PRANEETH SAI CHUNCHU

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

SUMMARY

AI Engineer with **2+ years** of experience building production-grade LLM applications and AI agent systems using Python, LangChain, and cloud platforms. Specialized in developing **RAG pipelines**, prompt engineering, and deploying scalable AI solutions on AWS and Azure. Proven track record in fine-tuning LLMs (GPT-4, Claude, Gemini), implementing vector databases, and delivering end-to-end AI workflows in Agile environments with robust monitoring and CI/CD integration.

WORK EXPERIENCE

Research Assistant | UAB EN4D2 Lab **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.
- Collaborated with radiologists to translate clinical requirements into technical specifications, delivering intuitive data-driven interfaces that improved research workflow efficiency by 40% (manuscript in preparation).

AI / ML Engineer & Co-Founding Team Member | LifeSizeAgents.ai, USA **Mar 2025 – Sep 2025**

- Architected and deployed **RAG (Retrieval-Augmented Generation)** systems using LangChain and vector databases (Pinecone, ChromaDB) for semantic search and context retrieval, processing 100K+ daily requests on **AWS Bedrock.**
- Engineered LLM-powered applications with prompt engineering and function calling patterns, implementing AI agents for task automation and workflow orchestration with 35% improvement in processing efficiency.
- Optimized LLM inference pipelines through quantization (INT8) and **LoRA fine-tuning** on NVIDIA GPUs, reducing latency by 81% and GPU memory usage by 40%.
- Built production AI systems with FastAPI and Docker, implementing **CI/CD pipelines** via GitHub Actions with automated monitoring using Prometheus, reducing infrastructure costs by 40%.

ML Engineer | KaarTech 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 95% on-time delivery through systematic testing and monitoring.
- Engineered scalable data pipelines with **Apache Spark** for distributed feature engineering, optimizing **ETL workflows** for efficient model training and inference serving 50M+ daily predictions.
- Collaborated with cross-functional teams to translate business requirements into technical solutions, delivering ROI driven AI applications with clear performance metrics and economic impact analysis.

PROJECTS

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

- Implemented modular 6-stage MLOps pipeline with Airflow from ETL to prediction.
- Created NGBoost and Random Forest models with automated feature engineering by TSFresh to improve maintenance forecasting.
- Iterable MLflow integration, including support for tracking, model registries, hyperparameters, and versions.
- Containerized with Docker, automated CI/CD pipeline with GitHub Actions, and served models with FastAPI.
- Monitored Data drift and model performance with Grafana and Prometheus.

Transformer-Based Neural Machine Translation | Github

- Created a Transformer encoder-decoder model with multi-head attention and positional encoding in PyTorch for English to German translation.
- Optimized training with FP16 precision, AdamW, gradient clipping, and LR scheduling for faster convergence.
- Fine-tuned T5-Small for comparative evaluation, analyzing BLEU and optimizing decoder KV-cache utilization.

Multi-Agent AI Music Generation System | Github

- Developed multi-agent AI systems with **LangChain** and Google Gemini 2.0 technologies, focusing on automated music production, billing systems, and social marketing.
- Integrated HuggingFace ACE-Step API with Flask backend for mood-based music generation across 6+ emotional categories.
- Music generation, posting, and charging, all automated with the help of a Python scheduler.

Custom Transformer-Based Storytelling Chatbot | Github

- Created a 3-layer transformer decoder with masked attention and positional encoding for TinyStories text generation.
- Training was optimized with DDP, FP16, AdamW, and gradient clipping to ensure convergence.
- Used supervised fine-tuning and KV-cache inference to achieve fluent, real-time story generation.

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**

TECHNICAL SKILLS

Programming & Scripting: Python, SQL, PySpark, Prompt Engineering, Shell Scripting, Github.
Machine Learning & AI: Linear Regression, Random Forest, Logistic Regression, XBoost, Classification Models, Hyper-parameter Tuning, Fine-tuned LLMs, Generative Al, Time Series Forecasting, PyTorch, TensorFlow, Scikit-Learn, ViTs, VLMs.
Data Engineering & ETL: AWS Glue, SQL (ETL), Apache Airflow, S3, MLFlow, Docker, MongoDB Atlas, Data Ingestion, Data Transformation, Feature Engineering.
Cloud & Deployment: AWS, Docker, MCP Server, Model Deployment, Production Integration, A/B Testing Infrastructure, Containerization, End-to-End MI. Workflow Automation.
Data Visualization & BI: Tableau, Power BI, Excel, Real-Time Daslab ards, KPI Monitoring.
NLP & Text Analytics: NLP, NLTK, spaCy, Transformers, Word Embeddings (Word2Vec, GloVe, BERT), Theme Extraction, Complaint Classification, NER, Prompt Workflows.

EDUCATION

Master of Science, Computer Science | GPA: 3.87 / 4.0 **Jan 2024 – Dec 2025**
University of Alabama at Birmingham, USA
Relevant Course Work: Data Structures and Algorithms, Database Systems, Machine Learning, Deep Learning, Cloud Computing, Data Mining, Advanced Web Applications, Advance Algorithms and Apps.
Bachelor of Technology, Electronics and Communication Engineering | GPA: 8.8 / 10.0
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.