G Shiv Santosh

shivsantosh04@gmail.com | +91 98924 58433 | LinkedIn | GitHub | Bengaluru, Karnataka

Computer Science Engineer with production experience in RAG pipelines, bilingual ASR/TTS voice systems, and parallel agentic API workflows — built and shipped across LangChain, FastAPI, and GPU infrastructure (H100, L40s) at a cloud AI company.

EXPERIENCE

AI Engineer Intern | **NxtGen Cloud Technologies** *Jan 2026 – May 2026*

RAG & LLM Infrastructure

- Built and maintained a production-grade RAG pipeline using **GPT-OSS 20B** via **LangChain**, served on an **NVIDIA L40s** GPU (single-GPU deployment alongside ASR/TTS and pilot agent workloads).
- Developed an **HR chatbot** for a client — integrated the RAG system with a structured knowledge base to answer policy queries with high contextual accuracy. Containerized the backend using **Docker Compose**.

ASR / TTS Voice Pipeline — English & Telugu

- Optimized a bilingual voice assistant pipeline — rewrote system prompts and voice prompts to significantly improve response quality in both languages.
- ASR: **Whisper Large V3** for speech recognition. TTS: **Magpie** for English, **Indic Parler** for Telugu — iterated on voice prompt design to improve naturalness.

API Testing & Agentic Routing

- Architected an intent-based agentic routing system over **AgriWatch** and **E-Nam** APIs — agent hits required endpoints in parallel and consolidates results, with filter-word streaming to cut perceived latency; APIs mapped end-to-end via **Postman**.
- Used **MongoDB** to cache API response dumps locally, cutting redundant live calls during development and enabling offline testing.

M-Vault — AI-Powered Cloud Storage

- Shipped a report-generation feature for **M-Vault** (a NotebookLM-style AI storage product) within a single sprint — feature was adopted and extended by the broader team.
- Studied **Redis** and **Celery** integration in the existing codebase — gained practical understanding of async task queuing in a production environment.

EDUCATION

B.E, Computer Science & Engineering — Jyothy Institute of Technology, VTU *GPA: 8.4 | 2026*

Pre-University, Computer Science — Sadhana PU College, Bengaluru *GPA: 9.5 | 2022*

TECHNICAL SKILLS

Languages	Python, C
AI / ML	RAG, LLM Integration, LangChain, Whisper ASR, Magpie TTS, Indic Parler, ChromaDB, scikit-learn
Models & Infra	GPT-OSS 20B, NVIDIA H100 / L40s
DSA	BFS / DFS, Sorting (Quick, Merge, Insertion), Recursion, Hash Tables, Trees, Graphs
Databases	MongoDB, MySQL, Supabase, Postgres
Tools & Frameworks	FastAPI, SQLAlchemy, Postman, Git, VS Code, Redis / Celery (familiar), Streamlit, Docker

PROJECTS

AI-Based Finance Tracker *Dec 2025*

- **Flutter** expense tracker — **Google ML Kit** OCR scans receipts and regex extracts structured transaction data; **FL_Charts** renders spend analytics. Messages are categorized to get trend analysis and budgeting the spends accordingly.
- User can scan or upload a receipt and the app will automatically collect useful details such as amount, name via **regex** patterns.

RAG Application *May 2026*

- Built a RAG-based academic assistant using **FastAPI**, vector databases, and embedding models. Investigated chunk-size, Top-K retrieval, and semantic search optimizations to improve answer relevance and reduce latency.
- Ingested 100s of pages of syllabus resources into a RAG pipeline — page-level chunking strategy, **ChromaDB** for vector storage, **Gemma 4** as reasoning model, and **Streamlit** frontend for **student Q&A.**
- Documents were ingested using the embedding model **BAAI/bge-small-en-v1.5** locally on the **edge device** and stored in vector storage accordingly. But this causes app to be slower due to each chunk being big and takes time in semantic search. This can be solved with reducing top k chunk to be minimal or implement Hybrid chunking or reducing the retrieval threshold.