Andy (Xiangyu) Cui

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

EDUCATION

Northeastern University Boston, MA
M.S. in Artificial Intelligence of Khoury College *Dec 2023*
P.H.D Program - Quit P.H.D Graduate as Master Degree - Major NLP

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

PROJECTS

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

- Designed and developed a multi-platform social application for ski and outdoor sport enthusiasts, currently serving over 3,000 active users in the U.S. The iOS app frontend was built using **Swift**, with UI/UX prototypes designed in **ComfyUI** to streamline user onboarding and club participation flows.
- The backend is powered by **FastAPI** and hosted on Supabase using a **PostgreSQL** document store, handling event registration, user profiles, geolocation tagging, and multilingual content sync. For deployment, I used Linux-based containers and CLI tools to manage versioning, rollout, and server logs.
- Built a companion **React-based** promotional website for SEO-friendly marketing, featuring static content, activity feeds, and club announcements. Integrated an AI-powered customer assistant (LLM-based, fine-tuned using LangChain + OpenAI API) to handle real-time user Q&A, membership support, and event recommendations.
- Currently experimenting with **SPARQL + RDF** knowledge graph prototypes to structure user interest taxonomies and cross-event connections. Also evaluating AWS Athena and QuickSight for performance dashboards to track club growth, regional popularity trends, and real-time RSVPs across cohorts.

Hobby-Based Outdoor Club Platform (U.S. Social Media Web App) Jun 2025

- Designed and built a social club platform targeting U.S.-based users with shared interests in snowboarding, skiing, hiking, and mountaineering, enabling community formation through event-based interaction and activity tracking. The platform currently supports over 3,000 active users and continues to grow organically.
- Developed a responsive web application using **React** (frontend) and **FastAPI** (backend) with a **MongoDB** document database, enabling fast iterations, easy deployment, and scalable data structures.
- Implemented a microservices-based architecture with **modular APIs**, including reserved **AI endpoints** for future integration of chatbot assistants to support event Q&A, user onboarding, and intelligent notifications.
- Deployed event hosting, **RSVP**, and location-based recommendation modules. Designed infrastructure to support cross-platform expansion and data sync with **WeChat Mini Programs** for bilingual user access and ecosystem integration.
- The platform serves as a hybrid between interest-based social media and outdoor club logistics, promoting real-world connections through technology.

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 the historical stock price and other financial assets data on the company of interest; Conducted data preprocessing by applying min-max scaling in **Sklearn** to normalize stock price values, ensuring consistency across the dataset.
- Implemented **LSTM**, **GRU**, and **Transformer models** in **PyTorch**, optimizing hyperparameters (e.g., number of layers, optimization methods) through grid search, increasing model accuracy by **20%**; Visualized opening and closing price trends to assess model performance in Python.

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

- Built a question-answering system using Amazon product review data (in **JSON** format) to compare the performance of two **LLM** architectures: **BERT** and **GPT-2**.
- Utilized PyTorch and Hugging Face Transformers to implement full training loops, including DataLoader with RandomSampler/SequentialSampler, and processed the dataset into tokenized input batches.
- Performed grid-based hyperparameter search (batch size, learning rate, epochs), applied K-Fold cross-validation (10-fold for **BERT**, 5-fold for **GPT-2**), and used AdamW optimizer for fine-tuning.

- Visualized training performance with matplotlib, and used metrics such as **CrossEntropyLoss**, **BLEU**, and Accuracy to evaluate and compare performance.
- Results showed **BERT** achieved higher factual accuracy, while **GPT-2** offered more natural and human-like output, providing empirical insights for dialog system design.
- Demonstrated strong industrial applicability by enabling automated product Q&A in e-commerce scenarios; reduced manual response time and labor cost by over **95%**, while maintaining answer accuracy above **85%**, meeting standard commercial requirements for customer support automation.

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 responsive web application using **React** for the frontend and **Node.js + FastAPI** for a modular backend architecture, supporting dynamic UI with custom **JavaScript** logic and **CSS** animations.
- Hosted static assets via GitHub, containerized the full-stack app with **Docker**, and deployed to **AWS Lightsail**, including kernel optimization to reduce resource overhead and improve runtime stability.
- Used **PostgreSQL** as the backend database to securely manage user data, and structured APIs with clean JSON responses for frontend integration.

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

- Developed and maintained a predictive pricing system for vehicles on CarGurus using **AWS serverless architecture**, enhancing market compatibility and streamlining operations. Leveraged key AWS services including **S3**, **Lambda**, **DynamoDB**, **SNS**, **CloudWatch**, and **Kinesis**, and used **Python** with **AWS CloudFormation** for scalable infrastructure deployment.
- Designed and implemented a **fully serverless** data pipeline to continuously monitor target data sources using **Kinesis streams** and **Lambda triggers**, eliminating the need for traditional polling. This approach reduced infrastructure and processing costs by **80%**, while maintaining high scalability and responsiveness.
- Integrated real-time monitoring to track market data fluctuations, enabling automated detection and adjustment of vehicle prices in response to deviations. This solution boosted daily operational efficiency by **80%** and improved pricing accuracy by over **50%** compared to industry standards.

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

- Built and managed a robust **AWS** streaming data pipeline to automate biomedical data ingestion from multiple public databases into **Kinesis Data** Stream. This system processed over **500,000** data entries daily, using **Lambda** Functions for real-time data transfer and **S3** and **DynamoDB** for efficient, scalable storage and retrieval. This architecture allowed seamless handling of high-volume data with minimized latency and reduced operational costs.
- Configured and optimized a structured **SQL** database to integrate and process data from diverse biomedical sources. Implemented an efficient tag-processing system for enhanced search and retrieval operations, reducing data retrieval time by **80%**. This improvement significantly boosted operational efficiency, making it easier to access and analyze critical information for downstream applications.
- **Designed and implemented an automated biomedical data acquisition and processing system analogous to the data pre-labeling and preprocessing pipeline in speech model training**, significantly boosting operational efficiency and enabling downstream applications in data analysis and modeling.