Nishant Sharma

Email: nishantsharma@nyu.edu
New York, NY

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

About

AI Software Engineer with experience building, deploying, and managing production-grade cloud-based services for 1K+ DAU for 10+ Clients at macverin.com. Currently building Agentic AI. Studying Computer Engineering focusing on Language and Multimodal Models training, finetuning and application in real-world software-based use cases. Seeking full-time roles in AI Engineering and Software Development starting May 2026.

Technical Skills

- **Programming:** Python, Java, C++, JavaScript, TypeScript, Bash
- **AI and ML:** Pytorch, Tensorflow, Model Evaluation, DSPy, RAG, Langgraph, LLMs, Supervised Fine-tuning, Reinforcement Learning (GRPO, PPO and DPO), ONNX, TensorRT, TorchServe, MLFlow, Prometheus, Grafana
- **Cloud:** AWS, GCP, Azure, Docker, Kubernetes, Terraform, Ansible, Jenkins, Github Actions
- **Development:** System Design, REST APIs, gRPC, Websockets, GraphQL, NodeJS, NextJS, Django, Flask, Node, Spring Boot
- **Databases:** MySQL, PostgreSQL, MongoDB, Redis, DynamoDB, Vector Databases (Qdrant, Chroma, Pinecone)

Education

- **New York University - Tandon School of Engineering** | New York, US | Sep 2024 - Expected May 2026
  Master of Science in Computer Engineering | GPA: 3.67
- **Dr. A.P.J. Abdul Kalam Technical University** | Delhi, India | Aug 2020 - Jun 2024
  Bachelor of Technology in Computer Science and Engineering | 1st Division with Distinction

Work Experience

Machine Learning Teaching Assistant | NYU Tandon School of Engineering | NY, USA | Sep 2025 - Present

- Architected and standardized PyTorch-based ML training pipelines (data loading, training, evaluation, visualization) used by 100+ students, improving code quality and experimental reproducibility.
- Led debugging and mentoring on deep learning systems (CNN optimization, backpropagation, autograd), helping students diagnose training failures and performance bottlenecks.
- Built high-impact ML learning content (videos and guides) reused across course offerings, significantly reducing common implementation errors and support overhead.

Co-Founder | Macverin Technologies | www.macverin.com | Jul 2022 - Aug 2024

- Led technical strategy and system design, architecting scalable distributed microservices using FastAPI for high-performance ML inference and Node.js for real-time websocket handling.
- Engineered full-stack architectures using JavaScript ecosystem (TypeScript, React, Next.js, Express), creating modular, type-safe interfaces backed by robust APIs to handle data-intensive workloads.
- Directed end-to-end product delivery and DevOps, managing complete SDLC from UI/UX design (Figma) to cloud deployment (AWS, Docker), and establishing automated CI/CD pipelines that ensured 99.9% system availability.

Projects

Deep Research Assistant (Hierarchical State Graph Architecture) | Nov 2025 – Feb 2025

- Designed a Hierarchical AI Agent with configurable inference depth (Deep/Broad/Fast), utilizing Task Decomposition to autonomously orchestrate research, synthesis, and writing sub-agents.
- Architected a Hybrid Search engine on Qdrant, utilizing Two-Stage Retrieval (Dense Semantic Search refined by BM25 Keyword Reranking) to maximize context precision.
- Developed Asynchronous Browser Agents using Headless Architectures (Playwright) to navigate complex single-page applications (SPAs) and handle dynamic user interactions.

Autonomous Multi-Agent Financial Research Platform (Finassistant) | live demo | code | Nov 2025 – Dec 2025

- Engineered a stateful LangGraph architecture simulating financial analyst workflows by orchestrating specialized Planner, Researcher, and Publisher agents to autonomously execute tools and self-correct.
- Implemented a Hierarchical RAG engine with Parent-Child Indexing to process massive SEC filings (10Ks), blending granular vector search with broader document context to prevent data hallucination.
- Built an event-driven asynchronous FastAPI backend with custom WebSockets to stream token-level reasoning to a React frontend, fully containerized via Docker and deployed on AWS.

Verifier-Guided Reasoning with Small Language Models | report | Sep 2025 – Dec 2025

- Developed a generator–verifier system to enhance mathematical reasoning, fine-tuning Phi-2 (2.7B) as a generator on GSM8K and training TinyLLaMA (1.1B) as a step-level verifier.

- Utilized QLoRA and Chain-of-Thought supervision to detect reasoning errors without modifying generator weights, achieving 64.1% Pass@1 accuracy with <0.3% trainable parameters.
- Demonstrated that optimized system design and step-level verification on the PRM800K dataset outperform larger models in reliable reasoning tasks under tight compute constraints.

Snap2Caption - ML Systems for Caption Generation | repo | Mar 2025 – May 2025

- Built an end-to-end image captioning service that generates Instagram-ready captions and hashtags via a FastAPI inference API with <2s P90 latency.
- Fine-tuned LLaVA-1.5/1.6 (7B) using LoRA on 100k urban images, enabling efficient multimodal training with reproducible, containerized workflows.
- Deployed and operated the system on GPU infrastructure, automating provisioning with Terraform and implementing MLflow-based experiment tracking and monitoring.