

# 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://arxiv.org/authorindex/Yao_Xiao)

## 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.

## PUBLICATIONS

AUTHORS WITH <sup>†</sup> ARE SORTED BY  $\alpha$ - $\beta$  ORDER, OTHERS ARE SORTED BY CONTRIBUTION

- [1] Yuejie Wang, Qitong Men, **Yao Xiao**, Yongting Chen, and Guyue Liu. 2024. ConfMask: Enabling Privacy-Preserving Configuration Sharing via Anonymization. In *Proceedings of the ACM SIGCOMM 2024 Conference (ACM SIGCOMM'24)*. Association for Computing Machinery, New York, NY, USA, 465–483. [doi:10.1145/3651890.3672217](https://doi.org/10.1145/3651890.3672217)
- [2] Shengbin Yue, Shujun Liu, Yuxuan Zhou, Chenchen Shen, Siyuan Wang, **Yao Xiao**, Bingxuan Li, Yun Song, Xiaoyu Shen, Wei Chen, Xuanjing Huang, and Zhongyu Wei. 2024. LawLLM: Intelligent Legal System with Legal Reasoning and Verifiable Retrieval. In *29th International Conference on Database Systems for Advanced Applications (DASFAA'24)*. [doi:10.1007/978-981-97-5569-1\\_19](https://doi.org/10.1007/978-981-97-5569-1_19)
- [3] Xinyu Li<sup>†</sup>, **Yao Xiao**<sup>†</sup>, and Yuchen Zhou<sup>†</sup>. 2023. Efficiently Visualizing Large Graphs. [doi:10.48550/arXiv.2310.11186](https://doi.org/10.48550/arXiv.2310.11186)

## WORKING EXPERIENCE

**Google** | Software Engineering Intern

2025.06 – 2025.08

SKILLS: Java, Dart, Flutter, Android/Pixel, Binder IPC, Protobuf, MCP, on-device AI agents

- Integrated Model Context Protocol (MCP) into Android, allowing apps to expose MCP services and participate in agentic interactions.
- Built an Android MCP proxy service that mediates communication of MCP clients and servers, enabling fine-grained security controls.
- Designed APIs for agentic AI apps to seamlessly discover and securely connect to on-device and remote MCP servers through the proxy.
- Showcased an MCP host utilizing multiple apps (map, calendar, email, etc.) for multi-step travel planning with minimal user intervention.

**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.

**Analyzing the Critical Behavior of Bernoulli Percolation in  $\mathbb{Z}^3$  via Invasion Percolation** | Capstone

2023.09 – 2024.01

ADVISOR: Professor Wei Wu, [ww44@nyu.edu](mailto:ww44@nyu.edu)

- Proved that in  $\mathbb{Z}^d$ , infinite cluster density at critical point  $\mathbf{P}_\infty(p_c) = 0$  if and only if invasion probability  $G_{\mathbb{Z}^d}(0, x) \rightarrow 0$  as  $|x| \rightarrow \infty$ .
- Verified via simulation that  $G_{\mathbb{Z}^3}(0, x) \leq G_{\mathbb{Z}^2}(0, x)$  for each  $|x|$ , thus proving that  $\mathbf{P}_\infty(p_c) = 0$  in  $\mathbb{Z}^3$  given the fact that it holds in  $\mathbb{Z}^2$ .
- Estimated the fractal dimension of the invasion percolation cluster in  $\mathbb{Z}^3$  numerically which was approximately 2.627.

**Efficient Distributed Serving System for Large Language Model Inference** | Capstone

2023.09 – 2024.01

ADVISOR: Professor Guyue Liu, [guyue.liu@gmail.com](mailto: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.

**Efficiently Visualizing Large Graphs** | Dean's Undergraduate Research Fund | [ArXiv](#)

2022.05 – 2022.08

ADVISOR: Professor Jie Xue, [jiexue@nyu.edu](mailto: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.

## TEACHING EXPERIENCE

---

- **Computer Networks**, COMPSCI.1450, Harvard School of Engineering and Applied Sciences, Teaching Fellow, Spring 2025
- **Linear Algebra**, MATH-SHU.0140, NYU Shanghai, Learning Assistant, Spring 2024
- **Calculus II**, MATH-SHU.0131, NYU Shanghai, Learning Assistant, Fall 2021, Fall 2023
- **Operating Systems**, CSCI-UA.0202, NYU Courant Institute, Teaching Assistant, Spring 2023

## PROJECTS

---

**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.

**Scikit-learn** | **Core Developer** | [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.

**Distributed Fault-Tolerant KV Store Using Raft** | Course Project | [GitHub](#) 2025.02 – 2025.05

SKILLS: Go, C++, Rust, RPC, distributed systems, consensus algorithms, fault tolerance, database sharding

- Built a sharded, fault-tolerant KV store that guarantees strong consistency and high availability using the Raft consensus algorithm.
- Implemented leader election, log replication, state machine updates, and snapshotting to tolerate node failures and network partitions.
- Delivered E2E implementations in Go, C++, and Rust, exposing identical APIs with consistent behavior, performance, and resilience.
- Achieved high throughput under strong consistency, sustaining >600/>100 ops/s under reliable/unreliable networks with 10 clients.

**Distributed Column-Store Relational Database System** | Course Project | [GitHub](#) 2024.09 – 2025.05

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.

**Comparitive Analysis of LAPACK Symmetric Tridiagonal Eigensolvers** | Course Project | [Paper](#) | [GitHub](#) 2024.12

SKILLS: Python, scipy, Fortran LAPACK

- Compared QR iteration, divide & conquer, bisection, and MRRR algorithms regarding performance, stability, and accuracy.
- Evaluated their LAPACK implementations on real-world and synthetic symmetric tridiagonal matrices with different characteristics.

**CampusHelper: WeChat / Alipay Miniprogram** 2023.12 – 2024.08

SKILLS: TypeScript, MongoDB, WeChat / Alipay cloud, miniprogram frameworks | **Full-stack**

- Utilized miniprogram cloud, including cloud functions, database, and storage to enhance data management and service reliability.
- Optimized miniprogram performance through cloud-based technologies, lazy loading, list virtualization, etc.
- Built a clean, consistent, accessible, and user-friendly interface, enhancing the overall user experience.
- Won the 2nd Prize (4th Place) in the [2023 Alipay Miniprogram Developers' Competition](#).

**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](#) | [GitHub](#) 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.

## HONORS AND AWARDS

---

- [1] **Magma cum laude**, NYU Shanghai, 2024
- [2] **Level I Certification**, [CRLA's International Tutor Training Program](#), 2024
- [3] **2nd Prize, 4th Place**, [Alipay Miniprogram Developers' Competition](#), 2023
- [4] **Meritorious Winner**, [Mathematical Contest in Modeling](#), 2023
- [5] **Dean's List of Academic Year**, NYU Shanghai, 2020 – 2021, 2021 – 2022, 2022 – 2023

## SKILLS

---

**Programming:** Python, Rust, C/C++, Go, JS/TS, Java, Dart, SQL, MATLAB, Julia; **Web & App Frameworks:** React, Tauri, Flutter, Android; **Database:** SQLite, MongoDB; **Cloud:** Google Cloud Platform, AWS; **HPC:** CUDA, SIMD/AVX, OpenMP/MPI; **DevOps:** Docker, Kubernetes, GitHub Actions, Ansible