

SHRAVYA REDDY AKMY

PROFESSIONAL SUMMARY

Senior Software Engineer with 5+ years building Python backend systems, agentic AI infrastructure, and evaluation/observability tooling across fintech, SaaS, and network infrastructure. Currently, building AI-agents and RAG platform components at Citizens Bank, including tool orchestration, multi-agent workflows (LangGraph, MCP), and reliability instrumentation for measuring agent behavior in production. Strong record of designing Python APIs and SDK-facing components, validating structured data with typed schemas, and turning runtime evidence into measurable product improvements.

CORE SKILLS

Languages & Frameworks: Python, JavaScript, Golang, FastAPI, Django, Next.js
AI/ML Engineering: Agent Orchestration, RAG, Embeddings, LangChain, LangGraph, MCP, Vector Databases (Pinecone), TensorFlow
Databases: PostgreSQL, MongoDB, Redis, Firebase
Cloud & Infrastructure: AWS (Bedrock, Lambda, EC2, DynamoDB, S3, SQS, CloudWatch), Docker, CI/CD, Containers (exposure to Kubernetes)

EXPERIENCE

Citizens Bank — Remote, USA Jan 2025 – Present

Senior Software Engineer

- Build Python/FastAPI backend services powering AI agent tool orchestration and tool calling for approval workflows, reducing manual steps in controlled action execution by an estimated 30–35%; develop RAG pipelines and reusable AI platform components — tool contracts, prompt orchestration, memory strategies, policy validation, and guardrails — using Pydantic-based typed schemas, extending into multi-agent orchestration with LangGraph and MCP.
- Strengthen reliability and observability through async execution, retries, idempotent processing, and structured logging, cutting failed-request retries by an estimated 20–25%; integrate AWS Bedrock for governed model inference and harden deployments with Docker, CI/CD, and monitoring tooling.

Headstarter — Remote, USA Jul 2024 – Dec 2024

Software Engineer

- Built core backend services for Club3, an AI-enabled social discovery platform, using Python and Next.js to power user-matching, messaging, and real-time community features; redesigned discovery from keyword matching to LLM-based semantic search using OpenAI embeddings and Pinecone, lifting match relevance from ~38% to ~75%.
- Shipped a location-based discovery feature end-to-end, from design discussion through rollout, driving a ~25–30% lift in weekly active users.

Intellisavvy — Irving, TX, USA Jun 2023 – Jul 2024

Full-Stack Developer

- Built and optimized RESTful APIs powering core transaction and job-matching workflows on the Traavo marketplace platform using Node.js and Express.js; engineered a job-opportunity notification system that meaningfully increased user engagement.
- Delivered role-specific access controls, including approval workflows for teenage users, supporting platform safety and compliance requirements; replaced manual data prep with Python-based ETL scripts for operational reporting.

Juniper Networks — Bengaluru, India Aug 2018 – Jul 2021

Python Developer

- Built an NLP-based script-quality ranking and evaluation tool achieving 96% classification accuracy; parallelized the workload to cut full-repository analysis time, later adopted company-wide as the standard QA benchmark.
- Built a Python automation tool with daily cron-driven ticket summaries and duplicate detection, lifting low-priority SLA adherence from ~30% to ~70–75% and cutting unassigned critical tickets by more than half; automated 200+ regression and performance test suites using Robot Framework.

PROJECTS

Camera Surveillance — Real-Time Video Intelligence Pipeline

- Own the face detection component of a real-time video intelligence pipeline, evaluating YOLO and RetinaNet for accuracy, latency, and GPU efficiency, and building the detection pipeline feeding downstream tracking and analytics.
- Currently extending the pipeline's vision-language model (VLM) layer for dense video captioning and anomaly detection, exploring NVIDIA Cosmos-Reason2-8B and other VLM backends against real-time performance constraints.

Object Detection for Visual & Mixed Reality Applications M.S. Capstone

- Fine-tuned a TensorFlow SSD MobileNetV2 FPN-Lite object detection model using transfer learning, custom data labeling, and TFRecord pipelines, achieving 0.88 mAP@0.5 IoU and 0.72 average recall after 20,000 training steps.

EDUCATION

M.S. Computer Science — University of Dayton, OH Aug 2021 – May 2023

B.Tech, Electronics and Communication — NIT Puducherry Jul 2014 – May 2018