NIKET JAIN

niketj@andrew.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

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

- 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 performant Java-based ETL pipelines using Kafka to aggregate daily credit risk transactions across 13 data sources, reducing data latency by 60%.
- Implemented schema harmonization and data quality checks within streaming workflows, improving credit exposure accuracy by 23% and ensuring consistency across regional data feeds.
- Designed and optimized Oracle 19c infrastructure supporting 10TB+ financial datasets, implementing time-series partitioning and composite indexing to accelerate query performance by 40%.
- Ensured 87% system uptime during UBS–Credit Suisse merger integration via multi-master replication, minimizing risk exposure.
- Developed core modules of the Nucleus document processing system integrating OpenAI GPT-3.5 APIs for OCR, information extraction, and summarization. Processed 500+ sensitive financial documents with 92% extraction accuracy, reducing manual review time by 87.5%.

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

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

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**).

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**

- Developed a recommendation microservice using Java Vert.x, computing scores based on user interactions, keywords, and hashtags from ~1 TB of Twitter data processed through Apache Spark ETL jobs written in Pyspark.
- Established seamless production pipeline using Terraform, Docker, AWS ECR, Helm Charts, and GitHub Actions.
- Optimized performance using SQL schema denormalization and indexing, asynchronous REST communication between microservices, and fine-tuning Kubernetes pod configuration, to get an average latency of under 13 ms.
- Achieved 42,000+ RPS using AWS EKS Kubernetes cluster with a node group of 4 m8g.xlarge EC2 instances.

Multi-Cloud Microservices with Kubernetes | CMU — **Apr 2025**

- Decomposed a monolithic chat application into Users, Messaging, and Auth microservices, containerizing them using optimized Dockerfiles for deployment on multi-cloud registries (GCR, ACR).
- Deployed fault-tolerant, high-availability clusters on GKE and AKS using custom Helm charts, configuring HPA, service mesh, and global load balancing (Azure Front Door).
- Implemented automated CI/CD pipelines via GitHub Actions to ensure continuous integration, testing, and multi-cloud deployment.

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, Ray, Accelerate, FastAPI, Flask, Django, OpenCV, NumPy, Pandas, scikit-learn, PyTorch Lightning, LangChain, MCP, CUDA, Kafka, Samza, Onnx