# Bharath Mandha

AI/ML Engineer, 4+ years in production ML; RAG and LLM evaluation at Goldman Sachs
Dallas, TX | bharathmandha.200@gmail.com | linkedin.com/in/bharathmandha | github.com/bharathkumardev1 | bharathkumardev1.github.io

## EXPERIENCE

**Goldman Sachs**, AI/ML Engineer — Jun 2025 to Present

- Co-designed and optimized the retrieval layer (metadata filtering, reranking) and the LangGraph orchestration layer of a LangChain, LangGraph and FAISS RAG platform over 3M+ internal financial documents (tens of millions of chunks) used weekly by 70+ credit and risk analysts, and built its FastAPI integration, working with platform, risk and compliance teams
- The platform's automated evidence discovery, document analysis and validation, with analyst review preserved, is reported to have cut average credit-memo preparation time by about 40% (project-level estimate, not my own time study)
- Built a 500+ query RAG evaluation benchmark (factual retrieval, financial comparisons, evidence-based reasoning) scoring retrieval with Hit Rate and MRR and generation with RAGAS faithfulness, groundedness and answer relevance (end-to-end workflow: 10 to 20 s); used it to compare chunk size, overlap, top-k and embedding choices experiment by experiment; groundedness above 90%
- Built LangGraph agents for retrieval, analysis, validation and delegation that took over repetitive credit-analysis reasoning steps, with analysts reviewing sensitive decisions
- Implemented document preprocessing, recursive and semantic chunking, metadata enrichment, embedding generation and incremental FAISS index updates in the team's PySpark and AWS S3 ingestion pipeline, so routine updates skip a full rebuild
- Evaluated quantization and LoRA/QLoRA approaches for open-source models with Hugging Face Transformers and PEFT, weighing GPU memory, inference latency, quality and serving cost, and concluded that infrastructure savings alone didn't justify moving complex credit reasoning to smaller models

**Cigna Healthcare**, Machine Learning Engineer — Mar 2022 to Aug 2023

- Built 30-day hospital readmission-risk prediction with XGBoost and LightGBM over 1.5M+ HIPAA-regulated encounters, reaching 0.91 ROC-AUC on a held-out cohort of 200K+ encounters from 70K+ patients with time-aware, patient-level splits (reported internal result on private data, not reproduced in my public projects)
- Engineered 150+ candidate features from labs, demographics, encounter history and clinical notes; about 100 kept in production for predictive value and stability; handled class imbalance and tuned the decision threshold
- Productionized training, batch inference, registration and retraining (MLflow, Airflow) on SageMaker and EKS through CI/CD, using Terraform for infrastructure as code, with PSI drift monitoring; SHAP explanations let clinical reviewers prioritize several thousand high-risk patients monthly

**Best Buy**, Associate Data Scientist — Jun 2020 to Feb 2022

- Cut weighted MAPE of SKU- and store-level demand forecasts by about 14% relative with gradient-boosted models benchmarked against time-series baselines on time-based holdouts, tracking RMSE alongside MAPE
- Built PySpark, SQL and BigQuery pipelines over 50M+ transactions across thousands of SKUs and hundreds of stores; raised response on comparable campaigns by about 12% with six customer segments from clustering and cohort analysis

## PROJECTS (CODE, RESULTS AND CI PUBLIC ON GITHUB)

**Agentic-Research-Assistant**, does a review loop beat a single pass? — 2026

- LangGraph loop that reviews its own cited answers, measured on 150 QASPER test questions with every model call recorded so CI replays the evaluation without an API key. It ties a single pass over 5 passages (58.7% vs 56.0% judged correct) and loses to a single pass over 10 or 15 passages (68.0%); the matched baselines were added after seeing test results. Judge checked against 40 answers I graded myself (87.5% agreement, one grader)

**finance-rag-eval**, retrieval over SEC filings — 2026

- Page-level retrieval on FinanceBench (150 questions, 360 filings, 53,399 pages) with a held-out split and paired bootstrap intervals: a company and period filter (+0.29) and statement-page routing (+0.28) take Hit@10 from 0.14 to 0.71; a MiniLM hybrid tied BM25 on test (0.73 vs 0.75) at about 5x the latency, so the service runs BM25 at 12 ms p95

**readmission-capacity**, outreach planning on public hospital data — 2026

- Readmission models judged at a fixed call capacity on 101,766 UCI encounters split by patient: a 10% LightGBM call list holds 25.0% of 30-day readmissions (95% CI 23.2 to 26.7%) vs 20.5% for prior stays and 15.8% for LACE; TreeSHAP reason codes; CI retrains and fails if headline metrics move

## SKILLS

**Languages and data:** Python, SQL, PySpark, BigQuery
**ML and LLMs:** PyTorch, scikit-learn, XGBoost, LightGBM, SHAP, Hugging Face, LoRA/QLoRA, LangGraph, LangChain, FAISS, RAGAS
**MLOps and cloud:** MLflow, Airflow, FastAPI, Docker, Kubernetes, Terraform, CI/CD, AWS (SageMaker, S3, EKS), GCP Vertex AI

## EDUCATION

**University of North Texas**, Master of Science in Computer Science — Aug 2023 to May 2025