Technical Advisor Intern - GenAI

Ashwin A. Sharma

+1 929-245-5681 | sharmaashwin4000@gmail.com | github.com/ashwin-2001 |
linkedin.com/in/ashwinsharma2001
Work authorization: F-1 (NYU) — CPT eligible Fall 2025/2026 • NYC / Remote

Summary

MS Computer Science candidate at NYU with hands-on experience building generative AI (GenAI) systems and LLM tooling. Applied ML eng with production experience deploying GPT-4, RAG, and custom RL/CV solutions, plus cloud/DevOps skills for scalable workflows. Seeking Te Advisor Intern - GenAI to contribute to cutting-edge GenAI projects and support technical evaluation and advising.

Professional Experience

Founding ML Engineer & Tech Lead — Nextap AI • Apr 2023–Jul 2024

- Published Android GenAI app Artific AI (GPT-powered text, image, and code utilities): 1,500+ installs and 100+ DA
- Scaled team to eight and delivered 300+ client ML/cloud deliverables with 98% customer retention.

Senior Machine Learning Engineer — ViaLYTICS Consulting (Remote) • Aug 2022–Apr 2023

- Cut AWS EC2 idle spend by 19% across three regions using a PPO/DQN autoscaler; developed custom Gym simulat CloudWatch triggers, and CI/CD pipelines.
- Deployed Faster R-CNN + OpenCV traffic service (18 FPS, 97% precision) with JSON feeds powering Power BI dashboards for 12 municipalities.
- Built async video-screening tool reducing recruiter review time by 75%, enabling 3× candidate throughput.
- Developed crypto-portfolio optimizer (Sharpe 0.99; +80% annualized backtested return) using PyTorch/RLlib and sh REST SDK for daily rebalancing.

Founding ML Engineer — Nextap AI • Sep 2021–Aug 2022

- Bootstrapped AI tooling studio to five-figure USD revenue in six months; shipped Codilarity.com (React/Node/Fireba adopted by 2,000+ developers.
- Formed five-person core team and delivered 120+ ML & cloud proofs-of-concept for ed-tech and fintech clients.

Open Source & Research Projects

SnapPhil — AI Job-Application Assistant (Chrome Extension) • TypeScript, OpenAI API, GPT-4

- Auto-fills job forms and cover letters, reducing effort from ~45 min to <1 min; secure prompt engine with auth and st
- LLM match engine + Apps Script backend processes 900+ applications/day in beta; includes BRD, Cypress E2E, and CI/CD.

RAG Assistant for ROS2 Robotics • TensorFlow, Hugging Face, Gradio

- Fine-tuned a 4-bit LLM with 1,700+ instruction examples, reducing robotics development time by 40%.
- Implemented Vector DB + MongoDB retrieval achieving 95% accuracy with <500 ms latency for dev queries.

Autonomous-Driving Simulator + RL Agent • Python, Pygame, RLlib

- 2D simulator with LiDAR and procedural tracks; PPO policy outperformed DQN by +160% reward and generalized unseen maps with 94% success.
- Packaged as a pip module used in NYU RL labs.

Technical Skills

ML/CV: PyTorch | TensorFlow | scikit-learn | OpenCV | Faster R-CNN | RLlib (PPO, DQN) | LLM: GPT-4 | RAG | LangCh | Hugging Face | Cloud/DevOps: AWS EC2/S3/CloudWatch | Docker | GitHub Actions | Databases: MongoDB, PostgreSQL | Languages: Python, JavaScript/TypeScript, C++, SQL

Education

New York University — MS Computer Science • Sep 2024–May 2026 • GPA 3.77/4.00

University of Mumbai — BE Computer Engineering • 2019–2023 • CGPA 8.21/10

Publications