Lenin Jacob Regi

B.Tech – Computer Science & Engineering (AI/ML)
AI/ML Engineer | LLM & RAG Specialist | Full-Stack AI Developer
Karunya Institute of Technology and Sciences, Coimbatore *(2023–2027)*

+91-8891201402
leninjacobregi891@gmail.com
GitHub | Portfolio | LinkedIn

Education

Degree	Institute	CGPA/%	Year
B.Tech. (CSE – AI/ML)	Karunya Institute of Technology and Sciences	6.0	2023–2027
Senior Secondary / Secondary	State Board	74% / 100%	2023 / 2021

Experience

- **AI/ML Engineer – Independent Developer & Open-Source Contributor**
  30+ production AI systems: LLMs, RAG pipelines, multi-agent architectures, computer vision, full-stack apps 2023 – Present
  – Architected **hybrid RAG systems** (semantic + FTS + RRF) with multi-DB backends (Milvus, pgvector, Neo4j, Elasticsearch); integrated **100+ LLMs** across 7 providers via LiteLLM abstraction.
  – Optimized LLM inference to run **70B+ models on 4GB GPU** via layer-wise loading & block quantization – **3× speedup**; achieved **100% defect detection** (CV) and **93.3% efficiency** (google solution challenge).
  – Built **multi-agent LangGraph systems** with tool use, sub-agent delegation; engineered real-time full-stack apps with React/Next.js, FastAPI/Node.js, WebSocket, and containerized GPU-enabled Docker deployments.
- **Gen-AI Head – Karunya Innovation & Design Studio (KIDS)**
  Leading LLM/SLM research, agentic AI workshops, and hackathon teams for 100+ students Jan 2025 – Present
  – Built and deployed **Karunya-SLM** from scratch – GPT-2 architecture, custom BPE tokenizer, PyTorch training, WandB tracking.

Projects

- **SurfSense – AI Research Agent & Knowledge Management** *2024 – Present*
  20+ source integrations (Notion, Slack, GitHub, YouTube); Deep Agent with 11 tools, 6000+ embedding models, RBAC GitHub
  – **Stack:** FastAPI, Next.js, TypeScript, PostgreSQL+pgvector, LangChain, LangGraph, LiteLLM, Redis, Celery, Docker | **2-tier hierarchical RAG** with RRF, Pinecone/Cohere rerankers, podcast generation in **<20 sec**.
- **A.E.G.I.S – AI Code Vulnerability Detection & Auto-Remediation** *2024 – Present*
  Pre-deployment security platform: SAST (Semgrep, CodeQL) + DAST + LLM code review in zero-trust Kata Containers GitHub
  – **Stack:** Python, Mistral 7B, CodeLlama 13B, Semgrep, CodeQL, Flask, React, Docker | CVSS-scored prioritization, automated multi-candidate fix generation, human-in-the-loop approval UI with risk heatmaps.
- **Scoratis – Agentic 3D Socratic Learning Platform** *2024*
  Three.js 3D museum + LangGraph RAG tutor: 11 tools, 5 sub-agents, 6-step Socratic scaffolding, 10 subject domains GitHub
  – **Stack:** React 18, Three.js, FastAPI, PostgreSQL+pgvector, LangGraph, LiteLLM, Manim, Docker | **50+ API endpoints**, short/long-term memory, Manim animations, multi-LLM with VRAM-based model selection.
- **QUERIX.AI – YouTube Semantic Chatbot with RAG** *2024*
  Video pipeline: yt-dlp → Whisper → chapter-aware chunking → Milvus IVF_FLAT → deepseek-r1 Q&A GitHub
  – **Stack:** Python, Whisper, yt-dlp, FFmpeg, Pydub, Milvus, Ollama (mxbai-embed-large, deepseek-r1), Docker | LLM-generated fallback chapters; 4096-char chunk storage with etcd/MinIO Milvus stack.
- **maestro-studio – 100% Offline AI Video Generator** *2024*
  LLM script → Manim animation → Piper TTS narration → FFmpeg compositing; zero cloud, full GPU acceleration GitHub
  – **Stack:** Python, llama.cpp (Llama 3.2 3B), Manim, Piper TTS, FFmpeg, Gradio, CUDA | Generates **1080p 2–5 min videos in 2–5 min**; CLI batch mode, multiple voice personas, cinematic Manim templates.
- **YATHRA – Real-Time Full-Stack Travel Companion** *2024*
  Trip planning, smart expense splitting, memory wall (Cloudinary CDN), real-time group chat via WebSockets GitHub
  – **Stack:** Node.js, Express, MongoDB, Socket.IO, React, TailwindCSS, Cloudinary, Passport.js (OAuth 2.0), Chart.js | Railway + MongoDB Atlas deploy; Helmet, **rate limiting (100 req/15 min)**, settlement optimization.

Technical Skills

- **Languages:** Python, JavaScript, TypeScript, SQL, C++, Scala
- **AI/ML:** PyTorch, TensorFlow, LangChain, LangGraph, LiteLLM, Hugging Face, RAG, Milvus, pgvector, Whisper, OpenCV, XGBoost, WandB
- **Backend & Infra:** FastAPI, Flask, Django, Node.js, Express, Docker, Kubernetes, Redis, Celery, Nginx, GitHub Actions, NVIDIA CUDA
- **Frontend & DB:** React.js, Next.js, Three.js, TailwindCSS, Socket.IO, PostgreSQL, MongoDB, Neo4j, Elasticsearch, SQLite

Positions of Responsibility

- **Gen-AI Head,** Karunya Innovation & Design Studio (KIDS) – AI workshops, LLM research, hackathon mentorship *Present*
- **Coordinator,** K-Hacks Hackathon Club, KITS – Inter-college hackathons and technical events *Jan 2025*

Achievements

- **SIH Shortlisted** – **PGR-SR**: Physics-Guided Residual Super-Resolution for Thermal IR Imagery; **RMSE: 0.088K**, **PSNR: 41.44dB**, **SSIM:**
- **CV Excellence** – **100% defect detection rate** (40/40) using Laws' Texture Filters; 9-class plant pest detector with MobileNetV2 *2024*
- **Open Source** – SurfSense & Scoratis: active community, Discord channels, external contributors; 30+ public AI repositories

Certifications

- DeepLearning.AI: Vector Databases – Embeddings to Apps
- DeepLearning.AI: LangChain for LLM Application Dev
- DeepLearning.AI: Tools & Agents with LangChain
- DeepLearning.AI: ChatGPT Prompt Engineering for Devs
- Google: Foundations of Cybersecurity
- PCAP: Programming Essentials in Python
- Coursera: Deep Learning with PyTorch – GradCAM
- Coursera: Life Expectancy Prediction with ML
- CLA: Programming Essentials in C