Danying Xu

(+1) 917 589 0569 | xdy0314@gmail.com | LinkedIn | Github

EDUCATION

New York University — **New York, United States**
Master of Science in Computer Engineering, Grade: 3.89/4.0 — 2023.09-2025.05

Southeast University — **Nanjing, China**
Bachelor of Engineering in Artificial Intelligence — 2019.09-2023.06

SKILLS

Professional Languages: **Python** (NumPy/Pandas/Scikit-learn/Pytorch/Tensorflow/Matplotlib/Seaborn/SQLAlchemy/OpenCV/ Beautiful Soup/PyGame/Django)**, C++/C, Linux** (Git), **SQL**, **Cypher, Java, Latex**
Tools: Google Colab, Azure CosmosDB, MySQL, PostgreSQL, Github, Gitee, Apache Spark, Apache Hadoop, Protégé, Neo4j
Methodologies: Machine Learning, Deep Learning, NLP, LLM, Database Management (SQL/NoSQL), Big Data, Statistics, Data Analysis, Predictive Modeling, Paper Reading &Writing

PROFESSIONAL EXPERIENCE

Global AI — **New York, United States**
Machine Learning Engineer — 2024.01-2024.04

- Established a **Postgres database** for 30k+ pieces of GDELT news data, boosting time efficiency by 20% on **DBeaver**.
- Implemented and visualized descriptive statistics on 1 million MSCI US Index stocks such as sparsity analysis and correlations.
- Cleaned the data using LOCF, NOCB and periodical average fillings.
- Developed **LSTM** improving stock forecasts by 19.7% with MSE score of 0.4 compared to the linear regression baseline.

Huawei Nanjing Research & Development Center — **Nanjing, China**
Software Development Engineer in Test, Data Communications Department — 2022.08-2022.09

- Conducted **Gray Box Testing** by examining 143 static path graphs with thousands of functions in **C/C++**.
- Conducted **White Box Testing** by **FUZZ test** technology for 872 code files by **Linux**, expected to improve product performance by 30%.

PROJECT EXPERIENCE

Text Gender Bias Rewriter (Research) — 2022.12-2023.06

- Proposed an NLP framework to reduce data gender bias via pattern transform, neural translation and data aggregation.
- Performed **Seq2Seq model** and **Seq2Seq attention model** (character/word level) on 148k+ Chinese sentences on **Pytorch**.
- Modified Word-Embedding Association Test to Chinese evaluated with **CBOW model**, reducing gender bias by 45.4%.
- Conducted **Coreference Resolution** downstream task on **wwm-RoBERTa Model** with consistent performance around 92% after reducing gender bias.

Recognition of Children's Autism — 2021.11-2023.05

- Extracted degree matrices and adjacent matrices of BOLD signals from 884 brain fMRI data.
- Built **GCN** and **GAT** models for determining autism, achieving the accuracy of 68.9% and 71.3%.
- Used the **Mixup** method for expanding the data size by three times, resulting in 3.4% accuracy improvement for GAT.

Deep Learning-based Explanatory Brain Science — 2020.11-2022.05

- Extracted 1.2 million images from 1297 videos of trained monkeys playing Pac-Man game using **Python (OpenCV)**.
- Developed a **ConvRNN model with AlexNet and LSTM** on **TensorFlow**, predicting players next move with 84.6% accuracy.
- Performed **Class Activation Map (CAM) heatmap** for activation layer visualizations.
- Modified the **Grad-CAM heatmap** for each layer to interpret the brain's decision-making mechanism with visualizations.

Knowledge-Based Question Answering (KB-QA) System — 2022.03-2022.04

- Crawled 1k+ La Liga game data from Wikipedia and FBref webs for fact extraction.
- Aligned the knowledge by regularization, Google translation API and Python (difflib).
- Built the ontology map on **Protégé** and knowledge database on **Neo4j**.
- Developed a **BiLSTM-CRF model** to segment natural language to generate queries in **Cypher**.
- Implemented a dynamic web page with the locations of La Liga football clubs for visualization using Python (pyecharts).

Predictions for Potential 5G Customers — 2021.05

- Used the **K-Means** on 140k+ China Mobile users to classify 5G user behaviors, resulting in the silhouette score of 0.29.
- Improved the silhouette score to 0.37 using **Gaussian Mixture Model (GMM)** by **Python (Scikit-learn)** on the same task**.**
- Leveraged multiple regression tree **CART** for **Gradient Boosting Decision Tree (GBDT)** to build a 5G consumer behavior prediction model, achieving the best accuracy of 91.41% with grid search.