Yutong (Beeth) Xue

Houston, TX|346-632-0550 | xue_yutong@outlook.com | linkedin.com/in/yutong-xue-01y07m08d

EDUCATION

Rice University – George R. Brown School of Engineering **12/2024**

Master of Computational Science and Engineering-Statistics (MCSE-STAT) **GPA:** *3.72/4.00* *Houston, TX*

Shanghai University of Finance and Economics **06/2023**

Bachelor of Laws *Shanghai, China*

SKILLS

Programming & Tools*: Python (NumPy, Pandas, Scikit-learn, PyTorch, TensorFlow, Matplotlib, Seaborn, XGBoost), SQL, C++, MATLAB, R, JAVA, Git,*

AI & ML Frameworks*: Hugging Face, LangChain, OpenAI API, SpaCy, NLTK, Transformers, FastAPI, Flask, AWS, Azure, GCP*

PROFESSIONAL EXPERIENCE

[Hegemonics] **[Houston, TX]**

[AI Engineer] *[12/2024] – [now]*

• Design and Deployment of a Multi-Agent AI Legal Assistant:

- Developed an AI-powered legal research and document analysis system using a modular **multi-agent** architecture with agents using **Langgpraph**, specialized in query parsing, **semantic search**, legal reasoning, and answer synthesis, reducing manual document review time by **60%**.
- Engineered a **LangChain**-based agent framework integrating **RAG**, **external legal search APIs**, and **LLM-based summarization** to automate legal knowledge workflows.
- Utilized **PDF parsing libraries** (PyMuPDF, Unstuctor) and **OCR tools** to extract structured and unstructured data—including clauses, tables, and figures—from legal documents with a **95%+ extraction success rate**.
- Indexed extracted content using an in-memory **Pinecone vector database**, **sub-100ms** average semantic retrieval latency, allowing for fast semantic retrieval of relevant clauses and legal provisions in response to user queries.
- Created tools to answer user questions by retrieving document segments, generating concise responses, and structuring output using legal reasoning formats such as **IRAC**.
- Built an interactive **CLI interface** for dynamic user-agent conversations and support for iterative legal query refinement, contract review, and real-time clause exploration with **Flask.**
- Integrated the system into a containerized environment using **Docker** and deployed it on cloud **infrastructure Azure** to ensure persistent access, scalability, and **API interoperability** with **Node.js**.
- Designed and tested **prompt-driven agents** capable of performing reasoning over retrieved text to deliver legally coherent summaries and clause-level insights.
- Enabled r**eal-time user interaction** via terminal interface, allowing continuous dialogue with the legal assistant for legal research, compliance checks, and **contract review**, supporting **20+ concurrent user sessions** during peak usage tests.
- Collaborated closely with **cross-functional** teams with **Agile**—including **backend** engineers, **frontend** developers, and legal experts—to define **API interfaces**, ensure seamless data flow, and align system outputs with **UI/UX** requirements, resulting in a **30% faster integration cycle** and improved end-user experience.

• Fine-Tuning & Deployment of Adaptive LLM Agents (LangChain, Docker, AWS EC2):

- Fine-tuned the **Open Source DeepSeek-R** model using **DPO** and Hugging Face RLHF pipeline, with the help of Spark, boosting decision-making accuracy to **95%.**
- Implemented **RLHF** loop from live interactions to drive continual self-supervised learning.
- Deployed agent pipeline in **Dockerized containers** using **Kubernetes on AWS EC2**, enabling scalable inference and agent-serving infrastructure with the help of **Git.**
- Used **Langchain** for low-latency conversational deployment and prompt robustness validation.

• Multi-Modal, Hierarchical RAG Agent System (LangChain + LangGraph):

- Built a **hierarchical agent system** using **LangGraph** to coordinate a planner agent and specialized sub-agents for retrieval, reasoning, and output formatting.
- Integrated multi-modal RAG pipeline combining **PDF/image parsing (Unstructured.io)**, **MySQL structured queries**, and **vector search (FAISS/Pinecone)**.
- Developed custom **LangChain tools** and **tool routers** for dynamic decision-making, recursive reasoning, and user query decomposition.
- Optimized latency and output quality using **Cohere rerankers**, prompt engineering, and asynchronous **LangGraph edges**.

[Traderverse] **[NYC, NY]**

[Machine Learning Engineer] *[12/2022] – [08/2023]*

• Quantitative LLM-Enhanced Strategy Development:

- Engineered data-driven trading strategies by integrating **transfer learning**, **LLM-based sentiment analysis**, **CNNs for candlestick pattern recognition**, and **feature engineering** on multi-modal datasets (text, price charts, news).
- Deployed **real-time prediction modules** using **Docker**, **scikit-learn**, **PyTorch**, and **Pandas**, enabling **automated backtesting**,

• Real-Time Market Briefing Automation (Autogen, GPT-4o, DevOps)

- Developed an **end-to-end** LLM-based briefing tool with **GPT-4o** and **Vertex AI, FMP AP**I, and **Autogen**, reducing report prep time by 50%.
- Automated data ingestion and transformation pipelines with **Python**, **Pandas**, and **HTML**, employing **DevOps workflows**, **CI/CD**, and **cron-based orchestration** to enhance analyst productivity and system reliability.

• Sentiment Analysis Pipeline for Financial NLP (Scikit-learn, Linux)

- Built NLP-driven sentiment models using **unsupervised learning techniques** with **mathematics** such as **K-Means**, **PCA**, and **factor analysis**, improving signal precision by **30%** for downstream **alpha generation** in investment strategies.

• End-to-End Predictive Modeling & Strategy Execution (LSTM, Transformers, GARCH)

- Built supervised models including **LSTM, transformer-based** architectures, and regression models to forecast market trends using features such as growth rates and volatility. Applied **Monte Carlo** simulation and **GARCH** to improve volatility modeling and enhance generalization using **Databricks.**
- Executed data-driven trading strategies by integrating **transfer learning, computer vision**, and LLM-based sentiment analysis, achieving high **return-to-risk ratio** through hyperparameter tuning and cross-validation.

• Index Construction & Portfolio Optimization (PCA, Time-Series Analysis)

- Collaborated cross-functionally to develop a multi-asset index using clustering and **PCA** using **Keras** for dimensionality reduction. Applied **time-series modeling** and **mean-variance optimization** to boost diversification and improve risk-adjusted returns. Visualization using Tableau and PowerBI.

  - Improved sentiment analysis accuracy by 18% and reduced processing latency by 40% through GPU optimization, model compression, and **CI/CD** integration. Applied advanced NLP preprocessing and feature extraction techniques to enhance downstream alpha signals.

- **Recommendation Engine & Real-Time Infrastructure (Reinforcement Learning, DevOps)**

  - Led the development of a recommendation system leveraging collaborative filtering and **reinforcement learning** and **deep learning**, improving user **engagement by 22%**. Built a real-time data pipeline with Docker and CI/CD automation, reducing data prep time by 40% and enabling continuous model updates.