Tingxuan Wu

New York, NY | tw3196@nyu.edu | Google Scholar | LinkedIn

EDUCATION

New York University (NYU) — **New York, NY**
M.S. in Information Systems — Sep 2025 – Exp 2027

- **Coursework:** Machine Learning, Deep Learning, Realtime and Big Data Analytics, Big Data Application Development, Fundamental Algorithms, Database Systems.

London School of Economics and Political Science (LSE) — **London, UK**
B.S. in Financial Mathematics and Statistics — Sep 2021 – Jun 2024

- **Coursework:** Probability, Distribution Theory and Inference, Regression and Generalized Linear Models, Stochastic Processes, Computational Methods, Operations Research Techniques.

PROGRAMMING SKILLS

Languages: Python, R, C++, Java, SQL
LLM & GenAI: PyTorch, Transformers (Hugging Face), OpenAI API, LangChain, RAG systems, multimodal models, vision-language models (VLMs), multi-agent systems, LLM-as-judge, RLHF concepts, reward modeling, prompt engineering, sentence-transformers, cross-modal fusion
Reinforcement Learning: Policy optimization concepts, reward function design, RLHF (Reinforcement Learning from Human Feedback)
ML Frameworks: scikit-learn, XGBoost, LightGBM, CatBoost, Keras, LSTM, autoencoders
Vector & Databases: FAISS, PostgreSQL, MongoDB, Redis, vector similarity search
MLOps & Cloud: Docker, MLflow, AWS (EC2, S3), distributed training (PyTorch DDP), GPU optimization, model deployment

EMPLOYMENT

Founder Securities — **Hangzhou, CN**
Machine Learning Engineer Intern — Jun – Aug 2024

LLM Applications RAG Systems Model Optimization Production ML

- Developed production RAG system for financial document analysis processing 10K+ analyst reports and Fed statements, implementing retrieval pipeline with sentence-transformers embeddings (all-MiniLM-L6-v2) and FAISS vector indexing achieving <100ms query latency, integrating GPT-3.5 for question answering and summarization, reducing manual document review time by 60% in pilot deployment with research team.
- Built end-to-end ML pipeline for time-series forecasting with automated model selection and hyperparameter optimization using Optuna, comparing 15+ architectures (LSTM, XGBoost, LightGBM, Prophet) across 500+ configurations, implementing walk-forward validation that improved forecast accuracy by 15% (RMSE) on out-of-sample test set, with systematic ablation studies validating feature importance and model choices.
- Designed production inference infrastructure deploying ensemble models via Flask REST API achieving p99 latency <100ms with 200+ QPS throughput in staging environment, implementing automated model retraining pipeline with MLflow experiment tracking, A/B testing framework, and SHAP explainability analysis for model interpretability and risk management.

Everbright Futures — **Hangzhou, CN**
Machine Learning Engineer Intern — Sep – Oct 2022

NLP Fine-Tuning Ensemble Learning Feature Engineering

- Implemented NLP pipeline for sentiment analysis processing 10K+ daily financial news articles, fine-tuning BERT-based models (FinBERT) on domain-specific corpus achieving 71% directional prediction accuracy (9% improvement over baseline), generating contextualized embeddings that fused with price data through attention mechanisms, improving out-of-sample performance by 12%.
- Built commodity futures prediction system with ensemble gradient boosting models (XGBoost, CatBoost, LightGBM) processing 1M+ daily predictions across 20+ contracts, achieving p99 inference latency <50ms through optimized feature caching and batched prediction serving, with systematic hyperparameter tuning and

cross-validation.

- Designed automated feature engineering framework with dimensionality reduction (PCA, autoencoders) reducing feature space from 300+ to 80 dimensions while maintaining 95explained variance, implementing recursive feature elimination with cross-validation to identify optimal feature subsets, improving model generalization and reducing overfitting.

Guosen Securities **Hangzhou, CN**
Data Engineer Intern July – Aug 2021

Data Mining ML Pipelines Statistical Modeling Feature Engineering

- Built automated data processing pipeline for healthcare sector analysis handling 4,000+ companies, implementing ETL workflow with Apache Airflow orchestration and PostgreSQL storage, designing automated data quality validation with anomaly detection algorithms that reduced manual data cleaning time by 90% (from 8 hours to 45 minutes daily).
- Developed machine learning model for stock screening and ranking integrating 50+ financial features (valuation metrics, growth indicators, profitability ratios), implementing ensemble methods with feature engineering (z-score normalization, polynomial features) and backtested validation achieving 85% precision in identifying top-performing stocks, demonstrating strong out-of-sample performance through walk-forward testing.
- Designed systematic feature selection framework comparing multiple approaches (forward selection, backward elimination, L1 regularization), implementing cross-validation with stratified sampling to ensure robust model performance, with statistical significance testing (t-tests, chi-square) validating feature importance and model reliability.

PUBLICATIONS

Multimodal Social Media Bot Detection Using Heterogeneous Information **LSE**
(**Paper Link**) ***Tingxuan Wu****, Zhaorui Ma, Yanjun Cui, Ziyi Zhou, Eric Wang* May – Oct 2024

Vision-Language Models Multimodal Learning Cross-Modal Fusion First Author Publication

- Paper accepted at AAAI W3PHIAI-25, to be published in the Springer/Nature in Studies in Computational Intelligence.
- Led research on detecting AI-generated social media accounts, designing novel Cross-Modal Residual Cross-Attention (CMRCA) fusion mechanism that improved detection accuracy by 8% over state-of-the-art baselines on 50K+ profile dataset.
- Implemented end-to-end multimodal pipeline integrating image encoders (CLIP-based), text encoders (RoBERTa), and user metadata, achieving 91% precision at 85% recall.

Learning Musical Representations for Music Performance Question Answering **Dartmouth College**
(**Paper Link**) *Xingjian Diao, Chunhui Zhang,* ***Tingxuan Wu****, Ming Cheng, Zhongyu Ouyang, Weiyi Wu, Jiang Gui* Feb – Jun 2024

Multimodal Question Answering Vision-Language-Audio Models Audio-Video-Text Alignment

- This project is supported by the Department of Defense's Congressionally Directed Medical Research Programs (DOD CDMRP) Award HT9425-23-1-0267.
- Paper published at EMNLP 2024, one of the top three conferences in natural language processing. (Ranked A*)
- Contributed to multimodal QA framework achieving state-of-the-art on Music-AVQA benchmarks, implementing cross-modal attention adapters and music-specialized encoders (rhythm, source) processing 9,288 performance videos.

SELECTED PROJECT

GPT-3-Inspired Transformer with Reinforcement Learning Exploration Jan – Feb 2026

- Implemented decoder-only transformer language model from scratch in PyTorch inspired by GPT-3 architecture (Brown et al., 2020), including custom BPE tokenizer, causal self-attention, learned positional embeddings, and end-to-end training pipeline, training 17M parameter model on TinyStories achieving 1.45 validation perplexity and generating coherent multi-paragraph text.
- Developed distributed training infrastructure with PyTorch DDP and mixed precision (FP16) enabling scaling to 125M parameters, conducting ablation studies on architectural choices (pre-norm vs post-norm: 0.3 perplexity

improvement; learned vs fixed positional encodings: 15% faster convergence).

- Explored reinforcement learning fine-tuning approaches implementing reward functions based on text quality metrics (coherence, fluency, factuality), experimenting with policy gradient methods on sample generation tasks, achieving 15% improvement in human preference alignment on evaluation set, demonstrating research methodology applicable to RLHF and model alignment techniques.

  Transformer Architecture   PyTorch   LLM Training   Distributed Training   Ablation Studies

Multi-Agent LLM Collaboration Framework Dec 2025

- Designed multi-agent system with specialized LLM agents (researcher, critic, synthesizer) coordinating through structured message passing to solve complex reasoning tasks, implementing agent orchestration framework with LangChain and OpenAI API.
- Implemented LLM-as-judge evaluation framework using GPT-4 as evaluator to assess agent outputs with structured rubrics, achieving 20performance improvement on multi-step reasoning benchmarks (GSM8K) through collaborative agent interaction.

  Multi-Agent Systems   LLM-as-Judge   Agent Orchestration