

SOURAB GAIROLA

AI Engineer — LangGraph · Agentic AI · RAG Systems · LLMOps
+91 73025 95984 • gairolasourab@gmail.com • [LinkedIn](#) • [GitHub](#)

SUMMARY

AI Engineer specializing in production-grade agentic systems and Retrieval-Augmented Generation. Built and deployed multi-agent pipelines with LangGraph/LangChain, RAG systems using ChromaDB/FAISS with cross-encoder reranking and MMR, and containerized these services on Azure with Docker and CI/CD. Experienced across the full lifecycle: agent orchestration, observability (token/cost tracking, LangSmith), human-in-the-loop approval workflows, and offline/local LLM deployment with Ollama.

TECHNICAL SKILLS

Agentic AI & Orchestration: LangGraph, LangChain, CrewAI, Multi-Agent Systems, MCP (Model Context Protocol) — Server & Client, Human-in-the-Loop Workflows, MemorySaver, n8n, Webhooks

RAG & Retrieval: ChromaDB, FAISS, Pinecone, Hugging Face Sentence Transformers, MMR, Cross-Encoder Reranking, Semantic Chunking, Embedding Models, RAGAS

LLMs & APIs: OpenAI, Azure OpenAI, Gemini, Claude (Anthropic), Groq, Ollama (local/offline inference), Prompt Engineering, Tavily

MLOps & Deployment: Docker, Kubernetes, CI/CD, Azure Cloud, FastAPI, Git, REST APIs, Pydantic, LangSmith, Execution Monitoring, Token/Cost Tracking

Languages & Backend: Python, JavaScript, SQL, React, PostgreSQL, MongoDB, Streamlit

ML Foundations: TensorFlow, PyTorch, Keras, Scikit-learn, NumPy, Pandas, CNNs, LSTMs, Reinforcement Learning

PROFESSIONAL EXPERIENCE

AI Engineer (Independent) | Bright Ripple Studios, Jaipur, India

Aug 2024 – Present

- **Multi-agent Market Intelligence System** — designed and built a LangGraph + Gemini + Tavily pipeline automating market research, competitor analysis, and strategy generation end-to-end.
- **Enterprise RAG Assistant** — built a FastAPI-backed knowledge assistant over company documents using ChromaDB and OpenAI embeddings for semantic search and contextual Q&A.
- **LLMOps & Deployment** — containerized agentic services with Docker and Kubernetes and shipped them on Azure via CI/CD pipelines; applied prompt engineering and structured Pydantic schemas, reducing LLM hallucinations by 35% and achieving a 94% grounding score in production; used LangSmith for tracing and observability.
- **Automation Pipelines** — built n8n workflows integrating Firecrawl and Apify for lead enrichment into HubSpot, automating 90% of manual data entry and processing 5k+ leads weekly; also automated event-driven email outreach triggered from Google Sheets.
- **Automated Medical Report System** — engineered custom Odoo modules to auto-generate structured patient PDF reports, owning the system from development through production deployment.
- **Full-Stack Ownership** — led React frontend integration and end-to-end delivery across all self-directed projects, from architecture through deployment.

Machine Learning Intern | IBM (Remote)

Jan 2024 – Jun 2024

- Computer vision — built a CNN handwritten-digit classifier in TensorFlow/Keras reaching 96% test accuracy.
- Reinforcement learning — developed a Q-learning maze-navigation agent achieving 94% path optimality in OpenAI Gym.

Freelance AI Automation (Part-Time / Concurrent with Studies) | Fiverr

2019 – 2026

- Delivered 500+ client engagements building automated content and growth-analytics pipelines, including generative AI video production workflows (HeyGen, ElevenLabs) taking scripts to finished renders end-to-end.

FEATURED PROJECTS — AGENTIC AI & RAG

[Autonomous Market Intelligence Platform](#) · [Demo](#)

Technologies Used: Python | LangGraph | LangChain | Google Gemini 2.5 Flash | LangSmith | Tavily Search API | Pydantic | ReportLab | Docker | Git | GitHub

- Production-ready LangGraph multi-agent platform automating market research, competitor analysis, and marketing content generation with parallel execution and memory persistence.

- Built FastAPI REST APIs + Streamlit dashboard; integrated Gemini 2.5 Flash, Tavily, structured Pydantic outputs, execution-time monitoring, and LLM cost/token tracking for observability.
- Added human-in-the-loop approval, Docker containerization, and automated PDF/JSON report generation via ReportLab.

[Enterprise Agentic RAG System \(Company Policy Agentic RAG\)](#) · [Demo](#)

Technologies Used: Python | Streamlit | LangChain | LangGraph | Ollama (Gemma 3) | ChromaDB | HuggingFace Sentence Transformers | PyPDFLoader | RecursiveCharacterTextSplitter | CrossEncoder (MS MARCO) | python-dotenv

- Architected a modular LangGraph RAG pipeline — retrieval, context construction, reranking, grounded generation — over enterprise policy documents.
- Improved retrieval precision using ChromaDB, Sentence Transformers, MMR, semantic chunking/deduplication, and cross-encoder reranking, reducing hallucinations.
- Deployed fully local/offline using LangChain + Ollama (Gemma 3) + Streamlit — no external LLM API dependency.

[Autonomous Self-Healing Infrastructure & Security Patching Engine](#) · [Demo](#)

Technologies Used: Python | FastAPI | n8n | Groq LLM (Llama 3) | JavaScript | REST APIs | Webhooks | Gmail Integration | Git | GitHub

- Built an AI incident-response system (FastAPI, n8n, Groq) with an Investigation Agent for root-cause analysis and a Repair Agent generating guarded code-fix recommendations.
- Implemented human-in-the-loop email approval before deployment, with automated crash logging and webhook alerting.

[Google Workspace AI Assistant](#) · [Demo](#)

Technologies Used: Python | Streamlit | n8n | Google Gemini | AI Agents | MCP (Model Context Protocol) Server & Client | Prompt Engineering | Google Workspace APIs (Gmail, Google Calendar, Google Tasks, Google Docs) | Webhooks | SerpAPI | Simple Memory | REST APIs | Git | GitHub

- Built a Gemini-powered assistant using tool-calling to operate Google Tasks, Calendar, Gmail, and Docs via natural language, orchestrated through n8n AI Agents and webhooks.
- Integrated an MCP (Model Context Protocol) Server and Client into the workflow, exposing tools (e.g., GitHub issue creation) to the AI Agent through a standardized protocol alongside native n8n tool nodes.

AI COMMUNITY & CONTENT

- **AI Content Creator, @aiwithsourab (YouTube)** — 2024–2026 — create videos on AI agents, LLMs, and automation, covering the latest tools, tutorials, and developments in the field.
- **Co-Manager, @artificialintelligencefacts (Instagram)** — 2020–2026 — grew an AI-focused educational community to 130K+ followers, publishing content on LLMs and emerging GenAI tools.

EDUCATION

Master of Computer Applications (MCA) — AI & Machine Learning UPES Dehradun, 2022 – 2024 | GPA: 7.9/10

- **Final-Semester Project — Hybrid Emotion Detection System:** Built a dual-modality deep learning model combining image and speech analysis — CNNs (TensorFlow, Keras, OpenCV) for facial emotion recognition and RNNs for speech-based emotion classification. ([GitHub](#))

Bachelor of Computer Applications (BCA) DIMR Rishikesh, 2020 – 2022 | GPA: 7.5/10