Andy (Xiangyu) Cui

xiangyucui@outlook.com | +1(402)-853-3000 | +8618698680522 | Linkedin | Portfolio

EDUCATION

Northeasten University Boston, MA
M.S. in Artificial Intelligence of Khoury College *Dec 2023*
Completed Ph.D-level coursework with a focus in NLP; conferred M.S. degree upon early graduation

University of Nebraska-Lincoln Lincoln, NE
B.S. in Computer Science of Arts Science College *May 2020*
Double Major Computer Engineer

SKILLS

Languages: Python, JavaScript, SQL, Bash
Frameworks & Libraries: PyTorch, Scikit-learn, FastAPI, React, LangChain, HuggingFace Transformers
Data & ML: Pandas, NumPy, LSTM, Transformer, TF-IDF, Recommendation Systems
Databases: PostgreSQL, MongoDB, DynamoDB
Cloud & DevOps: AWS (Lambda, Kinesis, S3, CloudWatch, QuickSight), Docker, GitHub Actions
Visualization & Analytics: QuickSight, matplotlib, SPARQL, Elasticsearch

PROJECTS

Hobby-Based Outdoor Club Platform (Ski & Hiking Social App) Aug 2025

- Designed and launched a platform serving **3,000+** users to organize ski/hiking events, improving user retention by **40%** through personalized group recommendations.
- Built **iOS** and web apps using **Swift**, **React**, and **FastAPI**, and managed multilingual content with **PostgreSQL** and **MongoDB** hybrid design.
- Integrated **LangChain** + **OpenAI API** to support AI-powered Q&A, reducing human admin workload by **70%** and improving onboarding efficiency.
- Leveraged **AWS Athena** + QuickSight to build a performance dashboard that helped identify high-ROI regions for new events.
- Designed interest-graph modeling using **SPARQL** for personalized event matching, increasing **RSVP** conversion.

Automated Tax Office AI Assistant Tool for Tax May 2025

- Built a **PyQt5/PySide6** desktop tool with **pywinauto** to automate W-2 and 1099 entry for tax preparation, supporting Excel uploads and real-time progress tracking via **QTableWidget**.
- Integrated **GPT-4** API to assist staff with data formatting and form guidance, reducing operational time cost by **80%** and cutting manual errors by **70%**.
- Implemented error logging and auto-organized user data folders using a unit format to improve traceability and file management.

Job Recommendation System Design Jan 2025

- Developed a user interface for job searching using **Axure RP 10**; Applied content-based filtering using **TF-IDF** and cosine similarity, achieving **82%** precision in matching user skills to job descriptions; Conducted **collaborative filtering** in **Python** with implicit user feedback, improving recommendation diversity by 18% via matrix factorization.
- Leveraged **deepseek API** to dynamically adjust recommendations based on real-time user feedback; Reduced cold-start bias by 30% through RL-driven exploration of niche roles.

Stock Price Prediction with Deep Learning Oct 2024

- Collected and preprocessed historical stock price data from multiple financial APIs, including open, close, volume, and volatility features for small/mid-cap equities.
- Engineered lag features, moving averages, and sector-based indicators to enrich input signals for sequential learning.
- Implemented deep learning models for time-series forecasting, including **LSTM**, **GRU**, and **Transformer** architectures in **PyTorch**, achieving **20%** improvement in **RMSE** over traditional **ARIMA** baselines.
- Constructed a modular hyperparameter search framework using Sklearn + GridSearchCV, enabling repeatable tuning of model depth, learning rate, and sequence length.
- Integrated a stock recommendation component using similarity-based filtering (based on return profiles and volatility clustering) to suggest alternative stocks under comparable risk profiles.
- Simulated risk-adjusted returns using Sharpe Ratio and drawdown analysis on predicted trends to inform portfolio construction strategies.

Amazon QA Bot: Comparative Evaluation of BERT and GPT-2 Models Sep 2023

- Designed a product Q&A system using Amazon review data (in **JSON** format) to compare **BERT** (encoder) and GPT-2 (decoder) architectures, focusing on factual accuracy, fluency, and generation control.
- Trained both models using **PyTorch** + **HuggingFace** Transformers, implemented tokenization, DataLoader batching, and applied **K-Fold Cross-Validation** (10-fold for BERT, 5-fold for GPT-2) for robust comparison.
- Tuned hyperparameters (batch size, learning rate, epochs) via grid search; used **BLEU**, Accuracy, and CrossEntropyLoss as core evaluation metrics.

- Findings showed **BERT** achieved 85% factual correctness, while **GPT-2** produced more human-like responses (+20% increase in user fluency ratings).
- The final model pipeline reduced manual response volume by 90% in simulated customer support scenarios and informed future NLP stack decisions.

Wind Tower Weld Depression Prediction via Supervised Regression Models Sep 2022

- Developed a machine learning pipeline to predict weld depression profiles in thin-walled wind turbine towers, which directly affect structural stability and sustainability. The system supports data-informed design and manufacturing decisions for renewable energy infrastructure.
- Processed over **6,000** structured data points using **3D laser** scans of scaled tower cross-sections, extracting radius deviations between actual and ideal circular columns within ±250mm weld zones.
- Explored and compared three supervised modeling approaches:
  A. **Maximum Likelihood Estimation** (MLE) based on Rotter-Teng theoretical models
  B. **Polynomial Regression** with feature orders ranging from 2 to 15
  C. **Feedforward Neural Network with ReLU** activation, two hidden layers, and 1,000 neurons each.
- Used **MSE, R², Pearson's r, and Kendall's Tau** to evaluate model performance on training and test sets. **Neural network** outperformed other models with the lowest error and strongest generalization ability, making it suitable for real-world industrial applications.

WORK EXPERIENCE

King 7 Club Corp Jan. 2025-Present
Senior Software Engineer, Full Time *Los Angeles, CA*

- Developed and deployed a full-stack web application using **React**, **Node.js**, and **FastAPI**, architected as modular microservices to reduce code coupling and enable rapid feature iteration.
- Containerized the app with **Docker** and deployed to **AWS Lightsail**, reducing deployment time by **40%** and ensuring consistent environments across staging and production.
- Designed and documented **RESTful** APIs with clear **JSON** schema, supporting multiple frontend clients and enabling third-party integration.
- Migrated data workflows to **PostgreSQL**, improving query reliability and security; implemented role-based access control (RBAC) to meet compliance standards.
- Integrated Prometheus and Grafana for real-time backend monitoring, reducing system downtime by **75%**.
- Introduced GitHub Actions for **CI/CD** pipeline, cutting manual deployment effort by **80%** and accelerating release cycles.

CAC Auto Group LLC Feb. 2024-Dec. 2024
Data Engineer, Full Time *Southborough, MA*

- Built a fully serverless **ETL** pipeline using **AWS Lambda**, **S3**, **DynamoDB**, and **Kinesis**, enabling real-time processing of vehicle pricing data across multiple dealership sources.
- Integrated predictive pricing models in Python based on historical trends and market volatility, increasing pricing precision by over **50%**, directly boosting sales conversion rates by **18%**.
- Leveraged CloudWatch and SNS for automated anomaly detection, alerting management of outlier pricing behavior with >**90%** detection accuracy.
- Supported pricing dashboard through **QuickSight**, enabling business teams to track inventory competitiveness and reduce manual reporting by **90%**.

AlpalifeBio LLC Dec. 2022-Jun. 2023
Data Engineer, Internship *Woburn, MA*

- Developed a biomedical **ETL** pipeline using **Kinesis** and **Lambda** to aggregate more than **500K** records per day, streamlining complex ingestion workflows and reducing delay from 2 hours to less than 10 minutes.
- Integrated datasets into **SQL-based** warehouse for downstream modeling teams, enabling cleaner joins and faster exploratory analysis.
- Built an internal search layer using Elasticsearch with advanced tagging logic to improve discovery of rare disease datasets, saving researchers more than 30 hours per month.
- Collaborated cross-functionally with data science and ops teams to ensure pipeline robustness and easy rollback strategies.