

Aryan Sanjay Patil

9344519501 | New York | ✉ arpatil@cs.stonybrook.edu | 🐙 [GitHub](#) | [in 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, R, C++, Git, Linux, Docker, Kubernetes

Frameworks: PyTorch, TensorFlow, Hugging Face, LangChain, OpenCV, ONNX, JAX, PySpark, CUDA, MLflow

AI Techniques: Transformers, Diffusion Models, GANs, LoRA, VLMs, Prompt Engineering, MLOps

Cloud and Data: AWS (EC2, SageMaker, S3), GCP (Hadoop, Spark), Azure, MySQL, PostgreSQL, Pinecone, FAISS

EXPERIENCE

AI Engineer Intern

Feb 2025 – May 2025

SteadFast AI

Remote

- Developed a scalable anomaly detection system for MCP servers powered by LLMs (Claude), achieving **90% accuracy** on enterprise-scale log data using secure prompt engineering techniques.
- Formulated a modular API and real-time ingestion pipeline integrated with a **RAG + Pinecone** retrieval system, reducing missed anomalies by **80%** through automated, context-aware detection.

Research Project Assistant

Aug 2024 – Dec 2024

Stony Brook University

Stony Brook, NY

- Increased quiz platform engagement by **30%** by deploying a real-time emotion recognition app using **multimodal** (image+text) RAG on **AWS EC2** with **FastAPI**.
- Engineered a **CNN3D** model using GAN-augmented training data for real-time video analysis, achieving **90%** accuracy and integrated **ViT**, **DeTR**, and **UNet** to enhance user tracking, improving model consistency by **5%**.

Machine Learning Researcher

Jun 2024 – Aug 2024

Brookhaven National Laboratory - Nuclear Physics | Publication in Progress

Upton, NY

- Reduced inference time on 3D heavy-ion collision data from **2 hours to 1 second**, improving **NERSC** cluster efficiency by **150%** via **CUDA**-optimized transformer models.
- Designed custom metrics, loss functions, and embedding layers to reduce prediction noise by **10x**.
- Published a [module](#) with **FP8 quantization**, achieving **3x faster inference** and reduced memory footprint.

PROJECTS

Reddit Analyzer App ([Demo](#)) 🐙 | [LLM Finetuning](#), [Hallucination Mitigation](#) | [Read Paper](#)

- Prototyped a real-time Reddit summarization app by fine-tuning **BART**, integrating RoBERTa-based sarcasm detection to **reduce hallucinations by 11%**.
- Deployed an end-to-end NLP pipeline on **AWS SageMaker** with **CI/CD** for training and real-time serving.
- Achieved **92% ROUGE-L** and **9% accuracy** boost on a custom Reddit benchmark post-finetuning.

RAG-based Recommendation System 🐙 | [Latency Optimization](#)

- Implemented a RAG-based microservice using FastAPI, LangChain, and FAISS to serve natural language recommendations with **85%+ Recall@k** using LLaMA 7B; tracked performance via **MLflow**.
- Achieved **150ms latency** using Redis caching; benchmarked retrieval quality via **A/B testing** of embeddings.

Dropwise – Predictive Uncertainty for Transformers 🐙

- Built [Dropwise](#), a toolkit for MC Dropout in **20+** Hugging Face Transformers with support for predictive entropy.
- Gained **5K+ downloads** in 2 weeks; recognized by a **Lightning AI** engineer who requested a TorchMetrics-style extension — delivered the module along with a Dockerized API and Lightning Hub demo.
- Integrated CI/CD via **GitHub Actions**; designed for active learning, reliability eval, and HF compatibility.

OPEN SOURCE CONTRIBUTIONS ([🌐 OSS PROFILE](#))

LocalityAI: Released an open-source [tool](#) to generate localized ads using GPT-4o and QLoRA-tuned Stable Diffusion.

Hugging Face: Contributed PRs to finetune tutorials with EarlyStopping, memory cleanup, gradient clipping, LoRA.

GitHub: Maintained 36 ML repositories(**100 stars**), and published 4 GenAI modules on [PyPi](#) (**9k downloads**).

Preswald (2.5k stars): Merged AI-integration PRs into the core repo, with work featured in the company's [blog](#).