# Yifei Xu

+1-310-728-5031 | yifeix00@gmail.com | www.yifeix.com | Google Scholar

Redmond, WA 98052, United States

## Research Interests

My research spans machine learning and systems, with a current focus on LLM post-training, including agent memory, reasoning, and alignment. I have a background in systems and algorithms, including ML for systems, data mining algorithms, VR streaming, and 5G systems.

## Education

- **University of California, Los Angeles** *Sep. 2021 - Mar. 2026*
  *Ph.D. in Computer Science* Los Angeles, CA
  - Advisor: Prof. Songwu Lu
- **University of California, Los Angeles** *Sep. 2021 - Mar. 2024*
  *M.S. in Computer Science* Los Angeles, CA
- **Peking University** *Sep. 2017 - Jul. 2021*
  *B.S. in Computer Science* Beijing, China
  - Undergrad Research Advisor: Prof. Tong Yang

## Experience

- **Microsoft** *Jun. 2024 - Present*
  *Research Intern, Researcher* Redmond, WA
  - Led research on RL from targeted human feedback **(ICML 2025)**
  - Led research on RL for reasoning on open-ended tasks
  - Leading research on optimizing rubric with agent memory for long-horizon adaptation
  - Founding member in the 0-to-1 launch of a new agent post-training org, leading efforts to operationalize research into products
- **Alibaba Cloud** *Jun. - Sep. 2022, 2023*
  *Research Intern* Sunnyvale, CA
  - Led research on benchmarking LLMs for cloud configuration **(MLSys 2024)**
  - Led research on large-scale VR streaming systems **(USENIX ATC '25)**
  - Contributed to research on anycast optimization **(USENIX NSDI '26)**
- **ByteDance** *Apr. - Jul. 2021*
  *Networking Research & Development Intern* Beijing, China

## Publications

*Equal Contribution

### Conference Papers

**[C.9]** Minyuan Zhou*, Yuning Chen*, **Yifei Xu**, Jiaqi Zheng, Guihai Chen, Wanchun Dou, Pan Hu, Yongping Tang, Wendong Yin, Jie Lin, Qingyan Yu, Yuanchao Su, Songwu Lu, Wan Du. (2026). AnyPro: Preference-Preserving Anycast Optimization based on Strategic AS-Path Prepending. To appear in *23rd USENIX Symposium on Networked Systems Design and Implementation* **(USENIX NSDI '26)**

**[C.8]** Zihan Gao, **Yifei Xu**, Jacob Thebault-Spieker. (2026). LocalBench: Benchmarking LLMs on County-Level Local Knowledge and Reasoning. To appear in *Proceedings of the 40th AAAI Conference on Artificial Intelligence* **(AAAI 2026)**

**[C.7]** **Yifei Xu**, Tusher Chakraborty, Emre Kıcıman, Bibek Aryal, Eduardo Rodrigues, Srinagesh Sharma, Roberto Estevao, Maria Angels de Luis Balaguer, Jessica Wolk, Rafael Padilha, Leonardo Nunes, Shobana Balakrishnan, Songwu Lu, Ranveer Chandra. (2025). RLTHF: Targeted Human Feedback for LLM Alignment. In *Proceedings of the 42nd International Conference on Machine Learning* **(ICML 2025)**

**[C.6]** **Yifei Xu**, Xumiao Zhang, Yuning Chen, Pan Hu, Xuan Zeng, Zhilong Zheng, Xianshang Lin, Yanmei Liu, Songwu Lu, Z. Morley Mao, Wan Du, Dennis Cai, Ennan Zhai, Yunfei Ma. (2025). Roaming Free in the VR World with MP2. In *2025 USENIX Annual Technical Conference* **(USENIX ATC '25)**

**[C.5]** **Yifei Xu***, Yuning Chen*, Xumiao Zhang*, Xianshang Lin, Pan Hu, Yunfei Ma, Songwu Lu, Wan Du, Z. Morley Mao, Ennan Zhai, Dennis Cai. (2024). CloudEval-YAML: A Practical Benchmark for Cloud Configuration Generation. In *Proceedings of Machine Learning and Systems 6* **(MLSys 2024)**

**[C.4]** Jinghao Zhao, Zhaowei Tan, **Yifei Xu**, Zhehui Zhang, Songwu Lu. (2022). SEED: A SIM-Based Solution to 5G Failures. In *Proceedings of the ACM SIGCOMM 2022 Conference* **(ACM SIGCOMM 2022)**

**[C.3]** Zhaowei Tan, Jinghao Zhao, Yuanjie Li, **Yifei Xu**, Songwu Lu. (2021). Device-Based LTE Latency Reduction at the Application Layer. In *18th USENIX Symposium on Networked Systems Design and Implementation* **(USENIX NSDI '21)**

**[C.2]** Hao He, Yulin Xu, Yixiao Ma, **Yifei Xu**, Guangtai Liang, Minghui Zhou. (2021). A Multi-Metric Ranking Approach for Library Migration Recommendation. In *2021 IEEE International Conference on Software Analysis, Evolution and Reengineering* **(IEEE SANER 2021)**

**[C.1]** Jizhou Li*, Zikun Li*, **Yifei Xu***, Shiqi Jiang, Tong Yang, Bin Cui, Yafei Dai, Gong Zhang. (2020). WavingSketch: An Unbiased and Generic Sketch for Finding Top-k Items in Data Streams. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* **(KDD '20)**

**Journal Papers**

**[J.2]** Yuhan Wu*, Shiqi Jiang*, **Yifei Xu***, Siyuan Dong, Kaicheng Yang, Peiqing Chen, Tong Yang. (2023). Unbiased Real-Time Traffic Sketching. *IEEE Transactions on Network Science and Engineering* **(IEEE TNSE)**

**[J.1]** Zhaowei Tan, Jinghao Zhao, Yuanjie Li, **Yifei Xu**, Yunqi Guo, Songwu Lu. (2023). LDRP: Device-Centric Latency Diagnostic and Reduction for Cellular Networks Without Root. *IEEE Transactions on Mobile Computing* **(IEEE TMC)**

**Preprints**

**[P.5]** Saeid Asgari Taghanaki, Rakshanda Agarwal, Bruce Sun, Rohan Jha, Elias Stengel-Eskin, Sara Malvar, Rui Ying, **Yifei Xu**, Guilherme Potje, Tusher Chakraborty, Leonardo de Oliveira Nunes, Ranveer Chandra, Emre Kıcıman (2026). Diagnosing Capability Gaps in Fine-Tuning Data. *arXiv preprint arXiv:2604.27547*

**[P.4]** **Yifei Xu**, Guilherme Potje, Shivam Shandilya, Tiancheng Yuan, Leonardo de Oliveira Nunes, Rakshanda Agarwal, Saeid Asgari, Adam Atkinson, Emre Kıcıman, Songwu Lu, Ranveer Chandra, Tusher Chakraborty. (2026). SibylSense: Adaptive Rubric Learning via Memory Tuning and Adversarial Probing. *arXiv preprint rXiv:2602.20751*

**[P.3]** **Yifei Xu***, Tusher Chakraborty*, Srinagesh Sharma, Leonardo Nunes, Emre Kıcıman, Songwu Lu, Ranveer Chandra. (2026). Direct Reasoning Optimization: Token-Level Reasoning Reflectivity Meets Rubric Gates for Unverifiable Tasks. *arXiv preprint arXiv:2506.13351*

**[P.2]** Aman Ganapathy Manvattira*, **Yifei Xu***, Ziyue Dang, Songwu Lu. (2025). DeepSpecs: Expert-Level Question Answering in 5G. *arXiv preprint arXiv:2511.01305*

**[P.1]** Prasoon Patidar, Alex Crown, Kevin Hsieh, **Yifei Xu**, Tusher Chakraborty, Ranveer Chandra, Yuvraj Agarwal. (2025). Orchestration for Domain-specific Edge-Cloud Language Models. *arXiv preprint arXiv:2507.09003*

**Workshop Papers & Posters**

**[W.3]** Zihan Gao, Jiaying Liu, **Yifei Xu**, Jacob Thebault-Spieker. (2025). From Clips to Communities: Fusing Social Video into Knowledge Graphs for Localness-Aware LLMs. In *Companion Publication of the 2025 Conference on Computer-Supported Cooperative Work and Social Computing* **(ACM CSCW 2025 Companion)**

**[W.2]** **Yifei Xu***, Yuning Chen*, Xumiao Zhang*, Xianshang Lin, Pan Hu, Yunfei Ma, Songwu Lu, Wan Du, Z. Morley Mao, Ennan Zhai, Dennis Cai. (2023). CloudEval-YAML: A Realistic and Scalable Benchmark for Cloud Configuration Generation. In *37th Conference on Neural Information Processing Systems (NeurIPS 2023) Workshop on ML for Systems* **(NeurIPS 2023 Workshop)**

**[W.1]** Zhuochen Fan, Zhoujing Hu, Yuhan Wu, Jiarui Guo, Wenrui Liu, Tong Yang, Hengrui Wang, **Yifei Xu**, Steve Uhlig, Yaofeng Tu. (2022). PISketch: Finding Persistent and Infrequent Flows. In *Proceedings of the ACM SIGCOMM 2022 Workshop on Formal Foundations and Security of Programmable Network Infrastructures* **(ACM SIGCOMM 2022 Workshop)**

## HONORS AND AWARDS

- **USENIX ATC '25 Student Travel Grant** *2025*
  *National Science Foundation*
- **MLSys '24 Student Travel Grant** *2024*
  *National Science Foundation*
- **Excellent Research Award** *2020*
  *Peking University*

## PROFESSIONAL SERVICES

- **Reviewer**
  *NeurIPS, ICML, ICLR, MLSys, IEEE International Conference on Parallel and Distributed Systems, IEEE Network*

## TEACHING

- **UCLA CS 35L - Software Construction** *Spring 2024*
  *Associate Instructor*
- **UCLA CS 31 - Introduction to Computer Science I** *Winter 2024*
  *Associate Instructor*
- **UCLA CS 118 - Computer Network Fundamentals** *Winter/Spring/Fall 2023, Winter/Spring 2025*
  *Teaching Assistant, Associate Instructor, Teaching Fellow*
- **UCLA CS 180 - Introduction to Algorithms and Complexity** *Fall 2022*
  *Teaching Assistant*