# NIKET JAIN

niketj@cs.cmu.edu | 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

## SKILLS

**Programming Languages & Databases** - Python, SQL, Java, R, Scala, C++, MySQL, MongoDB, PostgreSQL, Neo4j, Redis
**Tools** - Git, Anaconda, Azure (Microsoft AZ-900, DP-900 & AI-900 Certifications), AWS, GCP Vertex AI, Databricks, Langgraph, Langsmith, Docker, Terraform, Helm, GitHub Actions, HPC (SLURM), Maven, Kubernetes, Docker
**Frameworks & Libraries** - PyTorch, TensorFlow, Hugging Face, vLLM, PySpark, MLFlow, Accelerate, Flask, Django, OpenCV, NumPy, Pandas, scikit-learn, PyTorch Lightning, LangChain, RAG, MCP, MCP Inspector, CUDA, Kafka, Samza

## 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 system 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

- Designed agentic workflow with QwenVL-32B + QwenCoder-32B (served with vLLM), speeding up UI edits by 40%.

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

- Engineered Java-based ETL pipelines processing daily credit risk transactions across 13 data sources, reducing data latency by 60% and enhancing risk assessment accuracy by 23%.
- Architected Oracle 19c data infrastructure supporting 10TB+ financial data with multi-master replication across 3 regions, achieving 87% uptime during UBS-Credit Suisse merger integration.
- Developed Nucleus document processing system integrating OCR and OpenAI APIs, processing 500+ financial documents monthly with 92% accuracy rate, reducing analysis time from 4 hours to 30 minutes per document.

*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**
*Summer Intern* May 2020 - Jun 2020

- Conducted data wrangling, power-spectral analysis, and visualization of solar data from the Global Oscillation Network Group, uncovering subsurface magnetic activity linked to solar cycle behavior and space weather events. [paper]

## PROJECTS

**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%; findings submitted to **NeurIPS 2025**. [paper]

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

- Engineered 3 microservices (Blockchain, QR Code, Twitter) to handle 100K+ real-time HTTP requests over 4-hour load tests, achieving target RPS benchmarks (e.g., 74,258 RPS for Blockchain Service).
- Processed and transformed >1TB of raw Twitter JSON data using distributed ETL pipelines on Azure Databricks.
- Designed and deployed a fault-tolerant web-tier + MySQL-backed architecture on a self-managed Kubernetes cluster using Helm, with CI/CD automation on GitHub Actions.

**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.