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*
- **Courses:** Machine Learning, Natural Language Processing, Algorithm Analysis, Reinforcement Learning, Unsupervised Learning, High Performance Machine Learning, Data Science, Computer Systems for Data Science, Probability, Statistical Inference, Scaling LLM Systems

Tongji University Sep 2019 – Jun 2023
B.S. in Bioinformatics and Computer Software Engineering | GPA: **4.61** / 5 *Shanghai, China*
- **Courses:** Machine Learning, Software Engineering, Database Foundations, Micro-service and Web Service, Numerical Algorithms

EXPERIENCES

Kaliber AI Jun 2025 – Present
Machine Learning Researcher Intern *Santa Clara, CA*
- Achieved 99% accuracy on finetuned transformer-based model for speech recognition and 1.3x speed by TensorRT triton inference server
- Finetuned ViT/ResNet-based models on image recognition tasks with 90+% accuracy while applying SAM2 to enhance tracking functionality
- Integrated 3D Object Detector and tool functionalities to develop AI-agentic systems enabling robotic decision-making and task execution

L'Oréal Jan 2025 – May 2025
Machine Learning Engineer Intern *New York City, NY*
- Led 4-team to engineer a knowledge graph RAG pipeline with 3 LLM-based methods combining OpenAI and Neo4j for product recommendation
- Designed NoSQL queries with KNN and graph community detection algorithms to enhance RAG workflows to customize customer support
- Attained 0.91 answer relevance and 0.58 faithfulness on question answering task when benchmarking GraphRAG pipeline using Llamaindex

Department of Computer Science, UAlbany Apr 2025 – Present
Research Engineer *New York City, NY*
- Improved 8% on Exact Match and 9% F1 for multi-hop QA tasks by fine-tuning small-scale LLMs with margin-aware preference learning method
- Customized 3 preference learning trainers variants based on DPO, ORPO, and CPO in trl (reinforcement learning library) for reasoning tasks

Data Science Institute, Columbia University Jan 2025 – May 2025
Research Scholar *New York City, NY*
- Tidied 1 million realistic images while utilizing OpenCV to improve data quality and MySQL on AWS platform to facilitate data management
- Applied unsupervised clustering algorithms KMean, DBSCAN with sklearn to assist labeling and train ResNet18 using PyTorch for classification

DitecT Laboratory, Columbia University Sep 2024 – Dec 2024
Computer Vision Graduate Researcher *New York City, NY*
- Fine-tuned a diffusion-based video generation model with 1.5k traffic scenarios video captioned by LLaVA on HPC with NVIDIA H100 GPU
- Evaluated text-to-video model built by vision transformer, LoRA and achieve 0.8 Contrastive Language–Image Pretraining (CLIP) metric

PROJECTS

Reinforcement Fine-Tuning for Reasoning Enhancement in LLM Feb 2025 – May 2025
- Incorporated 3 Reinforcement Learning algorithms such as GRPO with Causal Language Model and LoRA in PEFT for reinforcement fine-tuning
- Enhanced 1.7% exact match accuracy performance of lightweight Qwen2.5 model on math reasoing task like GSM8K with trl implementation

Efficient Knowledge Distillation for Knowledge-based Tasks Feb 2025 – May 2025
- Increased 9% performance of small models by distilling knowledge from BERT/Qwen2.5 for classifications, language modeling, summarization
- Accelerated 19% running speed by Flash Attention, mixed precision, PyTorch Dynamo for training and vLLM (Page Attention) for inference

Controlling Generative Diffusion Models with Unsupervised Machine Learning Algorithms Sep 2024 – Dec 2024
- Steered 3 teammates to applied 5 unsupervised learning algorithms for dimension reduction (PCA, ICA, MDS, Random Projection, tSNE) to analyze latent representations within diffusion models, enhancing model interpretability and feature insights by extracting 6 semantic dimensions
- Optimized 56% running time by integrating DDIM scheduler for image generation in U-Net model with analysis on low-dimension latent space

Recommendation System for Skin Care Product Oct 2024 – Dec 2024
- 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

SKILLS

***Machine Learning*:** sklearn, Regression, Bagging, Boosting, Supervised Learning, Featuring Engineering, Deep Neural Network (DNN)

***Deep Learning*:** PyTorch, Tensorflow, HugingFace, accelerate, Megetron, distributed training, Ray, Lightning, Deepspeed, Optuna, PEFT, HPC

***Natural Language Processing*:** transformers, RNN, LSTM, BERT, GPT, T5, LLaMA, LangChain, LlamaIndex, RAG, Agent, nltk, spaCy

***Computer Vision*:** diffusion, torchvision, diffusor, OpenCV, Pillow, CNN, ResNet, YOLO, UNet, DDPM/DDIM, ControlNet, ViT, CLIP, VLM

Others: numpy, pandas, PySpark, Hadoop, Docker, Kubernetes, Rabbit MQ, Kafka, Flink, Neo4j, AWS, GCP, Azure, Distributed system