# ARYAN SANJAY PATIL

Stony Brook, NY | +19344519501 | arpatil@cs.stonybrook.edu| linkedin.com/in/aryanpatil01 | github.com/aryanator

## EDUCATION

**Master of Science in Computer Science** **August 2023 – May 2025**
Stony Brook University, New York **(GPA: 3.87/4)**
Relevant Coursework: Natural Language Processing, Machine Learning, Computer Vision, Reinforcement Learning

**Bachelor of Technology in Information Technology** **August 2019 – July 2023**
University Of Mumbai, Mumbai, India **(CGPA: 9.68/10)**
Relevant Coursework: Data Structures & Algorithms, Database Systems, Operating System, Computer Networks

## SKILLS

- **Languages and Tools:** Python, R, C++, Git, Linux, Docker, Kubernetes
- **Frameworks:** PyTorch, TensorFlow, Hugging Face, LangChain, OpenCV, ONNX, JAX, TensorRT, MLflow
- **Cloud and Database:** AWS (EC2, SageMaker, S3), GCP (Hadoop, Spark), Azure, MySQL, PostgreSQL, Pinecone, FAISS
- **AI/ML:** Transformers, Diffusion Models, RAG, GANs, LoRA, VLMs, Prompt Engineering, MLOps

## PROFESSIONAL EXPERIENCE

**SteadFast AI | Artificial Intelligence Intern** **January 2025 – May 2025**

- Deployed a scalable fraud detection pipeline processing **15M+** transaction logs across 9 enterprise clients
- Led end-to-end ML development by integrating **RAG** retrieval to filter suspicious logs using Flask, Azure, Docker
- Collaborated with cybersecurity team and SDEs to develop a real-time log ingestion system with RabbitMQ queues
- Eliminated **Claude hallucinations** by **80%**, cut **API cost** by **18%**, and achieved 95% accuracy in classification
- Reduced fraud cases by 50%, saved 25+ hours/week of Cloud Ops, and delivered $5M annual cost savings for clients

**Stony Brook University | Machine Learning Intern** **September 2024 – December 2024**

- Formulated a quiz platform using FastAPI, EC2, and a finetuned CNN3D for live webcam emotion recognition
- Boosted engagement by 30% (500 users tested) using **multimodal** RAG (emotion + text) to generate questions
- Augmented training with **GAN**-generated data, achieving 90% model accuracy for emotion classification
- Explored ViT, DeTR, and UNet models for improved classification and tracking, gaining 5% higher accuracy in testing

**Brookhaven National Laboratory | Machine Learning Intern** **May 2024 – August 2024**

- Replaced physics simulations with transformers trained in JAX, exported via ONNX, and accelerated with TensorRT and FP8 quantization, accelerated inference time by **7200×** and boosting NERSC cluster efficiency by 300%
- Designed physics-aware metrics and loss functions to reduce **prediction noise 10×** and improve reliability at scale
- Published scientific modules now used by **1.5k+ researchers and 2 National Labs** by coordinating with researchers

## PROJECTS

**Dropwise : Uncertainty Quantification for Transformers** | *Toolkit*

- Built Dropwise, a toolkit for MC Dropout compatible with 20+ Hugging Face Transformers for model calibration, flagging risky inputs to safeguard model from adversarial attacks
- Integrated CI/CD via **GitHub Actions** and **Docker,** used Git for version control, published via PyPi
- Gained **5K+ downloads** in 1 week, recognized by a **Lightning AI team**, and soon to be adopted to TorchMetrics

**RAG-based Product Recommendation System** | *GitHub*

- Implemented a FastAPI + LangChain + FAISS **microservice** that parses user queries to retrieve products semantically matched across **100K+ items**, and maintained SQL databases for structured product details and metadata
- Integrated **LLaMA 7B** as an agent to generate natural language explanations combining the user prompt with matched product data, improving trust and relevance (**85%+ Recall@k**)
- Optimized end-to-end latency to 150 ms via **Redis** caching; tracked versions and performance with **MLflow**

**Reddit Analyzer App** | *Demo*

- Prototyped a text summarization app for Reddit by fine-tuning BART (92% ROUGE-L), using Flask + React stack
- Merged RoBERTa-based **sarcasm detection** to cut hallucinations by 11% on a **custom Reddit benchmark**
- Deployed a real-time API on **AWS SageMaker**, versioned and CI/CD-managed for autoscaled inference and training

## OPEN SOURCE CONTRIBUTIONS - *OSS Portfolio*

**LocalityAI**: Released an ads generation agent used daily by **1k+ startups** with **2M+ users** across the world- *Demo*
**Hugging Face**: Collaborated PRs to finetune tutorials, wrote clean code and documentation, and raised issues
**Preswald (2.5k stars)**: Merged AI-integration PRs into the core repo, with work featured in the company's *blog*
**GitHub**: Maintained 36 ML repositories(**100 stars**), and published 4 GenAI modules on *PyPi* (**9k downloads**)
**Research**: Reproduced DeepMind's and Microsoft AI research papers from scratch to **assist a MIT based startup**