Aryan Sanjay Patil

9344519501 | New York | ✉ arpatil@cs.stonybrook.edu | 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: Reinforcement Learning, Natural Language Processing, Data Science, Probability and Statistics.

University of Mumbai Mumbai, India
Bachelor of Engineering (BE) in Computer Engineering GPA: 9.7/10.0 *Aug 2019 – Aug 2023*

- Courses: Data Structures, Computer Architecture, Operating Systems, Big Data Analytics, Cloud Computing.

Skills

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

Experience

AI Engineer Intern *Mar 2025 – Present*
SteadFast AI *Remote*

- Built and deployed LLM-powered anomaly detection agents for cybersecurity pipelines using **Claude** and privacy-aware RAG systems, supporting enterprise-scale monitoring.
- Designed production-scale API pipelines integrating self-supervised RAG modules (Pinecone) for real-time anomaly detection, improving detection rates by **20%** and reduce False Negatives by **65%**.

Research Project Assistant *Aug 2024 – Dec 2024*
Stony Brook University *Stony Brook, NY*

- Deployed a real-time emotion detection system for an e-learning platform on **AWS EC2** with **FastAPI**, using **multimodal RAG** to boost user engagement by **30%**.
- Enhanced 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 - Nuclear Physics | Publication in Progress *Upton, NY*

- Led GPU-accelerated predictive modeling for Heavy-Ion Collisions on **3D data**, scaling inference across **NERSC** to cut analysis time from **1 hour to 2 seconds** and replace physical lab experiments.
- Developed a Python module using RandomForest Regressor, reducing error by **80%** (1e-4 to 2e-5) and optimizing model inference time by **50%** via **quantization** (FP32 to FP8), enabling efficient analysis.
- Finetuned 12 transformers with Gaussian embeddings for particle physics, reducing prediction noise to 1e-5 (**10x lower than RNNs**) and accelerating training by 40% via **CUDA** optimized **distributed clusters**.

Machine Learning Engineer Intern *Oct 2022 – May 2023*
Dharmanandan Techno Projects Pvt Ltd *Mumbai, India*

- Innovated a CNN3D variant with temporal attention for HAR in CCTV (**93% accuracy, 25% faster**), deployed via Docker on AWS using MLFlow for model deployment, monitoring, and optimization (**50ms latency**).

Projects

Reddit Analyzer App (Live Demo) | *Read Paper* | *Multimodal, Boto3, Full-stack*

- Built a real-time Reddit summarization app by fine-tuning **BART and Mistral 7B** and incorporated sarcasm detection (**LLaMA 7B**) to reduce hallucinations by **11%** and improve generalization.
- Formulated an end-to-end CI/CD NLP pipeline on **AWS SageMaker** (data prep, training, evaluation), reducing summarization time by **30%**, deployment time by **40%** and improving sentiment analysis by **10%**.

LocalityAI (Live Demo) | *Recommendation System, Vector Database, React, Flask*

- Prototyped a real-time ad recommendation engine using **GPT-4o** and multimodal prompts composed with a **text-to-image** workflow powered by a **QLoRA** fine-tuned **Stable Diffusion** model.
- Orchestrated a **RAG pipeline** (LangChain + FAISS) with hybrid search achieving a **25%** relevance boost.
- Reduced the latency of the system by 25% with **Redis** caching, running a modular Flask server on AWS EC2.

Open Source Contributions (OSS Profile)

Published ML modules for LLM evaluation on PyPi with version control via Git, surpassing 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.