Anay Gupta

469-486-2271 | anaygupta2004@gmail.com | https://www.linkedin.com/in/a-gupta123/ | https://anaygupta2004.github.io

Education

Georgia Institute of Technology Atlanta, GA
Masters of Science in Computer Science [Machine Learning Specialization] *Aug 2025 – May 2026*

Georgia Institute of Technology Atlanta, GA
Bachelor of Science in Computer Science [Machine Learning Specialization]: 3.93/4.00 *Aug. 2022 – May 2025*

Skills

Languages: Python, SQL, Java, JavaScript
Frameworks: PyTorch, DeepSpeed, React.js, Node.js, Next.js
Developer Tools: Git, Matplotlib, Jupyter, Pandas, Weights & Biases, AWS
Recognitions: 2022 Barry Goldwater Scholar, 2023 Google CSRMPa Recipient, 2024 Bessemer Fellowship Finalist

Experience

Microsoft Research Cambridge, MA
Research Intern *May 2024 – Aug. 2024*

- Engineered custom PyTorch dataloaders to process 200GB datasets, increasing training efficiency by 125%.
- Implemented data-driven tokenization algorithms, tokenizing 150B tokens, resulting in 1000+ hours of cloud compute time saved and over $25,000 in cost reduction.
- Developed pretraining code for a 650M parameter transformer model for internal purposes with distributed training and GPU optimization.

Delta Air Lines Atlanta, GA
AI Software Engineer Intern *Jan. 2024 – Apr. 2024*

- Developed the first evaluation system for LLM answer generation in a Retrieval Augmented Generation (RAG) pipeline on AWS via Python and TypeScript, serving 20,000 customer service agents internationally and handling 100,000 AI responses.
- Improved LLM effectiveness with prompt engineering techniques, reducing OpenAI operational costs by $5,000.

Genentech San Francisco, CA
Artificial Intelligence / Machine Learning (AI/ML) Software Intern *May 2023 – Oct. 2023*

- Applied advanced AI architectures, including Language Models and Convolutional Neural Networks, to develop a novel software framework for gene therapies. Currently used by 15+ ML scientists.
- Increased evaluation accuracy by 25% through statistical methods and advanced machine learning techniques.

Broad Institute MIT and Harvard Cambridge, MA
Research Intern *Jun 2021 - May 2023*

- Spearheaded a single-cell bioinformatics pipeline for 1M datapoints to analyze cancerous tissue in 20 patients on Google Cloud Platform with Python and R.
- Identified 100+ cancer-resistance-causing genes, pathways, and transcription factors to determine cancer drug targets for treatment.

Projects

Anay-ssistant | *Python, Flask, Node.js, React.js, Docker, OpenAI* August 2023 – October 2023

- Developed a personalized LLM-powered voice assistant with a React.js user interface, utilizing OpenAI API to interpret and respond to user queries, integrated with Google Workspace APIs / Canvas APIs to manage 300+ monthly events and academic deadlines
- Implemented Hark and Whisper for real-time voice recognition and translation, allowing the assistant to effectively process and respond to spoken commands in <3.5 seconds

arXiv Paper Recommendation | *Beautiful Soup, Firebase, Next.js, Git, OpenAI* Aug 2024 – Present

- Developed a Next.js website where users can fine-tune their paper references by annotation to receive personalized recommendations using LLMs
- Implemented a pseudo-active learning approach to improve the recommendation system by leveraging user feedback and adaptive algorithms for more tailored research suggestions