萧尧

(+86) 186-2182-3612 | yaoxiao@g.harvard.edu | charlie-xiao.github.io | Charlie-XIAO | yao-xiao-200073244

教育背景

哈佛大学 | 理学硕士 | 计算科学与工程 2024.09 – 2026.05 (预期)

- GPA: 3.92/4.00, 相关课程包括：计算机网络、数据系统、高性能计算、分布式系统 (MIT) 等。

纽约大学 | 理学学士 | 荣誉数学、计算机科学 上海 | 纽约 | 2020.09 – 2024.05

- 荣誉数学 GPA: 4.00/4.00, 相关课程包括：线性代数、抽象代数、数学建模、概率论、数值分析、实/复分析、随机分析等。
- 计算机科学 GPA: 3.97/4.00, 相关课程包括：数据结构、算法、随机化算法、计算机架构、操作系统、机器学习、软件工程等。

出版物

标注为 † 的作者按字母顺序排列，其他作者按贡献排序

[1] Yuejie Wang, Qiutong Men, **Yao Xiao**, Yongting Chen, and Guyue Liu. 2024. ConfMask: Enabling Privacy-Preserving Configuration Sharing via Anonymization [ConfMask: 通过匿名化实现隐私保护的网络配置共享]. 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 **[CCF-A]**

[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 [LawLLM: 具备法律推理和可验证检索的智能法律系统]. In *29th International Conference on Database Systems for Advanced Applications (DASFAA'24)*. doi:10.1007/978-981-97-5569-1_19 **[CCF-B]**

[3] Xinyu Li†, **Yao Xiao**†, and Yuchen Zhou†. 2023. Efficiently Visualizing Large Graphs [大规模图的高效可视化]. doi:10.48550/arXiv.2310.11186

工作经历

谷歌 (Google) | 软件开发 (SWE) 2025.06 – 2025.08

技能: Java, Dart, Flutter, Android/Pixel, Binder IPC, Protobuf, MCP, 设备端 AI 代理

- 在安卓平台上集成模型上下文协议 (MCP), 使安卓应用应用能够声明并提供 MCP 服务, 以参与 AI 代理的交互。
- 开发了安卓 MCP proxy 服务, 用于调度 MCP 客户端与服务器之间的通信，实现高细粒度的安全管控。
- 设计了面向 AI 代理应用的 API, 使其能够轻松发现并通过安卓 MCP proxy 安全地连接到设备端和远程的 MCP 服务。
- 实现了一套 MCP 样例, 整合地图、日历、邮件等多款应用, 自动完成多步骤旅行规划，最大限度减少用户操作的需要。

复旦大学 DISC 实验室 | 实验室助理 | **发表于 DASFAA'24 (CCF-B)** | GitHub 2023.05 – 2023.08

技能: Python, PyTorch, HuggingFace, 大语言模型 (LLM), 指令微调, 检索增强

- 构建 40 万条法律知识指令数据，并采用法律三段论提示等技巧以提高模型回复的专业性。
- 微调 DISC-LawLLM (基于 Baichuan 13B Chat 的司法领域大模型)，性能超越 GPT-3.5 Turbo (当时最先进的通用模型)。
- 参与设计可验证知识检索模块，引入外部知识库进行检索增强，提高模型输出的真实性并减轻其幻觉。
- 推动了司法领域大模型系统评测基准的实现，全面覆盖了多个客观和主观的评测维度。

科研经历

通过匿名化实现隐私保护的网络配置共享 | **发表于 SIGCOMM'24 (CCF-A)** | GitHub 2022.10 – 2024.08

指导教授: 刘古月教授, guyue.liu@gmail.com

- 提出并实现了 ConfMask 框架，系统性地对网络配置中的拓扑与路由信息进行匿名化处理。
- 设计适用于不同网络路由协议 (OSPF, BGP, EIGRP 等) 的网络配置匿名化算法，在降低去匿名化风险的同时，保持关键网络功能。
- 严格证明 ConfMask 能够保证等效的网络路由，并维持可达性、多路径一致性、路径长度等性质以确保共享的网络配置的可用性。

通过入侵渗流分析 $\mathbb{Z}^3$ 中伯努利渗流的临界行为 | 毕业设计 2023.09 – 2024.01

指导教授: 吴炜教授, ww44@nyu.edu

- 证明了在 $\mathbb{Z}^d$ 中，临界点处无限簇密度 $\mathbf{P}_\infty(p_c) = 0$ 当且仅当入侵概率 $G_{\mathbb{Z}^d}(0, x)$ 随着 $|x| \to \infty$ 时趋于 0。
- 通过模拟验证，对于任意 $|x|$，均有 $G_{\mathbb{Z}^3}(0, x) \le G_{\mathbb{Z}^2}(0, x)$，从而在已知 $\mathbb{Z}^2$ 中 $\mathbf{P}_\infty(p_c) = 0$ 的前提下证明了 $\mathbb{Z}^3$ 中也满足该性质。
- 用数值方法估算了 $\mathbb{Z}^3$ 中入侵渗流簇的分形维数，结果约为 2.627。

高效的分布式大语言模型推理服务系统 | 毕业设计 2023.09 – 2024.01

指导教授: 刘古月教授, guyue.liu@gmail.com

- 观察到只有自注意力层依赖 KV 缓存，从而针对除自注意力层之外的层，实现了突破 KV 缓存限制的更大规模的批处理。
- 在自注意力层中对预填充和解码进行动态批处理，以缓解由于 Transformer 输入长度变化而产生的流水线空泡问题。
- 将多个短注意力计算与最长计算进行打包，同时并行交换 KV 缓存以最小化开销。

大规模图的高效可视化 | 院长本科研究基金 (DURF) | ArXiv 2022.05 – 2022.08
指导教授: 薛杰教授, jiexue@nyu.edu
- 设计了专门针对图的 t-SGNE 方法，利用节点间的邻接关系，实现了 6.7 倍的计算效率提升。
- 提出了 SPLEE, 一种基于拉普拉斯特征映射与最短路径的图嵌入方法，旨在适应 t-SGNE。
- 将 SPLEE 与 t-SGNE 相结合，对拥有 30 万节点和 100 万边的图进行可视化，实现了 10% 的视觉效果提升。

教学经历

- **计算机网络**, COMPSCI.1450, 哈佛大学工程与应用科学学院, 高级助教 (Teaching Fellow), 2025 春季
- **线性代数**, MATH-SHU.0140, 上海纽约大学, 助教 (Learning Assistant), 2024 春季
- **微积分 II**, MATH-SHU.0131, 上海纽约大学, 助教 (Learning Assistant), 2021 秋季, 2023 秋季
- **操作系统**, CSCI-UA.0202, 纽约大学柯朗数学科学研究所, 助教 (Tutor), 2023 春季

项目经历

Deskulpt: 跨平台的桌面定制工具 | GitHub 2024.03 – 至今
技能: Rust, TypeScript, Tauri, React, Vite, 跨平台桌面应用, 组件打包工具, 插件系统 | **全栈开发**
- 主导开发 Deskulpt, 一款基于 Tauri 的跨平台桌面定制工具，支持用户使用 React 编写桌面小组件。
- 利用 IPC 和自定义的通信协议，设计了一套插件系统，使后端同时保持轻量化和高度可扩展性。
- 使用 Rust 构建基于 Rolldown 的桌面组件打包工具，支持组件热重载、外部依赖管理、组件间共享 React 运行时等。
- 利用 Rust 异步确保 UI 响应流畅，实现桌面组件打包、渲染等多任务的高效并发执行。
- 在 Deskulpt 内集成丰富的开发工具，支持桌面组件和插件的创建、检索、调试、打包和分发。

Scikit-learn | **核心开发者** | GitHub (60K+ Star) | 128+ 贡献 2023.04 – 至今
技能: Python, Cython, JavaScript, Sphinx, scikit-learn, numpy, scipy, pandas, polars, CI/CD
- 负责维护任务，如测试套件覆盖率、代码重构、开发者 API 改进、自动化 GitHub 工作流等。
- 增强对稀疏数组和 polars 的支持、改进估计器的表达方式、优化指标函数的可视化、支持多标签数据交叉验证等。
- 优化增量主成分分析在稀疏数据上的性能 (速度 10x, 内存占用 3%); 改进半正定矩阵生成器 (内存占用 10%) 等。
- 对 scikit-learn 官方文档和网站的进行了整体改版，并负责协调后续文档内容和 UI/UX 的改进工作。

基于 Raft 共识算法的高容错键值存储系统 | 课程项目 2025.02 – 至今
技能: Go, RPC, 分布式系统, 共识算法, 系统容错
- 使用 Go 语言开发了一个分布式键值存储系统，并通过 Raft 共识算法确保系统的强一致性。
- 实现了领导节点选举、日志复制和状态机更新等机制，可在节点故障和网络分区等情况下继续正常运行。
- 利用 goroutines 和 channels 实现并发高效的 I/O 处理、RPC 通信及容错机制。
- 在 MIT 6.5840 提供的测试框架下验证该系统设计，确保其在多种故障场景下具备正确性、可靠性与高性能。

VeritasTrial: LLM 驱动的临床试验搜索与数据解读 | 课程项目 | GitHub 2024.09 – 2024.12
技能: TypeScript, React, 指令微调, 检索增强, RESTful API, Google Cloud, Kubernetes, Ansible
- 开发了 VeritasTrial, 一款 LLM 驱动用于简化临床试验搜索和数据解读的应用。
- 采用向量嵌入数据库提升搜索和筛选能力，实现高效的语义分析与精准的临床案例匹配。
- 设计并实现了直观的用户交互界面，优化试验信息的探索与解读体验。
- 通过 Google Cloud, Kubernetes, Ansible, GitHub Actions 进行部署，实现自动化部署与弹性扩展。

对不同 LAPACK 对称三对角特征值求解器的比较分析 | 课程项目 | 论文 | GitHub 2024.12
技能: Python, scipy, Fortran LAPACK
- 对比了 QR 迭代、分治法、二分法和 MRRR 算法在性能、稳定性和精度方面的表现。
- 在具有不同特征的实际及合成对称三对角矩阵上评估了它们的 LAPACK 实现。

CampusHelper: 微信/支付宝小程序 2023.12 – 2024.08
技能: TypeScript, MongoDB, 微信/支付宝云, 小程序框架 | **全栈开发**
- 利用小程序云，包括云函数、数据库和存储，提升了数据管理和服务可靠性。
- 通过云技术、懒加载、列表虚拟化等手段优化了小程序性能。
- 构建了简洁、一致、无障碍且用户友好的界面，提升了整体用户体验。
- 在 2023 年支付宝小程序开发者大赛中荣获二等奖（第四名）。

YouTube 界面定制浏览器插件 | 课程项目 | GitHub 2023.02
- 开发了一个火狐浏览器插件，支持更换颜色主题、重新排列和定制 YouTube 界面的各个元素。
- 编写了功能文档及贡献指南，并在 Mozilla 插件商城中发布了 v1.0 自分发版本。

模拟经济体中的不平等过程 | 课程项目 | 论文 | GitHub 2022.12
- 通过细致设计的随机交易函数模拟了经济体中的不平等过程，反映了现实中的不平等经济现象。
- 发现现实经济体中的财富最终分布呈现出伽马分布或贝塔素分布的形态。

模拟垂直堆叠的陀螺仪塔 | 课程项目 | 论文 | GitHub 2022.10

- 将陀螺仪建模为由弹性元件构成的网络，通过微分方程构建陀螺仪系统模型，并采用欧拉法进行求解。
- 模拟了垂直堆叠的陀螺仪，并发现当中间轴具有一定柔性时，陀螺仪塔遵循与单个陀螺仪类似的陀螺进动规律。

荣誉奖项

[1] **极优等毕业 (Magma cum laude)**, 上海纽约大学, 2024
[2] **一级认证 (Level I Certificate)**, CRLA 国际助教培训项目 (CRLA's International Tutor Training Program), 2024
[3] **二等奖 (第四名)**, 支付宝小程序开发者大赛, 2023
[4] **M 奖 (Meritorious Winner)**, 美国大学生数学建模竞赛 (美赛, Mathematical Contest in Modeling), 2023
[5] **院长学年度优秀学生名单**, 上海纽约大学, 2020 – 2021, 2021 – 2022, 2022 – 2023

技术能力

编程语言: Python, Rust, C/C++, Go, JS/TS, Java, Dart, SQL, MATLAB, Julia; **Web & 应用框架:** React, Tauri, Flutter, Android; **数据库:** SQLite, MongoDB; **云平台:** Google Cloud Platform, AWS; **高性能计算:** CUDA, SIMD/AVX, OpenMP/MPI; **DevOps:** Docker, Kubernetes, GitHub Actions, Ansible