Pei Tian

(+1) 646-469-9928 | tptrix29@outlook.com | GitHub | LinkedIn | Website

EDUCATION

Columbia University, The Fu Foundation School of Engineering and Applied Science Sep 2023 – Dec 2025 (Expected)
M.S. in Data Science | GPA: **4** / 4 *New York City, NY*

- **Core Courses:** Machine Learning, Natural Language Processing, Algorithms for Data Science, Reinforcement Learning, Unsupervised Learning, High Performance Machine Learning, Data Science, Computer Systems for Data Science, Probability, Statistical Inference, Modern Analysis

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++, C#, shell, HTML/CSS/JavaScript
Data Science: sklearn, PyTorch, TensorFlow, numpy, pandas, scipy, OpenCV, PySpark, HuggingFace, Transformers, PEFT, tidyverse, Shiny
Concepts: Machine Learning, Deep Learning, Natural Language Processing, Computer Vision, Transformers, Diffusion, Object-Oriented Programming, Data Structure, RESTful API, RDBMS, NoSQL, Agile Development, Google Cloud

EXPERIENCES

Data Science Institute, Columbia University Jan 2025 – Present
Research Scholar *New York City, NY*

- Processed a realistic phytoplankton image dataset while utilizing OpenCV for segmentation to obtain 1 million cell items across 200+ stations
- Applied unsupervised clustering algorithms including K-Means, Spectral Clustering, and DBSCAN to classify diverse phytoplankton species cluster based on both physical attributes and ResNet50-generated image embeddings

DitecT Laboratory, Columbia University Sep 2024 – Dec 2024
Graduate Researcher *New York City, NY*

- Fine-tuned a diffusion-based video generation model with 1.5k traffic collision scenarios video preprocessed by OpenCV and captioned by LLaVA
- Employed 2 phrases training procedure to enhance domain relatedness and temporal consistency separately with LoRA technique on GCP
- Evaluated performance of collision text-to-video generation model and achieve 0.8 Contrastive Language–Image Pretraining (CLIP) metric

AlQuraishi 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 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

PROJECTS

Automated Knowledge Graph Creation for GraphRAG Jan 2025 – Present

- Engineered a pipeline with 3 teammates by using LLMs to construct knowledge graph from an Amazon product dataset and store entities in Neo4j
- Designed Cypher queries and applied Leiden community detection algorithms to distill context for RAG workflows developed by LangChain
- Executed rigorous performance evaluation using DeepEval library, benchmarking GraphRAG pipeline against key LLM task metrics, attaining 0.58 faithfulness and 0.91 answer relevance, demonstrating improved contextual accuracy and reliability in generated responses

Exploration of Semantic Latent Spaces in Diffusion Models Sep 2024 – Dec 2024

- Undertook literature review to explore latent space (h-space) of DDIM model and its properties to accommodate semantic manipulation
- Applied 5 linear and non-linear dimension reduction algorithms (PCA, ICA, MDS, Random Projection, tSNE) to interpret and analyze latent representations within diffusion models, enhancing model interpretability and feature insights by extracting 6 main semantic dimensions

Recommendation System for Skin Care Product Oct 2024 – Dec 2024

- Embedded 100k+ review texts using advanced text embedding models (BAAI) to capture nuanced customer sentiment and contextual details
- Harnessed review embedding alongside product information to train 3 models (Linear Regression, Random Forest, XGBoost) to predict rating
- Incorporated collaborative filtering methods including explicit/implicit/hybrid matrix factorization to build recommendation systems while realizing 0.83 recall@5 and 0.78 precision@5

Custom LLM Chatbots with Character-Specific Tone Aug 2024 – Sep 2024

- 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 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