Xiaoyu Wang

Florida State University, Tallahassee, FL

+1 301-648-8554 xw22e@fsu.edu linkedin.com/in/xiaoyuwang98 github.com/drizzle98

Education

Florida State University **Expected Graduate: 05/2026**
Ph.D. Candidate in Statistics *GPA: 3.78*

University of Southern California **08/2020 - 05/2022**
Master of Science in Applied Data Science *GPA: 3.82*

University of Maryland **08/2016 - 05/2020**
Bachelor of Science in Applied Mathematics *Major GPA: 3.82, GPA: 3.77*

Relevant Coursework

- Linear Algebra & Calc
- Statistics Applications
- Statistics Inference
- Probability Theory
- Data Management
- Data Mining
- Machine Learning
- NLP

Projects

OneFlorida HIV Prediction Model Project **01/2024 - Present**
Researcher *Tallahassee, FL*

- **Data Pre-processing:** Utilized MySQL to store, filter, and explore datasets, ensuring efficient data handling and preparation for subsequent model training.
- **Variable Selection:** Applied dimension reduction and feature selection algorithms to identify key variables, optimizing model complexity and performance.
- **Model Training:** Trained machine learning models using Autogluon, evaluating model performance based on recall, precision, and F1 score to identify the best-fit models for specific outcomes.

LabGenie Project **01/2024 - Present**
Researcher *Tallahassee, FL*

- **Prompt Design:** Develop interactive prompts for laboratory results analysis (QA) system, enhancing user experience and system efficiency.
- **Database:** Implement vector database for efficient storage and retrieval of extensive lab test information.
- **Retrieval-Augmented Generation (RAG):** Connect the QA system and database using ChromaDB to enhance accuracy and authority.
- **Evaluate RAG System:** Compared the performance of a Retrieval-Augmented Generation (RAG) system with original LLMs, and submitted findings to the AMIA 2025 Informatics Summit.

BioCreative 8 Challenge **07/2023 - 03/2024**
Researcher *Tallahassee, FL*

- **Relation extraction:** Extract complex relationships from PubMed abstracts, significantly increasing data processing speed.
- **Finetune Models:** Finetune transformer-based models for NER task on medical domain, reach a F-1 score of 92%.
- **Finetune LLM:** Finetune Llama2 using Alpaca-LoRA to predict novelty among the relations, contributing to a 3.5% increase in F1 score.
- **NER via LLM:** Explore the method to outperform the ceiling of NER task using LLM.

Internships

Data Scientist Intern **05/2024 - 08/2024**
Insilicom LLC *Tallahassee, FL*

- **Data Extraction and Analysis:** Extracted data from knowledge graphs and generated comprehensive statistical tables to support data-driven decisions.
- **Pipeline Development:** Built and implemented a pipeline to create an update system for ClinicalTrials data, ensuring timely and accurate data updates.
- **Prompt Design and Testing:** Designed and tested prompts for Large Language Models (LLM) to enable the generation of comprehensive tables from clinical trial abstracts, improving data presentation and accessibility.

Skills

Languages: Mandarin (Native), English (Fluent)
Programming Language: Python, Matlab, SAS, R
Database: MySQL, MongoDB, Firebase, ChromaDB
Machine Learning: scikit-learn, Pytorch