

# Rajeev Gundavarapu

rajeevgu97@gmail.com · +1 949-245-4772 · LinkedIn · GitHub · Legal Status: US Citizen

## Engineering Manager | Applied AI & Generative AI Architecture | Machine Learning

Technology leader and Engineering Manager with 7+ years of experience driving enterprise technology transformation, architecting scalable data platforms, and leading complex engineering initiatives from strategy through delivery. Currently spearheading the architecture and development of Capital One's conversational GenAI reporting platform to transform financial reporting and analytics. Further advancing expertise in artificial intelligence through graduate studies in Generative AI and Large Language Models at Carnegie Mellon University.

### Experience

---

#### Software Engineering Manager (Lead Software Engineer)

Jun 2026–Present

*Capital One*

*New York, NY*

- Personally architect DataAI, an in-development enterprise conversational reporting solution integrating LangChain, RAG, vector retrieval, and financial data to support grounded natural-language reporting and account, transaction, and balance insights; guide evaluation of GPT-OSS-120B, Gemma 4 26B, and Nemotron 3 Nano 30B.
- Lead the engineering team delivering DataAI and reporting capabilities; set engineering priorities and technical direction, coordinate implementation, resolve design challenges, and own delivery of engineering initiatives.
- Architected streaming ingestion from relational databases and the data lake into AWS OpenSearch; optimized mappings and Elasticsearch queries for filtering, sorting, and retrieval, reducing p95 retrieval time by more than 50%.
- Direct development of reusable Python reporting infrastructure and AWS Lambda processing components for enterprise reporting workflows.

#### Principal Associate Software Engineer

Jun 2023–Jun 2026

*Capital One*

*New York, NY*

- Developed NLP sentiment-classification models to identify customer complaints in call transcripts.
- Architected the enterprise Python/FastAPI Transaction View API on PostgreSQL and AWS Fargate as the single source of truth for transaction and balance data, adopted by five mobile, web, and analytics teams.
- Pioneered reporting architecture and PostgreSQL schemas; led legacy Oracle migration to PostgreSQL using AWS DMS and Lambda transformations.

#### Senior Associate Software Engineer

Jun 2021–Jun 2023

*Capital One*

*New York, NY*

- Built a Python ingestion service that transformed upstream files into byte streams and loaded them into the data lake with AWS Lambda, S3, Step Functions, and Boto3.
- Designed the target architecture and led data-pipeline migration to a fully serverless AWS platform, removing server-managed infrastructure and reducing operating costs; automated sensitive-data tokenization and detokenization.

#### Global Technology Analyst

Jul 2019–Jun 2021

*Bank of America*

*New York, NY*

- Built a Python/Pandas/Scala data-lineage application to trace dataset relationships and visualize lineage in a web interface.
- Prepared regulatory data for Federal Reserve CCAR reporting; designed Scala/RDF metadata schemas and publishing capabilities, and built R metrics dashboards.

## Additional Experience

---

### Data Science Teaching Assistant

Aug 2020–Feb 2021

*Fu Foundation School of Engineering and Applied Science, Columbia University*

*New York, NY*

Supported fintech instruction in time-series analysis, machine learning, NLP, AWS, applied statistics, and exploratory data analysis.

### Software Engineering Intern

Jul 2018–Aug 2018

*Cytocybernetics*

*Buffalo, NY*

Improved mathematical modeling algorithms and contributed to SC1/DC1 interfaces using GTK.

## Projects

---

### Pythia-1B Fine-Tuning and Tool-Augmented Reasoning

Fine-tuned EleutherAI Pythia-1B on ASDiv with rank-16 LoRA and integrated a constrained arithmetic calculator. Accuracy was 63.1% with calculator access versus 39.5% without (+23.6 points); this measures tool access, not fine-tuning alone.

### Decoder-Only Transformer Language Model

Implemented GPT-style decoder blocks with causal multi-head attention, residuals, layer normalization, and GELU feed-forward layers; trained an autoregressive LM and implemented BPE tokenization, temperature sampling, and perplexity comparison with Pythia-1.4B.

### Common Crawl Language-Model Data Pipeline

Built a streaming Hugging Face dataset loader for Common Crawl WARC pages; extracted visible HTML text, filtered low-quality content, masked phone numbers and SSNs, and yielded accepted URL/text records.

### Airline Passenger Satisfaction Classification

Prepared 103,904 records with one-hot encoding and 5-neighbor KNN imputation; compared Logistic Regression, AdaBoost, and Random Forest. Notebook reports Random Forest test accuracy of 96.36% and F1 of 0.958.

### IPL Auction Player Valuation and Performance Segmentation

Joined IPL batting/bowling statistics with salary data; engineered performance features, compared regression models using cross-validation, and segmented players with standardized K-Means and PCA visualization.

### Political Tweets and News Text Analysis

Collected candidate tweets and New York Times articles; removed retweets and duplicates, cleaned/stemmed text, and analyzed word frequencies and co-occurrence.

### Housing Recommendation System

Built an interactive county recommender ranking locations by cosine similarity to user-weighted preferences across seven factors, with supporting visualizations.

## Education

---

### Graduate Specialization in Generative AI and Large Language Models, GPA: 4.0

*Carnegie Mellon University (In Progress)*

Curriculum: LLM methods and applications, multimodal machine learning, and LLM systems.

### Master of Science in Analytics, GPA: 3.83

Jan 2020–Dec 2022

*Georgia Institute of Technology*

### Bachelor of Science in Computer Science, GPA: 3.82, Summa Cum Laude

Aug 2017–May 2019

*University at Buffalo, SUNY*

## Technical Skills

---

**Languages:** Python, SQL, Scala, R, C.

**ML and LLMs:** PyTorch, scikit-learn, Hugging Face Transformers, PEFT, LoRA, NLP, classification, model evaluation, tokenization.

**Generative AI:** LangChain, retrieval-augmented generation (RAG), embeddings, vector databases and search, tool use.

**Backend and data:** FastAPI, Pandas, ETL, data modeling.

**Cloud:** AWS Lambda, S3, EC2, Fargate, Step Functions, Boto3, DMS, Amazon Aurora PostgreSQL, AWS OpenSearch.

**Databases/search:** PostgreSQL, Oracle, Elasticsearch.