

# YAO XIAO

(+86) 186-2182-3612 | ✉ [yaoxiao@g.harvard.edu](mailto:yaoxiao@g.harvard.edu) | 🏠 [charlie-xiao.github.io](https://charlie-xiao.github.io) | 🗣️ Charlie-XIAO | 📄 [yao-xiao-200073244](https://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, Parallel Computing, Distributed Systems (MIT), 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, JavaScript/TypeScript, Go; Familiar with C, C++, SQL, MATLAB, Java, 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** | 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 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](mailto: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

**Deskulpt: A Cross-Platform Desktop Customization Tool** | [GitHub](#) 2024.03 – present

SKILLS: Rust, TypeScript, Tauri, React, Vite, 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.

**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.

**VeritasTrial: LLM-Driven Clinical Trial Search and Interpretation** | Course Project | [GitHub](#) 2024.09 – 2024.12

SKILLS: TypeScript, React, instruction tuning, augmented retrieval, RESTful API, Google Cloud, Kubernetes, Ansible

- Led the development of VeritasTrial, a LLM-driven application streamlining clinical trial searches and data interpretation.
- Enhanced searching and filtering with a database of vector embeddings for comprehensive semantic analysis and efficient matching.
- Designed and implemented an intuitive user interface for trial exploration and data interpretation.
- Deployed the application on Google Cloud with Kubernetes, Ansible, and GitHub Actions for automated deployment and scaling.