Xinyu Tian

📞 (+1)919-717-6944   ✉️ xinyu.tian.1104@gmail.com   in linkedin.com/in/tian-xinyu/   ⌂ https://github.com/XinyuTian1104

EDUCATION

Duke University   **Aug. 2023 – May 2025 (Expected)**
M.S. in Computer Science   Durham, NC, USA

Duke University / Duke Kunshan University   **Aug. 2019 – May 2023**
B.S. in Interdisciplinary Studies (subplan: Data Science, Duke)   Durham, NC, USA
B.S. in Data Science (Duke Kunshan)   Suzhou, Jiangsu, China

WORKING EXPERIENCES

Machine Learning Engineer Intern | Baidu, Inc.   May 2024 – Aug. 2024, Beijing China

- Developed a **document processing pipeline** for RAG LLMs in a large-scale **AI-APaaS** platform (Baidu AppBuilder), integrating text detection, strategic slicing, and knowledge retrieval optimization using **Python**.
- Optimized **RAG pipeline** using **LlamaIndex** and **Langchain** and accelerated reference by **KV Cache**.
- Boosted **ERNIE-Speed** model (LLM) performance by **5%** on a comprehensive business dataset through **RAGT fine-tuning** and optimized a BERT reranker by **distilling knowledge** from GPT-4.
- Collaborated across distributed teams of product managers, front-end/back-end engineers, and architects to ensure successful product releases and authored 19 technical documents to support continuous updates.

Software Engineer Intern | Baidu, Inc.   Jun. 2022 – Aug. 2022, Beijing, China

- Worked on an **MLOps pipeline** with **MLFlow** for model tracking, **Docker** for containerizing models, and **Triton Inference Server** deployed on **Kubernetes** for scalable model inference. Leveraged **MLFlow** to manage the full model lifecycle, enabling seamless updates and version control in production environments.
- Applied the MLOps pipeline in an OCR task using the **COCO dataset** and **PaddlePaddle** framework, ensuring its effectiveness in a real-world scenario.

Big Data Intern | Yonyou Software Co. Ltd.   May 2021 – Aug. 2021, Hybrid

- Developed machine learning models leveraging **regression algorithms** and **time-series analysis** on data from **5000+** test soil fields, **8000+** experimental fields, and **18000+** plantations for crop growth prediction.
- Optimized a **Soil Big Data Pipeline** for a client agricultural company with **MongoDB** and **PostgreSQL** for multi-source data management, **integrating the machine learning models** for crop prediction.
- Developed an interactive **data visualization dashboard** utilizing **GIS** systems to visualize multi-source agricultural data for agricultural clients, using **ggplot** and **geoplotlib**.

PROJECTS

Web3Env: An Open-source ML-based Platform on Web3 ⌂
Supervised by Dr. Luyao Zhang

- Developed an interactive platform with Flask that embedded ML environment for Web3, offered RESTful API for databases and algorithms interaction, and supported non-coding programming via a designed UI.
- Encoded the Web3 stakeholders as RL agents and used our platform to work out the best staking strategies using the customized ML environment and building a data visualization dashboard.
- **First-authored a paper**, supervised by Dr. Luyao Zhang, **_Web3Env_** (shortly named), presented on SIGKDD'24 poster session, accepted by SDBD'24 (SIGKDD'24 workshop) and ChainScience'23.

RL-based Staking Mechanism Design on Ethereum 2.0 ↗ ⌂
Funded by Ethereum Academic Grants | Supervised by Dr. Luyao Zhang & Dr. Yulin Liu

- Encoded the dynamic financial environment on Ethereum blockchain as an RL model with Gym, designed a punishment scalar that regulates on-chain stakeholders' financial behaviors. Optimized the model with A2C/DDPG/PPO to find efficient dynamics that enforce good on-chain stakeholders' financial behaviors.
- Built a data analysis dashboard with retrieved 40,000+ groups of FinTech data from Ethereum API.
- **First-authored a paper**, supervised by Dr. Luyao Zhang & Dr. Yulin Liu, presented on 33rd International Conference on Game Theory, GAMES 2024, CCFCE'24, DOCS'24, and published on IEEE DOCS.

SKILLS

Programming Languages: Python, SQL, JAVA, C++, R, C#, C, Go, JavaScript
ML Frameworks/Tools: PyTorch, TensorFlow, LlamaIndex, Langchain, MATLAB, R Studio, Tableau