Jishnu Warrier

+1-(929)-735-9747 | jishnu.warrier@nyu.edu | Linkedin | Github

Education

New York University, Tandon School of Engineering — New York City, USA
Master's in Computer Engineering (GPA - 4.0 / 4.0) — *Aug 2024 – May 2026*

BITS Pilani, Goa Campus — Goa, India
B.E. Electronics and Communication Engineering — *Aug. 2019 – May 2023*

Technical Skills

Courses: Machine Learning, Machine Learning Operations, Artificial Intelligence, Data Structures & Algorithms, Operating Systems, Efficient AI, Responsible Data Science, Object Oriented Programming, Computer Architecture
Languages: Python, C, C++, Java, SQL
Technologies: PyTorch, FastAPI, Databricks, Git, Kubernetes, Docker, Azure, GCP, MLFlow, Airflow, Terraform, Ansible, ArgoCD
Libraries: Tensorflow, PyTorch, Matplotlib, OpenCV, Scikit-learn

Work Experience

Research Assistant — June 2025 – Present
SAI Lab - NYU — *New York City*

- Designing a graph-constrained, multi-agent RL + LLM pipeline that tackles an NP-hard scheduling problem, accelerating post-disaster electrical grid restoration while providing explainable, constraint-aware plans.

Software Engineer — July 2023 – Aug 2024
Chubb — *Hyderabad, India*

- Architected and implemented the entire backend infrastructure of a product recommender system using FastAPI to generate **100+** personalized assets (eg. taglines and images) through LLMs, increasing campaign engagement by **25%**
- Designed and implemented SQL-based workflows to create, clone, and manage prompt templates, supporting structured storage and retrieval of prompt configurations and generated outputs. Leveraged relational database design to optimize data integrity, querying efficiency, and integration with LLM interfaces.
- Designed a concurrency management wrapper that improved Databricks job throughput by **30%** in high load periods.
- Developed a real-time monitoring system for travel policies, enabling **risk based alerting with 1min latency**, reducing incident response time by **60%**
- Integrated machine learning models to develop dynamic pricing strategies, enabling real-time price surging, which empowered users to optimize profitability through data-driven experimentation.
- Streamlined deployment processes using Docker and Azure Kubernetes Service (AKS), ensuring seamless, scalable, and reliable cloud-based operations.

Research Intern — July 2022 – June 2023
Nanyang Technological University — *Remote*

- Developed a deep learning pipeline to classify student engagement levels using webcam video data, aimed at enhancing adaptive online learning platforms.
- Processed and integrated **19,000+ video samples** from EngageNet and EmotiW datasets; applied label correction, temporal smoothing, and data augmentation to address class imbalance.
- Built an ensemble model combining ViViT (Video Vision Transformer) and Temporal Convolutional Networks (TCNs) to capture both spatial and temporal patterns in classroom behavior.

Projects

SpotifyBuddies: Organic Playlist Recommender GitHub — Jan 2025 – May 2025

- Built a 40 GB data-prep pipeline (Pandas + SciPy sparse) that merged Million Playlist and Echo Nest datasets, computed **100M+** user–playlist overlaps in minutes, and produced **20M** contrastive triplets for model training.
- Implemented a FastAPI micro-service (multi-stage Docker, `uv` installer) delivering **1000-user batch recommendations <300ms on CPU**, shrinking image size 6GB → 2GB and build time by 85 %.
- **Deployed an end-to-end MLOps pipeline on Chameleon Cloud** with Terraform & Ansible: Kubernetes for orchestration, MLflow + MinIO for experiment tracking, Airflow DAGs for scheduled retraining, and Prometheus/Grafana dashboards for real-time latency, throughput, and cold-start drift.

EfficientDebates: Pruning-Based Multi-Agent Reasoning with Small LLMs GitHub — Feb 2025 – May 2025

- Designed an adaptive pruning-based multi-agent debate protocol that boosts reasoning accuracy of small LLMs (LLaMA 3B/8B, Phi 3 Mini) on GSM8K to **90.3%**, rivaling LLaMA 70B while reducing compute cost by **4.3x**.
- Reduced average inference to **3.27 LLM calls** per sample by asynchronously triggering debates only when agents disagree, avoiding unnecessary debating rounds in **92%** of cases.
- Benchmarked against Self-Consistency baselines, demonstrating that model diversity and early-exit gating lead to cost-efficient accuracy improvements in multi-step reasoning.

License Plate Recognition — Jan 2022 – May 2022

- Curated and pre-processed a custom dataset of Indian license-plate images; applied augmentation and OpenCV transforms to handle diverse lighting and viewing angles.
- Built a YOLOv5 detector plus contour-based character segmentation and CNN OCR, topped with a novel rule-based state-code auto-correction layer that significantly boosted end-to-end accuracy.