Gokul Sree Chandra Polavarapu

polavarapugokul@gmail.com — Ph: 6300405040 — LinkedIn — GitHub

Portfolio - GokulAIx

Technical Skills

- **Backend & APIs** Python, FastAPI, REST APIs, SQL
- **LLMs & Agentic Systems** Retrieval-Augmented Generation (RAG), LangChain, LangGraph, Agent Orchestration, Prompt Engineering, LLM APIs (Gemini, OpenAI, Groq), RAG Evaluation (Ragas, TruLens)
- **Retrieval & Vector Systems** Sentence-Transformers, Vector Databases, Embedding Models
- **Machine Learning Deep Learning** Supervised Learning, Feature Engineering, Transformer Architectures, Fine-Tuning, PyTorch
- **ML Tooling** NumPy, Pandas, Scikit-learn

Experience

Quant AI Engineer - Internship **Fidura AI** Remote, India | October 2025 to Present

- Designed and owned a production-grade RAG extraction pipeline converting tender PDFs into 19 schema-validated fields, using hybrid retrieval (sentence-transformers + pgvector), Docling/PyMuPDF ingestion, bounded-concurrency LLM execution, and page-level provenance for traceability.
- Architected and implemented a LangGraph-based agent orchestration system to automate quantitative workflows from research to simulated execution.
- Built and maintained 5+ FastAPI services exposing AI agents and financial tools as external APIs, handling request validation, concurrency, and downstream integrations.

Head of Technical Operations **Young Compete** Remote, India | July 2024 to January 2025

- Led 3+ tech projects (e.g., Prof Connect) to streamline networking for 1,500+ students and teachers.
- Automated data scraping from 5 sources for the off-campus placements page, utilizing APIs, and improved access rate by 30%.
- Led a 10-member technical team to build community-focused engagement solutions.

Projects

VidQuery - Low-Latency RAG System for Long-Form Video QA

- Designed and built a sub-5s latency RAG system for question answering over long-form YouTube videos, explicitly trading off chunk size, retrieval recall, and end-to-end response time. Implemented hybrid search (dense + keyword) with multi-query retrieval to improve semantic coverage while enforcing strict top-k and token caps to keep the final LLM context small and cost-efficient. Added URL-based transcript caching to avoid redundant ingestion and retrieval for repeated videos, significantly reducing latency and compute cost for subsequent queries.
- Tech: Python, LangChain, Gemini, Hugging Face, Sentence-Transformers, ChromaDB.
- **GitHub** | **Live Demo**

Blaze - Real-Time Web Page Intelligence System (Chrome Extension)

- Designed a real-time web intelligence system for webpage summarization and contextual Q&A using a Map–Reduce RAG pipeline with aggressive context filtering to maintain interactive latency in browser-constrained environments. **Tech:** Python, LangChain, Gemini, ChromaDB, Flask, JavaScript. **GitHub** | **Demo Video**

Not - Jarvis (V1, In Progress) Agentic AI Assistant Framework

- Currently designing and implementing a planner–executor agent system using LangGraph to decompose user intent into multi-step, tool-driven execution flows. Introduced schema-validated planning with Pydantic to enforce interpretable and debuggable agent behavior during execution. Focused on explicit state transitions and controlled tool invocation rather than open-ended conversational autonomy.
- Tech: Python, LangGraph, LangChain, Gemini API, Pydantic.

Education

Gandhi Institute of Technology and Management (GITAM) Visakhapatnam, India
Bachelor of Technology in Computer Science and Engineering (3rd Year) August 2023 to Present