RAMACHARAN REDDY KASIREDDY

✉ kramacharanreddy@gmail.com · ☎ +1 716-292-6481 · ramacharanreddy-k · ramacharanreddy-k

AI Engineer with 2+ years of experience designing and delivering production AI systems – multi-agent LangGraph workflows, RAG pipelines with hallucination reduction, vector search, and data ingestion engines – across both AWS and Azure, with end-to-end ownership from Terraform infrastructure to client-facing delivery.

SKILLS

AI/ML & LLMs	LangChain, LangGraph, MCP, RAG (hybrid k-NN + BM25), Prompt Engineering, Fine-Tuning, Multi-Agent Systems, ReAct, Tool Calling, Structured Outputs, Guardrails, Langfuse, LangSmith, Semantic Search, Conversational AI
LLM Platforms	Azure AI Foundry, Azure OpenAI (GPT-4, Embeddings), AWS Bedrock (Claude, Titan Embeddings), Google Gemini
Deep Learning	PyTorch, TensorFlow, Hugging Face Transformers, LoRA, PEFT, QLoRA, Quantization, Distillation
Programming	Python, SQL, JavaScript, Java, React, FastAPI, Uvicorn, Pydantic, Jinja, Playwright, pytest
Data & APIs	PostgreSQL, Cosmos DB, OpenSearch, SQLAlchemy, RESTful APIs, OpenAPI/Swagger, OAuth 2.0, JWT, MSAL, JWKS, SSE, WebSockets
AWS	Bedrock, Bedrock Guardrails, OpenSearch, EKS, ECR, Lambda, SQS, S3, Transcribe, DynamoDB, CodePipeline, KMS, IAM (IRSA)
Azure	OpenAI Service, Cosmos DB, AKS, ACR, Key Vault, Blob Storage, Entra ID, Managed Identity, Monitor, Logic Apps
DevOps & Tools	Docker, Kubernetes, Helm, Terraform, Airflow, Git, GitHub Actions, Jenkins, GitLab CI/CD, GitHub Copilot, Cursor, Claude Code

WORK EXPERIENCE

Feuji INC — AI Engineer | Aug 2024 – Present

Customer Analytics AI – PLZ Corp

Python, FastAPI, LangGraph, Azure OpenAI, Azure AI Foundry, LangSmith, Cosmos DB, PostgreSQL, Terraform, React

PLZ needed to track customer growth, monitor industry news, and identify the best time to close deals. The client requested a static reporting dashboard – I proposed and delivered a conversational AI platform that lets the sales team query all of that data in natural language, replacing weeks of manual research with real-time insights.

- **Architected a full-stack AI analytics platform** across 4 microservices (data engine, SSO gateway, conversational agent, React SPA) deployed on AKS, eliminating 48 hours/week of manual analysis and saving $150K/year.
- **Engineered an AI agent with LangGraph and Azure OpenAI** – 20 specialized tools, 7-layer middleware stack, and dual-domain RAG over Cosmos DB vector search (DiskANN, 1536-dim embeddings).
- **Grounded all responses in verified data** to reduce hallucinations; added summarization and short-term memory for conversational coherence.
- **Built a multi-stage async data pipeline** (parse, clean, chunk, embed, index, validate, store, sync) ingesting structured and unstructured data into a 12-table PostgreSQL warehouse and a Cosmos DB vector store.
- **Designed a centralized SSO gateway** with FastAPI, MSAL, OAuth2 authorization code flow, Entra ID group-based access control, JWKS S2S validation, and zero hardcoded secrets via Key Vault and Managed Identity.
- **Implemented observability and cost monitoring** – LangSmith for agent trace debugging and prompt regression testing; Azure Monitor for token-level cost tracking and content safety logging.
- **Provisioned the full Azure environment with Terraform** across 7 modules with passwordless RBAC, covering compute, data, AI, identity, storage, scheduling, and communication services.
- **Recognized by PLZ Corp's Director and CIO** for exceeding project expectations and delivering a platform that transformed their sales decision-making process.

Ask the Prospect (ATP) – TriNet

Python, FastAPI, LangGraph, AWS Bedrock, Bedrock Guardrails, OpenSearch, Aurora, Langfuse, Airflow, EKS

TriNet's sales reps typically take 2 years to reach full proficiency. The client asked for a searchable training knowledge base – I expanded the scope to a multi-mode AI agent combining RAG Q&A, persona-based role-play coaching, sales analytics, and gamified learning to compress that ramp-up across the entire sales organization.

- **Built a multi-mode LangGraph agent** powering 4 capabilities – RAG chatbot (dual-index: training docs + Gong transcripts), persona-based sales role-play with AI coaching, analytics chat, and automated Jeopardy question generation.
- **Enabled memory continuity** via LangGraph PG checkpointer and store for short-term and long-term memory across sessions.
- **Engineered a sales analytics pipeline** – LLM enriches each Gong transcript with derived columns (win/loss, objections, competition mentions), then generates OpenSearch queries from natural language for aggregate analysis.
- **Enforced PII protection with AWS Bedrock Guardrails** – configurable filters that detect and redact sensitive data (names, emails, SSNs) in both inputs and outputs before reaching the client.
- **Integrated Langfuse for observability and cost monitoring** – LLM-as-judge evaluations for quality scoring, trace-level debugging, latency monitoring, and per-call cost tracking in production.
- **Served as solution architect and client-facing lead**, owning end-to-end technical vision across 4 microservices and 2 Airflow-orchestrated data pipelines on EKS.

- **Recognized by TriNet's CTO, Executive Director, and VP** for expanding scope beyond original requirements and delivering measurable impact to the sales organization.

ReQon – Automated Testing Tool

Python, LangChain, LangGraph, Google Gemini, Playwright MCP, Aurora PostgreSQL, S3, Docker, pytest

Feuji is a QA-heavy company where test engineers were spending the bulk of their time writing and maintaining low-level test cases. I led an internal initiative to build an AI platform that auto-discovers application structure and generates test cases and regression scripts from a single browser recording, freeing engineers to focus on high-value exploratory testing.

- **Engineered a LangGraph agent with Playwright MCP** that replays a recorded browser flow, navigates every page, and extracts locators, functional elements, and behavioral data into S3 and Aurora PostgreSQL.
- **Automated the full QA authoring workflow** – from captured browser behavior the agent generates test cases, executes them via Playwright MCP, and outputs verified pass/fail results, eliminating manual test writing.
- **Built a code generation engine** using Jinja templates that converts page metadata into pytest-compatible regression scripts with Page Object Model classes for every discovered page.
- **Designed and dockerized 3 microservices** – Recorder (Playwright codegen), MCP Service (AI agent + test generation), and Execution Service (isolated runner) – saving 100+ team-hours weekly and cutting QA costs by $230K/year.

AI-Powered IVR Automation – Cotiviti

Java, Python, LangGraph, AWS Bedrock, Transcribe, DynamoDB, Lambda, SQS, KMS, Twilio

Cotiviti's Medicare IVR verification ran on legacy Jabber VMs, requiring a human to manually review and update navigation paths every morning. The team migrated to a fully serverless, horizontally scalable AWS architecture and introduced an AI agent that autonomously discovers the correct IVR path daily, eliminating the manual review step entirely.

- **Designed an AI pathfinding workflow with LangGraph and AWS Bedrock** that navigates Medicare's IVR daily – recording calls via Twilio, transcribing with AWS Transcribe, and writing the correct DTMF path to DynamoDB for 200,000+ daily production calls.
- **Built a mock Medicare IVR simulator** in Java as a regression harness, validating AI agent accuracy against simulated decision trees before production deployment.
- **Configured the AWS CI/CD pipeline** (CodePipeline) with KMS-encrypted PII handling across all environments.
- **Part of a cross-functional team** that saved 730+ engineer-hours annually and cut infrastructure costs by ~60% through serverless auto-scaling.

University at Buffalo AI Research Assistant | Jan 2024 – May 2024

Energy-Efficient LLM Inference and Carbon-Aware Optimization

Python, PyTorch, TensorFlow, Hugging Face, NumPy, Pandas

- **Benchmarked parameter-efficient fine-tuning (PEFT) techniques** – LoRA, knowledge distillation, and post-training quantization (4-bit/8-bit) – on GPT-2, measuring GPU energy consumption and carbon emissions across task configurations.
- **Achieved 19.8% CO2 reduction** with only 6% perplexity increase using 8-bit quantization, demonstrating significant energy savings with minimal quality degradation.
- **Demonstrated 45.2% emissions reduction** by combining distillation with quantization, mapping the accuracy-efficiency trade-off curve for carbon-aware deployment.

Flable AI Machine Learning Engineer | Mar 2023 – Aug 2023

Flable Digital Assistant

Python, LangChain, OpenAI, PostgreSQL, FastAPI, Docker, Jenkins, GitLab

- **Built a LangChain conversational AI with OpenAI** that classifies user intent via transformer embeddings, maps queries to PostgreSQL data categories, and generates natural-language responses grounded in real results.
- **Developed an on-the-fly visualization engine** that dynamically queries PostgreSQL, renders charts, and delivers them inline via Telegram – giving non-technical users instant access to analytics.
- **Deployed the full stack** (FastAPI backend + Telegram bot) using Docker with GitLab CI/CD and Jenkins pipelines, reducing client's weekly analysis time by 60%.

OPEN SOURCE

langgraph-checkpoint-cosmos | github.com/LangModule/checkpoint-cosmos

- Published a LangGraph checkpoint saver for Azure Cosmos DB with sync/async support, tip-document optimization, transactional consistency, and DefaultAzureCredential keyless authentication.

EDUCATION

University at Buffalo, The State University of New York Buffalo, New York
Master of Science, Computer Science, CGPA: 3.7/4 Aug 2023 – Dec 2024
Coursework: Data Intensive Computing, Computer Vision, Machine Learning, Deep Learning