BHANUJA AINARY
+1 (940) 843-8539 | ainarybhanuja@gmail.com | LinkedIn | GitHub

SUMMARY

Results-driven full-stack developer with 2+ years of experience building web and AI-powered applications. Proficient in React, Java, Spring Boot, Node.js, MongoDB, MySQL, Python, Flask and leveraging AWS cloud services to build scalable and cost-effective solutions. Experienced in researching and working with Large Language Models (LLMs) and MLLMs for AI-driven applications. Seeking full-stack roles to drive innovative, high-impact solutions.

PROFESSIONAL EXPERIENCE

HPCC Lab | Denton, Texas | Research Assistant **January 2024 - Present**

Audo-Sight | Multimodal Large Language Models

- Developed Audo-Sight**,** an assistive system integrating Multimodal Large Language Models (**MLLMs**) for context-aware interaction and guidance for Blind and Visually Impaired (BVI), improving navigation and environmental understanding by over 70% compared to traditional white canes.
- Implemented latency-aware routing, optimizing response time by 25% and enabling optimal LLM selection based on task urgency.
- Integrated **NeMo Guardrails** to ensure respectful, BVI-aware, and safe language generation, reducing inappropriate or unclear responses by 85%.
- Utilized **LangChain**, and quantized **LLaVA** to enable edge-deployable multimodal reasoning for real-time assistive interaction.
- Conducted latency benchmark on 5 LLMs; results showed pure text LLMs were 96.4% faster on GPU and 97.7% faster on CPU compared to MLLMs.

Presidio Cloud Solutions | Coimbatore, India | Software Engineer **January 2022 - July 2023**

Ardent Mills - Client Project | C#, **.NET Core 8**, **Azure** (APIM, App service, Key Vault, SQL Server), Moq, xUnit, Swagger, Redis

- Designed and developed scalable REST APIs to handle complex pricing logic, enabling real-time margin and price calculations for each contract
- Optimized API performance by ~80% using **Redis cache** and advanced asynchronous programming techniques such as **Task.WhenAll**, **Semaphore** control, and parallel loops, significantly improving application responsiveness and reducing thread-blocking operations.
- Conducted thorough unit testing using **xUnit** and **Moq**, increasing test coverage by 35% and ensuring robust validation of business logic.

EventBuzz - Event Management for Presidio Employees | **React**, **Node.js**, **AWS** (API Gateway, Lambda, Cognito, CloudFront, S3, RDS, CI/CD)

- Led frontend development using React.js and Redux, including a drag-and-drop form builder that improved event organization efficiency by 30%.
- Implemented middleware for Authorization and Authentication in ExpressJs framework. Developed scalable backend APIs for Leaderboard flow and integrated them with the Frontend systems.
- Utilized **PostgreSQL** database hosted on **AWS RDS** for efficient data management and seamless integration with the Event Management System
- Designed and implemented end-to-end **CI/CD pipelines** using AWS CodePipeline, CodeBuild, and Code Commit, automating build, test, and deployment workflows

PROJECTS

Finance Fraud Detection | Transformers, Chroma DB, OpenAI APIs, Python, Streamlit

- Built an intelligent fraud detection system for phone call transcripts, leveraging LLMs with **RAG** to identify and classify scam in real time.
- Utilized vector-based semantic search (Chroma DB) to retrieve scam-related patterns and phrases from a curated database of known fraudulent scripts.
- Integrated RAG pipeline with **OpenAI** to provide context-aware risk analysis of incoming conversational data.
- Achieved over 90% precision in detecting scam attempts by comparing real-time call content with retrieved scam templates

Zhuttle - Commute app | Node, React Native, Typescript, AWS EC2

- Developed backend using Node and **MongoDB**, and deployed on AWS EC2, ensuring scalability, reliability, and efficient scheduling management.
- Designed and implemented **React Native** screens with Redux for seamless, responsive user experiences and real-time state management
- Integrated **Google Cloud Messaging (GCM)** for push notifications and **websockets** for live driver location tracking, enhancing communication.
- Implemented MFA and integrated Stripe for secure payment processing. Secured user authentication with **JWT** for stateless session management.

TECHNICAL SKILLS

- **Programming & Scripting:** C++, C#, Java 11+, Python, JavaScript, Typescript, HTML5, CSS, SCSS
- **Frameworks:** React, React Native, Node.js, Express, Spring Boot, .Net Core 8, Flask, Streamlit
- **Cloud Platforms:** AWS (EC2, EMR, S3, ECS, DMS, Lambda), Azure (Key Vault, Development & DevOps, Serverless)
- **Database Technologies:** MySQL, MongoDB, DynamoDB, AWS RDS
- **DevOps:** Docker, Git, Gitlab CI/CD
- **Libraries:** PyTorch, TensorFlow, Scikit-learn, Keras, Hugging Face, LangChain, OpenCV, Pandas, NumPy, SciPy, NLTK

EDUCATION

Master of Science in Computer Science | University of North Texas (UNT), Denton, Texas. | GPA - **4/4**
Bachelor of Technology in Electronics and Communication Engineering | Sastra University, Thanvajur, India. | GPA - **3.7/4**

ACHIVEMENTS

- **Top 3%** - Recognized as Distinguished Computer Science Student at University of North Texas
- Won **special award** in HackUNT - Goldman Sachs Problem Statement, 2024
- Won **2nd Place** in SMU Datathon, 2024
- **Top 2%** of Dean's merit list in SASTRA University, 2022

RESEARCH

- Audo-Sight: Enabling Ambient Interaction For Blind And Visually Impaired Individuals
- HE2C: A Holistic Approach for Allocating Latency-Sensitive AI Tasks across Edge-Cloud

CERTIFICATIONS

- **AWS Solutions Architect-Associate Certification,** 2023