Pei Tian

(+1) 646-469-9928 | tptrix29@outlook.com | Fort Lee, NJ, 07024 | GitHub | LinkedIn | Website

EDUCATION

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

Tongji University Sep 2019 – Jun 2023
BS in Bioinformatics | GPA: **4.58** / 5, *Minor in Software Engineering* | GPA: **4.81** / 5 *Shanghai, China*
- **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++, C#, shell, HTML/CSS/JavaScript
Data Science: Numpy, Pandas, Scipy, sklearn, pyspark, PyTorch, TensorFlow, HuggingFace, Transformers, Peft, tidyverse, ggplot2, Shiny, Power BI
Software Development: SpringBoot, FastAPI, Flask, MySQL, SQLServer, MongoDB, Docker, React.js, Axios.js, Node.js, Bootstrap, JUnit
Concepts: Machine Learning, Deep Learning, Natural Language Processing, Computer Vision, Graph Neural Network, Object-Oriented Programming, Data Structure, RESTful API, RDBMS, NoSQL, Agile Development, Cloud Computing (AWS, Google Cloud, Azure, Alibaba Cloud)

EXPERIENCES

AIQuraishi Laboratory, Columbia University Apr 2024 – Aug 2024
Graduate Researcher (NLP + MLOps) *New York City, NY*
- Collected and tidied 600k+ peptide datasets and 35 protein property datasets, ensuring high-quality data for model training and benchmark
- Trained transformer-based language models with PyTorch Lightning on Slurm-supported HPC to generate peptide/protein representation
- Conducted benchmark with various methods like Neural Network, Query Attention and Contrastive Learning on tasks such as motif detection

Radical AI Inc. May 2024 – Aug 2024
AI Engineer Intern (Langchain + FastAPI) *New York City, NY*
- Engineered a chat-based learning assistant using the Google Gemini model, featuring automated quiz generation and customized instruction
- Developed a robust FastAPI backend to process diverse file formats (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 (SQL + shell) *Shanghai, China*
- Designed relational database architecture (ER diagrams) and managed unstructured data sources (OSS) on Alibaba Cloud
- Wrote shell scripts for database access permissions and backup operations, ensuring stability in production and development environments

PROJECTS

Custom LLM Chatbots with Character-Specific Tone | *LLM, LoRA, NLP* Open Source Project, 2024 Summer
- Automated the collection of chat datasets from public wiki websites using a web scraper built with BeautifulSoup and Selenium in Python
- Fine-tuned state-of-the-art LLM like LLaMA using LoRA technique on the HuggingFace platform with customized dataset to acquire specific tone
- Developed RESTful API with FastAPI as backend and a multi-page app with Streamlit as frontend for interactive usage of cutomized models

Billionaire Omics | *R shiny, Github Pages* Data Science Course, 2023 Fall
- Conducted exhaustive exploratory data analysis (EDA) with tidyverse to uncover patterns and insights of billionaires assets dataset
- Developed a Shiny App for interactive data exploration featuring dynamic visualizations in longitudinal and geographic prospective
- Developed a Bootstrap-based website on GitHub Pages, showcasing comprehensive findings and insights about billionaires worldwide

Course Management System | *SpringBoot, React* Software Engineering Course, 2022 Fall
- Led the development of microservice-based system using SpringBoot and React.js as framework including requirement specification, system design, implementation and testing, resulting in a multi-functional and user-friendly web service application
- Designed a hybrid database structure with MySQL for relational data and MongoDB for archival data, maintaining isolation with Docker
- Implemented and tested RESTful APIs using SpringBoot and interactive website using React, Node.js, Axios, Bootstrap, Webpack

Neurodegenerative Diseases Onset Prediction | *Python, sklearn* Bioinformatics Algorithm Course, 2022 Summer
- Completed data collection and feature engineering on open-source patient data about the onset of Alzheimer's disease and Parkinson's disease
- Trained the prediction models with SVM, decision tree and delivered a website with Flask for project demonstration and interactive prediction

Solid Waste Composition Prediction with Neural Network | *PyTorch, Tensorflow* Innovative Project, 2022 Summer
- Collected and processed 36 solid waste datasets from Zhejiang Province to establish a robust foundation for model training and analysis
- Predicted solid waste composition 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 waste processing schedule

PlantDB Desktop App | *SQL Server, VS.NET* Database Course, 2022 Spring
- Delivered desktop app with interactive interface for plant information retrieval and note-taking with VS.NET framework
- Designed and deployed a relational database on SQL Server platform to set up schema for user accessibility, note storage and plant searching