PRANEETH SAI CHUNCHU

+1 (659) 346-6106 | praneethsaichunchu@gmail.com | Linkedin | Portfolio

EXPERIENCE

AI Intern & Co-Founding Team Member - LifeSizeAgents.ai, USA **Mar 2025 - Present**

- Architected and maintained state-of-the-art **LLM** platform using Python backend applications (**RESTful APIs and queue-based systems**) with LangChain, CrewAI, and vector databases, serving 4000+ queries/day with 80% performance improvement.
- Built and deployed **LLM servers** and model management systems for production environments, implementing prompt engineering strategies and real-world LLM use cases for interactive 3D avatar applications
- Developed RAG pipelines with embeddings, semantic search, and knowledge graphs, increasing contextual recommendation accuracy by 75% for 200+ users.
- Optimized multi-agent orchestration, prompt tuning, and embedding models for scalable, low-latency AI, reducing response time from 5s to 1s.

ML Engineer - KaarTech Solutions, India **Jan 2023 – Dec 2023**

- Built and maintained production ML platform with Python backend applications, implementing RESTful APIs and queue-based processing systems using MLflow, Docker, Kubernetes.
- **Deployed and tested ML models** on cloud infrastructure, developing stack knowledge for troubleshooting distributed systems across multiple environments, reducing deployment cycles by 50%.
- Applied machine learning and deep learning theory to build scalable recommendation and personalization systems using regression, classification, clustering, and ensemble methods.
- **Enhanced performance through code improvements** in Python and Bash scripting. Applied Apache Spark ecosystem knowledge with hands-on experience in YARN-managed cluster environments.
- Designed real-time monitoring and anomaly detection using **Prometheus & Grafana**, increasing production reliability by 40%.

Research Assistant - GVPCE, India **June 2022 – Nov 2022**

- Designed CNN-based feature extraction pipelines for image and video data, supporting automated attendance and recognition systems.
- Leveraged **transfer learning (ResNet, EfficientNet)** and optimization techniques (pruning, quantization, TensorRT) to enable **real-time inference** on large video datasets.
- Applied advanced preprocessing to handle noisy or incomplete data, ensuring robust feature extraction under varying conditions.
- Leveraged transfer learning, model distillation, and inference acceleration to achieve real-time performance on large datasets.

PROJECTS

End-to-End MLOps Pipeline for Predictive Maintenance of Turbines | Github **June 2025 – Aug 2025**

- Developed modular 6-stage MLOps pipeline with Airflow from ETL to prediction.
- Integrated MLflow for experiment tracking, model registry, automated hyperparameter tuning, and model versioning.
- Containerized with Docker, automated CI/CD via GitHub Actions, and served models through FastAPI.
- Monitored data, drift, and model performance with Grafana and Prometheus.

Attendance Management System Using CNN and HOG | Github **May 2024 – July 2024**

- Built a real-time attendance system with Django, OpenCV, and CNN/HOG, using 128-D face encodings for automated check-ins.
- Containerized with Docker and deployed on AWS EKS using Kubernetes and Terraform for scalable orchestration.
- Developed an end-to-end CI/CD pipeline with IaC, face capture, and an admin dashboard for attendance analytics.

CERTIFICATIONS &ACHIEVEMENTS

1. AWS Certified ML Engineer - Assosiate
2. Databricks Certified Associate Data Engineer
3. Complete MLOps Bootcamp by Udemy
4. ML in Production by Andrew Ng
5. AWS Certified Cloud Practitioner - Foundational
6. Solved 350+ Python (DSA) problems and 70+ SQL problems on LeetCode, earning 7+ LeetCode badges. Leetcode Profile

SKILLS

Software Development & Cloud: Knowledge Graphs, Embeddings, Python, C, JavaScript, Django, Flask, FastAPI, React.js, React Three.js, AWS, Databricks, Docker, Kubernetes, Terraform, Git, Jenkins, Airflow, Lambda, Glue, MySQL, PostgreSQL, MongoDB.

Machine Learning & AI: TensorFlow, PyTorch, Scikit-learn, Hugging Face, OpenAI APIs, BERT, GPT, Dlib, OpenCV, MLflow, DVC, Prometheus, Grafana, Agile/Scrum, AI Interoperability (MCP, A2A) awareness.

EDUCATION

Master of Science, Computer Science | GPA: 3.87 / 4.0 **Jan 2024 – 2025**

University of Alabama at Birmingham

Bachelor of Technology, Electronics and Communication Engineering | GPA: 8.8 / 10 **June 2019 – April 2023**

Gayatri Vidya Parishad College of Engineering, Visakhapatnam, India

Relevant Course Work: Data Structures and Algorithms, Database Systems, Machine Learning, Deep Learning, Cloud Computing, Data Mining, Advanced Web Applications, Advance Algorithms and Apps.