Nikhil Soni

New York, NY | ns6062@nyu.edu | +1 (934)-233-1695 | linkedin.com/in/nikhil-soni-7a8645190/

Education

New York University, New York, NY Expected May 2025
Master of Science in Computer Science (*Recipient of Merit-based scholarship*) GPA: 3.96/4.00

- **Relevant coursework:** Data Structures and Algorithms, Data Science, Machine Learning, Artificial intelligence, Deep Learning, Big Data Analytics, Database Systems, Web Search Engines, Programming Languages
- **Graduate Teaching Assistant** for CS 6953 / ECE 7123 - Deep Learning

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

Experience

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

- Developed LLM-powered validation tool automating error detection in text-based code reports, reducing manual review effort
- Applied chain of thought prompting on T5 and BERT improving recognition accuracy to 91% and overall tool efficiency
- Designed data pipeline, processing 10,000+ text files efficiently 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
- Programmed automated log and report generation with highlighted error sections, enhancing stakeholder data traceability
- Conducted POC on Retrieval-Augmented Generation (RAG) techniques to further enhance domain-specific text retrieval

Data Science Intern, Junglee Games India – Gurugram, India Jan 2023 – Jul 2023

- Extracted and analyzed Fraud users data using SQL and Python-based EDA to build and optimize predictive models at scale
- Executed testing and deployment of the "Problem Gamer" model to catch game addicts in a pool of 100 million users
- Streamlined deployment and monitoring through MLOps pipelines using AWS Lambda, reducing model update time by 18%
- Implemented 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, India Jun 2022 – Jul 2022

- Built a full-stack application with Django backend and HTML/CSS/JavaScript frontend for intra-team issue reporting
- Crafted a real-time analytic dashboard to visualize key trends and prioritize actionable insights
- Deployed the system on AWS using EC2 instances and VPC, ensuring scalability, security, and high availability

Projects

GenVision: Personalized Image Generator [Live Demo|Code] Nov 2024

- Created full-stack app fine-tuning Stable Diffusion XL with DreamBooth and LoRA for personalized prompt based images
- Built 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
- Achieved high CLIP scores (36.408) for personalized datasets, showcasing high output quality on brief training times

BoOgLe: The Web Search Engine [Github] May 2024

- Developed a web crawler and inverted index system, processing 12,000+ web pages to enable large-scale data retrieval
- Optimized storage with VarByte compression and index sharding, reducing overhead by 30% and improving query speed
- Engineered a query processor using BM25 scoring, designing ranking algorithms to handle complex queries with precision

OopsFix: 311 Service Optimization for NYC Boroughs [Github] Jan 2024

- Designed scalable ETL pipelines to process 34 million records, integrating multimodal data for efficient processing
- Constructed and fine tuned ensemble learning models, achieving 87% accuracy in predicting borough-specific service delays
- Conducted spatiotemporal analysis to identify inefficiencies, integrating multimodal data to optimize resource allocation

Technical Skills

Languages: Python, SQL, C/C++, Java, HTML/CSS, Javascript

AI / ML: Scikit-learn, TensorFlow, PyTorch, Hugging Face, LLM, RAG, Model Deployment, Keras, Pandas, Matplotlib, NLP

Cloud / DevOps: AWS (EC2, S3, Lambda, API Gateway, SQS, SageMaker), GCP, Azure, Docker, Kubernetes, Jenkins, Airflow

Databases & APIs: MySQL, PostgreSQL, MongoDB, DynamoDB, Redis, Snowflake, Elasticsearch, Kafka, REST/GraphQL APIs