Aryan Sanjay Patil

9344519501 | New York | arpatil@cs.stonybrook.edu | Portfolio | GitHub | LinkedIn

Education

Stony Brook University — Stony Brook, NY
Master of Science (MS) in Computer Science GPA: 3.85/4.0 — *Aug 2023 – May 2025*
- Courses: Machine Learning, Natural Language Processing, Data Science, Probability and Statistics, Visualization.

University of Mumbai — Mumbai, India
Bachelor of Engineering (BE) in Computer Engineering GPA: 9.7/10.0 — *Aug 2019 – Aug 2023*
- Courses: Artificial Intelligence, Deep Learning, Big Data Analytics, Blockchain, Cloud Computing.

Skills

Programming and tools: Python, C++, SQL, R, C#, React, Docker, Git, Linux, CI/CD, Kubernetes
Frameworks: PyTorch, TensorFlow, Hugging Face, LangChain, OpenCV, ONNX, PySpark, CUDA, MLflow, FastAPI, Flask
AI Techniques: Transformers, LLMs, VLMs, GANs, VAEs, RAGs, Diffusion Models, LoRA, Prompt Engineering, RL, MLOps
Cloud and Data: AWS (EC2, SageMaker, S3), Azure, GCP, MySQL, PostgreSQL, Pinecone, FAISS, Tableau, Power BI

Experience

Research Project Assistant — *Aug 2024 – Dec 2024*
Stony Brook University - Advisor: Prof. Xiaojun Bi — *Stony Brook, NY*
- Deployed a real-time emotion detection system for an e-learning platform on **AWS EC2** with **FastAPI**, using **multimodal RAG** to power adaptive difficulty adjustment (boosted engagement by 30%).
- Boosted CNN3D-ConvLSTM2D-based emotion recognition accuracy to **95%** with **23ms detection time** using **GAN**-augmented optical flow and linear algebra optimizations (**32% faster inference**).
- Integrated **SOTA** architectures (ViT, DeTR, UNet) for user tracking, improving real-world classification by 5%.

Machine Learning Research Assistant — *Jun 2024 – Aug 2024*
Brookhaven National Laboratory - High Energy Nuclear Physics | Publication in Progress — *Upton, NY*
- Led **predictive modeling** for BK Evolution and Heavy-Ion Collisions using 3D data, reducing analysis time from **1 hour to 2 seconds** by replacing lab experiments with GPU-accelerated predictions.
- Developed a Python module using RandomForest Regressor, reducing error by **80%** (1e-4 to 2e-5) and optimizing model size/inference time by **50%** via **quantization (FP32 to FP8)**, enabling efficient analysis.
- Engineered 12 transformers with Gaussian embeddings for particle physics, reducing prediction noise to 1e-5 (10x lower than RNNs) and accelerating training and inference by 40% via CUDA optimized **distributed clusters**.

Machine Learning Engineer Intern — *Oct 2022 – May 2023*
Dharmanandan Techno Projects Pvt Ltd — *Mumbai, India*
- Innovated a custom **CNN3D + temporal attention** model for real-time human activity recognition in CCTV (**93% accuracy**, **25% faster training**), containerized with **Docker** and deployed on **AWS** with **MLFlow**-based model tracking and optimization (**50ms inference latency**).

Projects

Reddit Summarizer Application | *Read Paper* | *Flask, Multimodal, AWS SageMaker, Boto3*
- Prototyped a Real-Time Reddit Web App for Summarization and Visualization, powered by fine-tuning LLMs (**BART-large, BERT, RoBERTa, XLNet**), reducing summarization time by **30%**, with Flask backend.
- Extracted Reddit data using PRAW API, performed sentiment analysis with Mistral 7B and **LLaMA 7B**, and integrated **sarcasm detection**, improving BART summarization accuracy by **11%**.
- Implemented an **End-to-End CI/CD NLP Pipeline**, combining data preprocessing, statistical analysis, model fine-tuning, and deployment on **AWS SageMaker**, reducing deployment time by **40%**.

RAG-based Recommendation System | *Vector Databases, Full-Stack Development*
- Formulated a RAG pipeline using **React, FastAPI**, **LangChain**, and **Claude**, enabling natural language queries with **85% retrieval relevance** and context-aware recommendations.
- Optimized **FAISS-based similarity search** to support follow-up queries with **150ms latency** across 10k+ entries, scaling to production-ready workloads.

AI Applications | *GitLab, LangChain, Flask, React, Vector Databases*
- **LocalityAI**: Created a real-time ad-strategy engine using **RAG + GPT-4o**, reducing ad latency by 40%.
- **AI Sports Analytics**: Improvised an interactive football tactic generator using **LangChain**-orchestrated **GPT-NeoX** and structured embeddings, achieving **90% accuracy** via an interactive **Plotly dashboard**.
- **MediChat AI**: **Q-LoRA**-trained **T5** model for healthcare (perplexity=18), running on an AWS EC2 instance.

Open Source Contributions

Published ML/GenAI modules for LLM evaluation on PyPi with more than **3k downloads**.
Managed 32 repositories (Transformers, RL, ML) on GitHub with more than **100 stars**.
Merged 3 Pull Requests, including Preswald (2.5k stars), with work featured in the company's blog.