic Initiatives (Lead / Core Contributor)

- Email SaaS Platform: Led backend architecture for AWS SES based system processing 60K+ emails/month, with real-time analytics via SNS → Lambda pipeline.
- ABDM-Compliant Healthcare System: Technical lead for compliance and integration; achieved all three ABDM milestone certifications (M1–M3) using FHIR R4 standards and blockchain-backed consent management.
- AI Booking Agent: Designed and led production rollout of a RAG-based conversational AI using LangChain and Pinecone, handling 800+ conversations per week with measurable reduction in manual booking effort.
- Real-Time Voice AI Platform: Built a LiveKit-based voice interaction system integrated with Google Vertex AI Gemini Live API for low-latency, AI-driven hiring and interview workflows.
- Internal HRMS: Architected and delivered Frappe + Next.js HRMS with WhatsApp Business API integration, serving 40+ employees.

Experience

Software Engineer – Full Stack | Incresco **July 2023 - December 2023**

Aceprep (Q3 2023) - AI-Powered Mock Interview Platform

- Built Aceprep from ground up, an AI-driven interview preparation platform with personalized coaching, mock interviews, and adaptive question banks across 12 industries
- Developed end-to-end AI feedback pipeline: implemented video capture with WebRTC, audio extraction, OpenAI Whisper integration for speech-to-text conversion, and structured feedback generation achieving 85% user satisfaction scores
- Created role-specific interview modules with 1,000+ tailored practice questions across SDE, Product, Marketing, and Finance roles, plus live coding environment with 50+ algorithm challenges supporting real-time execution

Outreach (Q4 2023) - Omnichannel Social Media Management Platform

- Engineered scalable architecture for Outreach, a unified platform for publishing and managing content across LinkedIn, X, Instagram, Facebook, and Google Business
- Integrated multiple third-party APIs (Meta Graph API, LinkedIn API, Google Business API) enabling one-click cross-platform publishing with approval workflows and AI-powered content ideation
- Developed performance intelligence dashboard for data-driven optimization, tracking 50+ engagement metrics across all connected platforms with systematized Google Reviews reply feature processing 200+ reviews/month

Software Engineer | Incresco **January 2022 - June 2023**

- Led React Native upgrade (v0.63 → v0.71) for Edvanza mobile app, resolving 30+ breaking changes and ensuring backward compatibility across 15 production modules serving 50K+ monthly active users
- Delivered production critical features including Jobs marketplace, Career portal, and Learning Management System (LMS) across web and mobile platforms, driving 25% increase in user engagement
- Rebuilt marketing website with Astro framework, achieving 60% improvement in page load times (3.2s → 1.3s) and 40% increase in organic traffic through SEO optimization and performance tuning
- Introduced React Query for global state management, reducing redundant API calls by 45% and improving cache hit rates from 30% to 75% across the application
- Implemented global search functionality and contributed to Cypress browser based testing framework, achieving 70% test coverage across critical user journeys

Software Engineer Trainee | Incresco **November 2021 - January 2022**

- Engineered the Social Module for Edvanza mobile app using React Native: Social Feeds, Discussion forum, and Network features enabling 10K+ monthly peer interactions
- Spearheaded the development of a component architecture with 20+ reusable components for the Edvanza mobile application, ensuring code consistency and reusability while integrating REST APIs for seamless data flow across social features

Education

Government College of Technology, Coimbatore **January 05, 2019 - September 15, 2021**
M.E. Engineering Design — Masters of Engineering
8.45 GPA

Panimalar Institute of Technology, Chennai **June 2015 - Mar 2019**
B.E. Mechanical Engineering — Bachelor of Engineering
7.7 GPA

Projects

Copilot Agents – Production Multi-Agent AI System

Production-grade multi-agent AI system enabling secure natural-language interaction with enterprise CRM and operations data in PostgreSQL, purpose-built for multi-tenant SaaS environments with strict tenant isolation, access control, and governance guarantees.

- Enabled secure natural-language querying of PostgreSQL-backed CRM and operations systems without exposing direct SQL access.
- Implemented LangGraph-based agent workflows translating user intent into validated, tenant-aware queries with structured outputs and visualizations.
- Designed domain-specific agents with schema awareness, scope enforcement, and deterministic response handling.
- Applied MCP-style context management to maintain consistent conversational state across agent steps and UI interactions.
- Containerized and deployed as a backend service with environment-driven configuration for production reliability and safe iteration.

Python, FastAPI, LangGraph, LangChain, Azure OpenAI, PostgreSQL, CopilotKit, Docker, Prompt Engineering

Support Chat – Real-Time AI-Augmented Ticketing System

Production-grade real-time support chat system combining conversational AI, ticket lifecycle orchestration, authenticated user context, and document-aware conversations.

- Built a real-time support chat UI backed by Azure Web PubSub, enabling low-latency bi-directional messaging between customers, agents, and an AI copilot.
- Integrated a CopilotKit-based AI assistant with tightly controlled prompts to enforce support workflows, ticketing rules, and knowledge-base boundaries.
- Implemented robust session and conversation persistence using session IDs and conversation IDs to support reloads, reconnections, and multi-tab continuity.
- Enforced a strict ticket lifecycle workflow (auto-create → associate contact → update → safe close) via AI instruction scaffolding and backend coordination.
- Added file-aware conversations with OCR-based text extraction, authenticated user context injection, custom CopilotChat UI components, typing indicators, and idempotent message handling.

Next.js, Azure Web PubSub (WebSocket), CopilotKit, TypeScript, REST APIs, Session Storage, OCR (PyMuPDF, Tesseract)

Real-Time Voice AI Platform – Gemini Live & LiveKit Integration

Production-grade voice AI platform for hiring and interview workflows, enabling natural, low-latency conversational interactions through Google Vertex AI Gemini Live API with LiveKit-based audio streaming.

- Built a scalable real-time audio backend using LiveKit integrated with Gemini Live API, enabling natural voice-based conversations with sub-second latency for AI-powered candidate screening and behavioral interview simulations.
- Implemented LiveKit room and session management to handle concurrent interview sessions with reliable participant lifecycle control and session continuity.
- Designed bi-directional WebSocket audio streaming to support real-time speech input/output with stable session handling.
- Deployed compute workloads on Azure Virtual Machines with GCP Cloud Storage for persistent audio/video artifacts and interview data.
- Achieved 99.5% uptime across concurrent real-time voice sessions, delivering a production-ready Voice AI platform capable of supporting AI-led interviews and interactive hiring workflows.

Python, LiveKit, Docker, Google Vertex AI (Gemini Live API), WebSockets, Azure VM, GCP Cloud Storage, Multi-Cloud Architecture

Projects

ContextOne – Multi-Tenant RAG Chatbot Platform

Enterprise-grade Retrieval-Augmented Generation platform that turns static documentation into context-aware AI chatbots deployable in under 2 minutes via a lightweight JavaScript widget for websites and SaaS products.

- Built a document ingestion and embedding pipeline that processes PDFs, DOCX, HTML, TXT, and MD into Qdrant vectors, enabling accurate RAG-based Q&A with source citations over private knowledge bases.
- Implemented secure multi-tenant isolation using Supabase Row Level Security and Qdrant payload filters, allowing each tenant to operate fully segregated chatbots on shared infrastructure.
- Developed a production-ready FastAPI backend integrated with IBM watsonx ai LLMs to deliver low-latency, grounded responses with configurable system prompts and safety controls.
- Shipped an embeddable, dependency-free JavaScript chat widget plus a React-based maker dashboard for managing projects, docs, and analytics, enabling rapid integration into SaaS products and marketing sites.

FastAPI (Python), React 18 + Vite + TypeScript, Qdrant (vector DB), IBM watsonx.ai (Granite)

Skills

Backend & APIs

NodeJS, WebSockets, Webhooks, GraphQL, OAuth 2.0, Prisma, Python, SQL, JavaScript, FastAPI

AI Platforms & Models

OpenAI (GPT-4 / GPT-3.5), Anthropic Claude (Sonnet / Opus), Google Vertex AI, IBM Watsonx, Hugging Face

Agentic AI & Frameworks

LangChain, LangGraph, LlamaIndex, CrewAI, OpenAI Agents SDK, Multi-Agent Systems, Agent Development Kits (ADK), Model Context Protocol (MCP)

Cloud & DevOps

AWS Cognito, AWS SES, AWS CloudWatch, AWS Amplify, AWS AppSync, Azure App Service, AWS EC2, Azure VM, Terraform, Docker, CI/CD Pipelines

Generative AI & LLMs

Large Language Models (LLMs), Small Language Models (SLMs), Prompt Engineering, Embeddings, Semantic Search, Retrieval-Augmented Generation (RAG)

AI Specializations

LLM Engineering, RAG Architecture, Multi-Agent Orchestration, Context Management, Prompt Evaluation, LoRA / QLoRA (exposure), MLOps & Observability

Databases & Storage

PostgreSQL, MySQL, Prisma ORM, Redis, Vector Databases (Pinecone, Chroma, PgVector, FAISS), MongoDb

Frontend Development & Architecture

React, React Native, Astro, React Query, Tailwind, Next.js, HTML, CSS, Design Systems, Micro-frontends

Data Engineering

ETL Pipelines, Data Migration, Multi-tenancy Architecture, BulkMQ

Third-Party APIs

Meta Graph API, LinkedIn API, Twitter API, Google Business API, WhatsApp Business API

Awards

Move the Mountain **November 09, 2022**

Incresco

For exceptional problem-solving and consistently meeting project deadlines to ensure timely delivery for clients.

Certificate of Appreciation **January 21, 2022**

Incresco

For Outstanding Contribution and demonstrating strong ownership