SHIVA SHANKAR S

AI Engineer · RAG Systems · LLM Agent Architecture · LangChain · LangGraph

s.shivashankar1304@gmail.com | +91 7676823681 | github.com/Shiva7147 | Bengaluru, India — Open to Remote / Relocating to Spain

PROFESSIONAL SUMMARY

AI engineer who builds and ships working AI systems, not just experiments — production RAG pipelines, LangGraph ReAct agents with tool-calling, and automated backend systems deployed to real infrastructure. Understands the full stack from LLM internals through deployment. Every project evaluated with documented accuracy metrics rather than demoed and forgotten. Open to AI/ML engineering internships and remote roles, with a longer-term focus on the Spanish and European market.

TECHNICAL SKILLS

AI/LLM Engineering: RAG pipelines, LangChain LCEL, LangGraph ReAct agents, ChromaDB, prompt engineering, tool calling, LLM-as-judge + rule-based evaluation

LLM APIs & Models: OpenRouter, Anthropic Claude, OpenAI API, Google Gemini, HuggingFace embeddings

Languages & Core: Python, TypeScript, JavaScript, SQL, Git/GitHub, REST APIs

Backend & Infra: FastAPI, Node.js, Fastify, Next.js, Supabase, PostgreSQL, Railway, Docker

Automation & Orchestration: n8n, workflow orchestration, Apify

PROJECTS

AthleteIQ — AI Sports Science RAG Agent *github.com/Shiva7147/athleteiq*

- Built a production-grade LangGraph ReAct agent with four specialized tools: sports-science RAG retrieval over ChromaDB, a deterministic ACWR injury-risk calculator, timestamped athlete metric logging, and a three-tier safety escalation system.
- Designed a 56-entry evidence-based knowledge base embedded with HuggingFace all-MiniLM-L6-v2 and stored in ChromaDB for semantic retrieval.
- Built an end-to-end evaluation harness across 20 ground-truth test cases in 9 categories, combining LLM-as-judge scoring with rule-based checks — improving measured accuracy from 85% to 95%.
- Implemented Responsible AI guardrails: the agent skips knowledge retrieval entirely for medical emergencies and escalates immediately.

Tech: Python, LangChain LCEL, LangGraph, ChromaDB, HuggingFace, OpenRouter API

AI Employee Platform *Enterprise Multi-Tenant SaaS*

- Built an enterprise AI employee platform enabling businesses to deploy AI-powered receptionists that answer phone calls, book appointments, qualify leads, retrieve business knowledge, and automate customer workflows.
- Developed a cloud-native multi-tenant architecture supporting configurable AI agents, business-specific knowledge bases, real-time voice conversations, analytics dashboards, CRM integrations, and appointment scheduling.
- Designed scalable backend infrastructure with secure webhook processing, conversation persistence, tool-calling, and workflow orchestration.
- Implemented production-grade deployment, API security, cloud hosting, and monitoring for reliable, low-latency voice automation.

Tech: TypeScript, Node.js, Fastify, Next.js, React, Supabase, PostgreSQL, Twilio, ElevenLabs, Google Gemini, Railway, REST APIs

Alpha Lead Engine *AI-Powered Outbound Sales Automation Platform*

- Built and deployed a live outbound sales automation platform that discovers leads, qualifies them with AI, generates personalized outreach emails, automates multi-step follow-ups, and classifies replies.
- Architected six focused AI modules (lead qualification, website analysis, email generation, follow-up generation, reply classification, campaign recommendation), each with a single responsibility for maintainability.
- Built a full-stack production system: Next.js frontend, FastAPI backend, Supabase (database + auth), OpenAI API for personalization, n8n for workflow orchestration, Apify for lead scraping, and Docker for deployment.
- Designed a full campaign lifecycle with automated day-based follow-up sequencing and a live analytics dashboard tracking leads, replies, and conversion rates.

Tech: Next.js, FastAPI, Supabase, OpenAI API, n8n, Apify, Docker, REST APIs, Tailwind CSS

Machine Unlearning in LLMs *IEEE Generative AI Challenge — Research Paper*

- Co-authored “Surgical Amnesia: A Framework for Targeted Machine Unlearning and Copyright Compliance in Large Language Models,” a parameter-efficient unlearning method (Llama-3.2-1B, LoRA) for GDPR/EU AI Act compliance without full retraining. Team reached the final round of the IEEE Generative AI Challenge.

EDUCATION

Master of Computer Applications (MCA) — St Joseph's University, Bengaluru 2025 – 2027

Focus: AI/ML engineering, RAG systems, agentic architectures, LLM evaluation.

Bachelor of Computer Applications (BCA) — Kristu Jayanti University 2022 – 2025

Internships in web development and data analysis. Coursework in big data analytics: Hadoop, MapReduce, NoSQL, Hive, Pig.

LANGUAGES

English (Full Professional) · Kannada (Native) · Hindi (Professional Working) · Tamil (Elementary) · Spanish (Elementary, actively learning)