Sumit Kumar

Bengaluru, India ✉ sklegacy789@gmail.com ☎ +91-7033241380 in/sumitkumar22 sumit22012004.github.io/Portfolio/

SUMMARY

AI Engineer specializing in multi-agent systems and memory-augmented architectures, enabling continuous learning and intelligent coordination. Focused on advancing adaptive AI solutions that integrate reasoning, retrieval, and scalability for enterprise innovation.

SKILLS

Large Language Models & AI Technologies: Python, LLMs (GPT-5, Claude, Gemini, Llama, Mistral, Nvidia), Prompt Engineering, Retrieval-Augmented Generation (RAG), Generative AI, Multi-Agent Systems, Natural Language Processing (NLP), Deep Learning, Machine Learning

AI Frameworks & Libraries: LangChain, LangGraph, LangSmith, TensorFlow, PyTorch, Hugging Face, Scikit-Learn, NumPy, Pandas, Matplotlib, OpenCV, SpaCy, Text-to-Speech (TTS), Speech-to-Text (STT)

Backend Development & API Integration: FastAPI, RESTful APIs, API Integration with LLMs, Object-Oriented Programming (OOP), Rule Engine Design, Chatbot Development, Testing & Debugging, Automation Workflows

Data Engineering & Management: Data Preprocessing, Feature Engineering, Model Evaluation, Data Pipelines, ETL Processes, Web Scraping, SQL & NoSQL Databases (MySQL, MongoDB), Vector Databases (Qdrant, Pinecone, LanceDB), Redis, Data Optimization

Cloud & DevOps: AWS, Azure, Git, GitHub, CI/CD Pipelines, Postman, API Monitoring, Tableau, Streamlit

Professional Competencies: Analytical Thinking, Strategic Decision-Making, Problem-Solving, Innovation, Team Collaboration, Effective Communication, Project Ownership, Time Management

EXPERIENCE

RAG/ Multi-Agent AI Engineer

VRVV Ventures | Bangalore, Karnataka **May 2025 - November 2025**

- **Developed an advanced conversational AI system** with **persistent memory**, built entirely from scratch to store, retrieve, and contextualize user interactions. Designed a **memory architecture** using **Qdrant**, **MongoDB**, and **Memgraph** for encrypted user data and lifelong memory visualization, achieving **95.2% accuracy in LongMemEval**, surpassing **Mistral's 86% benchmark**.
- Engineered a **multi-agent system** and **secure, scalable backend** integrating **Gemini 2.5 Flash**, **GPT-5**, and **offline LLMs**, all **trained and deployed locally** to ensure maximum data privacy. Built **custom APIs** for retrieval, decryption, and LLM orchestration, implementing robust **encryption protocols** and ensuring intelligent, context-aware responses.
- Designed and managed **AI personality modules** leveraging **Zero-shot**, **Few-shot**, **Role prompting**, **Chain-of-Thought (CoT)**, and **Prompt Chaining** techniques. Automated **memory updates, data cleanup, and system health monitoring** through optimized **cron jobs** scheduled at multiple intervals (6h, 12h, etc.) to maintain consistent performance and reliability.

Software Development Trainee

Udyat Technologies | Mohali, Punjab **September 2024 - March 2025**

- Developed and optimized **backend APIs** using **Python, FastAPI, and LangChain**, integrating **GPT-5 and Claude LLMs** with **Retrieval-Augmented Generation (RAG)** and **prompt engineering** to enhance AI content generation, accuracy, and response efficiency by **30%**.
- Designed and deployed **multi-agent workflows** for **AI-driven automation**, enabling **session management**, **fine-tuned conversational responses**, and scalable **chatbot applications** through **LangChain**, **MySQL**, **MongoDB**, and **web scraping** pipelines for **data preprocessing and model training**.
- Leveraged **AWS (Lambda, EC2)** and **CI/CD pipelines** to ensure seamless deployment, scalability, and performance optimization of **AI-powered systems**, while integrating **Tesseract OCR**, **custom NLP models**, and **vectorized insights generation** for automated document parsing.

PROJECT

Intelligent Multi-Agent Research Assistant

March 2025 - May 2025

- **Architected a collaborative multi-agent framework** enabling specialized agents (Researcher, Summarizer, and Analyst) to autonomously scrape, parse, and synthesize academic literature using RAG pipelines with **Qdrant** and **Pinecone** for semantic retrieval.
- **Implemented long-term context sharing and persistent memory** via LangGraph and custom vector embeddings, allowing agents to maintain topic continuity and cross-reference previous research sessions with **95% retrieval relevance** across multi-turn tasks.
- **Deployed an interactive Streamlit interface** powered by **FastAPI-based backend orchestration**, enabling real-time research queries, insight visualization, and multi-agent dialogue — reducing manual literature review time by **40%**.

EDUCATION

Bachelor of Engineering – Computer Science (AI & ML Specialization)

Chandigarh University | 2021 – 2025