G SHIV SANTOSH

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

Computer Science engineer who has shipped production LLM systems end to end — a bilingual English/Telugu ASR–TTS voice pipeline, retrieval chatbots, and a parallel agentic API router — on FastAPI and self-hosted GPU infrastructure (H100, L40s). Comfortable owning a feature from retrieval design through Docker deployment. Currently building a foundation in offensive security through TryHackMe's Jr Penetration Tester path, with the aim of growing into AI security and building AI applications that are safe and reliable by design.

EXPERIENCE

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

Bilingual Voice Pipeline — English & Telugu

- Owned system- and voice-prompt design for a bilingual voice assistant: Whisper Large V3 for ASR, Magpie (English) and Indic Parler (Telugu) for TTS — iterated on prompt phrasing to improve response quality and speech naturalness in both languages.
- Served alongside RAG and pilot agent workloads on a single NVIDIA L40s, working within shared GPU memory and latency constraints.

M-Vault — AI-Powered Cloud Storage (NotebookLM-style product)

- Designed a dynamic, intent-gated inference path: a greeting resolves in a single LLM call, while a substantive query is first decomposed and spell-normalised by the same model, then sent for retrieval and answering — removing wasted calls on trivial turns.
- Shipped a report-generation feature to production within one sprint, including an optional tone parameter (technical / academic) that conditions GPT-OSS 20B's structure and register per use case; the feature was adopted and extended by the wider team.
- Prototyped a reranking stage and kept it behind a flag after measuring its latency cost — an explicit relevance-versus-latency trade-off rather than a default-on feature.

HR Policy Assistant — Client RAG Chatbot

- Chose page-level chunking with overlap so that large travel- and leave-policy tables stayed intact within a single chunk; tables split across chunk boundaries had been the main source of incomplete answers.
- Diagnosed and fixed a systematic output-fidelity bug where city and country names were returned with stray spaces and symbols (notably on tier-classification queries), by adding a post-generation validation pass — deterministic filters plus a second LLM check before the response reached the user.
- Persisted user- and session-scoped chat history in MongoDB so follow-up questions retained context; containerised the stack with Docker Compose using named volumes for durable storage.

Agentic API Routing

- Architected an intent-based router over the AgriWatch and E-NAM APIs that calls the required endpoints in parallel and consolidates results, with filter-word streaming to cut perceived latency; endpoints mapped end-to-end in Postman.
- Cached API response dumps in MongoDB to eliminate redundant live calls and enable offline development and testing.

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

CERTIFICATIONS, LEARNING & EVENTS

TryHackMe — Jr Penetration Tester path (in progress, Premium) — working through the networking and Metasploit modules; AI security planned as the next focus.

Databricks × v4c.ai Data + AI Summit — attended at the Databricks office, Bengaluru (May 2026); sessions on Lakehouse architecture and applied data science workflows.

TECHNICAL SKILLS

Languages: Python, C, SQL

AI / ML: RAG pipelines, LLM integration, LangChain, prompt & context design, Whisper ASR, Magpie / Indic Parler TTS, ChromaDB, scikit-learn

Backend & Infra: FastAPI, SQLAlchemy, Docker & Docker Compose, Redis / Celery (familiar), Streamlit, Postman, Git, Linux

Databases: PostgreSQL, MongoDB, MySQL, Supabase

Platforms: NVIDIA H100 / L40s, GPT-OSS 20B, Databricks (Lakehouse fundamentals)

Security (learning): Networking fundamentals, Metasploit, penetration-testing methodology

PROJECTS

AI-Based Finance Tracker — Flutter, Supabase, Google ML Kit *Dec 2025*

- Gated OCR ingestion behind an intent classifier so only genuine receipts and financial documents are stored, keeping unwanted captures out of the user's transaction history; regex and heuristic rules extract structured fields from the scanned document.
- Supabase for sign-up, auth and per-user storage; fl_chart for categorised spend analytics, with budget notifications and a roadmap toward personalised budgeting guidance driven by user history.

RAG Academic Assistant — FastAPI, ChromaDB, Streamlit *May 2026*

- Page-level chunking, ChromaDB vector store and Gemma as the reasoning model behind a Streamlit Q&A frontend; tuned chunk size, Top-K and retrieval thresholds (including hybrid chunking) to raise answer relevance and reduce latency for edge deployment.