NIKET JAIN

niketjaina4@gmail.com | LinkedIn | Google Scholar | +14122848345

EDUCATION

Carnegie Mellon University (CMU) | School of Computer Science — **Pittsburgh, PA**
Master of Computational Data Science - GPA: 3.96/4.0 — Aug 2024 – Dec 2025
Selected Coursework: Introduction to ML, LLM applications, Cloud Computing, Deep RL, LM Inference, DL Systems
Teaching Assistant: Mathematical Foundations of ML, Computational Foundations of ML, Interactive Data Science

Vellore Institute of Technology (VIT) — **Vellore, India**
Bachelor of Technology in Computer Science and Engineering - GPA: 8.96/10.0 — Jul 2018 - May 2022

EXPERIENCE

Honeywell Inc. — **Atlanta, GA**
Machine Learning Intern | Honeywell Forge Team — Jun 2025 – Aug 2025

- Optimized sentence-transformer inference 2× via ONNX export, knowledge distillation, and Kubernetes HPA with PVC caching, accelerating client onboarding to the data fabric platform for building sensor analytics.
- Built multimodal RAG agent with MCP server for native image-doc processing, replacing chunking/OCR workflows and improving generation quality by 35%, enabling industry engineers to better process and understand maintenance documents.
- Designed random forest–based agitator fault detection agent, winning company hackathon and cutting maintenance costs 41% through predictive scheduling optimization.

Carnegie Mellon University — **Pittsburgh, PA**
Research Assistant | Language Technologies Institute | Advisor: Prof. Carolyn Rose — Jan 2025 – May 2025

- Authored and benchmarked a novel 150-instance dataset for image-guided webpage code editing, establishing state-of-the-art baselines using GPT-4o/4.1, QwenCoder, and QwenVL-32B (0.7686 similarity).
- Designed agentic workflow with QwenVL-32B + QwenCoder-32B (served with vLLM), strategically decomposing UI modification tasks to target a 40% speedup in code editing.

UBS — **Mumbai & Pune, India**
Software Engineer | Credit Risk Insights Team — Jul 2022 - Jul 2024

- Developed core modules of the Nucleus document processing system integrating OpenAI GPT-3.5 APIs for OCR, extraction, and summarization, reducing manual review time by 87.5% across 500+ financial documents.
- Engineered performant, fault-tolerant Java-based ETL pipelines using Kafka to aggregate daily credit risk transactions, reducing data latency by 60%.

Software Engineer Intern | Business Automation Team — Jan 2022 – Jul 2022

- Built RPA bots with Python and Alteryx, automating data workflows and improving integration with data science pipelines.

National Solar Observatory — **Boulder, CO**
Research Intern — May 2020 - Jun 2020

- Conducted data wrangling, power-spectral analysis, and visualization of solar data, uncovering subsurface magnetic activity linked to solar cycle behavior and space weather events. Work published in **The Astrophysical Journal Letters**. [paper]

PROJECTS

Neural Network Backend Accelerator (needle) | CMU — **Sept 2025**

- Engineered a full deep learning system (PyTorch clone) with autodiff, standard modules (Linear, Conv, TransformerLayer), and optimized low-level backends (Python, C++, CUDA, **XLA/TPU**).

LLM-Based Data Rewriter for DCLM Pre-training | Advisor: Prof. Chenyan Xiong | CMU — **Sept 2025**

- Designed a data-centric LLM rewriter to progressively replace high-impact heuristic cleaning filters (e.g., page length, repetition) in the DCLM pre-training data construction pipeline.
- Set up and benchmarked the DCLM training/evaluation pipeline on cloud accelerators for comparison and analysis baseline filter methods.

Data Attribution Benchmark for LLMs | Advisor: Prof. Chenyan Xiong | CMU — **Apr 2025**

- Benchmarked 8+ data attribution methods (LESS, MATES, gradient-based) across 3 LLM tasks (training data selection, toxicity filtering, factual attribution) with modular pipeline supporting models from Pythia-1B to Llama-3.1-8B.
- Ran large-scale evaluation showing no method dominates; simple baselines matched gradient methods at significantly lower computational cost (up to 11× reduction in FLOPs). Released Hugging Face leaderboard with community submissions and pre-trained checkpoints, cutting evaluation burden 70%. Work accepted at **NeurIPS 2025 Datasets and Benchmark Track**. [paper]

Cloud-Native Scalable Microservice for Twitter Analytics (1TB+ ETL, Kubernetes, REST) | CMU — **Apr 2025**

- Designed, built, and deployed a fault-tolerant microservice architecture on a self-managed Kubernetes cluster (Helm, CI/CD), processing >1TB of raw data and sustaining 100K+ real-time HTTP requests.

Enhancing LLM Math Solving Abilities via Code Generation | CMU — **Sept 2024**

- Implemented distributed QLoRA PEFT of Llama (3B) & Qwen (7B) models using data parallelism across four GPUs, achieving 4× faster training speed with increased batch sizes and accelerated training. Outperformed base models with improved perplexity (1.28) and accuracy (0.68) vs. zero/3/10-shot baselines.

SKILLS

- **Programming Languages & Databases** - Python, SQL, Java, R, Scala, C++, MySQL, MongoDB, PostgreSQL, Redis
- **Tools** - Git, Anaconda, Azure, AWS, GCP Vertex AI, Databricks, Langgraph, Langsmith, Terraform, Helm, GitHub Actions, HPC (SLURM), Maven, Kubernetes, Docker, MCP Inspector, UV, SonarQube
- **Frameworks & Libraries** - PyTorch, TensorFlow, Hugging Face, vLLM, PySpark, MLFlow, Onnx, Ray, Accelerate, FastAPI, Flask, Django, OpenCV, NumPy, Pandas, scikit-learn, PyTorch Lightning, LangChain, MCP, CUDA, Kafka, Samza