SHANTAM SHARMA

National University of Singapore, Singapore
shantam.s@u.nus.edu ◇ LinkedIn ◇ Github ◇ +65 84696935

EDUCATION

- **National University of Singapore, Singapore** 2025 - Present
  Master's of Technology, AI Systems
- **Indian Institute of Technology, Delhi (IITD)** 2016 - 2020
  Bachelor's of Technology, Industrial and Production Engineering — First Class Honours

TECHNICAL SKILLS

Technologies	Tableau, Git, Docker, Snowflake, AzureML, AWS Sagemaker, Triton
Programming Languages	Python, C++, SQL, MATLAB
Relevant ML Libraries	Pytorch, Langchain, Mlflow, NLTK, OpenCV, Sklearn, Pandas, Numpy

WORK EXPERIENCE

Research Intern, Singapore-MIT Alliance for Research and Technology (SMART) Mar 2026 - Present
- Experimented with online bandit/expert strategies for adaptive policy selection in stochastic contextual bandit setting.
- Implemented a co-optimization algorithm for online selection of policy and arm, consistently reducing cumulative regret w.r.t. any fixed baseline policy for various non-linear reward functions.

Research Intern, I2R - Astar: *Improved DinoV3 model inference throughput by 33%* Nov 2025 - Jan 2026
- Conducted quantization experiments (dynamic, static, QAT) on meta's pretrained Dino-V3 model fine-tuned for classification. Reduced memory footprint by 60% with INT8 precision static quantization using NVIDIA's modelopt library.
- Converted the quantized model to tensorrt engine and deployed it using NVIDIA's triton inference server.

Senior Data Scientist, World Wide Technology Inc. July 2020 - July 2025
Received employee excellence award and certificate for delivering cutting edge ML solutions in mining industry

- **GenAI for Asset Integrity and Process Safety**: *Mitigated risk of 10M$/y due to safety incidents at ADNOC*
  - Developed a RAG agent with automated graph generation on top of Text2SQL engine to visualise latest trends
  - Utilized LLMs for trend analysis and key insights on hazardous events using prompt engineering & data manipulation
  - Devised control KPIs to indicate the likelihood of hazardous events at each equipment and location in various assets
  - Performed benchmarking of pre-trained LLMs and fine-tuned BERT models for text-classification on Oil&Gas data
  - Deployed the dockerized applications on AzureML and presented a demo of the final product at ADIPEC 2024
- **Recommendation System for Retail**: *Recommending top 3 relevant products to consumors at JITB (USA)*
  - Reconciled datasets collated by tools like customer engagement, loyalty and marketing, to build a data warehouse
  - Evaluated various techniques based on collaborative filtering and matrix factorization for recommendation model
  - Profiled customers using KMeans Clustering and DBSCAN; integrated customer profiles with recommendation model
- **Vision Models for Real-Time Inspection**: *Minimized safety risks and operation downtime for PTFI*
  - Explored efficacy of features derived using GLCM, Gabor filters, power spectral density and optical flow like algorithms on images of mineral ore to depict the properties of its texture, volume and flow rate
  - Trained and fine-tuned anomaly detection models using architechtures including Autoencoder and Siamese network
  - Deployed the final model utilizing gpu's on edge computing devices in TensorRT for optimized latency of 4 fps
- **Physical Modelling for Product Quality Forecast**: *Increased production of molybdenum mineral by 3 %*
  - Developed Python-based digital-twin model of three chemical plants to predict impurity concentration in the product
  - Automated digital-twin performance optimization using methods like L-BFGS-B, Nelder-Mead and trust-region
  - Analyzed ML techniques like Random Forest, Boosted Trees, RNNs and CNNs to forecast product quality
- **Mining Operation Optimization**: *Increased production hours of mining operations by 8%*
  - Implemented and fine-tuned Kalman Filter to accurately estimate mineral concentrations in the absence of sensors
  - Trained a CNN-based image classification model to detect obstruction in mineral ore crushers with 98% accuracy
  - Deployed these solutions across edge devices and integrated AzureML toolkit for real-time monitoring

RESEARCH PUBLICATIONS

- **WWT R&D**: An article on "Power Usage Effectiveness(PUE) Optimization" for datacenters Jul, 2023
- **WWT R&D**: Presented on "A Modular approach to AIOps" at NVIDIA GTC'22 Nov, 2022