Yashvardhan Gupta

+14085910621 ◇ gupta.yashv@northeastern.edu ◇ San Jose, California, United States ◇ LinkedIn ◇ Github ◇ Website

EDUCATION

MS in Artificial Intelligence, Northeastern University (GPA: 4.0/4.0) — Sep '25 — Present, San Jose, CA, United States

- **Related Courses :** Machine Learning, Foundations of Generative AI, Reinforcement Learning
- **Teaching Assistant :** Reinforcement Learning (CS 5180) | Spring 2026

B.Tech in Mechanical Engineering, Delhi Technological University (GPA: 7.84 (out of 10)) — Aug '19 — Jul '23, Delhi, India

- **Related Courses :** Computer Vision, Machine Learning, Engineering Economics

TECHNICAL KNOWLEDGE

Languages : Python, TypeScript, Java, C++, SQL
ML & AI Frameworks : PyTorch, TensorFlow, JAX, Flax, Keras, Scikit-Learn
Generative & Agentic AI : Transformers, Hugging face, Diffusion models, RAG pipelines, FAISS, LangChain, CrewAI
Systems & Devops : Docker, FastAPI, Flask, CI/CD, multithreading/asyncio, GCP, AWS, CUDA, ONNX Runtime
Certifications : ML Specialization - Stanford University, TensorFlow: Adv. Techniques Specialization

WORK EXPERIENCE

Digital Solutions & Technology Engineer — Apr '23 — Mar '25
Biowolk Healthcare — Delhi, India

- Engineered a scalable analytics microservice using **Python (FastAPI)** and **Docker** to process daily sales and inventory data, effectively reducing manual reporting time by **15% (~7 hours/week)** via **end-to-end data pipeline orchestration**
- Utilized **Meta Business Suite** to execute high-precision targeted advertisement campaigns for > 65 pharmaceutical products, optimizing audience reach and conversion metrics through data-driven performance analysis.

Machine Learning Research Intern — Dec '21 — May '22
Tvishtryon Solutions Pvt. Ltd

- Architected the backend for a "Virtual Teacher" MVP, utilizing **Flutter** and **Python** to deliver interactive lessons with dynamic content generation. Optimized real-time video/audio streaming pipelines, achieving low-latency response flows.
- Deployed pre-trained transformer models for on-the-fly media generation, reducing inference latency by **20%** through model quantization techniques. Delivered a cost-efficient MVP for educational solutions aligning with ROI objectives.

PROJECTS

Read my lips (Visual Speech-to-Avatar Interface) , Northeastern University Link — Sep '25 — Dec '25

- Architected a multimodal assistive pipeline converting silent lip movements into synthesized speech and synchronized avatars using **Auto-AVSR**, **Qwen 0.6B** (for semantic error correction), and **FLOAT** (Flow Matching).
- Engineered the system for **Apple Silicon (MPS)** by porting CUDA-centric generative models and optimizing tensor operations, achieving **~200ms** VSR inference latency **while maintaining high-fidelity output for real-time applications**.
- **Security Protocol:** Implemented enterprise-grade security protocols including air-gapped local processing, confidence-based user verification thresholds (<0.45), and granular interaction auditing to protect sensitive user intent.

Real Time Speech, Northeastern University Link — Sep '25 — Oct '25

- Developed a low-latency **WebRTC** and **Python** pipeline for real-time, full-duplex browser-to-server audio streaming.
- Integrated a **VAD** and **ONNX**-optimized speech enhancement model to ensure real-time inference on consumer hardware.
- **Security & Monitoring:** Developed a real-time telemetry dashboard to monitor per-stream latency, packet loss, and **Signal-to-Noise Ratio (SNR)** while ensuring containerized isolation via **Docker**.

AI Lawyer Link — Mar '25 — Jun '25

- Developed a **Google Gemini** and **FAISS**-based RAG system, increasing legal answer relevance by **30%**.
- Optimized retrieval performance through hybrid search and batch processing, successfully reducing search latency by **50%** for complex legal document queries **while ensuring high precision through re-ranking and metadata filtering**.
- **Security & Privacy:** Engineered a secure "document-vault" microservice featuring **end-to-end encryption**, **AES-256**standards, and **Role-Based Access Control (RBAC)** to protect sensitive legal drafts and audit logs.

PUBLICATIONS

Vision Language Models : A complete survey (Ongoing) — Dec '25

This survey reviews Vision–Language–Action models that combine visual perception, language grounding, and action generation for robotics. The paper identifies strengths, gaps, and opportunities for building next-generation embodied systems.

An End to End Solution to Automated Hiring — Dec '22
IEEE

Proposed and evaluated a GAN-, NLP-, and CV-driven e-recruitment platform automating resume short-listing, deepfake-simulated interviews with dynamic question generation, and CV-based proctoring to accelerate hiring.
https://ieeexplore.ieee.org/document/10060436