RAMAKRISHNA REDDY VENNAM | **Phone**: +1 656-208-4865 | **Email**: vennam | **GitHub**: RamaKrishnareddy17 | **LinkedIn**: VENNAM

PROFESSIONAL SUMMARY

Junior Software Engineer with 3 years of experience designing and implementing scalable APIs and efficient backend systems. Skilled in Python, Django, Flask, and cloud platforms (AWS, Azure, GCP), with a strong focus on AI-driven solutions. Experienced in utilizing advanced tools like LangChain, Retrieval-Augmented Generation (RAG), and Docker to deliver innovative, impactful results in AI-powered projects. Passionate about system design, debugging, and enhancing user experiences through robust and scalable solutions. Adept at working collaboratively in diverse teams, demonstrating leadership, discipline, and a commitment to excellence.

EDUCATION

University of South Florida, Tampa, **Master of Science in Computer Science,** August 2023 – May 2025, **GPA**: 3.45

CORE COMPETENCIES

- **Programming Languages**: Python, C, C#, C++, Java
- **Generative AI**: Azure OpenAI, Azure AI search GPT-4, Fine-Tuning, Text Generation, Retrieval-Augmented Generation (RAG), Large Language Models (LLM), Gemini API, AutoGen
- **Machine Learning**: Supervised and Unsupervised Learning, Classification, Regression, Clustering, K-Means Clustering, KNN, Decision Trees, Ensemble Methods
- **Web Technology**: HTML, CSS, JavaScript, AngularJS, ReactJS, NextJS, NodeJS and jQuery
- **Databases**: MySQL, Azure SQL, MongoDB
- **Data Visualization**: Matplotlib, Seaborn, Power BI
- **Cloud Platforms**: AWS, Azure, Google Cloud Platform (GCP)
- **Tools**: Anaconda, Git, Docker, Version Control, Developer Tool, Microsoft Office Suite

WORK EXPERIENCE

Social AI Development Project | SuperINTRO | NOV 2023 Present

- Designed and developed scalable APIs using Python, Django Rest Framework, and Flask to power AI-driven agent functionalities for networking. Integrated tools like LangChain, Retrieval-Augmented Generation (RAG), and OpenAI technologies to enhance natural language understanding and intelligent decision-making. Optimized backend systems with MongoDB, Firebase, and Node.js to ensure high reliability, performance, and scalability. Collaborated with cross-functional teams, utilizing Docker and GCP to efficiently deploy, monitor, and scale AI systems while adhering to SLA and coding standards.

Generative AI Project | SEP 2022 NOV 2023

- Developed an AI-powered system using OpenAI and Azure AI search to extract and retrieve relevant information from databases, PDFs, and Word documents. Tuned GPT-4 models to achieve 90% accuracy in natural language processing tasks, optimizing automation and text generation. Embedded Azure AI search to enhance retrieval efficiency and deployed the system using Azure Function App for scalability and performance. Applied Quality Assurance practices to ensure the system's reliability, incorporating stakeholder feedback for continuous improvements.

Deep Learning & OCR Number Plate Detection | *SEP 2022 – NOV 2023*

- Trained a custom YOLO model for vehicle number plate detection, achieving 95% accuracy in object detection from video streams. Developed a custom OCR model to extract alphanumeric characters from detected plates, improving text recognition performance. Deployed the system as an API service, enabling real-time number plate recognition and seamless integration with other applications. Enhanced the system's efficiency and user experience through iterative improvements and precise optimization techniques.

Machine Learning Targeted Article Recommendation System | MARCH 2022 SEP 2023

- Built a recommendation engine using supervised learning algorithms to suggest articles based on user survey responses. Achieved 98% prediction accuracy through data preprocessing, hyperparameter tuning, and model optimization. Deployed the system in a production environment to deliver personalized content recommendations for an online platform. Demonstrated strong planning, documentation, and communication skills to ensure alignment with project specifications and client expectations.

Internship Experience

AI Full-Stack Engineer Intern | *SuperINTRO* | Nov 2023 Present

- Designed and developed scalable APIs and backend systems using Python, Django, Flask, MongoDB, and Firebase. Integrated advanced AI tools like LangChain, RAG, and LLMs to enhance AI-driven functionalities. Deployed solutions using Docker and GCP, ensuring scalability and reliability. Actively tested systems and developed test cases, maintaining high performance and coding standards

Python Developer Intern | *SHIASH INFO SOLUTIONS* | Nov 2020 Jan 2021

- Built a real-time data pipeline for e-commerce analytics, processing millions of records efficiently. Improved sales forecasting accuracy by 30% through actionable insights and dynamic reporting. Leveraged Python for data analysis and visualization, optimizing data collection and decision-making.

CERTIFICATIONS

- Artificial Intelligence Foundations: Machine Learning (LinkedIn)
- Microsoft Certified: Azure Fundamentals
- Data Science and Machine Learning Bootcamp with Python (Udemy)
- Workshop in Machine Learning – Pantech E-Learning

Availability: Open to internships, full-time opportunities, and collaborations, particularly in product engineering, healthcare, finance, and supply chain sectors.