Gokul Sree Chandra Polavarapu

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

Portfolio - **GokulAIx**

Technical Skills

- **Backend & APIs** Designing and operating backend services and APIs; Python, FastAPI, REST, SQL
- **LLMs & Agentic Systems** Agentic workflows, RAG pipelines, agent orchestration, evaluation and control; LangChain, LangGraph, LLM APIs
- **Retrieval & Vector Systems** Embedding-based retrieval, vector databases, hybrid search, semantic indexing
- **Machine Learning & Deep Learning** Supervised learning, feature engineering, Neural Network architectures (CNNs, RNNs), transformer-based models, fine-tuning, PyTorch

Experience

Quant AI Engineer Intern **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 a 10-member team to build community engagement and managed 3+ tech projects (e.g., Prof Connect) streamlining networking for 1,500+ users.
- Automated data scraping from 5 sources for the off-campus placements page using APIs, improving platform access rates by 30%.

Projects

Not-Jarvis (V1) - Stateful AI Agent Framework

- Built a stateful, multi-turn AI agent using LangGraph with an iterative single-step planning loop, executing one action per iteration with explicit completion checks.
- Designed a hybrid Python + LLM architecture where Python performs deterministic URL extraction and normalization, achieving zero URL hallucination and duplicate action prevention.
- Implemented persistent conversation memory (PostgreSQL checkpointer), Server-Sent Events (SSE) streaming, and schema-validated planning (Pydantic) to ensure reliable, debuggable agent execution.

Tech: Python, FastAPI, LangGraph, LangChain, Gemini API, PostgreSQL (Supabase), Pydantic **GitHub** — **Demo**

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

- Designed and built a sub-5s latency RAG system for long-form YouTube videos, implementing URL-based transcript caching to reduce compute costs and latency for repeated queries.
- Implemented hybrid dense + keyword search with multi-query retrieval while enforcing strict top-k and token caps to keep LLM context small and cost-efficient.

Tech: Python, LangChain, Gemini, Hugging Face, Sentence-Transformers, ChromaDB.
GitHub — **Demo Video**

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

- Designed a real-time web intelligence system for webpage summarization and contextual Q&A.
- Used 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**

Education

GITAM University Visakhapatnam, India
B.Tech in Computer Science and Engineering - 3rd year Aug 2023 – Present