Yao Xiao

(+86) 186-2182-3612 | ✉ yx2436@nyu.edu | 🏠 charlie-xiao.github.io | Charlie-XIAO | in yao-xiao-200073244

Education

Partially taken at NYU Courant, courses marked with * are at graduate-level

NYU Shanghai | Bachelor of Science | Honors Mathematics | Computer Science | **GPA: 3.90/4.00** 2020.09 – present

- **Honors Mathematics GPA: 4.00/4.00**, including: Honors { Analysis, Theory of Probability, Numerical Analysis, Algebra }, Partial Differential Equations, Modeling and Simulation, Complex Analysis*, Stochastic Calculus*, Probability Limit Theorems, etc.
- **Computer Science GPA: 3.97/4.00**, including: Data Structures, Computer Architecture, Algorithms, Operating Systems, Open Source Software Development, Randomized Algorithms, Machine Learning*, Computer Networking, Software Engineering, etc.

Research Experience

Efficient Distributed Serving System for Inference of Large Language Models 2023.09 – present

Advisor: Professor Guyue Liu, guyue.liu@gmail.com

- Enabled larger batch sizes beyond KV cache limit for layers except self-attention, observing that only self-attention relies on KV cache.
- Batched prefills and decodes dynamically in self-attention to mitigate pipeline bubbles caused by varying transformer input lengths.
- Packed multiple short attention computations with the longest one, while concurrently swapping KV cache to minimize overhead.

Privacy-Preserving Network Configuration Sharing via Anonymization | Submitted for Publication 2022.10 – 2023.04

Advisor: Professor Guyue Liu, guyue.liu@gmail.com

- Anonymized network topology and routing paths via a twin network approach, which prevailing anonymization methods overlook.
- Designed a re-routing algorithm to avoid potential violations in routing utilities caused by anonymization.
- Formulated and mathematically proved that our solution preserves essential routing utilities (e.g. multipath consistency).

Efficiently Visualizing Large Graphs | Dean's Undergraduate Research Fund (DURF) | Publication 2022.05 – 2022.08

Advisor: Professor Jie Xue, jiexue@nyu.edu

- Designed t-SGNE specialized for graphs, leveraging the neighboring relations between nodes and achieving 6.7x computation efficiency.
- Proposed SPLEE, a graph embedding method based on Laplacian eigenmaps and shortest paths, intended to suit t-SGNE.
- Combined SPLEE and t-SGNE for visualization of graphs with 300K nodes and 1M edges, achieving 10% improvement in visual effect.

Working Experience

DISC Lab, Fudan University | Core developer of DISC-LawLLM | Publication 2023.05 – 2023.08

- Constructed 403K instruction data and fine-tuned DISC-LawLLM, a large language model specialized in Chinese legal domain.
- Built a retrieval module for DISC-LawLLM and constructed its knowledge database with 800+ Laws and 24K+ legal examinations.
- Designed an evaluation framework from objective and subjective perspectives. DISC-LawLLM outperformed the base model by 23% and GPT-3.5 Turbo by 9%, and has currently 9K+ users.

NYU Courant / NYU Shanghai | Tutor / Learning Assistant Fall 2021, Spring 2023, Fall 2023, Spring 2024

- Tutored CSCI-UA.0202 Operating Systems (supervised by Professor Yang Tang) during Spring 2023 at NYU Courant.
- Tutored MATH-SHU.0140 Linear Algebra during Spring 2024 at NYU Shanghai.
- Tutored MATH-SHU.0131 Calculus during Fall 2021 and Fall 2023 at NYU Shanghai.

Projects

ml3m: Multilevel Evaluation Framework for Large Language Models | GitHub 2023.05 – present

- Leveraged GPT's natural language ability for evaluating a breadth of natural language tasks (e.g. multiple choice questions, Q&A).
- Utilized asynchronous I/O to enable 300x efficiency, making large-scale evaluation via GPT possible.
- Designed powerful API for data generation and evaluation, used by DISC-LawLLM, a powerful law LLM of Fudan University.

scikit-learn | Core Developer | Rank **#43** Contributor | GitHub 2023.04 – present

- Participated in maintenance and testing, as well as bug fixes and new features in preprocessing, decomposition, metrics, tree models, etc.
- Redesigned the whole scikit-learn website, along with improvements in the documentation and various UX enhancements.
- Contributed 88 pull requests to its codebase and documentation, and was nominated into the scikit-learn Triage Team.

pandas | Rank **#66** Contributor | GitHub 2023.03 – present

- Fixed bugs in a breadth of data analysis operations like groupby, missing data interpolation, resample, etc.
- Contributed 29 pull requests to its codebase and made tutorials and examples in the pandas API documentation.

YouTube Interface Customizer | Course Project | GitHub 2023.02

- Built a Firefox extension that supports changing color themes, rearranging, and customizing elements of the YouTube interface.
- Created the documentations of features and contribution guides, and released (self-distributed) v1.0 at Mozilla Add-ons.

Inequality Process Simulation | Course Project | Paper | Demo 2022.12

- Simulated inequality process in economic systems via nuanced random transactions functions, reflecting on real-world economy.
- Discovered that the final distribution of wealth in a real-world economic system fits the shape of a gamma or beta prime distribution.

Gyro-Tower Simulation | Course Project | Paper | GitHub 2022.10

- Modeled gyroscopes as networks of springs, formulated the system with differential equations, and solved it via Euler's method.
- Simulated vertical stacks of gyroscopes, and found that they obeyed gyroscopic precession assuming a flexible middle axle.

PUBLICATIONS

AUTHORS WITH † ARE SORTED BY α-β ORDER, OTHERS ARE SORTED BY CONTRIBUTION

[1] Y. Wang, Q. Men, **Y. Xiao**, Y. Chen, and G. Liu, “ConfMask: Privacy-Preserving Network Configuration Sharing via Anonymization,” *submitted for publication*.

[2] X. Li†, **Y. Xiao**†, and Y. Zhou†, “Efficiently Visualizing Large Graphs,” October 2023. doi:10.48550/arXiv.2310.11186

[3] S. Liu†, C. Shen†, **Y. Xiao**†, S. Yue†, Y. Zhou†, W. Chen, S. Wang, B. Li, S. Yun, X. Huang, and Z. Wei, “DISC-LawLLM: Fine-Tuning Large Language Models for Intelligent Legal Services,” September 2023. doi:10.48550/arXiv.2309.11325

HONORS AND AWARDS

[1] **Meritorious Winner**, Mathematical Contest in Modeling, 2023

[2] **Most Appointment Award for Learning Assistants**, NYU Shanghai, Fall 2021, Fall 2023

[3] **Dean’s List of Academic Year**, NYU Shanghai, 2020 – 2021, 2021 – 2022, 2022 – 2023

SKILLS

[1] Programming: Proficient in Python; Intermediate in C, Java, HTML, CSS, JavaScript; Familiar with PHP, C++, Julia, MATLAB.

[2] Others: Git, GitHub, Docker, numpy, pandas, scikit-learn, pytest, asv benchmarks.