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, CUDA, MLflow
- **Cloud and Database:** AWS (EC2, SageMaker, S3), GCP (Hadoop, Spark), Azure, MySQL, PostgreSQL, Pinecone, FAISS
- **AI/ML:** Transformers, Diffusion Models, 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 and reduce **Claude hallucinations** by **80%**, cutting **API cost** by **18%**, and achieving 95% accuracy in classification
- Developed a real-time log ingestion system with RabbitMQ queues, enabling robust and on-demand data streaming
- Collaborated with cybersecurity team and SDEs to ensure secure deployment and low-latency inference at scale

Stony Brook University | Research Project Assistant **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 CUDA-trained transformers, cutting inference time from 2 hours to 1 second
- Designed **Physics-aware** metrics and loss functions to reduce **prediction noise 10×**, improving model reliability
- Optimized NERSC cluster usage, boosting compute **efficiency by 300%** and working with interdisciplinary team
- Published scientific modules with FP8 **quantization**, currently being used by **1.5k+ researchers** and 2 National Labs

PROJECTS

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

- 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
- Gained 5K+ downloads in 1 week, recognized by a Lightning AI team, and soon to be adopted to TorchMetrics

Reddit Analyzer App | *Demo* | *Read Publication*

- Prototyped a real-time Reddit summarization app by fine-tuning BART, integrating RoBERTa-based **sarcasm detection** to cut hallucinations by 11%; achieved 92% ROUGE-L on a **custom Reddit benchmark**
- Deployed an end-to-end NLP pipeline on AWS SageMaker with **CI/CD** for training and inference

RAG-based Product Recommendation System | *GitHub*

- 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

Locality AI - Image Generation Engine | *Demo*| *GitHub*

- Formulated a multimodal GenAI agent for ads generation using GPT-4o and QLoRA-tuned Stable Diffusion, via real-time location, and trends using structured prompt orchestration
- Scaled to 20+ industry-specific use cases, deployed on AWS EC2 with a Dockerized Flask + React stack

OPEN SOURCE CONTRIBUTIONS - *OSS Portfolio*

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**