Jayraj Pamnani

+1(646)642-1043 | jmp10051@nyu.edu | LinkedIn | GitHub | Portfolio

Education

New York University — New York, NY
Master of Science, Computer Engineering — *Expected: May 2026*

Parul University — Vadodara, India
Bachelor of Technology, Computer Science and Engineering — *May 2024*

Technical Skills

Languages: Python, SQL, C/C++, Java, JavaScript
ML/AI Frameworks: TensorFlow, PyTorch, Matplotlib, Seaborn, Scikit-learn, Neural Networks, Computer Vision, NLP
Data & Cloud: Distributed Systems, AWS, CI/CD, MongoDB, PostgreSQL, Hadoop, Spark, ETL Pipelines, Tableau Dashboards
Tools: Git, Docker, Kubernetes, Jupyter, Gradio, Hugging Face, WandB, n8n, Postman, LangChain

Experience

GBCS Group — Calgary, Canada
Software Architect Intern — *Nov 2025 – Present*

- Spearheaded the architectural optimization of the group's software ecosystem, implementing resource efficiency strategies that reduced hosting and maintenance costs by 28%.
- Reengineered automated CI/CD pipelines and deployment workflows, accelerating release cycles by 40% while ensuring 99.9% system availability.
- Defined technical standards and architectural roadmaps, guiding development teams in building scalable microservices that reduced technical debt and improved maintainability by 30%.

New York University — New York, NY
Teaching Assistant – Machine Learning — *Sept 2025 – Dec 2025*

- Guided 50+ graduate students through ML fundamentals, including preprocessing pipelining, supervised/unsupervised learning, Deep Learning, and model optimization techniques.
- Conducted weekly office hours to debug Python code, explained algorithms, taught ML topics, and assisted with PyTorch implementations.

Projects

HexDrop (Secure File Transfer) | *Next.js, TypeScript, Prisma, PostgreSQL, AWS, Docker, K8s*

- Built a secure file-sharing app where users upload files, receive a shareable 6-digit key, and download via that key; implemented end-to-end flow with client-side encryption, Prisma-managed metadata in PostgreSQL, and object storage in AWS S3.
- Designed and deployed a full-stack DevOps pipeline: multi-stage Docker builds, GitHub Actions CI/CD for build and deploy, EKS orchestration with HPA and ALB Ingress, and RDS PostgreSQL with External Secrets for configuration management.

EyeConnect | *WebRTC, Supabase, OpenRouter AI, React, TypeScript*

- Developed an end-to-end accessibility platform that connects blind users with sighted volunteers using WebRTC for peer-to-peer video streaming and Supabase Realtime for instant signaling and volunteer matching; awarded 2nd place at NYU Hacks 2025.
- Architected a hybrid assistance system integrating OpenRouter AI for automated vision descriptions and a React/TypeScript frontend with shadcn/ui to ensure a high-contrast, accessible UI for low-latency visual support.

Model Merging (LLMs) | *Python, PyTorch, Hugging Face, Google Colab*

- Implemented and evaluated advanced LLM merging techniques (TIES, SLERP) to combine Mistral-7B variants into unified, high-performance models; engineered robust PyTorch pipelines to manage GPU memory constraints and SafeTensors-based checkpointing for 7B-parameter architectures.
- Designed automated experimental workflows using Hugging Face Transformers to optimize hyperparameters like sparsity density and merge ratios, validating model stability and cross-task performance through downstream inference testing.

ViT Optimization | *PyTorch, bitsandbytes, FlashAttention-2, LoRA*

- Optimized Vision Transformer (ViT-L/16) classifiers using 4-bit/8-bit quantization and efficient attention mechanisms (Flash Attention-2, SDPA), achieving a 4× reduction in model size and 40% lower latency with minimal (<2%) accuracy degradation.
- Engineered an end-to-end HPML pipeline for fine-tuning via QLoRA and profiling CUDA-kernel performance; conducted layer-wise ablation studies to identify critical encoder layers for mixed-precision execution, enabling high-throughput inference at ∼449 images/second.