Chloe Zhu

chloezxm56@gmail.com | Evanston, Illinois

EDUCATION

Northwestern University — **Evanston, IL**
M.S. in Machine Learning and Data Science, GPA: 4.00/4.00 — Expected Dec 2027

- **Relevant Courses:** Data Engineering, Machine Learning, SQL & Data Management, Database Schema Design, Databases & Information Retrieval, Data Analytics & Optimization, Generating Business Value with Data Science

University of California, San Diego — **La Jolla, CA**
B.S. in Mathematics (Statistics & Probability), Provost Honors — Sep 2021 – Mar 2025

- **Relevant Courses:** Applied Linear Algebra, Mathematical Statistics I & II, Probability, Data Analysis & Inference

TECHNICAL SKILLS

- **Programming & Platforms:** SQL, Python, R, TypeScript, Pandas, PySpark, PostgreSQL, Snowflake, Databricks
- **Data Engineering & Cloud:** ETL, Data Modeling, Data Validation, Apache Airflow, Docker, AWS (EC2), CI/CD, FastAPI
- **Analytics & BI:** Tableau, Power BI, Excel, Streamlit, A/B Testing, Cohort Analysis, Causal Inference, Forecasting
- **ML & NLP:** scikit-learn, PyTorch, XGBoost, LightGBM, FinBERT, RoBERTa, RAG, SHAP, OCR

PROFESSIONAL EXPERIENCE

Progentec Diagnostics — **Evanston, IL**
Data Scientist Practicum — Oct 2026 – Present

- Built **PySpark ETL pipelines** with **Apache Airflow** and **Docker** to process 180K+ clinical records, validating Optical Character Recognition (OCR) output, normalizing lab values and dates, and linking each record to its source PDF page
- Implemented **hybrid RAG retrieval** using **FAISS** and **BM25** to combine semantic and keyword search, filtering by patient and encounter metadata and linking retrieved passages to source PDF pages for evidence-backed LLM answers
- Developed an **LLM evaluation pipeline** comparing chunking and retrieval strategies on a fixed QA set, tracking **Recall@K**, answer faithfulness, and citation accuracy to isolate retrieval gaps, unsupported claims, and citation mismatches

SentiEdge Investment — **Boston, MA | Remote**
Data Scientist — Jul 2025 – Feb 2026

- Built **incremental ETL pipelines** using **Python** to scrape and ingest 150K+ monthly crypto news and social media records from multiple platforms, implementing **deduplication and idempotent loads** to support sentiment analysis and backtesting
- Analyzed 20K+ monthly user events using **SQL CTEs and window functions**; designed and evaluated an **A/B test** of Sentiment Signal UI, informing score visualization improvements followed by ~60% growth in feature WAU
- Optimized **Databricks/Spark** pipelines deployed on **AWS EC2**, reducing processing runtime by ~40%; fine-tuned **FinBERT** against RoBERTa and baseline models, achieving a ~15% relative gain in accuracy

Broad U.S.A. — **Hackensack, NJ**
Data Scientist Intern — Jul 2024 – Sep 2024

- Built an **LLM-powered RAG chatbot** indexing **10K+ documents into 130K+ chunks**; evaluated Top-5 vs. Top-8 retrieval using **Precision@K** and QA accuracy, improving accuracy by ~25%
- Designed **SQL Server (T-SQL)** data models for 250+ customers and 400+ HVAC units, integrating equipment and service histories while resolving duplicate records and join-induced overcounting to ensure accurate maintenance KPIs
- Developed **Tableau dashboards** and automated weekly **Excel reports** to identify maintenance trends and rank the top 10 high-maintenance units, helping engineers prioritize equipment reviews

Zoomlion Heavy Industry — **Hunan, China**
Data Analyst Intern, Construction Machinery Division — Jun 2023 – Aug 2023

- Built **PostgreSQL ETL** workflows for 180K+ records across 15+ machines and **interactive Tableau dashboards** with calculated KPIs for utilization, faults, and repair time, reducing issue-triage time by up to 30%
- Evaluated **ARIMA, Prophet,** and **Gradient Boosting** with time-based validation, reducing **MAPE** by 8–10% versus historical forecasts to support spare-parts planning

PROJECT & RESEARCH EXPERIENCE

NASA Space Explorer — **Evanston, IL**
Data Engineering Project — Sep 2026 – Nov 2026

- Engineered an **end-to-end Python ETL pipeline** integrating **NASA APOD and Image Library REST APIs**; orchestrated extraction, transformation, validation, and loading with **Apache Airflow DAGs** in a **Dockerized** environment
- Designed a **relational SQL database schema** for normalized image metadata, with primary keys and **data-quality checks** for duplicates, missing fields, and invalid image links to support reliable querying
- Performed **SQL/Python EDA** on topic and year coverage, finding an ~85× Mars–Nebula search-result gap and a 2004 Mars-image peak; built an interactive **Streamlit** search and visualization app

Quantifying the Impact of State-Level Opioid Policies on Overdose Mortality — **La Jolla, CA**
Research Assistant — Mar 2024 – Jun 2024

- Integrated **CDC mortality**, opioid-policy, and socioeconomic data into a **state-year panel**; engineered policy timing and lagged features for **mixed-effects models and XGBoost**, achieving ~0.76 out-of-sample R^2
- Estimated policy effects using **Difference-in-Differences (DiD)** and **event-study models** with state and year fixed effects; checked pre-treatment trends and built a **Streamlit dashboard** to compare policy effects and mortality trajectories