

LAVA KUMAR KOYI

lavakumarkoyi@gmail.com | +1 812-778-4652 | linkedin.com/in/lavakumarkoyi17 | Orlando, FL | lavakumar.com

PROFESSIONAL SUMMARY

Forward Deployed AI Engineer with 6+ years embedding directly with Fortune 500 clients to take LLM-powered applications from whiteboard to production — identifying high-value use cases, prototyping, building the full stack, deploying to the cloud, and owning long-term support. Shipped secure, grounded AI assistants across financial services (a 251,000-member credit union) and consumer electronics (JBL/Harman, Epson), pairing hands-on technical execution with the stakeholder-facing communication that turns ambiguous business problems into working systems. Deep expertise in Retrieval-Augmented Generation and agentic architectures — hybrid retrieval, cross-encoder re-ranking, LLM fine-tuning (QLoRA / PEFT), synthetic-data generation, RAG evaluation, and safety guardrails — delivered with Python, modern LLM frameworks (LangChain, MCP), React front-ends, REST APIs, and AWS / Azure CI/CD. AWS Certified Cloud Practitioner.

TECHNICAL SKILLS

LLMs & Generative AI: LLMs (GPT-4o / 4.1, Mistral-7B), Prompt Engineering, Function / Tool Calling, Structured Outputs, Chain-of-Thought, Guardrails, PII Redaction, Prompt-Injection Defense

RAG & Retrieval: RAG & Agentic RAG, Hybrid Retrieval (Dense + BM25), Reciprocal Rank Fusion, Cross-Encoder Re-ranking, Semantic Chunking, Query Rewriting & Decomposition, Metadata Filtering, Grounding & Citations

Fine-Tuning & Embeddings: QLoRA / LoRA / PEFT, 4-bit (NF4) Quantization, Synthetic Data Generation, Instruction Tuning, Sentence-Transformer & Multilingual Embeddings (E5 / BGE), FAISS (Flat / IVF / HNSW)

Evaluation & Observability: RAG Evaluation (faithfulness, context precision / recall, answer relevancy), Golden Datasets, Regression Testing, Latency & Cost Monitoring, Structured Logging

Frameworks & Platforms: LangChain, Model Context Protocol (MCP), Agentic Workflows, OpenAI & Azure OpenAI, Kore.ai (Agent AI, SearchAI, BotKit SDK)

Languages & Full-Stack: Python, JavaScript, TypeScript, Node.js, React.js, Vite, Streamlit, FastAPI, Express.js, REST APIs, WebSockets, OAuth 2.0 + PKCE, JWT / JWE

Cloud & DevOps: AWS (Lambda, API Gateway, S3, EC2, CloudFront, WAF, ALB), Azure (OpenAI, APIM), Docker, GitLab CI/CD, Bitbucket Pipelines, PM2, RunPod (A100 GPU)

Databases & Integrations: PostgreSQL, MongoDB, MySQL, Redis; Salesforce CRM & Commerce Cloud (SFCC), Microsoft Dynamics 365, Genesys, ServiceNow ITSM

PROFESSIONAL EXPERIENCE

Lead Forward Deployed AI Engineer — Kore.ai

Jan 2026 – Present

- Lead architecture and delivery of production LLM applications across multiple Fortune 500 deployments, owning the full lifecycle from use-case discovery and prototyping through full-stack development, cloud deployment, and production support.
- Translate stakeholder requirements into technical specifications, mentor implementation engineers, guide external contractors, and drive C-suite adoption of the AI platform across enterprise accounts.

Truliant Federal Credit Union — Production Banking AI Assistant

[Live bot • truliantfcu.org](https://truliantfcu.org)

- Architected and shipped a production banking AI assistant (live on truliantfcu.org) for a 251,000-member credit union — owning end-to-end design from the React / WebSDK front-end through AWS Lambda back-end services and external integrations.
- Built the RAG knowledge pipeline: crawled and semantically chunked truliantfcu.org content, embedded it into a vector index, and grounded every response with source citations to keep answers accurate and auditable for a regulated banking context.
- Added compliance-focused guardrails — PII redaction, prompt-injection defenses, and intent / scope classification — so unsafe or off-topic queries are caught before reaching the LLM or the member.
- Engineered secure multi-layer JWT / JWE authentication with progressive-delay brute-force defense on stateless Lambda + Redis, and closed a critical account-enumeration vulnerability surfaced in penetration testing.
- Integrated the Salesforce CRM post-conversation summary API and deployed via GitLab CI/CD + PM2 with structured logging and monitoring for production observability and ongoing support.

JBL / Harman — CLaRa LLM Retail Assist Bot

- Architected and delivered an end-to-end LLM platform (product search, feature-based recommendations, multi-product comparison, compatibility checks, and troubleshooting) spanning JBL/Harman's full consumer-electronics catalog.

- Fine-tuned Mistral-7B with QLoRA — 4-bit NF4 quantization with low-rank PEFT adapters on RunPod A100 GPUs — injecting product-domain knowledge while preserving the base model's general reasoning at a fraction of full fine-tuning cost.
- Built a GPT-4o synthetic-data pipeline generating question-answer pairs, paraphrases, and verification sets across 4,000+ product records, then applied a two-stage strategy (retriever priming, then joint retriever-generator alignment) to remove the need for hand-labeled data.
- Designed the retrieval layer over a FAISS dense index with cross-encoder re-ranking for fast, source-attributed answers, wrapped in an agentic orchestration layer that routes each query (search vs. compare vs. troubleshoot) before LLM synthesis.
- Built a real-time Salesforce Commerce Cloud (SFCC) pipeline to keep the index synchronized with the master catalog, and embedded CLaRa into the contact center via Kore.ai Agent AI for live agent use.

Senior Forward Deployed AI Engineer — Kore.ai

Jan 2025 – Dec 2025

Epson — Multilingual Agentic RAG at Scale (LATAM + NAM)

[Live bot · epson.com](#)

- Designed and deployed a production Agentic RAG support assistant across 20,000+ products in 19 LATAM and NAM countries, engineering the full knowledge stack from ingestion through grounded answer synthesis.
- Built a multi-source ingestion pipeline over 400K+ documents (PDFs, manuals, spec sheets) and 200K+ crawled URLs with layout-aware parsing, normalization, deduplication, and product-aware semantic chunking — enriching each chunk with product, category, country, and language metadata for precision filtered retrieval.
- Implemented hybrid retrieval fusing dense vector search with sparse BM25 keyword matching via reciprocal rank fusion, so both natural-language questions and exact model-number lookups resolve accurately against a unified index.
- Added a cross-lingual embedding layer routing Spanish, Portuguese, and English queries against one shared index, plus a cross-encoder re-ranking pass over top-K chunks to reduce hallucination on technical product questions before synthesis.
- Wrapped retrieval in an agentic orchestration layer performing query classification, rewriting, and dynamic strategy selection (product lookup vs. troubleshooting vs. market-specific retrieval), returning grounded, source-cited answers with fallback handling for low-confidence queries.

United Federal Credit Union — Contact Center AI + CRM Integration

[Live bot · unitedfcu.com](#)

- Built a BotKit integration connecting Kore.ai conversational AI to Genesys and Microsoft Dynamics 365 (via Azure APIM), automating live work-item creation at login and transferring full conversation context to agents on handoff.
- Implemented Redis session-scoped activity tracking to support multi-session users, and diagnosed and resolved a critical go-live routing incident through BotKit WebSocket flow analysis.

Graduate Research & Teaching Assistant — AI / NLP | Indiana University Bloomington

Sep 2023 – Jan 2025

- Built an end-to-end RAG chatbot for IU's online course platform — transcribing lecture audio, chunking and embedding course material (syllabi, slides, readings) into a vector store, and serving retrieval-augmented Q&A over natural-language student queries.
- Applied Named Entity Recognition to detect and mask PII (names, IDs, contact details) before indexing for FERPA-compliant handling, and trained CNNs for satellite flood-region classification on IU's Big Red 200 HPC cluster.
- As Teaching Assistant, led lab sections, graded assignments, and mentored graduate students across machine-learning and data-mining coursework.

Software Engineer — Kore.ai | Hyderabad, India

Jul 2020 – Jul 2022

- Built Node.js API adapters for IT Assist and HR Assist bots (ServiceNow ITSM), cutting ticket resolution time by 40%; developed live-agent transfer over WebSockets and optimized Redis chat-queue management, reducing message-delivery latency by 50%.

Junior Software Developer — Stratemis HR Technologies | India

Oct 2019 – Jul 2020

- Developed a Django-based chatbot platform integrating rule-based and ML-trained bots, and engineered a JavaScript web SDK for widget deployment with JWT authentication and session management.

EDUCATION

M.S., Computer Science — Indiana University Bloomington | GPA 3.83 / 4.0

Aug 2022 – May 2024

Relevant coursework: Applied Machine Learning, Deep Learning, Natural Language Processing, Data Mining, Computational Statistics, Software Engineering, Computer Networks.

B.Tech., Computer Science & Engineering — Lovely Professional University, India | GPA 3.66 / 4.0

Aug 2016 – Jul 2020

CERTIFICATIONS

AWS Certified Cloud Practitioner