QIRUN DAI

(+1) 530-709-2773 ⋄ qirundai@illinois.edu/daiqirun19@gmail.com ⋄ Website

EDUCATION

Fudan University, Shanghai *Sept. 2021 - Present*
B.Eng. in Artificial Intelligence (Honor Class, Data Science Track) **GPA:** 3.89/4.00; **Rank:** 1/25
Course Highlights: Mathematical Analysis I/II/III (A), Advanced Linear Algebra (A), Methods of Optimization (A), Stochastic Processes (A), Data Fusion and Assimilation (A), Set Theory and Graph Theory (A), Data Structures and Algorithm Design (A), Artificial Intelligence (A), Natural Language Processing (A)

University of California, Davis *Sept. 2023 - Dec. 2023*
Exchange Student in Computer Science **GPA:** 4.00/4.00
Course Highlights: Operating Systems (A), Numerical Linear Algebra (Graduate, A), Advanced Statistical Learning (Graduate, A+)

RESEARCH INTERESTS

My research interests broadly span **natural language processing** and **machine learning**. These days, I have an intense interest in advancing the capabilities and our understanding of **large language models (LLMs) from a data-centric perspective**, with a specific focus on the following topics:

- Data Curation and Synthetic Data.
- Data Attribution and Interpretability of Model Behavior.
- Data-efficient Learning Algorithms.
- Datasets for Trustworthy and Efficient Evaluation.

PUBLICATIONS

Improving Influence-based Instruction Tuning Data Selection for Balanced Learning of Diverse Capabilities
Qirun Dai, Dylan Zhang, Jiaqi W. Ma, Hao Peng.
Under Review

Demonstration Distillation for Efficient In-Context Learning
Tong Chen, **Qirun Dai**, Zhijie Deng, Dequan Wang.
Submitted to ICLR 2024

RESEARCH EXPERIENCE

University of Illinois Urbana-Champaign *Apr. 2024 - Present*
Research Intern, Department of Computer Science

- **Improving Influence-based Data Selection for Multi-task Instruction Tuning**
  Advisor: Prof. Hao Peng and Prof. Jiaqi W. Ma *Apr. 2024 - Oct. 2024*
  - When instruction tuning LLMs for learning multiple diverse tasks, we identified the poor performance of data selection methods built upon gradient-based influence estimation techniques, and attributed this problem to an inherent bias in cross-task influence.
  - We then proposed BIDS, a simple and effective influence-based **D**ata **S**election algorithm that addresses this problem and selects **I**nfluential data for **B**alanced capability learning.

- When training on UltraInteract, a SOTA high-quality dataset designed to enhance diverse reasoning capabilities, we showed that a 15% subset selected by BIDS can outperform full-dataset training in terms of the overall performance on 7 benchmarks spanning coding, math, STEM, logical reasoning and instruction following. We further provided in-depth analyses on what might be the good properties of a balanced set of influential data.
- This work resulted in a first-authored paper currently under review.

Shanghai Jiao Tong University *Jan. 2023 - Mar. 2024*
Research Intern, Qing Yuan Research Institute

- **Demonstration Selection for Knowledge-Intensive In-Context Learning**
  Advisor: Prof. Dequan Wang and Prof. Zhijie Deng *Nov. 2023 - Mar. 2024*
  - Proposed an ICL demonstration selection method built upon sparse retrieval techniques such as BM25, targeting knowledge-intensive QA tasks where existing methods based on pre-trained embedding models can be easily confused due to a lack of domain-specific knowledge.
  - Achieved more than 4% accuracy improvement in multiple challenging domains including medicine and college mathematics, which were rarely explored by previous ICL research.
  - Explored in depth what makes a good demonstration for domain knowledge-intensive ICL.
- **Demonstration Distillation for Efficient In-Context Learning**
  Advisor: Prof. Dequan Wang and Prof. Zhijie Deng *July. 2023 - Sept. 2023*
  - In order to optimize context efficiency for LLM in-context learning, we developed DGS, a multi-LLM-agent framework that iteratively compresses the given demonstrations with **D**istillist, **G**eneralist and **S**pecialist.
  - DGS achieved up to a 4.3x compression ratio and a 5% accuracy improvement on various QA tasks, reducing inference overheads while eliciting stronger reasoning capabilities.
  - This work resulted in a second-authored paper submitted to ICLR 2024.
- **The Forward-Forward Algorithm: Some In-Depth Investigations**
  Advisor: Prof. Dequan Wang *Jan. 2023 - May 2023*
  - Generalized the Forward-Forward learning Algorithm (FFA) proposed by Geoffrey Hinton to modern CNNs and Vision Transformers (ViTs).
  - Inspired by contrastive learning paradigms, explored different methods of generating positive and negative training data, and how they boost the learning efficiency of FFA.
  - To tackle the inherent training instability of FFA, conducted ablation studies on various factors including residual connection, normalization methods, loss function design and learnability of threshold values, resulting in an FFA training recipe especially for large modern vision models.

TEACHING EXPERIENCE

Introduction to Computer Systems (DATA130025) *Fall Semester 2024*
Teaching Assistant at the School of Data Science, Fudan University, with Prof. Jiaqing Liang

- Prepared tutorials for linux basics, and took charge of two lab assignments related to operating systems: ShellLab and MallocLab.
- With the ever-increasing importance of computer system knowledge in LLM research, focused on not only laying the foundation of system education, but also guiding my students into the fascinating world of LLMs and MLSys.

HONORS & AWARDS

Deans' Honors List (Top 2% in the College of Letters and Science)
University of California, Davis, Fall Quarter 2023

Academic Excellence Scholarship (Ranked first in the major)
Fudan University, 2022 - 2023

Panasonic Scholarship (Ranked first in the major)
Fudan University, 2021 - 2022

SKILLS

Programming: C/C++, Python, CUDA, MATLAB, R, HTML/CSS/JavaScript, SQL, LaTeX
Tools and Frameworks: Pytorch, HuggingFace, OpenMP, MPI, Vue
Languages: English (TOEFL iBT 107, speaking 24), Chinese (native)