Rajbasheer Baig Mogal (He/Him)

Harrison,NJ,07029 | mrajbasheer@gmail.com| (862) 405-2051 | linkedin.com/in/rajbasheerbaig-mogal/

SUMMARY

Senior Engineer with 6+ years of experience in building scalable, cloud-native applications and LLM-powered AI systems. Proficient in **Python**, **FastAPI**, **TensorFlow**, **PyTorch**, and **LangChain**, with hands-on expertise in **Generative AI**, **Retrieval-Augmented Generation (RAG)**, and **NLP**. Designed and deployed high-performance APIs and intelligent automation workflows, reducing **inference latency by 40%** and improving model accuracy by **25%**. Skilled in **AWS (Lambda, S3, API Gateway, DynamoDB, RDS)**, **Azure**, and **GCP Vertex AI** for deploying MLOps pipelines and microservices. Experienced in integrating **AI chatbots**, **speech-to-text/text-to-speech**, and fine-tuned **LLMs** into production systems. Passionate about combining backend engineering and GenAI to deliver robust, real-time solutions that scale across high-traffic environments.

EDUCATION

Pace University, Seidenberg School of Computer Science and Information Systems New York City

Masters in Computer Science

EXPERIENCE

Vitel Global New Jersey,USA

Backend Engineer(Gen-AI) Feb 2024 – Present

Project- Voice Agent(https://callingagent.pranathiss.com/)

Tech Stack: Python, AWS (Lambda, EC2, S3, API Gateway), GCP, FastAPI, Docker, Kubernetes, Twilio, Deepgram, Eleven Labs, Whisper, Tacotron, VITS, OpenAI, Gemini, LangChain

- Built an **AI-powered Voice Agent** using **Python** that automates outbound calling and appointment scheduling, achieving **3× response efficiency** and **60% reduction in manual work**.
- Developed **modular Python microservices** to handle voice input, transcription, LLM querying, and voice synthesis with **real-time streaming** capabilities.
- Deployed core services on **AWS Lambda** and **EC2** with **API Gateway** for real-time interactions; stored audio logs and analytics data in **AWS S3**.
- Integrated **Twilio voice APIs** and **Deepgram/Eleven Labs** using Python wrappers for high-accuracy transcription and natural-sounding voice responses.
- Achieved **95%+ transcription accuracy** and **30% faster response time** through optimized speech pipelines using GCP.
- Enhanced AI output with **LangChain + RAG**, increasing contextual accuracy by **40%** and reducing repetitive answers by **30%**.
- Packaged the solution with **Docker**, orchestrated with **Kubernetes**, and ensured **99.9% uptime** under **10,000+ calls.**

Project- Avatar Chat Bot(https://app.avatarchatbots.ai/)

Tech Stack: Python, AWS (EKS, CloudWatch), Synthesia, FastAPI, LLMs (OpenAI, Gemini), LangChain, NLP, Docker, Kubernetes

- Engineered an **interactive AI Avatar Chatbot** using **Python** and **LLMs**, delivering **45% higher user engagement** and **50% lower latency** in user queries.
- Utilized **Synthesia** avatars with Python-based message processing for animated, speech-enabled human-like chatbot interfaces.
- Implemented **contextual memory** and **intent recognition** via LangChain and NLP libraries to improve response relevance by **40%**.
- Integrated **text-to-speech** with expressive voice synthesis and avatar synchronization for seamless communication.
- Deployed on **AWS EKS (Elastic Kubernetes Service)** with **FastAPI containers**, enabling **high availability** and real-time autoscaling.
- Leveraged **CloudWatch** for performance monitoring and alerting, and automated health checks via Python scripts.
- **Deployed** the chatbot in a **containerized FastAPI & Kubernetes environment**, ensuring **99.9% uptime** and supporting **thousands of users concurrently**.

Project- Multi-LLM Code Generation Platform

Tech Stack: Python, FastAPI, LangChain, OpenAI, Claude, Gemini, Google Cloud (GCP), CI/CD, Docker, LLM APIs

- Contributing to the development of a **Generative AI platform** inspired by **Bolt**, designed for automated **code generation** using multiple **Large Language Models (LLMs)** including **OpenAI**, **Claude**, and **Gemini**.
- Architecting a **modular backend system** using **Python** and **FastAPI**, enabling intelligent **prompt routing**, **token management**, and **LLM fallback mechanisms** for enhanced reliability and response optimization.
- Building a **cloud-native infrastructure** on **GCP**, integrating **CI/CD pipelines** for scalable deployments, real-time request processing, and seamless **multi-model orchestration** of **LLM APIs** in a production-ready environment.

Accenture Hyderabad,India
Full-Stack Developer Sep 2021 – Jan 2023
Project- Toyota Material Handling(https://www.toyotaforklift.com/)

- **Developed scalable RESTful APIs** using **Python (FastAPI)** for real-time **inventory tracking**, **equipment management**, and **order processing**.
- Designed and maintained robust **Python backend services**, ensuring high performance and reliability across all customer and dealer-facing endpoints.
- Integrated APIs with **AWS API Gateway** and deployed backend logic using **AWS Lambda**, reducing infrastructure overhead by **30%**.
- Built responsive frontend interfaces using **React.js** and **Tailwind CSS**, enhancing user experience across devices.
- Implemented **JWT authentication** and **role-based access control (RBAC)** to manage access for customers, dealers, and administrators.
- Managed data with **PostgreSQL** and **MongoDB**, optimizing database queries for faster service history retrieval and product availability.
- Boosted system performance using **Redis caching**, **lazy loading**, and optimized API queries—achieving a **40% improvement** in load times.
- Automated deployments with **GitHub Actions** and **Jenkins**, enabling continuous integration and smooth production rollouts.
- Utilized **AWS S3** for media asset storage and **CloudWatch** for real-time application monitoring and logging.

PSS IN Hyderabad,India
Associate Jan 2019 – Aug 2021

- Developed and integrated **RESTful APIs** in **Python**, enabling smooth communication between frontend systems and third-party tools.
- Assisted in building a **rule-based chatbot engine** using Python and NLP libraries, which improved **customer response efficiency by 35%**.
- Wrote **Python scripts** for processing and storing customer interactions in **AWS DynamoDB** and **PostgreSQL**, ensuring fast, reliable data access.
- Helped implement **serverless functions** using **AWS Lambda**, improving scalability and reducing infrastructure costs.
- Contributed to **API performance optimization**, including rate limiting, logging, and security enhancements to support **99.9% uptime**.

TECHNICAL SKILLS

- **Programming Languages:** Python, Java, R, C++, JavaScript, TypeScript
- **AI/ML Frameworks:** TensorFlow, PyTorch, Hugging Face, LangChain, OpenCV, Keras, Vision Transformers.
- **MLOps & Deployment:** AWS SageMaker, Azure ML, GCP, Docker, Kubernetes, MLflow, Kubeflow, Jenkines.
- **CI/CD & DevOps:** Jenkins, Azure DevOps, GitHub Actions, GitLab CI/CD**.**
- **Web Development:** React.js, Next.js, FastAPI, Flask, Node.js, Tailwind CSS, Spring Boot.
- **Databases:** SQL, MongoDB, Hadoop, Spark, BigQuery.
- **IDEs & Development Tools:** IntelliJ IDEA, Eclipse, PyCharm, Jupyter Notebook, Visual Studio Code,Visual Studio, Postman.
- **Tools & Methodologies:** Agile, Scrum, JIRA, Selenium, Katalon, Apache Kafka, RESTful APIs

ACADEMIC PROJECTS / PERSONAL PROJECTS

MediLink Sep 2024-Dec 2024

- MediLink is Healthcare Record Management Platform for Patients and Doctors to access Medical Records at Oneplace.
- Developed a **AI-driven healthcare platform** integrating **React.js, Spring Boot, and MySQL** for managing patient records.
- Implemented an **LLM-powered search system** to enhance medical record retrieval, reducing search time by **50%**.
- Designed **role-based authentication** and **real-time document handling**, ensuring compliance with healthcare standards.

Emotion based Music Player

- Developed a music player with emotional recognition to personalize listening experiences.
- Built an **AI-powered music recommendation system** using **Facial Emotion Recognition (FER)** with **TensorFlow and OpenCV**.
- Developed a **Flask-based backend** to analyze emotions and dynamically generate personalized playlists.
- Optimized the **emotion classification model**, improving accuracy by **25%** for real-time user experience**.**