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 MongoDB, Firebase, Express.js, Next.js, and Python, alongside frameworks such as Django Rest and Flask. Currently interning at SuperINTRO, I am contributing to an ambitious project to democratize Social AI agents for networking by leveraging advanced tools like LangChain, Retrieval-Augmented Generation (RAG), and Large Language Models (LLMs). Proficient in cloud platforms like AWS, Azure, and Google Cloud Platform (GCP), I specialize in building robust, scalable systems and integrating AI-driven solutions. My expertise also includes REST API development, enterprise integration, Data Integration, and production support, ensuring stability and performance optimization. Passionate about innovative technologies, I bring a strong foundation in system design, debugging, troubleshooting, and stakeholder-focused project execution, with a keen ability to assist in articulating user requirements and enhancing user experience. Additionally, I exhibit strong leadership skills, ensuring clarity in communication and a commitment to discipline and integrity. My work ethic embraces a global perspective, working collaboratively with diverse teams. Demonstrating a respectable work ethic, I am dedicated to fostering a positive corporate culture and ensuring compliance with industry standards. I possess a wide-ranging technical engineering knowledge and always seek growth opportunities within the technology platform landscape.

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*

- **Responsibilities:** Designing and developing scalable APIs, building robust backend systems, integrating advanced AI functionalities, and deploying AI-powered solutions leveraging cloud technologies and frameworks. Additionally, ensuring quality through Code Review processes and maintaining coding standards.
- **Description:** Currently contributing to an ambitious project aimed at democratizing Social AI agents for networking. The project focuses on developing impactful AI-driven solutions by harnessing cutting-edge technologies like Generative AI, NoSQL databases, and scalable system architecture.
- Designed and implemented scalable APIs using Python, Django Rest Framework, and Flask to support AI-powered agent functionalities.
- Utilized tools like LangChain and Retrieval-Augmented Generation (RAG) to enhance AI-driven decision-making and natural language understanding.
- Integrated large language models (LLMs) and OpenAI technologies to power intelligent, automated interactions for seamless user networking experiences.
- Optimized backend systems using MongoDB, Firebase, and Node.js to ensure high performance, reliability, and scalability while fulfilling service operation expectations.
- Leveraged Docker and Google Cloud Platform (GCP) for efficient deployment, monitoring, and scaling of the AI systems, adhering to SLA (Service Level Agreement) guidelines and enhancing telemetry capabilities. This includes implementing crucial infrastructure that supports correctional processes and implementation of system updates.
- Actively collaborated with cross-functional teams to align project goals, focusing on the architectural design, ensuring seamless execution of deliverables, including facilitating continuous improvement efforts and providing guidance to inspire team performance. Regularly sought feedback from customers and stakeholders to enhance the user experience.

Generative AI Project | *SEP 2022 – NOV 2023*

- **Responsibilities**: Data gathering, preprocessing, model tuning, embedding AI search capabilities, and deploying a fully functional AI system into Azure Function App.
- **Description**: Led a project aimed at building an AI-powered system to extract information from multiple sources including databases, PDFs, and Word documents. The project involved using OpenAI and Azure AI search to optimize data retrieval and relevance, followed by tuning the AI model to achieve a 90% accuracy rate before deploying it in a production environment via Azure cloud services.

- Utilized Azure OpenAI's GPT-4 for fine-tuning natural language understanding and text generation, showcasing strong coding skills and a focus on automation.
- Embedded Azure AI search to improve information retrieval efficiency from large datasets.
- Deployed the system using Azure Function App for scalability and performance, ensuring alignment with best practices in Quality Assurance, maintaining strong accountability, and processing any relevant customer feedback for future iterations.

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

- **Responsibilities**: Trained custom YOLO models, implemented custom Optical Character Recognition (OCR) models, and deployed them as an API service.
- **Description**: Developed a system to detect vehicle number plates in video streams using a custom-trained YOLO model for object detection. The project aimed to improve the speed and accuracy of recognizing objects from video feeds and extracting text from number plates.
- The YOLO model was customized and trained to detect number plates with an accuracy of 95%.
- A custom OCR model was built to extract alphanumeric characters from the detected number plates, improving the text extraction accuracy and overall system performance.
- The system was deployed as an API service, allowing easy integration with other applications that require real-time object detection, effectively supporting various medical condition assessments and enhancing the overall user experience.

Machine Learning – Targeted Article Recommendation System | *MARCH 2022 – SEP 2022*

- **Responsibilities**: Data collection, preprocessing, model training, and deployment.
- **Description**: Built a machine learning-based recommendation engine to suggest targeted articles to users based on survey responses. The system processed survey data to identify user preferences and used supervised learning algorithms to recommend content with high relevance. This project involved a rigorous planning and documentation phase to ensure compliance with project specifications and customer expectations.
- Achieved 98% accuracy in recommendation predictions through hyperparameter tuning and model optimization.
- Deployed the system in production to serve personalized content recommendations for an online platform, ensuring strong communication and organizational skills were applied.

INTERNSHIP EXPERIENCE

AI Full-Stack Engineer Intern
SuperINTRO | November 2023 – Present
Project: Social AI Agent Development for Scalable Networking
Responsibilities: Designed and developed scalable APIs and backend systems using Python, Django, Flask, MongoDB, and Firebase. Integrated advanced AI tools like LangChain, Retrieval-Augmented Generation (RAG), and Large Language Models (LLMs) to enhance AI-driven functionalities while adhering to coding standards. Actively involved in testing and developing test cases to ensure the high quality and performance of the systems through a disciplined approach toward feedback integration.
Description: Contributed to the development of AI-powered Social Agents for networking by building robust backend systems and scalable APIs. Leveraged state-of-the-art tools and frameworks to improve system efficiency and reliability. Successfully deployed the solution using Docker and Google Cloud Platform (GCP), enabling seamless operations and scalability while supporting continuous improvement initiatives, and maintaining adherence to compliance standards.

Python Developer Intern
SHIASH INFO SOLUTIONS | November 2020 – January 2021
Project: Real-time Data Pipeline for E-commerce Analytics
Responsibilities: Designed and implemented a scalable data pipeline for processing real-time e-commerce data, enabling dynamic reporting and visualization. Conducted data analysis to refine data collection processes and improve accuracy, addressing potential technical issues through robust planning.
Description: Developed a real-time data processing system real-time, providing actionable insights and improving decision-making for business strategies. Throughout this process, I demonstrated a strong using Python, integrating it with a visualization tool for e-commerce analytics. This system improved the forecasting accuracy of sales by 30% by enabling rapid access to live data and analysis. Leveraged Python's robust data handling libraries to build a pipeline that processed millions of records in near commitment to corrective practices and sought guidance from team leads to enhance my technical solution skills while adhering to a creative approach in problem-solving. My contributions were aligned with the company's color and branding strategies to ensure a cohesive visual impact.

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.