Zhantao Wang

Tel: (+86) 156-6611-7867 | Email: 649251318@qq.com | WeChat: wanglewc2001
Current Status: Master's Degree Candidate

EDUCATIONAL BACKGROUND

University of Aberdeen **UK**
Master's Degree in Artificial Intelligence **2023.09-Present**

- **GPA:** 3.8/4.0.
- **Core Courses:** Machine Learning, Data Mining and Visualization, Artificial Intelligence Engineering Systems, Multi-Agent Systems, etc.

University of Liverpool **UK**
Master's Degree in Business Analytics and Big Data **2020.09-2022.11**

- **GPA:** 3.67/4.0.
- **Core Courses:** Data Mining and Machine Learning, SAP E-Business Enterprise Systems, Big Data Analytics for Business, etc.

INTERNSHIP EXPERIENCE

KPMG (Klynveld Peat Marwick Goerdeler) **2023.07-2023.10**
Machine Learning Engineer

- Conducted comprehensive data quality assessments, applied Python and SQL for data cleaning and anomaly detection, identified and resolved anomalies (e.g., special characters, security vulnerabilities), ensuring the integrity and consistency of the data for analysis.
- Utilized machine learning tools such as clustering analysis, Pandas, and Scikit-learn to extract key insights from large-scale customer data, performed wealth segmentation to support KPMG's strategic marketing plans, optimizing customer classification and targeting.
- Created dynamic displays using Tableau and Matplotlib, based on real estate valuation data, to forecast asset distribution, optimize regional customer asset allocation, and improve decision-making efficiency by 30%.

JPMorgan Chase & Co **2023.06-2023.07**
Software Engineer

- Contributed to the open-source community by enhancing JPMorgan Chase & Co.'s Perspective framework, using JavaScript and TypeScript to fix critical bugs, add new features, and improve documentation. Managed version control with Git, participated in code reviews through GitHub's pull request collaboration mechanism, and addressed open issues in the project repository, improving the overall code quality of the project.
- Integrated JPMorgan Chase & Co.'s Perspective framework to develop and deploy a real-time stock price data processing and visualization system. Utilized WebSocket technology to enable real-time data streaming and created a dynamic visualization interface using the React front-end framework to display stock price trends, ensuring efficient data updates and enhanced user interaction.
- Developed a custom stock price alert system using Python and JavaScript, applying algorithms to analyze real-time data, and set trigger conditions based on price thresholds and volume spikes. When conditions were met, the system instantly alerted traders through the front-end interface and notification system, optimizing the trading decision process and improving response speed and trading accuracy.

TATA Group **2023.02-2023.05**
Data Analyst Engineer

- Performed data preprocessing and cleaning using Pandas and NumPy, combined with SQL queries, to complete market research and data mining. Developed complex business scenario mappings to identify growth opportunities and consumer trends, providing data-driven insights to optimize new product strategies.
- Designed advanced data visualizations using Matplotlib and Seaborn, creating bar charts, heatmaps, and density plots to accurately convey customer satisfaction analysis results. This supported precise improvements in targeted services and enhanced the customer experience.
- Built dynamic and interactive dashboards using Tableau to consolidate sales data and conduct in-depth analysis. Applied regression analysis and time series forecasting techniques to develop data-driven pricing strategies, which significantly boosted the company's overall revenue.

PROJECT EXPERIENCE

Alphabet Credit Card User Default Risk Prediction Project **UK**
Project Leader **2024.03-2024.05**

- Applied data preprocessing, feature engineering, and model training techniques to complete the prediction of credit card user default risk.
- Built and optimized models using Random Forest, XGBoost, and LGBM algorithms, significantly improving prediction accuracy.
- Proposed recommendations for optimizing risk control, including real-time monitoring of large transactions and geolocation verification, with an estimated 25% reduction in fraud losses.

University of Liverpool Rocket Design Simulation Project **UK**
Data Analysis Engineer **2022.09-2022.12**

- Collaborated with Dr. Ahmed Al-Irhayim to integrate various component datasets and applied Long Short-Term Memory (LSTM) and Generative Adversarial Network (GAN) models for rocket engine fault detection and diagnosis.
- Implemented LSTM and GAN-based Fault Detection and Diagnosis (FDD) methods, significantly enhancing the fault elimination capabilities of large liquid oxygen/kerosene rocket engines.
- Using this method, the accuracy of fault detection in rocket engines increased by 30%, fault elimination time was reduced by 25%, significantly lowering engine testing and maintenance costs, and advancing the project timeline by two months.

SKILLS & CERTIFICATIONS

- **Technical Skills**
  - Familiar with deep learning frameworks such as PyTorch and TensorFlow, capable of building and applying Deep Neural Networks (DNN) and Recurrent Neural Networks (RNN).
  - Proficient in Python, with extensive experience using tools like Pandas, NumPy, Scikit-learn, and XGBoost to accomplish machine learning and data processing tasks.
  - Familiar with common Natural Language Processing (NLP) libraries, including Transformers, NLTK, and Hugging Face Datasets, with an understanding of their applications in tasks like text classification and sentiment analysis.
  - Familiar with MLOps tools like MLflow and DVC, skilled in managing model versions and data versions to support continuous integration and deployment in machine learning projects.
  - Proficient in data visualization tools like Plotly, able to use Python for visualizing and interactively analyzing data.
  - Familiar with Python-based API request and data processing tools such as Requests, Replicate, and SnowNLP, capable of using these tools for data collection, processing, and analysis.
  - Proficient in MySQL, NoSQL, SAP HANA, and database design. Highly skilled in SPSS, RapidMiner, Tableau, and Power BI, and proficient with Microsoft Office software.
  - Knowledgeable in Linux, decision trees, Kubernetes, Kata, ACRN, Apache MySQL, Zabbix, NGINX, and FTP.
- **Language Skills:** English (Proficient), Russian (Familiar), Mandarin (Native).
- **Certifications:** IBM Certification in Building Trusted AI for Enterprises, Bloomberg Market Concepts Certificate, National Intermediate Vocational Qualification for Computer Network Equipment Assembly and Debugging.

Personal Statement

I am excited to apply for a Ph.D. in Computer Science at Columbia University, focusing on **Applications of Large Language Models to Computational Social Science: Improvements in Predictive Models of Social Phenomena**. Under the mentorship of Professor Daniel Hsu, I aim to explore the intersections of machine learning, computational social science, and algorithmic fairness. This aligns closely with Professor Hsu's work in **algorithmic learning theory** and **machine learning foundations**. My previous research in large language models (LLMs), bias reduction in AI, and my recent work on virtual digital human systems place me in an excellent position to contribute meaningfully to this field.

Motivation and Research Interests

The potential of large language models (LLMs) to transform the study of social science phenomena is what drives my passion for this field. When appropriately designed and applied, LLMs provide a powerful tool to understand human behaviour, model social dynamics, and predict social phenomena at a previously unachievable scale and precision. I am particularly interested in how these models can **predict social behaviors**, **analyse sentiment across populations**, and **detect societal trends** from complex data sources like social media and clinical texts. Columbia's strong emphasis on data science and machine learning makes it the ideal place to pursue this research.

Research Experience with Large Language Models

In my previous research, I have worked extensively with **state-of-the-art large language models,** including **Gemma 7b Instruct**, **LLaMA 2 70b**, **LLaMA 3.8b Instruct**, **LLaMA 3.7b Instruct**, **Mistral 7b Instruct**, and **Megatron-BERT Cased-345m**. I applied these models in various domains, including **computational social science**, to predict patterns in human behavior and analyze large-scale text data. I explored **natural language processing (NLP)** tasks such as **sentiment analysis**, **topic modelling**, and **social trend detection**, using these models to predict social phenomena from platforms like **Twitter** and **Reddit**.

I fine-tuned these models using **Low-Rank Adaptation (LoRA)** and **Bayesian optimisation** to improve their performance for specific social science applications. For instance, I optimised **Megatron-BERT** for **sentiment prediction in political discourse**, significantly enhancing its ability to capture nuanced societal trends from sparse data. These experiences have equipped me with the necessary tools to improve LLMs for the **prediction of social behaviours** and the **analysis of societal shifts**.

Internship at Xiaoice (Microsoft Asia) – Virtual Digital Humans

During my **Xiaoice (Microsoft Asia) internship**, I applied advanced machine learning models to generate **virtual digital humans** using **StyleGAN** and **GANs (Generative Adversarial Networks)**. These digital avatars were designed to interact with real users in real time, providing a human-like experience in healthcare, entertainment, and education. This project enhanced my understanding of **real-time AI systems** and **multimodal data processing**, skills directly transferable to the **computational modeling of social phenomena**. I see a direct parallel between **virtual human interactions** and **social behaviors**, and I am eager to apply this knowledge to improve predictive models of human behavior.

Focus on Bias and Fairness in AI Models

I have also worked extensively on reducing bias in AI systems, a key challenge in **computational social science**. During my research on **career advisory systems**, I applied techniques such as **data augmentation** and **contextual embeddings** to reduce gender and racial bias in models like **Gemma 7b** and **LLaMA 2**. I believe this experience is crucial for building socially responsible AI systems, particularly when predicting social phenomena. My work aligns with Professor Hsu's research on **fair decision-making** and **causal fairness** in machine learning models.

Medical Diagnostics and Data Fusion

In addition to my work on social phenomena, I have applied **Convolutional Neural Networks (CNNs)** like **ResNet-50** and **VGG16** in **medical diagnostics**. I worked on multimodal data integration, combining **medical imaging**, **clinical data**, and **genomic information** to improve diagnostic accuracy. These multimodal fusion techniques are relevant to my Ph.D. project, as understanding **social phenomena** often requires integrating data from multiple sources—such as **social media**, **clinical texts**, and **economic indicators**.

Skill Set and Certifications

My technical certifications further strengthen my ability to contribute to the project:

- **MATLAB Certifications**: My certifications in **Deep Learning**, **Signal Processing**, **Reinforcement Learning**, and **Computer Vision** provide a solid foundation for **algorithm design** and **data modeling**, both crucial for improving LLMs.

- **TensorFlow and Python**: I have extensive experience with **TensorFlow** and **Python**, which I have used in various NLP tasks and predictive modeling, making me well-prepared for advanced research in machine learning.

- **Bayesian Optimization**: I have successfully applied **Bayesian optimization** in tuning large models, allowing me to refine the architecture and hyperparameters of models for **social science predictions** and **language model tasks**.

Alignment with Professor Hsu's Research

Professor Hsu's work on **algorithmic learning theory** and **machine learning foundations** strongly aligns with my research focus on **computational social science**. His research on **causal fair decision-making** is especially relevant to my work on bias reduction in AI systems. Additionally, my hands-on experience with LLMs, virtual digital humans, and social science applications will allow me to contribute meaningfully to ongoing research in the **Foundations of Data Science** group at Columbia University.

Broader Importance of My Research

The societal implications of this research are profound. As large language models become integral to **predictive social models**, there is an increasing need to ensure that these systems are both **accurate** and **fair**. My goal is to build AI systems that can **accurately predict**

social phenomena—such as **public sentiment** during elections or **social movements**—while minimizing bias and ensuring transparency. This work has the potential to influence **policy-making**, **public health strategies**, and **educational outreach** by providing insights based on large-scale, unbiased data analysis.

Conclusion

I am eager to contribute to Columbia University's innovative research environment, where I can further explore the applications of large language models to **computational social science**. I believe my background in **large language models**, **bias reduction**, and **virtual digital human systems** uniquely positions me to succeed in this research endeavor. Under Professor Hsu's mentorship, I am confident that I can make meaningful contributions to the advancement of **predictive models of social phenomena**. I look forward to the opportunity to discuss how my skills and experiences can support the research efforts at Columbia.

Thank you for considering my application.

Sincerely,
Zhantao Wang