Pei Tian

📞 (+1) 646-469-9928 | ✉ tptrix29@outlook.com | ⓖ GitHub | in LinkedIn | 🌐 Website

EDUCATION

Columbia University, The Fu Foundation School of Engineering and Applied Science   Sep 2023 – May 2025 (Expected)
M.S. in Data Science | GPA: **4** / 4   *New York City, NY*
- **Core Courses:** Machine Learning for Data Science, Natural Language Processing, Algorithms for Data Science, Applied Machine Learning, Data Science, Computer Systems for Data Science, Probability, Statistical Inference

Tongji University   Sep 2019 – Jun 2023
B.S. in Bioinformatics | GPA: **4.58** / 5, *Minor in Software Engineering* | GPA: **4.81** / 5   *Shanghai, CN*
- **Core Courses:** Data Structures (C++), Machine Learning Theory, Software Engineering, Foundation of Database, Micro-service and Web Service, Calculus, Linear Algebra, Discrete Math, Numerical Methods and Algorithms

SKILLS

Programming: Python, R, SQL, Java, C#, shell, HTML/CSS/JavaScript
Data Science: sklearn, PyTorch, TensorFlow, numpy, pandas, scipy, PySpark, HuggingFace, Transformers, PEFT, tidyverse, Shiny, Power BI
Concepts: Machine Learning, Deep Learning, Natural Language Processing, Computer Vision, Object-Oriented Programming, Data Structure, RESTful API, RDBMS, NoSQL, Agile Development, Cloud Computing, EC2

EXPERIENCES

AIQuraishi Laboratory, Columbia University   Apr 2024 – Aug 2024
Graduate Researcher   *New York City, NY*
- Collected and tidied 600k+ peptide datasets and 35 protein datasets with Python to ensure high-quality data for model training and benchmark
- Trained transformer-based language models by masked sequence modeling on Slurm-supported HPC to generate protein representation
- Conducted benchmark pipeline with 5 models including Neural Network, Query Attention and Contrastive Learning with PyTorch Lightning

Radical AI Inc.   May 2024 – Aug 2024
AI Engineer Intern   *New York City, NY*
- Engineered a chat-based course assistant leveraging the Google Gemini model, displaying quiz generation and personalized learning instruction
- Established a robust FastAPI backend to process diverse files (YouTube videos, Microsoft documents, etc.) with LangChain and ChromaDB
- Ensured high performance through meticulous unit testing with Pytest and comprehensive integration testing within Docker environments

Shanghai Foxhub Network Technology Company   Aug 2022 – Oct 2022
Data Engineer Intern   *Shanghai, CN*
- Formulated relational MySQL database architecture (ER diagrams) and managed unstructured data sources (OSS) on Alibaba Cloud
- Crafted shell scripts for database access permissions and backup operations, ensuring stability in production and development environments

PROJECTS

Custom LLM Chatbots with Character-Specific Tone   Aug 2024 – Present
- Scraped 100+ collections of chat datasets from public wiki websites by operating a web scraper built with BeautifulSoup and Selenium in Python
- Fine-tuned 3 state-of-the-art LLMs like LLaMA leveraging LoRA technique on the HuggingFace/PEFT platform to tailor specific tone of chatbot
- Constructed RESTful API with FastAPI as backend and a multi-page app with Streamlit as frontend for interactive usage of customized models

Billionaire Omics   Nov 2023 – Dec 2023
- Arranged 4 modules to perform exploratory data analysis (EDA) with tidyverse to uncover patterns of 10-years billionaires assets dataset
- Formed a Shiny App with 3 panels for interactive data exploration featuring dynamic visualizations in longitudinal and geographic prospective
- Created 9-entries Bootstrap-based website on GitHub Pages, showcasing comprehensive findings and insights about billionaires worldwide

Course Management System   Nov 2022 – Jan 2023
- Led 4-members group to construct a micro-service system utilizing Java and React with engaging in agile development process including requirement specification, system design, implementation and testing, achieving a web service application with 4 main functionalities
- Built a hybrid database structure with MySQL for relational data and MongoDB for archival data maintained separately with 2 Docker containers
- Implemented and tested 34 RESTful APIs with SpringBoot framework and interactive website with React, Node.js, Axios, Bootstrap, Webpack

Neurodegenerative Diseases Onset Prediction   Jun 2022 – Jul 2022
- Completed data collection and feature engineering on open-source patient data about the onset of Alzheimer's disease and Parkinson's disease
- Launched predictive models achieving 80%+ accuracy based on SVM, decision tree via sklearn and provided Flask website for interactive usage

PlantDB Desktop App   May 2022 – Jun 2022
- Delivered desktop app with 13 interactive interfaces and 3 roles for plant information retrieval and note-taking based on C# and VS.NET
- Devised and deployed a relational database on SQL Server platform to set up schema for user accessibility, note storage and plant searching

Solid Waste Composition Prediction with Neural Network   Mar 2022 – Oct 2022
- Collected and processed 36 solid waste datasets from Zhejiang Province to establish a robust foundation for model training and analysis
- Realized about 21% increase on metrics like Pearson coefficient of solid waste composition prediction task using neural network model with machine learning technologies including L2 regularization, Adam optimizer and dropout, batch normalization via PyTorch library
- Visualized data features and model evaluation results via matplotlib and Tensorflow library to give instructions for garbage processing schedule