PRANEETH SAI CHUNCHU

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

SUMMARY

Machine Learning Engineer with **2+ years** of production experience building GenAI/AI solutions, scalable data pipelines, and MLOps systems on cloud platforms. Skilled in NLP, computer vision, document processing, and LLM deployment, with a strong focus on data engineering, monitoring, and production-grade MLOps. Proven ability to deliver end-to-end ML solutions on Azure and AWS.

WORK EXPERIENCE

Research Assistant | UAB EN4D2 Lab **Sep 2025 – Present**

- Architected and developed NeuroStage, an **AI-powered neuroimaging platform** with modular Python design to streamline brain MRI preprocessing and quality validation.
- Integrated multiple **LLMs (GPT, Claude, Gemini)** for automated DICOM metadata extraction and classification, benchmarking model performance to identify optimal solutions for production deployment with 92% accuracy.
- Built intelligent agentic workflows that automatically convert, classify, and validate brain imaging data through multi-stage pipelines, reducing manual analysis time by 80% while maintaining research-grade quality standards.
- Partnered with radiologists and researchers to define AI features, translate clinical needs into technical specs, and deliver intuitive, data-driven user experiences (manuscript in preparation).

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

- Designed and deployed generative AI technologies including multimodal **LLM agents** using LangChain, implementing autonomous reasoning systems for production applications serving 100K+ daily interactions.
- Developed large-scale ML training pipelines using PyTorch with distributed training across **GPU clusters**, implementing memory-efficient strategies and mixed-precision optimization for sub-200ms inference.
- Conducted experimentation with foundation model architectures, implementing **post-training** techniques including fine-tuning, alignment, and distillation for domain-specific applications.
- Built scalable data processing infrastructure with automated **ETL pipelines**, implementing data curation and quality validation for training generative AI models at scale.

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

- Built distributed training infrastructure for 7B+ parameter models using DeepSpeed ZeRO-3 across **32-GPU clusters**, achieving 88% scaling efficiency and 4.2x training speedup.
- Optimized models for production through **INT8 quantization**, 60% pruning, and **knowledge distillation**, achieving 5.8x inference speedup while serving 50M+ daily predictions with <1% accuracy loss.
- Developed custom **CUDA/Triton kernels** for efficient attention mechanisms, enabling 16K context windows with 70% memory reduction and 3.2x throughput in production pipelines.
- Implemented **MLOps** platform with automated **CI/CD**, experiment tracking, and A/B testing for 25+ production models, reducing deployment cycles from 2 weeks to 18 hours.
- Productionized research models through distributed training optimizations and performance profiling, reducing infrastructure costs by 65% while maintaining model quality and research velocity.

PROJECTS

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

- Developed modular 6-stage MLOps pipeline with Airflow from ETL to prediction.
- Developed NGBoost and Random Forest models with automated feature engineering via TSFresh to enhance maintenance forecasting.
- 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.

Transformer-Based Neural Machine Translation | Github

- Built a Transformer encoder–decoder in PyTorch with multi-head attention and positional encoding for English-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

- Built multi-agent AI system using **LangChain** and Google Gemini 2.0 for autonomous music generation, billing management, and social media marketing.
- Integrated HuggingFace ACE-Step API with Flask backend for mood-based music generation across 6+ emotional categories.
- Automated music generation, social media posting, and billing using Python scheduler for full system orchestration.

Custom Transformer-Based Storytelling Chatbot | Github

- Built a 3-layer Transformer decoder with masked attention and positional encoding for TinyStories text generation.
- Optimized training with DDP, FP16, AdamW, and gradient clipping for stable convergence.
- Applied supervised fine-tuning and KV-cache inference for 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 AI, 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.