

Nikhil Soni

Brooklyn, NY | nikhilsoni700@gmail.com | +1 9342331695 | [linkedin.com/in/nikhilsoni15/](https://www.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

- Orchestrated cron-driven ingestion with BigQuery, scaling to 500+ financial articles/day across equities and ETFs.
- Built a Vertex AI sentiment analysis stage (Gemini + fine-tuned LLMs) with 92% accuracy on structured scores (–10 to +10).
- Deployed self-hosted LLM endpoints on Vertex AI for retrieval-on-demand and automated hedge-fund-grade reports.
- Delivered an end-to-end market sentiment and trading-signal platform; cut compute cost 90% vs prior baseline via scheduled batch and serverless execution.

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
- Collaborated to fine-tune 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

GenVision: Personalized Image Generator | *Diffusion XL, DreamBooth, LoRA, Flask, Gradio* [\[App|Code\]](#)

- Constructed full-stack app fine-tuning Stable Diffusion XL with DreamBooth and LoRA for personalized prompt-based images
- Created a Flask backend and Gradio-based frontend for real-time, user-specific image generation along with a feedback slider
- Applied optimizations like gradient checkpointing and mixed-precision training, to enhance performance on limited GPUs

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
- Refined storage with VarByte compression and index sharding, reducing overhead by 30% and improving query speed
- Established a query processor using BM25 scoring, designing ranking algorithms to handle complex queries with precision