Sujeet Gund

+91-7620420850 | sujeetgund@gmail.com | sujeetgund.in | github.com/sujeetgund | linkedin.com/in/sujeetgund

Summary

Agentic AI Engineer specializing in autonomous multi-agent workflows, production RAG architectures, and scalable LLM infrastructure. Hands-on experience engineering LangGraph stateful loops with HITL checkpoints, Model Context Protocol (MCP) servers, hybrid vector retrieval (pgvector, FAISS), and high-throughput FastAPI backends on AWS and GCP. Proven track record reducing retrieval latency by 10x, cutting hallucinations, and deploying low-latency real-time voice and data pipelines.

Technical Skills

Languages & Databases: Python, TypeScript, SQL, PostgreSQL, pgvector, Redis, MongoDB, FAISS
Agentic AI & LLMs: LangGraph (StateGraph, Checkpointing, HITL), LangChain, RAG (Hybrid RRF, CRAG), Model Context Protocol (MCP), NLI DeBERTa-v3, Prompt Engineering, Structured Outputs (Pydantic v2)
Machine Learning & Frameworks: PyTorch, Hugging Face Transformers, scikit-learn, NLP, Responsible AI Guardrails
Backend, Cloud & DevOps: FastAPI, Next.js, Docker, AWS (EC2, S3, Lambda), GCP (Cloud Run, Vertex AI), GitHub Actions (CI/CD), LiveKit WebRTC, Linux, Git

Experience

Divam Technologies — Remote, India
GenAI Developer Intern — *Apr 2026 – Sep 2026*

- Engineered a hybrid RAG pipeline coupling pgvector semantic search with indexed file-based lookup, boosting factual query precision and slashing retrieval latency by **10x (50–100ms down to <5ms)** for high-frequency queries.
- Built an asynchronous user context ingestion pipeline parsing resumes, PDFs, and portfolio links via rolling summarization, compressing **∼15k raw input tokens** into structured **2–3k token** windows with zero context truncation.
- Developed a real-time WhatsApp voice agent integrating Deepgram, OpenAI, Cartesia, and LiveKit WebRTC with sentence-level streaming and a Redis caching layer, cutting **time-to-first-audio from ∼4s to <700ms** and lookup silence to **<1s**.
- Configured production AWS cloud infrastructure with **GitHub Actions CI/CD** across isolated staging/production environments, integrating LinkedIn OAuth, Razorpay payment capture, and automated WhatsApp alert webhooks.

Projects

GroundedAI: Self-Correcting Agentic RAG | *LangGraph, FastAPI, pgvector, Redis* — GitHub

- Architected a Corrective RAG (CRAG) system using LangGraph StateGraph, combining Reciprocal Rank Fusion (RRF) hybrid search (**92.8% recall**) with dynamic web search fallback when document context fails relevance thresholds.
- Integrated local CPU-based *DeBERTa-v3* cross-encoder NLI model for real-time hallucination checking (<45ms latency), cutting ungrounded claims and achieving **94.2% grounding precision** across 250 dense domain queries without LLM token cost.

rzp Merchant: Autonomous Commerce Engine (MCP) | *LangGraph, Next.js, Redis, Razorpay* — GitHub

- Implemented an official Model Context Protocol (MCP) server exposing JSON-RPC 2.0 endpoints for external AI agents to autonomously discover catalogs, check live stock, and query basket recommendations.
- Designed a 3-Tier Security Protocol with programmable spending mandates (*X-Buyer-Agent-Token*), achieving **100% mandate adherence** across 150 simulated AI orders with dual-signature Razorpay HMAC-SHA256 verification.

MailMind: Event-Driven Agentic Email Engine | *LangGraph, FastAPI, PostgreSQL, FAISS* — GitHub

- Built an event-driven email triage agent on LangGraph StateGraph with PostgreSQL persistence and interrupt-driven Human-in-the-Loop (HITL) queues, reducing operator triage time by **88% (from 15 min to <1.8 min per email)**.
- Engineered Pydantic v2 structured tool-calling for intent routing (**96.5% accuracy**), streaming real-time status via Server-Sent Events (SSE) and grounding draft replies using FAISS vector search across verified documentation.

Education

VIT Bhopal University — Bhopal, India
Integrated M.Tech in Artificial Intelligence, ***CGPA: 9.33 / 10.0*** — *Sep 2023 – Expected Mar 2028*

Certifications

Building RAG Agents with LLMs – NVIDIA (Oct 2026) | **Complete Data Science, ML, DL, NLP Bootcamp** – Udemy (Dec 2025) | **Applied Machine Learning in Python** – University of Michigan (Dec 2024) | **Supervised Machine Learning** – DeepLearning.AI (Jul 2023) | **AWS Cloud Essentials** – Amazon Web Services (Jun 2023)