JAYRAJ M. PAMNANI

jmp10051@nyu.edu | linkedin/in/jayrajpamnani | +1(646) 642-1043 | New York City, NY | Portfolio

EDUCATION

New York University | Master's in Computer Engineering | May 2026
Relevant Coursework: High Performance Machine Learning, Machine Learning Operations, Deep Learning, Database Systems, Big Data

Parul University | Bachelor's in Computer Science & Engineering with Specialization in AI | May 2024
Relevant Coursework: Data Structures and Algorithms, Natural Language Processing, GPU computing, Pattern Recognition, Object Oriented Programming, Computer Networks

SKILLS

AI/ML & Data Science: Python, R, TensorFlow, PyTorch, Scikit-learn, Keras, LangChain, LLMs, VLMs, NLP, Computer Vision, ONNX, TensorRT, CUDA, Quantization, Pruning, Optimization
MLOps & Development: MLflow, DVC, BentoML, TorchServe, CI/CD (GitHub Actions, Jenkins), REST APIs, Django, Flask, Kafka, Airflow, OOP, Data Structures & Algorithms, API design
Big Data & Analytics: Spark, Hadoop, Hive, BigQuery, PostgreSQL, MongoDB, MySQL, Pandas, NumPy, Tableau, PowerBI, Matplotlib, Seaborn
Tools & Platforms: GitHub, VS Code, Jupyter, HuggingFace, Ollama, AWS, GCP, Azure, Vertex AI, Docker, Kubernetes, n8n, Postman, Insomnia

PROFESSIONAL EXPERIENCE

GBCS Group - Software Architect Intern, Calgary, Canada *11/2025 - Present*
- Currently architecting the real-time data integration layer for the **Orion-Duo** platform, synchronizing telemetry between aircraft sensors and ground operations to reduce data latency by **30%**.
- Leading the structural redesign of the **Komet** flight data engine, transitioning legacy manual processes into an automated **Machine Learning pipeline** that has improved data ingestion accuracy by **25%**.
- Developing a microservices-based blueprint for the **Orion** lifecycle management platform, ensuring the system can scale to handle 1,000+ global aviation assets while reducing cloud overhead by **20%**.
- Authoring comprehensive **Architecture Decision Records (ADRs)** and security frameworks for multi-tenant aviation environments, ensuring all designs meet strict international aerospace data compliance standards.

Swaroop.ai - AI Intern, Ahmedabad, India *03/2024 - 08/2024*
- Enhanced Text-to-Speech model performance by **25%** through **fine-tuning Coqui TTS** across multiple Indian languages, optimizing real-time synthesis pipelines.
- Improved Speech-to-Text accuracy by **18%** by refining Coqui STT models with domain-specific audio datasets and custom preprocessing scripts.
- Integrated **OpenAI Whisper** into the production workflow, increasing transcription accuracy on noisy datasets by **30%** and cutting i**nference latency by 20%**.
- Contributed to the end-to-end AI pipeline (data cleaning, model training, deployment), helping reduce model turnaround time from **3 days to under 24 hours**.

Robotskull - Data Scientist Intern, Vadodara, India *10/2023 – 03/2024*
- Processed and standardized **100K+** sales and inventory records, improving data accessibility and analysis speed by **40%**.
- Built demand forecasting models that achieved **92% prediction accuracy**, driving **15%** more efficient inventory restocking decisions.
- Automated weekly analytics dashboards for the procurement team, reducing manual reporting time by **70%**.
- Identified **top 10 high-demand SKUs**, directly influencing procurement priorities and saving **~10%** in inventory holding costs.

Freelance Software Developer, Vadodara, India *07/2023 – 09/2023*
- Engineered and deployed a Django-based web app that automated data transfer between **QuickBooks and KatanaMRP**, cutting the client's manual bookkeeping workload by **~85%**.
- Built a robust **REST API integration** using Django (backend) and JavaScript (frontend), enabling seamless synchronization of **10,000+ financial and inventory records** with zero data loss.
- Designed and implemented **5+ custom verification layers**, achieving **99.8% data accuracy** before syncing to KatanaMRP.
- Deployed on **AWS EC2** with automated error logging and uptime above **99.9%**, ensuring secure, continuous operation for the client's business processes.

PROJECTS

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.

ViT Optimization | *PyTorch, bitsandbytes, FlashAttention-2, LoRA*
- **Optimized Vision Transformer (ViT-L/16) classifiers** using 4-bit/8-bit quantization and efficient attention mechanisms (FlashAttention-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.

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.