Anuj Gupta

412-224-3076 | anujg2@cs.cmu.edu | linkedin.com/in/anuj-gupta-2k | github.com/Anuj-G-06 | anuj-m-gupta.vercel.app

EDUCATION

Carnegie Mellon University, School of Computer Science Pittsburgh, PA
Master of Computational Data Science *August 2025 – December 2026*
Courses: ML (PhD), LLM Systems, Cloud Computing, AI Venture Studio; Swartz Fellow; Applied ML Teaching Fellow.

Dwarkadas J. Sanghvi College of Engineering Mumbai, India
Bachelor of Engineering in Electronics and Telecommunication *August 2018 – July 2022*
Specialization in Artificial Intelligence and Machine Learning - I.B.M. (GPA: 9.52/10)

TECHNICAL SKILLS

Languages: Python, C++, SQL, R, JavaScript (Node.js, TypeScript), HTML5, CSS3, Java
Libraries: verl, vLLM, Jax, PyTorch, Tensorflow, HuggingFace, Networkx, Dask, Openpyxl, FastAPI, Pandas, NumPy
MLOps / DevOps: Azure, Snowflake, Docker, W&B, Apache Kafka, Git, GitHub
Cloud Certifications: AWS Certified Solutions Architect - Associate (link), GCP Certified Associate Cloud Engineer (link)

PROFESSIONAL EXPERIENCE

TEEL Lab, *Research Assistant,* Pittsburgh, Pennsylvania *August 2025 – Present*

- Formulated multidomain benchmark with reinforcement learning for semantic annotation of long documents, introducing real-world evaluation criterions missing in academic baselines, enabling stronger alignment for LLMs.
- Designed a modular CLI evaluation to standardize LLM benchmarking, implementing automated prompt optimization and majority vote ensembles to validate model reasoning capabilities.

Equifax, *Data Scientist,* Mumbai, India *August 2023 – July 2025*

- Automated financial analysis reports in LangGraph, crawling large scale client centric documents and generating multi step research with RAG microservices to improve reporting time by 90% across 300+ projects.
- Evaluated Random Forest, GBDT, and XGBoost architectures, deploying a 87.6% accuracy SHAP-enhanced pipeline on Vertex AI to resolve black-box interpretability in credit risk, impacting $45.5B across portfolios.
- Managed 4 contractors for system transformation on Linux-based GCP environments, optimizing workflows with Bash, BigQuery, Git, and Cloud Dataflow, reducing analytics costs by 40% and speeding up runtimes by 27%.

Quantiphi, *Machine Learning Engineer,* Mumbai, India *August 2022 – July 2023*

- Customized Document AI with YOLOv8 on distributed PyTorch DDP, improving multilabel classification and segmentation accuracy by 25% on 1,000+ financial tax documents.
- Engineered scalable ETL data cleaning pipeline using Selenium, and Airflow, reducing 2TB of raw enterprise data to 73GB for entity extraction pipelines with a throughput of 100K requests/sec (RPS).

Expify Pvt Ltd, *Founding Software Engineer,* Mumbai, India *July 2021 – July 2022*

- Developed a Word2Vec clustering algorithm on proprietary psychometric test datasets with Gensim, NLTK and Scikit-learn, boosting the customer academic inclination prediction metrics to 79% precision.

Nippon Asset Management, *Data Analyst Intern,* Mumbai, India *March 2021 – July 2021*

- Customized a 340M parameter pre-trained BERT transformer with NER embeddings in Tensorflow for automated stock price to sentiment correlation analytics, improving equity research signal quality by 1.5%.

PROJECTS

NeuraLume (link) *December 2024 – Present*

- Integrated multimodal agentic platform on AWS EC2, utilizing Langflow and FastAPI to orchestrate complex generative workflows, creating scalable infrastructure for the $150B AI content market.
- Engineered low friction product modules by conducting A/B testing on VFX interaction patterns, securing initial commercial traction from 5 to 50 paying customers.

EqRAG (github link) | *Advisor: Dr. Jaromir Savelka* *August 2025 – December 2025*

- Finetuned Qwen with LoRA on AWS A100 EC2 to engineer a RAG trading agent, achieving 51.7% profitability by benchmarking fine-tuning strategies against DeepSeek to enable real-time "buy/sell/hold" reasoning.

RedViz (github link) | *Advisor: Adam Perer* *August 2025 – December 2025*

- Constructed a unified red-teaming framework using Streamlit and Altair to accelerate safety alignment, combining attention-map interpretability with realtime multilingual harm detection across diverse foundation LLMs.