Jirui DAI

Phone:86+ 13955124219 & Email: Jdai27@jh.edu

EDUCATION

Nanchang Hangkong University(NCHU) Nanchang, China
Bachelor of Engineer in Software Engineering; Major GPA: 3.329/4.0 *09/2020-06/2024*
Key Courses: Software Engineering (A), Software Modeling (A), Software Project Management (A), Software Quality Assurance and Testing (A), Software Design and Architecture (A), Programming Training(A)
Johns Hopkins University(JHU) Baltimore, America
Master of Science in Engineering in Computer Science *08/2025-06/2027*
TOEFL: 103
GRE: 332(verbal:162; quantitative:170; analytical writing:4.5)
Computer Skills: C(2 yrs), C++(2 yrs), Java(4 yrs), Arkts(1 yrs), Python(1 yrs), C＃(1 yrs), etc.

PUBLICATION

1.Dai, Jirui. "Comparative analysis of federated learning algorithms under non-IID data." Applied and Computational Engineering (2024) DOI: 10.54254/2755-2721/86/20241581
https://www.researchgate.net/publication/382753798_Comparative_analysis_of_federated_learning_algorithms_under_non-IID_data

RESEARCH EXPERIENCE

Project 1:Clinical Validation Framework for Traditional Chinese Medicine (TCM) Practice
Research Assistant supervised by PhD. Zhi Liu,a postdoctoral researcher at Nanjing University of Chinese Medicine & Stanford Scholar *Upcoming*

- Designing an AI-driven validation system to quantify TCM clinical efficacy using biological data analysis, addressing long-standing gaps in evidence-based TCM research
- Integrating heterogeneous datasets: clinical cases, herbal formulae, TCM theory networks, and pharmacology profiles for cross-modal analysis
- Collaborating with hospitals to curate proprietary clinical datasets (in progress)

Project Two:TCM Heritance Framework via Structured LLM Training
Research Assistant supervised by PhD. Zhi Liu,a postdoctoral researcher at Nanjing University of Chinese Medicine & Stanford Scholar *Dec. 2024 – Aug. 2025*

- Developed China's first comprehensive TCM LLM framework using QWEN-2.5, creating 600K+ structured medical datasets from textbooks, clinical records, and expert knowledge
- Pioneered RAG-SFT (novel fine-tuning paradigm): Integrated expert knowledge retrieval into CoT data via GTE embeddings, improving dialectical reasoning accuracy by 32%
- Engineered full training pipeline: CPT → Cold Start SFT → GRPO (rejection sampling) → RAG-SFT → KTO alignment
- Built exclusive expert knowledge base covering 5 renowned TCM physicians' lifetime clinical insights
- Established industry-first TCM benchmark using Delphi method + GTE/GPT-4 similarity scoring
- Output: Preparing SCI Q1 journal paper (first-author); open-sourcing framework/training code

Project Three: Training and Evaluation of a Traditional Chinese Medicine (TCM) and Western Medicine Vertical Domain Model
Researcher supervised by PhD JiaxiYang, an associate researcher at Columbia University
Jun.2024-Dec.2024

- Preprocessed the open-source datasets to compile two high-quality datasets of 300,000 entries, each suitable for incremental pre-training and supervised fine-tuning, and manually annotated a third dataset with 10,000 samples for preference alignment
- Chose the Llama-3.1-8B-zh model and Llama-Factory framework for training on the AutoDL platform
- Incremental pre-training: using Low-Rank Adaptation (LoRA) to fine-tune the model by modifying parameters, which helped reduce the loss value from 14.0965 to 5.472
- Supervised fine-tuning: also utilizing the LoRA technology to train the model and achieved loss value decreasing from an initial 7.2746 to 4.400 in the mid-phase, and finally to 3.43

- Preference alignment: leveraging the Direct Preference Optimization (DPO) to align model outputs with user preferences (in progress)
- Next, I plan to evaluate the model's effectiveness in handling tasks related to TCM and Western Medicine on the OpenCompass platform

Project Four: Comparison of Federated Learning Algorithms for Predicting Results Based on the Fashion-MNIST Dataset in Non-IID Data Environments

Research Leader supervised by Prof. Soummya Kar from Carnegie Mellon University

Mar.-May 2024

- Collaboratively decided four federated learning algorithms for comparison, i.e., FedAvg, FedSGD, SCAFFOLD, and FedProx
- Used the FedSGD algorithm to implement the four models, i.e., CNN, FCNN, LSTM, ResNet, ultimately selecting CNN as the best-performing model for text classification tasks; implemented four federated learning algorithms in the CNN model, identifying FedProx as the most effective one among them
- Designed an experimental framework based on the Fashion-MNIST dataset, including data preprocessing, model architecture design, and evaluation metrics selection
- Simulated a non-independent and identically distributed (non-IID) data environment
- Responsible for coding and optimizing two specific algorithms, i.e., FedProx and SCAFFOLD
- Wrote distributed computing scripts to allocate experimental tasks across multiple machines for parallel execution, overcoming computational resource limitations
- Collected experiment data, interpreted data analysis results, and across all evaluation metrics, found that FedProx outperformed all, followed by SCAFFOLD and FedAvg, with FedSGD performing the worst.
- Achievement: based on this project, I individually composed a research paper published at the 6th International Conference on Computing and Data Science.

Project Five: LMSYS - Chatbot Arena Human Preference Predictions (Kaggle Competition)

Research Leader of a Four-person Team

Jun.- Aug.2024

- Split 20% of a competition dataset(user interactions from the ChatBot Arena) as a training validation set
- Trained two LLMs, i.e., gemma-2-9b and llama-3.1-8b, using the optimal configurations determined through adjusting the parameters (learning rate, frozen layers, prompt lengths, etc.)
- Leveraged the validation set to assess the two models' performance
- Utilized ensemble learning techniques to assign weights to the outputs of two models and then combine these outputs through weighted summation
- Achievement: Upon evaluating the logarithmic loss between the predicted probabilities and the actual values, our team achieved a silver medal in the competition.

Project Six: Designing a Text-CNN Model for Identifying and Classifying Social Hot Topics

Undergraduate Thesis, the 1st Author

Dec.2023- Mar.2024

- Collected news text from major news websites, and performed data cleaning and annotation
- Chose the Text-CNN Model to handle news text features and optimized the model structure through adjusting the number of convolutional layers and the size of convolutional kernels
- Incorporated dropout technology to enhance the model's generalization ability and prevent overfitting
- Used the cleaned and annotated dataset for model training, enabling to identify and classify hot topics in real time automatically
- Achievement: The model exhibited high accuracy on the training set and showed good generalization ability on the validation and test sets.

Project Seven: Improving the YOLOv7 Model for Object Detection in Drone Imagery(A National Project)

Undergraduate researcher supervised by Prof. Yun Ge from School of Software, NCHU

Mar.2022- Mar.2023

- Collected 6,471 training images, 548 validation images, and 3,190 test images with ten kinds of annotated objects using various high-definition drone cameras

- Integrated the C3STR module based on Swin Transformer, C3DCN module based on Deformable Convolution, and MP-InceptionNeXt Downsampling module into the YOLOv7 model to optimize the performance of drone image object detection
- Achievement: On the testing imagery dataset, the optimized model achieved a 3.1% increase in mAP value compared to the base model (YOLOv7) and demonstrated great improvements in detection performance across various drone imaging scenarios.

INTERNSHIP EXPERIENCE

Byte Dance Beijing, China

NLP Algorithm Intern in the Algorithm Department of TikTok *Sept. 2024-Dec.2024*

- Collected and preprocessed video comment data from TikTok, progressively increasing the data volume to 100,000 while implementing various techniques to achieve a clean and well-structured dataset
- Leveraged the SnowNLP library to perform sentiment analysis and computed the ratio of positive comments to negative comments
- Conducted a literature review on sentiment analysis and presented my findings during group seminars
- Reproduced a Named Entity Recognition (NER) algorithm from a reputable conference paper through annotating my dataset into BIOES format, developing the algorithm's components, and integrating them into a cohesive NER model for training and evaluation
- Optimized the NER algorithm by updating the pre-trained model, enhancing attention mechanisms, performing data augmentation, and implementing the mixed precision training and Dropout for model training
- Reproduced another NER algorithm published in a different conference paper using the same dataset
- Compared both models' performance based on their loss values and F1 scores on the validation set and test set

AWARDS, SCHOLARSHIPS & LEADERSHIP

- Won the Silver Medal in the LMSYS - Chatbot Arena Human Preference Predictions (Top 2%) *Aug.2024*
- Obtained the Third Prize in the Lanqiao Cup National Software and Information Technology Professional Talent Competition (Top10%) *May.2022*
- Awarded the Third-Class Scholarship at NCHU(Top7%, Three Times) *Mar.2022&2023&2024*
- Monitor of Class 15 of 2020 at School of Software, NCHU *Sept.2021-Jun.2024*