Xixian Liu

✉ xixianliu0@gmail.com · ✉ 21307110316@m.fudan.edu.cn · ☏ (+86) 188-174-43009

Education

Fudan University, Shanghai, China   Sept 2021 – June 2025 (Expected)

Major: Computer Science and Technology (Bachelor of Science)
GPA: 3.61 / 4.0
Related Courses:

Machine Learning (A), Probability Theory and Mathematical Statistics (A), Linear Algebra (A), Artificial Intelligence (A-), Computer Visualization (A), Programming (A), Object-Oriented Programming (A), Data Structures (A)

Research Experience

Material Foundation Model (MatterSim - Microsoft Flagship Model for Materials Design) Jan 2024 – Present

Supervisor: Ziheng Lu; Microsoft Research
Overview: **Core developer and trainer of MatterSim Microsoft's flagship foundation model for materials and molecular design.**

- Developed a scalable Transformer-based equivarnt architecture (up to 1B pretrained from scratch) for large scale atomistic data (up to 35M sampled with active learning).
- Explored scaling law of deep learning models for diverse atomistic simulation tasks on both model size and data size.
- Achieved SOTA on almost all materials tasks with a large margin including Matbench and Matbench Discovery

Status: Preprint of version 1 available at https://arxiv.org/abs/2405.04967; version 2 submitted to *Nature portfolio.* **GitHub:** https://github.com/microsoft/mattersim

Large Scale Foundation Model for Molecules, Materials, and Proteins   Jan 2024 – Present

Supervisors: Chang Liu, Shi Yu; Microsoft Research
Overview: Developed a multimodal foundation model integrating molecules, proteins, and materials, for both prediction and generation tasks

- Leveraged a Graphormer-based architecture incorporating diffusion model designs (e.g., DiT) for simultaneous property prediction and structure generation across multimodal data.
- Achieved unified intelligent perception through balanced training across diverse modalities (molecules, proteins, materials).

Status: Manuscript in preparation.

Content Awareness Tokenizer for LLM   Oct 2024 – Present

Supervisor: Jian Li; Tsinghua University

- Mainstream tokenization methods for LLM, such as BPE, WordPiece, and Unigram, have inherent limitations. We developed a language-model-driven automatic tokenization method to address these.
- The method begins with the smallest tokens of a word, performing encoding iteratively. Merge probabilities are derived from model predictions, and the Viterbi algorithm is used to determine the optimal encoding sequence.
- The proposed approach significantly reduces model perplexity compared to traditional tokenization techniques.

status: Experiments ongoing.

iFuCore: Superscalar Out-of-Order CPU   Jun 2023 – Sept 2023

Achievement: **1st Prize in the 2024 National College Student Computer System Ability Competition (Loongson Cup).**

- Designed an out-of-order superscalar CPU based on the LoongArch32-Reduced (LA32R) instruction set, incorporating a Fetch Buffer for front-end/back-end decoupling.
- Achieved the highest IPC among competition teams.
- Successfully passed all functionality and performance tests on the Loongson experimental platform; capable of running the Linux OS via u-boot.

GitHub: https://github.com/Anon-233/iFu

Language Model for Feature Engineering   Sept 2023 – Jan 2024

Supervisor: Qian Liu; Sea AI Lab
Collaborator: Shaowen Wang; Tsinghua University
Overview: Investigated the use of large language models (LLMs) and reinforcement learning (RL) for automated feature engineering on tabular data.

- Transformed table and feature information into textual formats for LLM training (e.g., Flan-T5, GPT-2).
- Enhanced model performance for specific table types using RL techniques (e.g., PPO, DPO) and advanced sampling methods (e.g., rejection sampling).
- Demonstrated significant improvements in downstream task performance post-feature selection.

Status: Manuscript in preparation.

INTERNSHIP EXPERIENCE

Research Intern, Microsoft Research   Jan 2024 – Present

- Contributed to the development and training of the flagship MatterSim model for materials design group, leveraging Transformer-based architecture and large-scale atomistic datasets.
- Worked on the Science Foundation Model group, integrating multimodal data (molecules, proteins, materials) for prediction and generation tasks.
- Conducted comprehensive literature reviews and supported academic writing for publications.
- Collaborated with interdisciplinary teams to advance AI solutions in science.

Teaching Assistant for Operating Systems, Fudan University   Aug 2024 – Present

- Designed labs on system calls, process switching, and file systems, inspired by Linux lab exercises.
- Conducted one-on-one lab interviews with students to evaluate and guide their progress.

Teaching Assistant for Introduction to Computer Systems, Fudan University   Aug 2023 – Jan 2024

- Designed the disassembly exercise lab, *Boom Lab*, and the final course project, *CPU Simulation*.
- Supervised 9 students, delivered 2 lectures, and conducted 2 exercise sessions, focusing on *program performance optimization* and *CPU design*.
- Graded all assignments and exercises for the class.

Lab Design GitHub: https://github.com/cjinfdu/ics24

HONORS AND AWARDS

Bronze Medal, 2022 ICPC Asia Regional Contest (ACM-ICPC)	Nov 2022
Silver Medal, 2022 China Collegiate Programming Contest (ACM-CCPC)	Sept 2022
Outstanding Student Scholarship (3rd Class, Top 20%), Fudan University	Dec 2022
Outstanding Student Scholarship (3rd Class, Top 20%), Fudan University	Dec 2021

SKILLS

- **Model scaling:** hands-on experience with large-scale model training with 256 A100 cluster.
- **Programming Languages:** Python, C/C++, SystemVerilog.
- **Frameworks and Systems:** PyTorch, Linux, Git.
- **Hardware Experience:** FPGA, Raspberry Pi (with GPIO).