Aksh Talati

Boston, MA | (857) 867 2050 | talati.ak@northeastern.edu | LinkedIn: akshtalati | GitHub: akshtalati

EDUCATION

Northeastern University, Boston, MA - Master of Science in Information Systems **Expected May 2026**

Dharmsinh Desai University, India - Bachelor of Technology in Information Technology **May 2024**

SKILLS

Programming: Python, SQL, Bash, Java

Data Engineering: Airflow (Cloud Composer), Azure Data Factory, Snowflake, Databricks (Bronze→Silver), Spark, DVC

AI & ML: RAG Pipelines, LLM Agents (ReAct, MCP), LangChain, LangGraph, Embeddings, Instructor/Pydantic, PyTorch, TensorFlow

Cloud: Google Cloud (Cloud Run, GCS, IAM), Azure, AWS (S3/EC2 basics)

Databases: Snowflake, PostgreSQL, MySQL, MongoDB, Qdrant, ChromaDB

Tools: Docker, FastAPI, Streamlit, Selenium, Git/GitHub, pdfplumber, Docling

Concepts: ETL/ELT, Dimensional Modeling, Schema Design, Vector Search, Data Pipelines, Observability

PROFESSIONAL EXPERIENCE

Research Assistant - Data Engineering | Northeastern University **Sep 2025 – Present**

- Engineer Python pipelines to process 600+ unstructured healthcare cases (4,000+ pages) into organized, analysis-ready outputs.
- Build a categorization & metadata engine that reduces manual review time by 50–60%.
- Develop RAG + agentic workflows for automated case lookup, summarization, and document intelligence.
- Implement statistical and rule-based validation to ensure accuracy, reproducibility, and dataset quality.

Teaching Assistant - Database Management | Northeastern University **Sep 2025 – Present**

- Teach SQL, schema design, joins, constraints, and normalization to 40+ students per semester.
- Create core course content — labs, quizzes, PPTs, SQL scripts, case studies, and project structure.
- Support student debugging, evaluate submissions, and hold weekly office hours for analytical problem-solving.
- Guide students in practical, engineering-first thinking for real-world database and BI workflows.

Research Intern | IRLAB (Dhirubhai Ambani University) **Jan 2024 – May 2024**

- Built multilingual retrieval pipeline (PyTorch/TensorFlow) across 4 languages and 501 samples.
- Evaluated using embeddings + R@1/R@100 (~73%+).
- Contributed to dataset and model framework showcased at FIRE Summit 2024.

PROJECT

ORBIT AI-50 Intelligence | Github

Python, Airflow, Cloud Composer, FastAPI, Streamlit, Qdrant, GCP, LLM Agents, LangChain, ChromaDB

- Engineered dual pipelines (RAG + structured extractors) to automate due-diligence profiles for all 50 Forbes AI-50 companies.
- Engineered Airflow ingestion pipelines and Qdrant vector indexes, powering agentic LLM workflows for real-time summaries and automated reporting.
- Deployed FastAPI + Streamlit stack on Cloud Run, delivering a scalable intelligence dashboard used for automated research and competitive analysis.

Aurelia Financial RAG | Github

Python, FastAPI, Streamlit, Airflow, Cloud Run, ChromaDB, text-embedding-3-large, Instructor, GCP

- Built a cloud-native RAG microservice that parses a 4,000-page Financial Toolbox, performs code-aware chunking, and generates embeddings for concept retrieval.
- Built an Instructor-based FastAPI service for structured financial notes with integrated citation validation and fallback retrieval logic.
- Automated the entire ingestion → chunking → embedding workflow via Cloud Composer to support scheduled weekly updates.

Report-Intelligence | Github

Python, Docling, pdfplumber, DVC, GCP, OpenAI, Metadata Validation, Chunking Frameworks

- Built a fault-tolerant parser that extracts text, tables, and metadata sections from SEC filings using Docling + pdfplumber with XBRL-aware validation.
- Created a modular chunking and validation framework that supports downstream RAG workflows, financial analytics, and automated narrative generation.
- Implemented data versioning (DVC) and validation checks to ensure audit-ready, reproducible financial document processing at scale.

Dow-30 IR Scraper | Github

Python, Selenium, Airflow, GCS, Metadata Normalization, Heuristic Recovery

- Built a resilient scraper that extracts investor-relations filings across all Dow-30 companies with fallback heuristics and anti-crash logic.
- Standardized metadata, removed duplicates, and automated ingestion through Airflow DAGs for scheduled refresh.
- Enabled centralized indexing of IR documents for downstream analytics, research pipelines, and reporting.