Sree Sai Preetham Nandamuri

preetham.vtf27@gmail.com +1(346) 430-3560 Portfolio LinkedIn GitHub

EDUCATION

University of Houston **Houston, USA**
Master of Science in Engineering Data Science • GPA: 4/4 *May 2025*

EXPERIENCE

Visual Technologies **Dallas, USA**
Data Scientist - Generative AI *July 2025 - Present*

- **Fine-tuned** domain-specific **LLaMA 2** for text summarization with memory-efficient **PEFT** techniques (**LoRA**, **QLoRA**) on company datasets, achieving 15% gain in relevance scores during internal testing.
- Led the design and optimization of a Retrieval-Augmented Generation (**RAG**) pipeline using **FAISS** for vector indexing and **OpenAI**'s embeddings API, achieving a 25% improvement in retrieval speed for internal knowledge queries.
- Collaborated on **LangGraph**-driven multi-agent system leveraging **OpenAI** function-calling, Model Context Protocol (**MCP**), and AWS Cost Explorer APIs to detect idle workloads, predict usage, and trigger scale-downs, reducing cloud costs by $30K.
- Boosted GPT-4 and Mistral efficiency by 20% via LangChain-based **prompt engineering** and dynamic context templating.

Pal AI **Dover, USA**
AI/ML Intern *September 2025 - Present*

- Architected transformer-driven semantic recommendation pipeline leveraging **e5-large-v2** sentence embeddings and multi-feature fusion, achieving 25% gain in match precision via cosine-similarity retrieval.
- Developed an **LLM-as-judge** re-ranking module achieving 83% alignment with counselor evaluations; integrated model feedback loops to refine ranking scores and produce context-aware fit summaries for each recommendation.
- Built **RESTful APIs** with **FastAPI** to serve embedding-driven semantic retrieval and LLM-based re-ranking, using **PostgreSQL** (pgvector) for vector similarity search and Redis for caching, achieving sub-2s end-to-end latency.

Yara Digital Farming **Bengaluru, India**
Associate Data Scientist - DevOps *September 2022 - July 2023*

- Automated classification of 5,000+ agricultural PDFs into 20 categories using fine-tuned **BERT** with custom **spaCy/NLTK** tokenization pipeline, achieving 92% accuracy and reducing manual sorting time by 89%.
- Deployed **FastAPI** classification system on **AWS EKS** with API Gateway and **SageMaker** inference with auto-scaling, cutting latency from 3s to 1.2s and implementing CloudWatch/SNS/Lambda monitoring for rapid anomaly detection.
- Reduced data collection costs by $50K+ and boosted classification accuracy by 12% by fine-tuning pre-trained **DDPM**-based synthetic crop imagery generator using **PyTorch,** creating 25,000+ high-fidelity images.
- Accelerated synthetic data generation by 3x by containerizing diffusion pipeline with **Docker** and **TorchServe** on AWS SageMaker, enabling **ECR**-managed scaling across 15+ crop varieties.

Associate Engineer - Data Science *August 2021 - August 2022*

- Developed a stacked ensemble for crop yield prediction combining **XGBoost** and **Random Forest**, reducing MAE by 18% and shortening training time by 45% via **Bayesian** hyperparameter optimization.
- Improved fertilizer dosage by 22% without yield loss across 4+ crops using an **ANN** on 15+ soil and climate features with cross-validation, hyperparameter tuning, and **MLflow** for model tracking.
- Optimized query execution time by 30% through advanced **SQL** joins, and aggregations on 5M+ agricultural records, enabling real-time insights and accelerating daily crop monitoring reports.
- Applied **PCA** and **t-SNE** for dimensionality reduction, cutting feature computation costs by 30% on large agronomic datasets.

Movidu (in partnership with IIT Bombay) **Mumbai, India**
Data Science Intern *August 2020 - February 2021*

- Designed a real-time COVID-19 safety compliance system using transfer learning with **ResNet-50** and custom CNN layers, achieving 94% accuracy in face mask detection.
- Delivered ML models with 90%+ accuracy, enabling personalized learning through automated content recommendation.

SKILLS

Languages: Python (NumPy, Pandas, Matplotlib, Seaborn, Scikit, PyTorch, TensorFlow, Keras), TypeScript, SQL, R, C
AI & Machine Learning: Deep Learning, NLP (Hugging Face, spaCy, NLTK), Computer Vision (OpenCV), Reinforcement Learning, Statistical Modeling, A/B Testing, MLflow
Generative AI & LLMOps: LangChain, RAG, LangGraph, LlamaIndex, AI Agents, Pinecone, AWS Bedrock, LLMs (OpenAI, Llama), Diffusion & GAN Models, Prompt Engineering, PEFT (LoRA/QLoRA), vLLM, Vertex AI, AutoGen, CrewAI
Data Engineering & Analytics: Data Mining, Big Data (PySpark, Hadoop), Data Visualization (Power BI, Tableau), Data Structures & Algorithms, Microsoft Excel, Streamlit, Airflow, RESTful APIs
Cloud, DevOps & Platforms: AWS, MLOps, Docker, Kubernetes, Terraform, CI/CD, Git

ACADEMIC PROJECTS

AI-Powered Social Media Caption Generation Engine GitHub

- Built an AI-powered caption generator achieving 0.75+ **CLIP** relevance on 100+ images by integrating **BLIP** for image description, **LangChain-**orchestrated Mistral-7B captioning, and CLIP for similarity ranking.
- Deployed the system using **Streamlit** with Distinct-2 (>0.95) and Self-BLEU (<0.05) for diversity optimization.

BertNet - Sales Prediction using Product Images and Descriptions GitHub

- Engineered a multimodal sales prediction system using **EfficientNetB5** for image feature extraction and **BERT** for text embeddings, achieving 75% accuracy, with a hybrid model handling missing data and class imbalance.
- Optimized feature fusions (text: 83%, image: 71%) using transfer learning and fine-tuning of pre-trained models.

MetroPT3 Failure Prediction GitHub

- Attained 100% accuracy in failure prediction using bidirectional feature elimination, and **LASSO-**driven feature selection.
- Enhanced prediction efficiency by 20% and cut maintenance downtime by 45% by leveraging **Extreme Learning Machine**, **XGBoost**, and Neural Networks, integrating them into a robust ensemble model to predict failures.