Nikhil Soni

Brooklyn, NY | nikhilsoni700@gmail.com | +1 9342331695 | linkedin.com/in/nikhilsoni15/ | nikhilsoni15.github.io

Education

New York University, New York, NY — Sep 2023 - May 2025
Master of Science in Computer Science (*Recipient of Merit-based scholarship)* — GPA: 3.97/4.00

- **Relevant coursework:** Data Structures and Algorithms, Data Science for Business (NYU STERN), Machine Learning, Artificial Intelligence, Deep Learning, Big Data Analytics, Database Systems, Web Search Engines, Programming Languages

Manipal University, Jaipur, India — Aug 2019 - Jul 2023
Bachelor of Technology (BTech) in Computer Science — CGPA: 9.16/10.00

Technical Skills

Languages: Python, SQL, C, C++, Java, HTML, CSS, JavaScript

LLM / ML: Hugging Face, PyTorch, Transformers, Scikit-learn, RAG, LoRA, Fine-Tuning, 8-bit Inference, vLLM, LangChain

Infra / Systems: Docker, Kubernetes, FastAPI, Ray Serve, GCP, AWS (Lambda, EC2, S3), Azure, Spark Streaming, Kafka, Airflow

Databases & APIs: SQL (Indexing, Query Optimization), FastAPI, Qdrant, MongoDB, Redis, Snowflake

Experience

AI Engineer, *Quant AI Research* – New York, NY — Jul 2025 – Present

- Built a cron-driven ingestion pipeline with BigQuery, scaling to 500+ financial articles per day across equities and ETFs
- Developed a FinBERT-based sentiment model producing –1 to 1 scores with confidence levels and achieving 92% accuracy
- Optimized and deployed Mistral-7B-Instruct on Vertex AI using 8-bit quantization, 30% prompt reduction, and tighter token limits, improving report latency by 50%, enabling faster financial insight delivery
- Designed a LangChain–Qdrant RAG with MiniLM-L6-v2 embedding 3K sentences/sec for low-latency, context-aware queries

Graduate Teaching Assistant - Deep Learning, *New York University* – New York, NY — Jan 2025 – May 2025

- Orchestrated and hosted 2 Kaggle-style competitions to evaluate student solutions in NLP, Generative AI and Transformers
- Mentored 400+ students through office hours, clarifying concepts in diffusion models, RL, and advanced deep learning

AI/ML Intern, *Emerson* – Pune, IN — Jun 2024 – Aug 2024

- Architected an end-to-end LLM-based tool to automate validation of DeltaV system control reports, reducing manual effort
- Elevated recognition accuracy to 91% by applying chain-of-thought prompting on T5 and BERT, increasing tool efficiency
- Streamlined data pipeline, processing 10,000+ text files saving around 25–30 human hours weekly when performed
- Fine-tuned LLMs, reducing model training time by 25% and improving alignment with domain-specific data

Data Science Intern, *Junglee Games* – Gurugram, IN — Jan 2023 – Jul 2023

- Extracted and analyzed Fraud users data using SQL and Python-based EDA to build and optimize predictive models at scale
- Led testing and deployment of the "Problem Gamer" model to catch game addicts in a pool of 100 million users
- Optimized deployment and monitoring through MLOps pipelines using AWS Lambda, reducing model update time by 18%
- Replicated a CNN research paper to calculate players Rummy skill score to predict game drop decision with 82% precision

Software Developer Intern, *Hewlett Packard Enterprise* – Chandigarh, IN — Jun 2022 – Jul 2022

- Developed a full-stack application with Django backend and HTML, CSS, JavaScript frontend for intra-team issue reporting
- Provisioned the system on AWS using EC2 instances and VPC, ensuring scalability, security, and high availability

Projects

CrisisCast: Real-Time Crisis Detection & Monitoring | *PySpark, LLM, Kafka, Qdrant, MongoDB* [Github]

- Built end-to-end real-time emergency detection system by ingesting Reddit data using Kafka and Spark Structured Streaming
- Integrated a locally hosted LLM-based classifier to tag posts by crisis type, storing enriched metadata into MongoDB
- Embedded 1,000+ Reddit posts into vector space using Sentence Transformers and stored semantic representations in Qdrant
- Designed an interactive Streamlit dashboard for real-time crisis trend monitoring and semantic search across incoming posts

PrimePicks: Personalized News App | *Django, React, Qdrant, LLM, Kafka, Pyspark* [Github]

- Built a web app with vector databases for personalized news summaries, handling 1000+ daily articles via a custom pipeline using MongoDB and Kafka, with CI/CD integration.
- Improved retrieval speed by 50% using Kafka, Celery; enhanced UX with LLM-generated summaries for personalized content.

Rent Raja - NYC Rental Price Prediction | *Scikit-learn, Flask, Dash, LLM, APIs* [Demo|Code]

- Crafted predictive ML model to estimate rental prices by mining and processing 300,000+ property listings from various APIs
- Devised hybrid classification-regression pipeline, 81% classification accuracy on 3 bins, reducing RMSE from $3,000 to $300
- Engineered 10+ predictive features along with Dash + Flask dashboard and LLM-generated reports for broker pricing insights

Web Search Engine | *TensorFlow, Python, C++* [Github]

- Implemented a web crawler and inverted index system, processing 12,000+ web pages to enable large-scale data retrieval
- Reduced index size from 13.74 GB to 8.56 GB using VarByte compression and sharding, improving query speed by 30%
- Established a query processor using BM25 scoring, designing ranking algorithms to handle complex queries with precision