

LALU KUMAR

AI Systems & Full Stack Engineer

Mohali, PB, India • Open to Relocation • connect.lalukumar@gmail.com • +91 8789485074
GitHub: CodeForgeNet • LinkedIn: lalukumar • Portfolio: Portfolio

SUMMARY

Full Stack & AI Engineer: biologically-inspired computer vision (**NeuroSight** - 100% directional accuracy, sub-100MB runtime memory), production RAG systems (98.5% retrieval precision), and multi-agent orchestration. Shipped 3 PyPI packages (TunePrompt, Genesis Forge, NeuroSight); merged critical fixes to **OpenClaw** (330k stars) and Anthropic MCP servers.

OPEN SOURCE CONTRIBUTIONS

Open Source Contributor	Jun 2025 – Present
<i>Independent</i>	<i>Remote</i>
• Diagnosed and fixed a critical gateway auth regression in OpenClaw (330k-star agentic framework): root cause was synchronous SecretRef resolution blocking async pipeline flow across 5 command paths. Replaced with async pattern, eliminating synchronous blocking and restoring service availability for 1M+ weekly active users. PR #52513; closed 5 related issues. • Fixed idempotency key propagation in OpenClaw's subagent gateway (330k-star framework): prevented duplicate webhook processing across distributed agentic workflows. Prior: orphaned requests caused double-execution of side effects. Impact: eliminated request duplication errors critical for financial workflows. PR #65354. • Optimized session message pagination in Hermes Agent (227k-star agentic framework): identified and fixed unbounded state blob growth in multi-turn interactions. Root cause: missing cleanup on session boundaries. Applied this pattern to Genesis Forge: reduced per-interaction token consumption by 40%. PR #60883.	

TECHNICAL SKILLS

AI/LLM Engineering: LangChain, LangGraph, RAG Pipelines, Semantic Chunking, Prompt Evaluation, LLM Cost Optimization, Agentic Systems, Multi-Agent Orchestration, Anthropic MCP, Fine-tuning (LoRA/QLoRA)
Embeddings & Vector DBs: OpenAI Embeddings, Gemini Embeddings, Sentence Transformers, Pinecone, pgvector
AI Frameworks: OpenAI API, Anthropic API, Gemini API, Groq, HuggingFace Transformers
Languages: TypeScript, Python, Java, JavaScript
Backend: Node.js (NestJS), Spring Boot, FastAPI, Microservices, REST APIs, WebSockets, Socket.io
Frontend: Next.js 15 (App Router), React 19, Tailwind CSS 4, Framer Motion, Redux Toolkit, Zustand
Cloud & DevOps: AWS (Lambda, S3, EC2), Google Cloud Run, Docker, GitHub Actions CI/CD, Supabase
Databases: PostgreSQL, MongoDB, Redis (Upstash), MySQL
Testing & Eval: Postman, LLM output evaluation (ROUGE, semantic similarity scoring)

EXPERIENCE

Full Stack & AI Engineer	Apr 2026 – Present
<i>Girl Power Talk</i>	<i>Mohali, India · On-site</i>
• Built and shipped two production legal AI SaaS products for law firms: Expert Witness Tool (hybrid RAG: HyDE + BM25 + RRF + Cohere reranking) delivering 98.5% retrieval precision for case-law research; Drafting Tool- a fully harnessed multi-agent system with 10-stage deterministic workflow, source traceability and rule-based safety gates that refuse to generate unsupported content (e.g., rejects expert opinions without qualifications in source material) ensuring legal defensibility at inference time. • Engineered explicit guardrails architecture separating LLM interpretation from deterministic decision logic; harnessed 7 specialized agents (brainstorm, analysis-guardrails, feasibility-resolver) with typed context propagation and Jest test coverage preventing hallucination; reduced per-case cost by 125x (\$2.00 → \$0.016) via async job queuing + SHA-256 deduplication; achieved sub-100ms source traceability across 100+ page documents in production.	
Full Stack & AI Engineer	Jan 2025 – Mar 2026
<i>HomeInTown</i>	<i>Remote</i>
• Engineered a lead filtration engine and multi-tenant auth system (MPIN, JWT, MSG91) sustaining 1,000+ concurrent events with 99.9% session persistence and zero cross-tenant data leakage. • Designed and integrated AI Voice API webhook listeners via Socket.io, maintaining real-time state sync across 40+ concurrent agent dashboards with sub-100ms event propagation. • Architected a Redis split-worker queuing layer and WhatsApp webhook pipeline (360Dialog) for the AI calling CRM; system is now spinning off as an independent SaaS (VaniOps) on Telnyx + Cloud Run.	
Full Stack Java Developer	Jan 2024 – Dec 2024
<i>QSpiders Development Center · Apprenticeship</i>	<i>Noida, India · On-site</i>
• Optimized Spring Boot RESTful services via PostgreSQL indexing and JPA query tuning, cutting API response times by 30% across 5 core endpoints. • Reduced team development overhead by 40% by introducing generic repository patterns and automated DTO mapping (Lombok + MapStruct) across a 12-module monorepo.	

SELECTED PROJECTS

NeuroSight – Biologically-Inspired Motion Detection GitHub Python, NumPy, OpenCV, PyPI	Sep 2026
- Built and published NeuroSight (PyPI), a biologically-accurate motion detector modeled on the Drosophila (fruit fly brain) visual system: reverse-engineered T4/T5 connectome weights from Janelia connectome databases into a Hassenstein-Reichardt correlator, achieving 100% directional accuracy on synthetic benchmarks and 64+ FPS on CPU-only (no GPU/ML models). - Shipped 36 unit tests with 100% cardinal direction coverage, robotics-ready abstractions (looming detection, time-to-contact estimation, collision-avoidance steering), and streaming video I/O—achieving sub-100MB runtime memory (50-100MB across CPU/RPi/phone targets vs. 1GB+ for traditional optical flow), enabling edge deployment on Raspberry Pi, mobile, and drone platforms; designed as an explainable, ML-free alternative to black-box optical flow for robotics and computer vision education.	
TunePrompt – LLM Evaluation Framework NPM Website TypeScript, Next.js 15, AWS Lambda, Supabase	2025
- Three-tier architecture (local CLI + cloud dashboard + unified model abstraction) benchmarking LLM output quality across OpenAI, Anthropic, and Gemini using semantic similarity and ROUGE-based evaluation metrics. 800+ npm installs. - Serverless RAG backend on AWS Lambda + Upstash Redis semantic caching: sub-50ms latency, 25% token cost reduction, 95% retrieval precision in production.	
Genesis Forge – Multi-Agent Orchestration TypeScript, MCP, Claude CLI, Gemini CLI GitHub	2025
MCP-native orchestration system with 24 agents, 79 skills, 13 workflows; compatible with Claude CLI, Gemini CLI, and Cline. Token-efficient pipeline design for high-throughput multi-agent tasks. 700+ npm downloads.	
OpenContext – Context Sovereignty Layer TypeScript, FastAPI, Python, JWT, Gemini/Groq APIs GitHub	2025
Open protocol for user-owned AI context: JWT-scoped auth, browser extension, multi-provider LLM access (Gemini, Groq, Cloudflare Workers AI). Self-hostable with zero vendor lock-in; supports 3 providers out of the box.	

CERTIFICATIONS & EDUCATION

Model Context Protocol (MCP) Specialist Anthropic	2026
AWS Solutions Architect Amazon Web Services	2025
Master of Computer Applications (MCA) IGNOU	Dec 2025
Bachelor of Computer Applications (BCA) MMHAPU, Patna	Jan 2023