SAI RAGHAVENDRA MADDULA
Dallas, TX, 75080 | 940-843-3560 | raghudo888@gmail.com
LinkedIn | GitHub | My Blog

SUMMARY

Highly motivated and results-driven **Computer Science and Information Technology** professional with a strong foundation in **AI/ML and DevOps.** Proven ability to independently **develop and deploy 10+ end-to-end AI/ML projects** within **tight deadlines**, earning praise from **over 15 satisfied clients** for exceptional communication and analytical skills. Passionate about integrating **Generative AI** practices to automate **workflows and deliver scalable**, efficient solutions. Adept at solving complex problems and **driving impactful innovations** within collaborative, fast-paced environments. Seeking a challenging role in the **United States** to contribute technical expertise and strategic thinking to cutting-edge projects.

TECHNICAL SKILLS

- **Programming Languages:** Python, Java.
- **Web Development:** HTML, CSS, React, Node.js, Flask, Axios, SaaS.
- **Automation Tools:** Docker, Kubernetes, Jenkins, Ansible, Terraform, AWS (incl. Amplify), Selenium.
- **Machine Learning & AI:** NLP, Deep Learning, Model Training & Fine-Tuning, AI Model Deployment.
- **Cloud Platforms:** AWS, Google Cloud Platform (GCP), EC2, S3, Google Analytics.
- **Database:** MySQL, FAISS, Chroma DB.
- **Security & Authentication:** OAuth 2.0, AWS Cognito.
- **AI Agents:** Autogen 360 / Lang graph
- **SEO Optimization:** Implementing effective strategies to enhance online visibility and search performance.

PROFESSIONAL EXPERIENCE

AI Engineer | Microsoft   *June 2020 – July 2021*

- Designed, developed, and deployed end-to-end **machine learning and deep learning models** for **classification, regression, and NLP**-based use cases using **Python, TensorFlow, and PyTorch.Build, train, and optimize machine learning** and **deep learning models** for classification, **prediction**, recommendation, and **automation tasks**. Automated infrastructure provisioning using **Ansible and Terraform**, improving **scalability**.
- Led **AI solution architecture** for **multiple client projects**, ensuring **scalability, security**, and performance optimization in production environments.
- Built and **fine-tuned transformer-based models** (e.g., GPT, BERT) for **text generation, summarization**, and **sentiment analysis** tasks.
- Integrated AI models with real-time systems using **REST APIs** and **microservices** (Flask, FastAPI, Node.js) in **cloud environments** (AWS/GCP/Azure).
- Applied advanced feature engineering techniques and **ML Ops practices** to **streamline model training**, versioning, and **deployment pipelines**.
- Collaborated with **cross-functional teams (Data Engineers, DevOps, Frontend)** to implement data-driven features and deploy AI capabilities into production.
- Utilized **vector databases** (e.g., **Pinecone, FAISS**) and **RAG architecture** to build **context-aware retrieval systems** and **chatbots.**
- Employed **AWS services** like **S3, Lambda, Rekognition, Amplify**, and **Cognito for scalable**, serverless **AI solutions.**
- Conducted **A/B testing** and **model evaluation** using metrics like precision, recall, **F1-score, ROC-AUC**, and **confusion matrices.**
- Delivered impactful **GenAI applications** including question **generation tools, intelligent chatbots**, and **document summarizers** tailored to industry needs

EDUCATION

Masters in Computer Science   *Dec 2024*
Christian Brothers University, Memphis, TN

Bachelors of Computer Science and Engineering   *May 2022*
KL University, India

PROJECTS

Project -1: Scorify AI
Technologies Used – Python, Flask, Axios, OpenAI, Google Vision API Integrations. Model Fine-Tuning, Google Cloud Deployment and GitHub Pages.
Links - GitHub

- **Real-Time Question Detection:** Utilized **Google Vision API** to extract **text from images** and **live camera feeds**, enabling instant question recognition for platforms like **Duolingo.**
- **AI-Driven Answer Generation:** Integrated **fine-tuned GPT-4** models to generate context-aware answers, summaries, and hints tailored to detected questions.
- **Interactive Learning Modules:** Built **Interactive Reading & Listening** modules using **SpeechRecognition and SpeechSynthesis** APIs for engaging **speaking and listening** test practice.
- **Secure User Authentication:** Implemented **OTP-based login** with **real-time countdowns** and alerts to ensure **secure and smooth onboarding**.
- **Live Feed AI Hub:** Unified all features into a **dynamic live interface** allowing **seamless transitions** between **reading, listening**, and practice modules using a responsive f**ront-end stack (HTML5, CSS3, JS, Bootstrap 5).**
- **OTP-Based Secure Login:** Implemented secure authentication via **OTP validation APIs** with **countdown timers and real-time alert mechanisms.**
- **Multimodal Learning Experience:** Supported both **text-to-speech and speech-to-text** for a fully **interactive learning loop,** enhancing user engagement and retention.
- **Tech Stack:** Built with **HTML5, CSS3, JavaScript (ES6), Bootstrap 5, Google Vision API, OpenAI GPT-4**, and browser-native speech APIs for a responsive, browser-based app.

Project -2: Zeez AI
Technologies Used - HTML, CSS, Python, JavaScript, OpenAI Integration, Model Fine-Tuning, LangChain, Flask, Axios, AWS EC2, S3 Bucket, OAuth2.0, AWS Amplify.
Links - GitHub

- **Smart Quiz Creation:** Developed **an AI-powered quiz platform** using fine-tuned **GPT-3.5 Turbo** to generate personalized, subject-specific **MCQs** from **real-time text, image, and audio inputs**.
- **Multimedia Integration:** Leveraged **Google Vision API** for **image-to-text**, **AWS Amplify for text-to-speech**, and a custom audio interface **for speech-driven interactions**.
- **Secure Authentication: Implemented user access control** and secure sign-in using **AWS Cognito**, ensuring **data integrity and user privacy.**
- **Educator Dashboard:** Enabled teachers to **monitor student quiz performance** with **analytics dashboards**, allowing **adaptive quiz assignments** based on **real-time scores.**
- **Scalable & Inclusive Design:** Designed with **multilingual and accessibility support**, built using **React, AWS services**, and **cloud-native architecture** for **seamless deployment**.
- **Secure User Authentication:** Ensured robust user authentication and **role-based access using AWS Cognito** for **students, faculty, and admins.**
- **Interactive Chatbot:** Implemented multilingual, context-aware chatbot using **GPT models** and **AWS Rekognition** for enhanced student support.
- **Tech Stack:** Developed with **React, AWS services** (Amplify, Cognito, Rekognition), **Google Cloud APIs**, **Python backend**, and **scalable cloud infrastructure**.

Project -3: LingoMate AI
Technologies Used - Python, JavaScript, Node.js, Flask, Custom Generative AI models for multilingual interaction and sentiment
Analysis, Local Python/Node.js environment with extensibility for cloud hosting.
Links - GitHub Repo

- **Multilingual Chatbot Development:** Built a **conversational AI** capable of **fluent interaction in Hindi, English**, and **code-mixed language,** catering to diverse user demographics.
- **Custom Generative AI Models:** Deployed **fine-tuned AI models** to generate **intelligent, real-time responses** using custom **APIs** and website content ingestion.
- **Sentiment-Aware Conversations**: Integrated **sentiment analysis** to detect **user emotions** and dynamically **adjust tone** and empathy in **AI responses.**
- **Context Retention & Smart Prompting:** Enabled the **chatbot** to remember user context **across conversations**, delivering a **seamless and personalized dialogue flow**.
- **Human Handoff with Admin Dashboard:** Designed an **admin interface** to allow **smooth handoff** to **human agents for** complex queries, and **backend customization** of **chatbot replies.**
- **Full-Stack Implementation:** Utilized **Python Flask** for backend logic and **Node.js** for the frontend, showcasing strong **full-stack development** capabilities**.**
- **Deployable Architecture:** Built in a **local environment** with **flexibility for cloud hosting**, making it easily extensible**, lightweight, and cost-effective**.
- **Tech Stack & Skills:** Hands-on expertise in **Python, JavaScript, Node.js, Flask**, custom **LLM integration**, **NLP, API** communication**, and web deployment** strategies.