Yao Xiao

(+1) 617-216-7284 | yaoxiao@g.harvard.edu | charlie-xiao.github.io | Charlie-XIAO | yao-xiao-200073244

Education

Harvard University | Master of Science | Computational Science and Engineering 2024.09 – 2026.05 (expected)

- GPA: 3.92/4.00, including: Computer Networks, Data Systems, Distributed Systems, Parallel Computing, etc.

New York University Shanghai | Bachelor of Science | Honors Mathematics | Computer Science 2020.09 – 2024.05

- Honors Mathematics GPA: 4.00/4.00, including: Linear Algebra, Math Modeling, Probability Theory, Numerical Analysis, etc.
- Computer Science GPA: 3.97/4.00, including: Data Structures, Algorithms, Operating Systems, Software Engineering, etc.

Skills

[1] **Programming:** Proficient in Python, Rust, TypeScript; Intermediate in C/C++, Java, Julia, MATLAB
[2] **Frameworks and packages:** CUDA; SIMD/AVX; OpenMP; Tauri, React; Numpy, Pandas, Polars, Scikit-learn, PyTorch
[3] **DevOps:** Docker; Git; AWS, GCP; Ansible; Kubernetes; CircleCI, GitHub Actions; Computer networks; Distributed systems; Linux

Working Experience

Scikit-learn | Open Source | Core Developer | 128 Merged Pull Requests 2023.04 – present

Skills: Python, Cython, JavaScript, Sphinx, scikit-learn, numpy, scipy, pandas, polars, CI/CD

- Managed maintenance tasks e.g., test suite coverage, code refactoring, developer API improvement, automated GitHub workflows, etc.
- Enhanced sparse array and polars dataframe support, estimator representation, metrics visualization, multilabel data cross-validator, etc.
- Optimized IncrementalPCA on sparse data (>10x faster, >30x less memory), SPD matrix generator (>10x less memory), etc.
- Led the redesign the entire scikit-learn main website and coordinated efforts in documentation improvements and UI / UX enhancements.

DISC Lab, Fudan University | Lab Assistant | DASFAA'24 | GitHub 2023.05 – 2023.08

Skills: Python, PyTorch, HuggingFace, LLM, instruction tuning

- Led the construction of 403K legal knowledge instruction data, curated with legal syllogism prompting for higher expertise.
- Fine-tuned DISC-LawLLM, an LLM specialized for legal services based on Baichuan 13B Chat, outperforming GPT-3.5 Turbo.
- Participated in designing a verifiable knowledge retrieval module to inject external knowledge and enhance output actuality.
- Drove the implementation of a comprehensive benchmark for legal systems evaluation in both objective and subjective dimensions.

Research Experience

Privacy-Preserving Network Configuration Sharing via Anonymization | SIGCOMM'24 | GitHub 2022.10 – 2024.08

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

- Proposed the ConfMask framework to systematically anonymize topology and routing information in network configurations.
- Designed the anonymization algorithm for different protocols that mitigated deanonymization risks yet preserved important utilities.
- Managed to rigorously prove the route equivalence and routing utility preservation properties of the anonymization framework.
- Led the implementation of the end-to-end network configuration anonymization system and the artifact evaluation.

Software Projects

Deskulpt: A Cross-Platform Desktop Customization Tool | GitHub 2024.03 – present

Skills: Rust, TypeScript, Tauri, React, Vite, SWC | **Full-stack**

- Led the development of a cross-platform system for highly customizable desktop widgets that can be written in React / TypeScript.
- Integrated rich development tools in Deskulpt, enabling streamlined widget creation and debugging, editor and type hints, etc.
- Built Deskulpt using Tauri to ensure system security and compatibility across Windows, macOS, and Linux environments.
- Utilized Rust's async capabilities in the backend to ensure responsive interactions between the UI and system resources.
- Implemented security measures, e.g., CSP protection, constraints on file system access, limiting frontend capabilities, etc.

Column-Store Database Management System | Course Project | GitHub 2024.09 – 2024.12

Skills: C, SIMD/AVX, database optimizations, cache-conscious algorithms

- Streamlined CSV parsing and cache-aware chunked loading, achieving >4x speedup over naive row-wise loading on 400M data.
- Implemented shared scan for batchable queries with parallelization, achieving >20x speedup for 100M data and 100 queries.
- Supported B+ tree indexes, with <20ms bulk loading overhead and >25x select query speedup over 100M data with 5% selectivity.
- Optimized and parallelized radix hash join, outperforming naive hash join by >15x when joining 100M$\times$100M data.