

Rihua Van Steenburgh

Data Scientist | Applied AI, NLP & Model Evaluation

Alpharetta, GA | 470-295-0545 | rihuavan@gmail.com

github.com/rihua-tech | linkedin.com/in/rihua | rihua.dev

Summary

Data Scientist with hands-on project experience in NLP classification, model evaluation, RAG retrieval, and applied machine learning. Skilled in Python, SQL, scikit-learn, PyTorch, Transformers, and statistical evaluation, with experience in leakage-safe validation, model comparison, error analysis, retrieval evaluation, and translating model outputs into human-in-the-loop decision policies.

Skills

- **Programming & Analysis:** Python, SQL, R, pandas, statistical analysis
- **Machine Learning & NLP:** scikit-learn, TF-IDF, Linear SVM, PyTorch, Transformers, DistilBERT, text classification
- **Model Evaluation:** Macro F1, accuracy, coverage, routed accuracy, misroute analysis, error analysis, leakage prevention, group-aware splits
- **AI / RAG:** embeddings, vector search, hybrid retrieval, RRF, PostgreSQL/pgvector, retrieval evaluation, grounded generation
- **Data Engineering & Cloud:** PostgreSQL, PySpark, Delta Lake, dbt, Databricks, Power BI, FastAPI, Docker, GitHub Actions, Azure, AWS, Airflow

Projects

- 1. Financial Complaint Auto-Routing with NLP | Python • scikit-learn • TF-IDF • Linear SVM • PyTorch • DistilBERT**
Jun 2026 - Aug 2026 [GitHub](#) | [Case Study](#)
 - Built a leakage-safe eight-class CFPB complaint classification study comparing TF-IDF + Linear SVM with a frozen DistilBERT challenger using group-aware 2024 splits with zero normalized-text overlap.
 - Improved Macro F1 from 0.7671 to 0.7949; analyzed accuracy, coverage, routed accuracy, and misroute behavior, then evaluated both frozen models on a 30,156-row retrospective 2025 cohort using Auto-Route/Human Review policies.
- 2. CivicLens RAG Operations Copilot | Python • Hybrid Retrieval • RRF • PostgreSQL/pgvector • RAG Evaluation**
May 2026 - Aug 2026 [GitHub](#) | [Live Demo](#) | [Case Study](#)
 - Built and evaluated a hybrid RAG retrieval system combining semantic search, PostgreSQL full-text retrieval, and Reciprocal Rank Fusion over curated NYC 311 knowledge.
 - Compared retrieval behavior and validated grounded responses with citation and abstention checks; hybrid retrieval achieved 83.9% Recall@5 and 92.9% expected-source retrieval.
- 3. NYC 311 Service Requests Lakehouse | Azure • Databricks • PySpark • SQL • Delta Lake**
Mar 2026 - May 2026 [GitHub](#) | [Case Study](#)
 - Built an Azure lakehouse pipeline using Azure Data Factory, ADLS Gen2, Databricks, PySpark, SQL, and Delta Lake to prepare Bronze, Silver, and Gold datasets.
 - Created analytics-ready fact tables and Gold marts using reusable cleaning, SQL validation, and star-schema modeling.
- 4. Cloud Flight Fare Pipeline | AWS • Python • SQL • S3 • Redshift • dbt**
Feb 2026 - May 2026 [GitHub](#) | [Case Study](#)
 - Built an AWS data pipeline using Python, S3, Redshift Serverless, SQL, and dbt to ingest and transform flight fare API data.
 - Developed dbt staging models, dimensional marts, and data quality tests to support reproducible downstream analytics.

Education

- **Master of Science (M.S.) in Information Technology, Data Analytics Concentration** - Middle Georgia State University, Expected Dec 2026 | GPA: 4.00
- **Associate of Applied Science (A.A.S.) in Website Design/Development** - Gwinnett Technical College, Dec 2024 | GPA: 3.96
- **Bachelor of Arts (B.A.) in Chinese Ethnic Language and Literature** - Guangxi University for Nationalities, China

Experience

IT Intern - Fulton County | Jan 2024 - Apr 2024

- Standardized IT asset and network inventory records, improving data consistency and audit readiness.
- Reconciled records across multiple systems to support inventory reporting and internal documentation.

Floral Assistant - Kroger | Dec 2022 - Aug 2023

- Reconciled inventory, pricing, and shrink records to improve reporting accuracy and exception tracking.