Gopala Krishna Abba

New York, NY | (929) 754-6845 | gopala.krishna.abba@gmail.com | linkedin.com/in/igopalakrishna | github.com/igopalakrishna

Education

New York University — New York, NY
Master of Science in Computer Engineering | GPA: 4.0/4.0 — *Sep 2024 - May 2026*
***Relevant Coursework**: Machine Learning, Big Data, Deep Learning, Database Systems (Distributed Systems), High Performance Computing, Computing Systems Architecture, Real-Time Embedded Systems*

National Institute of Technology Rourkela — Rourkela, India
Bachelor of Technology in Electronics and Communications Engineering | GPA: 8.03/10 — *Nov 2020 - May 2024*

Technical Skills

Languages: Python, SQL, C/C++, TypeScript, JavaScript
ML/AI: PyTorch, TensorFlow, scikit-learn, Hugging Face Transformers, NLP, LLMs, Generative AI, Agentic AI, LangChain, LangGraph, RAG, Vector Databases, Embeddings, Semantic Search, Recommendation Systems, Prompt Engineering, OpenAI API
Data: Pandas, NumPy, Apache Spark, PySpark, ETL, Data Pipelines, Feature Engineering, Model Evaluation, Statistical Modeling
Full-Stack (Backend & Frontend): FastAPI, Flask, Django, Node.js, REST APIs, React, Next.js, Streamlit, HTML, CSS
DevOps/MLOps: Docker, Kubernetes, Linux, Git, GitHub Actions, Jenkins, MLflow, Airflow, DVC, Weights & Biases, CI/CD
Cloud: AWS (EC2, S3, Lambda, SageMaker, Bedrock), Google Cloud Platform (GKE, GCR), Databricks (Delta Lake, Unity Catalog)
Databases: MySQL, MongoDB, PostgreSQL

Experience

Data Science Intern (Emerging Tech) — Feb 2026 - Apr 2026
Fox Corporation (FOX Tech) — *New York, NY*
- Built a 9-stage, 31-parameter production Databricks pipeline using PySpark, SQL, Delta Lake, Unity Catalog, and BGE embeddings to automate nightly prior-day primetime editorial recaps across 4 programs and generate 3 bronze tables.
- Developed a 2-stage ML ranking system with hybrid semantic agglomerative clustering, key-phrase and entity penalties, dynamic distance threshold, 5-factor story scoring, and 4-factor clip scoring to select up to 3 stories and 2–3 clips/story.
- Deployed a Streamlit Databricks App with SSO-aware UI, story explanations, clip rationale, LLM-generated SEO metadata, embedded video playback, and validated 32-42s runs across 14 dates, 41 stories, 110 clips, 14 playlists, and duplicate-free writes.

Graduate Research Assistant, Chunara Lab — Jan 2026 - Present
New York University — *New York, NY*
- Built a reproducible text data pipeline across 10 major U.S. newspapers, processing 7k+ articles per week with a 0.95 deduplication threshold to produce analysis-ready corpora with provenance and rerunnable jobs.
- Designed compliant data acquisition and storage architecture with schema and metadata, enabling scalable ETL and efficient longitudinal querying through a 7-table SQL schema, 8 Typer CLI commands, and 23 publication normalization rules.
- Prototyped LLM-assisted labeling to extract entities, topics, frames, and sentiment signals using JSON Schema validation, checkpointing every 50 articles, and rate limits of 30 requests per minute and 100k tokens per minute.

Software Engineer Intern – AI/ML & Data Infrastructure — Sep 2025 - Dec 2025
Global Futures Group — *New York, NY*
- Designed and deployed a production-grade recommendation & content-ranking system powering an AI expert-matching platform (Next.js, TypeScript, FastAPI, Python, Postgres, Prisma, Docker), serving 12k+ profiles with p95 search latency under 350 ms.
- Implemented a hybrid retrieval-ranking pipeline using sentence-transformers all-MiniLM-L6-v2, FAISS ANN, BM25, geo filters with runtime-tunable weights, BM25-only fallback mode, and automated index rebuilds under 60s.
- Containerized a 3-service stack with health checks (web, API, DB), seed/admin CLI, and 20+ tests (Vitest, Playwright, pytest), added profiling, structured logging, and performance tracing for debugging ranking failures and slow queries.
- Deployed the frontend on Vercel and API/database on Railway, resolving build and orchestration issues and stabilizing deployments to achieve 99%+ uptime, while cutting cold-start latency by 30%+ via ranking-path, caching, and query-plan optimizations.

Graduate Research Assistant, DICE Lab — May 2025 - Dec 2025
New York University — *New York, NY*
- Fine-tuned DyT-modified DistilGPT2 and Pythia models (17M to 410M) for generative NLP tasks using PyTorch and LoRA, reducing trainable parameters by 3.2x and validation loss by over 87%.
- Designed an evaluation pipeline with HuggingFace Trainer and Weights & Biases that revealed a 21% generalization gap for DyT under PEFT compared with LayerNorm baselines.
- Tuned DyT scale parameter to reduce the gap by approximately 28% and improve convergence consistency across model scales.

Projects

FuseRank – Hybrid Recommendation System (Collaborative and Content-Based)
Tech stack: TensorFlow, Flask, GCP, Docker, Kubernetes (GKE), DVC, CometML — May 2025 - Jun 2025
- Implemented a hybrid recommender system on 50M+ interactions combining matrix factorization and TF-IDF similarity, trained a TensorFlow model with early stopping and LR scheduling to reach MAE 0.1863 and MSE 0.0727 in under 20 minutes.
- Deployed a scalable Flask-based prediction service on GKE using Docker and Horizontal Pod Autoscaling, reduced p95 serving latency by 30%, and enabled autoscaling from 2 to N pods based on load.
- Versioned data and models with DVC and GCS, and tracked experiments in CometML, improving recommendation quality by 17% over a KNN baseline (A/B testing on held-out users)

Open Source

SBSim (google/sbsim) – Open Source Contributor *(PR #113, #114 – Python, CI/tests, RL simulation)* Sep 2025 – Present