Siddhant Mohan

linkedin.com/in/siddhant-mohan-1110 | github.com/siddhantmohan1110

+1 646-305-8138

siddhantmohan1110@gmail.com

Education

- **New York University Tandon School of Engineering** — New York City, USA
  Master of Science (MS) - Electrical Engineering; CGPA: 3.7/4 — *Aug. 2024 - May 2026*
  Courses: Efficient AI, Advanced Computer Vision, Natural Language Processing, *ML Systems Engineering & Operations*
- **Indian Institute of Technology (IIT) Tirupati** — Tirupati, India
  Bachelor of Technology (B. Tech) - Electrical Engineering; CGPA: 8.0/10 — *Jul. 2016 - Jun. 2020*
  Courses: Deep Learning, *Industrial Data Science, Medical Imaging, Digital Signal Processing*

Technical Skills

- **High-Performance Computing & GPU Systems**: Slurm job scheduling, Multi-GPU training, CUDA environments, Mixed Precision (AMP), Docker & Singularity, Linux (Ubuntu) system setup, NVIDIA driver/CUDA/cuDNN installation
- **AI/ML Frameworks**: PyTorch, HuggingFace, TensorFlow, Keras, OpenMMLab, OpenCV, ONNX, Scikit-Learn
- **Large-Scale ML & Optimization**: Transformers, Large Language Models (LLMs), Deep & Convolutional Neural Networks, Parameter-Efficient Fine-Tuning, Distributed Data Parallel (DDP), Fully Sharded Data Parallel (FSDP), Model Quantization, Statistics & Probability for ML
- **MLOps & Infrastructure**: Git, Docker, Kubernetes, ArgoCD, MLflow, Ray, Prometheus, MinIO, Terraform, GCP
- **Programming Languages**: Python (expert), C++, C, MATLAB

Work Experience

- **Toshiba Software India - R&D Division** — Bengaluru, India
  Senior Research Engineer — *Jul. 2023 - Jul. 2024*
  Research Engineer — *Jul. 2020 - Jul. 2023*
  - **Incremental Learning:** Engineered an **incremental learning** pipeline for a single-shot object detector model, using **knowledge distillation** and **weight importance-based regularization**, which enabled adaptation to new object classes without forgetting old classes, improving mAP by 15% on COCO and VOC.
  - **Unsupervised Anomaly Localization:** Directed a team in developing an unsupervised approach for localizing anomalies in industrial images using **feature extractors** and **k-nearest neighbor statistics**. Outperformed baselines on the **MVTec-AD** dataset by 22%.
  - **Systems & Infrastructure:** Configured and deployed NVIDIA GPU workstations from scratch to set up an R&D lab, including **Ubuntu installation**, **CUDA/cuDNN** and driver setup, and **PyTorch environments** for deep learning, reducing compute costs by 30% annually compared to equivalent commercial cloud usage.
- **UnboundX** — Short Hills, NJ
  AI Engineering Intern — *Jun. 2025 - Aug. 2025*
  - Led a team to develop **feed-ranking recommender systems** for an investment-focused **social media** app, combining a time-decay based scoring model with BERT-based content classification to generate personalized, category-driven feeds. Utilized vectorized operations to reduce ranking latency by 10%.
- **Tech Mahindra (formerly ZEN3 Infosolutions)** — Hyderabad, India
  Machine Learning Intern — *May 2019 - Jul. 2019*
  - Devised, trained and deployed a ML-based customer support **chatbot** for mobile data plans using **Recurrent Neural Network (RNNs)**, within 8 weeks. Facilitated the creation of a **custom Q&A dataset** for training.

Academic Research Projects

- **Efficient Data Pipelines for Vision-Language Models**: Improved computational efficiency of a **Hybrid Autoregressive Transformer (HART)** for vision-language generation by performing semantic clustering of visual-text prompts and reusing low-resolution intermediate image representations within clusters. Reduced redundant forward passes, optimized multi-GPU inference workflows, and lowered compute and memory overhead in large-scale visual generation pipelines. *(Sep 2025).*
- **Distribution-Aware Companding Quantization (DACQ)**: Developed a **LLM post-training quantization framework** for transformer models that models layer-wise weight distributions and applies non-uniform companding to enable **efficient low-bit inference** with minimal accuracy degradation. Carried out large-scale experiments on **NYU Greene HPC** using Slurm-orchestrated multi-GPU jobs to benchmark perplexity, memory footprint, and throughput. *(Sep 2025).*
- **Intelligent Multimedia Processing Module (IMP)**: Built an end-to-end pipeline to extract text and metadata from multimodal streams (audio, video, documents) and integrated a **Retrieval Augmented Generation (RAG) system with LLMs** for enterprise query resolution. Automated data extraction, semantic indexing and CI/CD workflows, along with vector storage and retrieval using **FAISS** for efficient similarity search across large-scale enterprise knowledge bases. *(May 2025).*
- **Diffusion-Based Image Compression (DBIC)**: Built a two-stage statistical image compression system that sends a simplified version of the image and uses **guided diffusion model based restoration** to enhance perceptual detail while preserving semantic fidelity. *(Sep 2025).*
- **Galaxy Morphology Classification (AI for Science)**: Developed a PCA-based dimensionality reduction pipeline for large-scale astronomical image data, performing eigenmode analysis to extract structural features and suppress noise in high-dimensional inputs, which improved classification accuracy by 4% while reducing computational complexity. *(Jun. 2020).*