PRANEETH SAI CHUNCHU

Dallas, TX, 75202 | +1 (659) 346-6106 | praneethsaichunchu@gmail.com | Linkedin | GitHub | LeetCode

PROFESSIONAL SUMMARY

AI Engineer with **2+ years** of experience building production-grade LLM applications, RAG pipelines, and scalable ML systems. Specialized in deploying end-to-end AI solutions using PyTorch, LangChain, and cloud platforms (AWS, GCP) that process 100K+ daily requests. Proven expertise in fine-tuning LLMs, architecting multi-agent systems, and implementing MLOps pipelines with 93% faster deployment cycles. AWS Certified ML Engineer with strong foundation in distributed training, model optimization, and delivering measurable business impact through AI-driven solutions.

SKILLS

Machine Learning & AI: Predictive Modeling, NLP, Computer Vision, Time Series, LLMs (Fine-tuning, RAG), AI Agents
Deep Learning: PyTorch, TensorFlow, Transformers, Hugging Face, CNN, RNN, LSTM, Generative AI (VAE, GPT-based)
MLOps: Docker, Kubernetes, Airflow, MLflow, CI/CD (GitHub Actions), Terraform (Certified Associate), FastAPI
Cloud Platforms: AWS Certified ML Engineer (SageMaker, EKS, Lambda), GCP (Vertex AI, GKE)
Data Engineering & Observability: Databricks Certified Data Engineer, Python, SQL, Apache Spark, Prometheus, Grafana

PROFESSIONAL EXPERIENCE

Research Assistant, UAB | Birmingham, AL **Sep 2025 – Present**

- Architected NeuroStage, a modular Python pipeline for DICOM preprocessing and metadata extraction, reducing manual processing time by 80% while maintaining research-grade accuracy.
- Benchmarked multiple LLMs (GPT-4, Claude, Gemini) for automated DICOM metadata extraction and classification, achieving 92% accuracy in production deployment and reducing radiologist review time by 65%.
- Engineered multi-agent workflows using LangChain for automated brain imaging data conversion, classification, and validation across 3-stage pipelines, processing 10K+ MRI scans with 94% accuracy.

AI/ML Engineer & Co-Founding Team Member, LifeSizeAgent.ai | Remote, TX **Mar 2025 – Aug 2025**

- Architected and deployed RAG (Retrieval-Augmented Generation) systems using LangChain and vector databases (Pinecone, FAISS) for semantic search and context retrieval, processing 100K+ daily requests on AWS Infrastructure.
- Engineered production LLM applications with multi-agent frameworks (CrewAI, LangChain) and prompt engineering patterns, implementing task automation workflows that improved processing efficiency by 35%.
- Optimized LLM inference pipelines through quantization (INT8) and LoRA fine-tuning on NVIDIA GPUs, reducing latency by 81% and GPU memory usage by 40% while maintaining model accuracy.

ML Engineer, Kaar InfoTech Solutions | India **Jan 2022 – Dec 2023**

- Architected end-to-end MLOps platform integrating MLflow experiment tracking, automated CI/CD pipelines, and performance monitoring, reducing model deployment time by 93% and enabling rapid iteration of 30+ production models.
- Built distributed ML training infrastructure on NVIDIA GPU clusters using PyTorch and TensorFlow, implementing mixed-precision training (FP16/FP32) and achieving 3.2x throughput improvement while processing 2PB+ training data.
- Deployed production ML models with Docker containerization and comprehensive A/B testing infrastructure, achieving 98% model performance retention and serving 50M+ daily predictions through optimized Apache Spark ETL pipelines.

EDUCATION

Master of Science in Computer Science **Jan 2024 – Dec 2025**
University of Alabama at Birmingham, Birmingham, AL
Coursework: Machine Learning, Deep Learning, Data Mining, Advance Algorithms and Apps, Computer Vision

PROJECTS

End-to-End MLOps Pipeline for Predictive Maintenance of Turbines

- Reduced unexpected turbine failures by building an automated MLOps pipeline (Airflow, MLflow, Docker) that predicts equipment lifespan with 90% accuracy, processing 100+ predictions/second with 99.9% uptime.
- Accelerated ML deployment from weeks to hours through automated CI/CD (GitHub Actions), real-time drift detection (Prometheus/Grafana), and containerized FastAPI model serving at 200ms average latency.

Transformer-Based Neural Machine Translation

- Enabled cost-effective multilingual communication by building English-to-German translation system that processes translations 40% faster and reduces model costs by 2x through custom architecture optimization, supporting scalable deployment for international applications.

Multi-Agent AI Music Generation System

- Delivered fully autonomous music production system automating 100% of business operations (content creation, billing, marketing) using multi-agent architecture (LangChain, Gemini 2.0), eliminating manual workflows and enabling scalable subscriber growth with automated daily music generation and distribution.