Nishant Sharma

Email: nishantsharma@nyu.edu
(646) 772-9142 | Brooklyn, NY

Portfolio: hellonish.dev
LinkedIn: linkedin.com/in/nishantsh20/
Github: github.com/nishant-ai

About

AI/ML Engineer and Researcher with 2+ years of experience architecting **Agentic AI systems** and scalable **Full Stack** infrastructure. Specializing in Language Models (LLMs, SLMs), RAG architectures, and scalable backend infrastructure. Co-founder of an Ed-Tech SaaS scaled to 5,000+ users. Seeking full time opportunities in Applied AI, ML Engineering, Data Science, Quant Researcher, AI Researcher starting May 2026.

Technical Skills

- **Programming Languages:** Python, Java, C++, Typescript, MATLAB
- **Applied AI and AI Research:** Pytorch, Tensorflow, Fine-Tuning, Model Evaluation, RAG, Language Models (LLMs/SLMs), Vision Models, Multi-Modal Models, Langgraph, Langsmith, TruLens, Arize, Vector Search, Embeddings
- **Software Development, DevOps and Systems Engineering:** NextJS, NodeJS, FastAPI, Django, Flask, SQL, NoSQL, Docker, Kubernetes, AWS, GCP, Prometheus, Grafana, MLFlow, Jira, Jenkins
- **Research & Documentation:** Git, LaTeX, Experimental Design, Research Methodologies, System Design

Education

- **New York University Tandon School of Engineering** | New York, US | Sep 2024 - Expected May 2026 | GPA: 3.7
  Master of Science in Computer Engineering
- **Dr. A.P.J. Abdul Kalam Technical University** | Delhi, India | Aug 2020 - Jun 2024 | 1st Division with Distinction
  Bachelor of Technology in Computer Science & Engineering

Work Experience

Graduate Course Assistant, Machine Learning | NYU Tandon Dept. of Computer Science | NY, USA | Sep 2025 - Present

- Serve as Technical Mentor for 50+ graduate students, debugging **PyTorch** implementations of Neural Networks and classical ML algorithms.
- **Created technical curriculum** and video lectures on core ML theory (e.g., Regularization, Optimization), bridging the gap between theoretical calculus and practical Python implementation.
- Conduct code reviews to optimize student projects, reducing training runtime by identifying inefficient vector operations.

Co-Founder and Lead Engineer | Ingelt Board | Delhi, India | Dec 2022 - Jul 2024

- Led a team of 6 engineers and launched SaaS Ed-Tech platform, scaling to **200+ enterprise clients** & **5,000+** active students.
- Architected a multi-tenant SaaS platform using **Django**, **Node** and **React**, scaling to 5,000+ active users and 200+ B2B clients.
- Deployed on **AWS (EC2, RDS, S3)** with a **CI/CD pipeline (Jenkins, GitHub Actions)**, achieving **99.9% uptime** and faster deployments.

Software Engineering Intern | Macverin Technologies | Hybrid (UP, India) | Jul 2022 - Dec 2022

- Delivered **Dockerized CMS/CRM** platforms to 8 clients, improving their content management efficiency by 40%.
- Built client-facing analytics dashboards with **Python and JavaScript (Chart.js, etc.)** to visualize user behavior and sales data.
- Contributed to feature development, documentation, and system uptime using **Agile methodologies (Jira, Git)**.

Applied AI Projects

SmolSolver - Mathematical Reasoning with SLMs | In-Progress

- Developing Generator and Verifier Small Language Models (SLMs) fine-tuned on PRM800K and GSM8K datasets using Python.
- Focusing on step-by-step mathematical reasoning and evaluation.

Cross-Domain Vision - Image Reconstruction Benchmark |In-Progress

- Benchmarking SoTA models (CNNs, Transformers, Diffusion) on super-resolution, denoising, and inpainting using Python.
- Analyzing cross-domain robustness and failure patterns across natural scenes, text, astronomy, and art.

General Projects

Wort.nyc – Personal Research Assistant |www.wort.nyc | In-Progress

- Developed an autonomous research agent using **LangGraph** and **RAG**, implementing a novel solution-tree algorithm for multi-step reasoning. Built a hybrid REST/WebSocket backend to handle real-time token streaming.
- Developing System based on a hybrid design (REST and WebSocket) with State Management, Data Persistence and Caching using Python, Typescript.
- Architecting a scalable-dockerized system to be hosted over AWS powered by CI/CD.

Snap2Caption - ML Systems for Caption Generation | code | 2025

- Built a complete ML pipeline to generate Instagram-ready captions and hashtags from photos in under 2 seconds.

- Fine-tuned LLaVA-1.5/1.6 (7B) vision models using LoRA on 100k urban images for efficient, high-quality training.
- Engineered a production setup handling 300+ requests/hour, fully monitored with MLflow, Prometheus, and Grafana.
- Automated infrastructure provisioning on GPU clusters using Terraform.

Finassistant - RAG-based Financial Agent | live demo | code | 2025

- Created a conversational financial analysis tool that pulls real-time data from equity, crypto, and macro-APIs.
- Utilizes RAG to search financial news and answer portfolio/risk questions in natural language.
- Currently fine-tuning on sentiment-labeled datasets to "read between the lines" of earnings calls.

Leadership & Extracurriculars

Technical Head | Computer Society of India (CSI) Student Chapter | Jun 2023 - Jun 2024

- Led a 15-member team and organized 12+ technical events.

Subject Matter Expert | Chegg India | Sep 2022 - Jul 2023

- Delivered 1,000+ academic solutions in Computer Science and Mathematics.

Community Contributor | Medium & LeetCode

- **Tech Blogger:** Author articles on Startups, AI, and Software Development at Medium.
- **LeetCode:** Solved 200+ problems, with 52 solutions posted and 4.6K+ community views.