Abhinav Gupta

New York, NY | ag10706@nyu.edu | linkedin.com/in/abg0148 | github.com/gptabhinav

Education

New York University **New York, NY**
Master of Science in Computer Science *Expected Graduation: May 2026*

Thapar Institute of Engineering and Technology **India**
Bachelor of Engineering in Electronics and Communication *Aug 2016 - July 2020*

Technical Skills

Programming & Scripting: Python, C++, Java, Scala, SQL, HTML, CSS
Machine Learning & AI: PyTorch, TensorFlow, scikit-learn, HuggingFace Transformers, LLMs, LoRA, LangChain, LangGraph, RAG, Vector DBs, ANN Search (HNSW, DiskANN, NSG, HCNNG, Parlay), OpenCV, YOLO
Data & Backend Systems: Apache Spark, Hadoop, ETL/ELT pipelines, REST/gRPC/GraphQL APIs, Distributed Systems
Databases & Cloud: Postgres, Snowflake, MongoDB, ChromaDB, Apache Airflow, DataHub, Kafka, AWS (EC2, S3, Lambda, DynamoDB, Elasticsearch, CloudWatch, CloudFormation), GCP, Docker, Kubernetes, Unix/Linux, Database Modelling, Schema Design

Experience

New York University **New York, NY**
Research Assistant *May 2025 – Present*

- Working on an ongoing paper to improve navigability in vector databases for ANN search, which enhances the recall-latency trade-off in vector databases.
- Wrote a high-performance C++ implementation of HNSW and Vamana, using SIMD vectorization and OpenMP to optimize ANN experiments, achieving state-of-the-art construction of 1M-point datasets in under 7 mins with 8 threads.

BlackRock **Gurugram, India**
Senior Data Engineer *Jan. 2023 – June 2024*

- Designed and deployed end-to-end Python backend APIs for time series ingestion and registration, migrating the metadata store from Snowflake to DataHub to support the next-generation enterprise metadata catalog.
- Designed Snowflake data warehouse schemas and data models for downstream analytics and reporting, improving query performance and reducing ETL complexity for 100+ workflows.
- Led development of bulk ticker ingestion pipelines using Postgres to onboard over 19k+ tickers for Haver's G10 dataset, enabling the first private preview of BlackRock's Quant SDK product.
- Collaborated cross-functionally to deliver liquidity data, managing ETL priority queues, meeting customer needs by ensuring timely data delivery, and effectively handling stakeholder communication.

Data Engineer *July 2020 – Dec. 2022*

- Built scalable ingestion platforms on Snowflake to support model feed growth, enhancing onboarding speed for various modeling teams.
- Optimized ETL processes using Python and Scala to efficiently handle TBs of daily data, reducing peak resource utilization.
- Migrated ETL job orchestration from legacy systems to Airflow, reducing failures by 25% and improving system monitoring.

Intern *Jan. 2020 – July 2020*

- Created a Python Flask server to automate file-triggered ETL jobs, accelerating data processing times.
- Developed APIs using gRPC, Protocol Buffers, and GraphQL, optimizing service-to-service communication in scalable solutions.

Projects

Probabilistic Baseball Trajectory Estimation | *YOLOv4, OpenCV, PyTorch, TensorFlow* *November 2025*

- Engineered a computer vision pipeline to predict baseball trajectories from MLB video, leveraging object detection, multi-frame tracking, and Kalman Filter–based state estimation for handling uncertainty.
- Trained and deployed a custom YOLOv4 strike-zone detection model, publicly released on Roboflow, with 98%+ recall, precision, and mAP@50, to enhance accessibility and usability.

Pinboard | *Flask, PostgreSQL, Data Modelling* *April 2025*

- Designed and implemented a Pinterest-style discovery engine, focusing on efficient database modeling and schema design for complex entities.
- Built features like repinning, liking, commenting, follow streams for personalized feeds and soft deletion to maintain data integrity.

FunnelPulse | *PySpark, Apache Kafka, GCP, Streamlit* *November 2025*

- Built a real-time streaming pipeline on GCP using Dataproc and Kafka, processing 8.7M+ e-commerce events to monitor conversion funnels and implementing a Lakehouse architecture for efficient data management.
- Implemented Z-score-based anomaly detection with an LSTM Autoencoder and Prophet, enhancing changepoint detection to handle cyclic patterns in conversion data.

Indexing for Scalable Sparse Neural Retrieval Systems

- Evaluated static pruning and tiering strategies on pruned DeepImpact indexes, improving retrieval latency on MS MARCO while preserving ranking quality.

CodeGuardian | *RAG, Static Analysis, On-Device Security*

- Developed an on-device vulnerability analysis tool combining static scanning with an LLM-powered RAG pipeline, producing interactive reports and chatbot-style explanations of risks in large codebases.

Smart Photo Album | *AWS Rekognition, ElasticSearch, AWS Lex*

- Built a web app using AWS Rekognition and ElasticSearch for image embeddings and semantic search, integrated AWS Lex for natural language queries to organize and retrieve photos at scale.

Dining Concierge | *AWS Lex, DynamoDB, SQS, Chatbot*

- Implemented a serverless restaurant recommendation chatbot using Amazon Lex, DynamoDB, and SQS, enabling scalable intent handling and automated email notifications.