Yao Xiao

(+86) 186-2182-3612 | 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, HPC, Data Systems, Distributed Systems (MIT), etc.

New York University | Bachelor of Science | Honors Mathematics | Computer Science Shanghai | New York | 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:** Python, Rust, JavaScript/TypeScript, C, C++, Go; SQL, Java, MATLAB, Julia
[2] **Frameworks and packages:** Tauri, React; Numpy, Pandas, Polars, Scikit-learn, PyTorch; CUDA; SIMD/AVX; OpenMP, MPI
[3] **DevOps:** Docker; Git; AWS, Google Cloud; Ansible; Kubernetes; GitHub Actions, CI/CD; Linux

Working Experience

Scikit-learn | Core Developer | **Open Source (GitHub 60K+ Star)** | 128+ Contributions 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 speedup, <3% memory usage), SPD matrix generator (<10% memory usage), 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, augmented retrieval
- 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.

Projects

Fault-Tolerant Key-Value Store Using Raft | Course Project 2025.02 – present
Skills: Go, RPC, distributed systems, consensus algorithms, fault tolerance
- Developed a distributed key-value store in Go, backed by the Raft consensus algorithm for strong consistency.
- Implemented leader election, log replication, and state machine updates to tolerate node failures and network partitions.
- Utilized goroutines and channels for concurrent and efficient I/O operations, RPC communication, and fault-tolerance mechanisms.
- Validated the design under MIT 6.5840 test suite, ensuring correctness, reliability, and high performance under various failure scenarios.

Distributed Column-Store Relational Database System | Course Project | GitHub 2024.09 – present
Skills: C/C++, SIMD/AVX, OpenMP, MPI, database sharding, cache-conscious algorithms
- Parallelized and vectorized complex select queries with OpenMP and SIMD, achieving >20x speedup on 100M data with 100 predicates.
- Supported B+ tree column index, with <20ms bulk loading overhead and >25x select query speedup over 100M data with 5% selectivity.
- Embarrassingly parallelized radix hash join, outperforming naive hash join by >15x when joining 100M×100M data.
- Implemented database sharding with MPI for distributed processing over multiple nodes, achieving near-linear speedup and scalability.

Deskulpt: A Cross-Platform Desktop Customization Tool | GitHub 2024.03 – present
Skills: Rust, TypeScript, Tauri, React, Vite, cross-platform desktop application, bundler, plugin system | **Full-stack**
- Led the development of Deskulpt, a cross-platform system built with Tauri that allows writing desktop widgets with any valid React code.
- Designed a plugin system with IPC and a custom communication protocol, keeping system backend lightweight yet highly extensible.
- Built a Rolldown-based widget bundler in Rust, supporting live reloading, external dependencies, shared React runtime, etc.
- Utilized async Rust to ensure UI responsiveness, concurrent widget bundling and rendering, and efficient execution of many other tasks.
- Integrated rich development tools in Deskulpt for widget and plugin creation or discovery, debugging, packaging, and distribution.