CHAITANYA YARRAMSETTI

Senior Software Engineer, Full Stack & AI

chaitanya.yarramsetti@gmail.com | +1 469-347-7710 | San Francisco, CA | LinkedIn | GitHub

TECHNICAL SKILLS

Programming Languages: Python, Go, JavaScript, TypeScript

Agentic AI: OpenAI, Claude, LangGraph, LangChain, LiteLLM, vLLM, multi-agent orchestration, MCP, prompt & context engineering

Full Stack: FastAPI, Node.js, Express.js, React, Redux, Next.js, PostgreSQL, MongoDB, Redis, REST, gRPC, GraphQL, Playwright

Distributed Systems & Cloud: Temporal, Kafka, SQS, AWS, Azure, Docker, Kubernetes, Terraform, CI/CD, Git, OpenTelemetry

Data Pipelines, RAG & Evaluation: Databricks, PySpark, Hybrid RAG, Amazon OpenSearch, Neo4j, DeepEval, RAGAS, LangSmith

AI Development Tools: OpenAI Codex, Claude Code

WORK EXPERIENCE

Satcon AI | Senior Software Engineer — Oct 2025 - Present | San Francisco, CA

- Co-designed and developed the **LangGraph supervisor/specialist multi-agent orchestration** and durable workflow layer for retail AI agent platform handling 10k+ requests per day, using Temporal for workflow state and retries and Kafka for asynchronous event processing and deployed conversational agents for D2C brands.
- Built **RAG data ingestion and relationship aware retrieval pipelines** for PDFs, DOCX files, and database records using Amazon OpenSearch for hybrid BM25 and semantic search with cross encoder reranking, and Neo4j for customer, product, case, and policy relationship traversal, improving retrieval precision by 20%.
- Created an **agent evaluation system** using LangSmith golden datasets, TAU2-bench simulations, and LLM as judge scoring, with OpenTelemetry and CloudWatch observability and production feedback loops that improved $Pass^4$ reliability from 40% to 70%.
- Optimized agent inference with query, semantic, tool response, and agent state caching alongside dynamic context handling, reducing LLM spend by 25% while maintaining p95 inference latency below 500 ms.
- Developed a **multi-tenant** refund workflow across a FastAPI Agent Runtime, Policy Service, and MCP Integration Gateway, validating schema-constrained tool calls through a centralized registry and using signed tenant context and human approval gates for authorization and auditability.

Pegasystems - Senior Software Engineer — Mar 2023 - Sep 2025 | Hyderabad, India

- Built an AI ticket triage and routing system on Pega GenAI, using Databricks pipelines to ingest and prepare **20 years of historical cases** for metadata filtered hybrid RAG, improving severity, category, and ownership decisions and reducing assignment time.
- Led a distributed team of 5 engineers building a customer support portal and microservices for a platform serving **1M+ DAU**, delivering ticket management, case resolution, SLA tracking, ownership workflows, and triage dashboards with React/Node.js.
- Developed the Pega React SDK, a reusable no-code UI platform across the Pega design system, powering enterprise workflow applications and complex business processes across SaaS implementations.

Caset | Founding Engineer — Aug 2021 - Feb 2023 | Hyderabad, India

- Led design and built a secure **revenue-based financing platform** across web, mobile, and backend systems, using React, Node.js, and PostgreSQL to deliver business onboarding, portfolio tracking, return calculations, analytics, and customer workflows.
- Partnered with founders on product decisions, customer feedback, operational workflows, and go-to-market execution, helping onboard 15 businesses and 10,000+ customers during early product market validation.

Dataphi Labs | Software Developer — Oct 2018 - Jul 2021 | Bengaluru, India

- Built and deployed a real-time **live commerce mobile app** with Flutter and the Agora SDK, enabling live popup shops and in-stream purchasing across product discovery, cart, and checkout flows, increasing live session conversion.

Open Source Projects:

- Agentic AI Lab: Built an open source learning platform for software engineers transitioning into AI engineering(1k+ Users).
- Agentic Support Platform: Created a platform that enables teams to create, deploy, & manage conversational AI agents.

EDUCATION

B.Tech, Electrical & Electronics Engineering - GMRIT, India, 2018