Sarim Ahmed

8582883122 sarimahmed3520@gmail.com Sarim Ahmed Sarim

Experience

Mira **Remote**

Senior Software Engineer *Sep 2023 – Present*

- **Pioneered the Mira Network**, a distributed LLM routing platform featuring a FastAPI router, Fastify service nodes, and a Go CLI, increasing request volume handled by 5x.
- **Built and scaled Klok**, the flagship RAG application on the Mira Network, to **over 300,000 registered users** with 1,000+ concurrent daily users.
- **Strengthened platform security and automated deployments** by provisioning a secure AWS VPC with private subnets for EC2 instances and implementing a full CI/CD pipeline with GitHub Actions.
- **Configured and optimized Milvus** as the primary vector database, enabling high-performance semantic search for the core RAG functionality.
- **Engineered a Python SDK and a YAML-based flow system** to accelerate development, leading to adoption by **over 100 hackathon participants**.
- Elevated team performance as a regular reviewer within the agile team, reducing code review turnaround time by 20% while improving collaboration and knowledge transfer with clear feedback.

WT Studios **Remote**

Founding Engineer *Jan 2023 – Jul 2023*

- **Spearheaded the implementation of a secure, multi-tenant architecture** utilizing Clerk for authentication and Supabase Row-Level Security with Drizzle ORM, ensuring zero data breaches.
- **Engineered an SDXL model training workflow** to generate fine-tuned models from user photos and created ComfyUI workflows to produce consistent, cinematic shots.
- **Automated internal content pipelines** with Python scripting, generating videos with AI-powered audio from ElevenLabs and creating posters from templates.

Skills

Languages: Python, TypeScript, JavaScript, Go, SQL

Backend & APIs: FastAPI, Node.js, Next.js, Kong API Management

AI/RAG: vLLM, OpenRouter, LlamaIndex, LangChain, RAG, SDXL, ComfyUI, MediaPipe, Milvus

Infra & Observability: AWS (VPC, ECS, EC2, Fargate, S3, Lambda, Glue, Athena, WAF), GCP, CI/CD (GitHub Actions), Docker, Grafana, Prometheus, Loki, Thanos, Cloudflare, NGINX

Data & Messaging: PostgreSQL/pgvector, Redis, Kafka, RudderStack, Drizzle ORM, Supabase

Mobile: Android, Expo, Jetpack Compose, Room, Hilt

Projects

Guardian Platform GitHub

- Devised a comprehensive full-stack service discovery platform to **manage 100+ services and multi-account AWS resources**, leveraging a '.guardian' YAML manifest system, enabling automated dependency mapping.
- Implemented an **AI-powered documentation chatbot using LlamaIndex** and pgvector. Built with Next.js 15, PostgreSQL, and Drizzle ORM, and containerized with Docker for easy deployment.

PennyWise AI GitHub

- Developed a privacy-first Android expense tracker using on-device AI (MediaPipe) to parse bank SMS, achieving **100+ GitHub stars and downloads** for privacy-centric approach.
- Engineered the application with a modern MVVM architecture using Jetpack Compose, Room, and Hilt, resulting in a robust, crash-free, and maintainable codebase.

Education

Visvesvaraya Technological University **Belagavi, India**

Bachelor of Engineering in Computer Science *2019 – 2023*