Nilesh Nayan

413-479-9836 nilesh.nayan42@gmail.com in/nilesh-nayan nnn007 portfolio

Education

University of Massachusetts, Amherst **Aug 2024 - May 2026 (expected)**
Master of Science in Computer Science *GPA: 3.9/4*
***Coursework**: Advanced Algorithms, Systems for Data Science, Machine Learning, Reinforcement Learning, Advanced NLP, Robotics*

Birla Institute of Technology and Science (BITS) Pilani, India **Aug 2017 - May 2021**
Bachelor of Engineering in Electrical and Electronics & Minor in Data Science *GPA: 8.74/10*
***Coursework**: Operating Systems, Neural Networks and Fuzzy Logic, Optimization, Applied Statistical Methods*

Work Experience

Software Development Intern, Applied Science MAIDAP (MIT) | Microsoft **Jan 2025 - Feb 2025**

- Optimized Microsoft's **GraphRAG** framework for **chatbots** and text search with custom **indexing**, reducing **latency** by **50%** and **token costs** by **90%** at a 0.2 accuracy tradeoff for **global** queries.
- Built a non-LLM **Knowledge Graph** generator with traversal support which outperformed **GraphRAG** and **LightRAG** by ∼2s/10k tokens docs in latency in custom Retrieval Augmented Generation (**RAG**) benchmarks using **DeepEval**.

ML Engineer 3, Applied AI | Comcast **Jul 2021 - Aug 2024**

- Designed backend for **AI4Ops** (AI for Operations) using **microservices** architecture on **AWS EKS** with **Airflow DAGs**.
- Reduced Mean Time to Resolve (**MTTR**) by ≥ **30 mins /incident** for **Olympics'24** broadcasting and streaming observability.
- Implemented **real-time inference** via **AWS SQS** and **Lambda** using State-of-the-Art (**SOTA**) **time-series** models (Meta's Prophet, N-BEATS, Transformer-based foundation models), improving seasonality-based **anomaly detection** by **50%**.
- Developed **auto-scalable APIs** with **FastAPI** on EKS, serving over **1M** requests/day and reducing **AWS MWAA** deployment costs by **80%** (∼$100k/month).
- Built an **event-driven architecture** for anomaly **alerting** and system **dependency graph-based** Root Cause Analysis (**RCA**) using Dynamic Time Warping (**DTW**) distance.
- Led R&D of a **log mining** pipeline using Drain3 with **LLM**-based triage (PEFT: QLoRA fine-tuned), adding **log trend alerts**, RCA, and **Q&A assistance**, cutting MTTR further by **25%**.

Software Development Intern | Jupiter (A Unicorn Fintech Startup) **Jul 2020 - Dec 2020**

- Improved cell phone **SMS NER** for **universal bank statements** and **credit card spends** detection by **20%** using SOTA **Flair** model (vs regex), and reduced **inference** time from **1.2s** to **300ms** per text via **embedding optimization** and **quantization**.
- Designed **Redis**-based data structures for Voice Annotation Platform (Indian accent **speech models**), modularized as a **Singleton** class, reducing read **latency** by **30%** over **SQL**.
- Built Big Data **KPI** reporting pipelines (**100GB/week**) using advanced SQL, **Airflow**, and **Spark**.

Projects

Scalable Forecasting Platform - Distributed Inference Service | *Comcast Innovation Lab Challenge (Internal Hackathon)*

- Implemented a scalable forecasting service with **FastAPI**, **Docker**, & **Kubernetes**, supporting inference for **10k** time series.
- Created clustering pipeline using **DTW** + **DBSCAN**, improving grouping efficiency by **40%** over KMeans, reducing redundant model runs and cutting inference cost by **5×** via global model deployment with batched inference using **Redis** queues.

Med-VQA KGRAG - Medical Visual Question Answering pipeline | *code*

- Improved existing medical **vision language model** on question-answering task by average of ∼**7%** by providing additional context using **multimodal BioMedCLIP RAG** and customized **GraphRAG FastAPI** pipelines without model finetuning.
- Used **Deepeval's** contextual relevancy and G-Eval (CoT) based latest NLP metrics showcasing significant improvement of **more than 10%** compared to Microsoft's **LLaVa RAD** and **Meta's Llama 3 visual instruct** model baselines.

Talent Acquisition (TA) helper framework | *Comcast Spring Labweek (Internal Hackathon)*

- Implemented LLM Agent using **LangChain** based ranking utilizing **GPT-3.5 embeddings** on **JD** and **candidate's profile**.
- Deployed **FAISS** with text embeddings to provide the relevant results based on queries' **cosine-similarity** to help the TA team optimize their normal filtering process, **speeding up** around **50 times**.

Technical Skills

- **Domain Expertise**: Software Development & Design (OOP), Databases, Cloud, Data Science, AI, Machine Learning, Networks
- **Languages and Frameworks**: Python, Java, C++, HTML, CSS, JavaScript, TypeScript, React, TensorFlow, PyTorch, FastAPI
- **Databases**: MySQL, Postgres, MongoDB, Redis, ElasticSearch, FAISS (FaceBook AI Similarity Score)
- **Dev Tools**: Git, Docker, Kubernetes, Linux, Bash, Shell Scripting, AWS, Airflow

Publications

- "NL2EQ: Generating Elasticsearch Query DSL from Natural Language Text Using LLM". *Springer Nature: Intelligent Computing*
- "AI for IT Operations (AIOps) - Using AI/ML for Improving IT Operations". In Proceedings of Society of Cable Telecommunications Engineers (SCTE) 2022 Fall Technical Forum, Philadelphia, United States of America. *NCTA*