Sai Kiran Saripilli — Hyderabad, Telangana, India

saikiran212001@gmail.com | +919000267793 | linkedin.com/in/sai-kiran21 | github.com/saikiran-devai

PROFESSIONAL SUMMARY

AI Engineer with 2+ years delivering production Agentic AI systems and data infrastructure LangGraph-based multi-agent pipelines, tuned vector retrieval (ChromaDB, hybrid search) for RAG systems, and cloud-deployed LLM applications on AWS. Shipped systems serving 100+ concurrent enterprise users, processing 100150 documents daily, and reducing manual workloads by 60

TECHNICAL SKILLS

AI & LLM Systems: LangChain, LangGraph, RAG, Agentic AI, Tool Calling, ChromaDB, FAISS, Claude API, Ragas
Languages & Backend: Python, FastAPI, REST APIs, Async Programming
Systems & Architecture: System Design, Microservices, Distributed Systems, Caching Strategy, DSA
Cloud, Data & Real-Time: AWS (EC2, S3, Textract), Docker, CI/CD, MySQL, Redis
Integrations: Google APIs, OAuth 2.0, Razorpay, Deepgram, Sarvam AI

PROFESSIONAL EXPERIENCE

Infyz Solutions *Jan 2025 – Present*
AI Engineer Hyderabad, India

- Designed a query-routing service with RBAC — enabled 100+ non-technical shipping port operations staff to query live databases directly in plain English, eliminating manual query bottlenecks and achieving 100% adoption across enterprise teams.
- Architected a document processing pipeline integrating structured extraction with post-processing — handles 100–150 documents daily, reducing manual processing workload by 60% across production operations.
- Architected a parallel task-orchestration system spawning concurrent workers via a fan-out/fan-in pattern, merging results through map-reduce, and streaming responses over SSE — 60% lower latency vs. sequential batch processing.
- Built and deployed a real-time management platform handling 11 yards and 7,000+ cargo slots — serves 10–15 concurrent enterprise users with live tracking, RESTful APIs, and automated inspection workflows.

Zessta Software Services *Nov 2023 – Dec 2024*
Software Engineer AI Hyderabad, India

- Designed a hybrid document processing pipeline combining AWS Textract with open-source models — resolved layout preservation failures on complex formats (handwritten, printed, mixed) that previously caused complete extraction loss on multi-column records, improving structured output reliability across production workflows.
- Built a frontend-integrated annotation tool that eliminated manual dataset labelling — reduced processing cycle time via scalable, automated labelling workflows for internal pipelines.

PROJECTS

ArtistOnGo — Conversational Booking Platform for Salon Space Rentals *FastAPI, Django REST Framework, Python, MySQL*

- Architected a two-stage routing pipeline serving 4 distinct user personas with per-persona permissions and session state managed across turns.
- Built a deterministic state-machine layer controlling flow position from session state rather than inferred logic — ensuring business rules (verification gates, licence checks) could never be skipped.
- Closed an account-takeover vector and eliminated bearer tokens persisted at rest in the session store, closing a cross-account data leak; maintained a 300+ test suite covering orchestration and latency-sensitive paths.

ZORA — Distributed Task Orchestration System *FastAPI, Distributed Systems, ReAct*

- Architected a parallel task-orchestration system spawning concurrent workers via a fan-out/fan-in pattern, merging results through map-reduce, and streaming responses via SSE — 60% lower latency vs batch response.
- Implemented runtime code synthesis with AST-validated sandboxed execution and a quantitative evaluation pipeline achieving a 9.2/10 reliability score — zero hardcoded handlers, systematic quality tracking across all runs.

Ayra — Real-Time Voice Assistant with Agentic Tool Use *Deepgram Flux, Claude Haiku, Sarvam AI, Google APIs*

- Built a real-time voice assistant: streaming speech-to-text (Deepgram Flux), replies from Claude Haiku, speech output via Sarvam AI, and barge-in so users can interrupt mid-response, with a live UI showing listening, thinking and speaking states.
- Designed a tool-calling layer separating read-only tools from write actions; implemented weather, reminders and Google Maps tools (geocoding, directions, nearby search) returning short, speech-friendly answers.
- Integrated Gmail and Calendar via OAuth 2.0 with least-privilege scopes and a spoken confirmation step before any action that sends or changes data.

EDUCATION

Centurion University *Aug 2020 – Jun 2024*
B.Tech in Computer Science | CGPA: 8.9/10