V S MONISH KUMAR

Software Engineer — AI/ML & Full Stack

kumarvsmonish@gmail.com | 8688470533 | Chennai, India | LinkedIn: linkedin.com/in/v-s-monish-kumar | GitHub: github.com/VSMONISHKUMAR | Portfolio: vsmonishkumarportfolio.netlify.app

PROFESSIONAL SUMMARY

AI/ML Engineer with production experience building LLMs, RAG pipelines, and computer vision systems. Fine-tuned a 37B parameter model and developed AI agents on Vertex AI. Proficient in Python, FastAPI, Flask, React Native, AWS, and GCP with end-to-end deployments across mobile and backend systems. Seeking AI/ML Engineer or Full Stack Software Engineer roles where deep learning meets real-world product impact.

TECHNICAL SKILLS

AI/ML & NLP: LLMs, RAG Pipelines, NLP, Generative AI, Stable Diffusion, U-Net, SAM, YOLO
Programming Languages: Python, Java, SQL
Frameworks & APIs: Flask, FastAPI, REST APIs, React Native
Cloud & DevOps: AWS, GCP, Git, JIRA

WORK EXPERIENCE

Software Engineer | CLUSTREX DATA PRIVATE LIMITED | *July 2025 – Present*

- Developed two Google Chat bots on Vertex AI: a Policy Assistant Bot (RAG-based retrieval from structured corpora) and an Office Management Bot automating attendance, leave tracking, and payroll summaries for employees.
- Designed and deployed scalable REST APIs using Flask and FastAPI with OTP-based authentication and Twilio integration for applications.
- Built cross-platform mobile applications using React Native with responsive UI and seamless backend API integration.
- Deployed and managed production services on AWS and GCP; implemented DevOps best practices including automated backups, database optimization, monitoring, and secure code reviews.
- Performed end-to-end testing including manual UI/UX validation and Playwright automation; tracked bugs and enhancements via Git and JIRA.

Software Engineer Intern | CLUSTREX DATA PRIVATE LIMITED | *Jan 2025 – June 2025*

- Fine-tuned a 37B parameter LLM-based resume parser using RAG pipelines, reducing manual parsing effort by ~80% and improving extraction accuracy across 500+ resumes.
- Built AI-powered computer vision systems using Stable Diffusion, U-Net, and SAM for high-precision image segmentation, processing 1000+ images across various tasks.

Research Intern | SSN College of Engineering | *May 2024 – Jul 2024*

- Built an IoT-based pedestrian detection system using YOLO and Arduino for real-time vehicle safety, achieving reliable detection performance in live traffic scenarios.

PROJECTS

IBM Watson Assistant for Facebook – AI-powered chatbot integrated with Facebook Messenger for automated customer interaction and support workflows. **Tech: IBM Watson Assistant, Facebook Messenger API**
Emotion Detection using Computer Vision – Real-time facial emotion recognition from live video streams using deep learning, achieving 89% accuracy across 5 emotion classes. **Tech: Python , CNN, TensorFlow**

AI Supported Traffic Management System – Built a prototype AI-powered Traffic Management System using a YOLO model trained on custom data for real-time violation detection, integrated number plate recognition (ANPR), and implemented Twilio API for sending OTP-based alerts to offenders. **Tech: Python, YOLO, Twilio**

AI-Powered RAG Chatbot using Vertex AI – Built a RAG chatbot that extracts documents (PDFs/docs) from Google Drive, converts them into a vector database stored in a RAG corpus, and answers queries strictly based on that content.**Tech: Google Script,Python ,Vertex AI**

EDUCATION

Bachelor of Engineering (B.E.) — Computer Science and Engineering | *Nov 2021 – Apr 2025*
Prathyusha Engineering College, Chennai | **CGPA: 9.08 / 10**

CERTIFICATIONS

- AWS Academy Cloud Foundations
- AWS Academy Machine Learning for Natural Language Processing
- Data Analytics (ICT Academy)
- Python for Data Science (IBM)