Yuhan Liu

https://yuhanliu11.github.io/ yuhanl@uchicago.edu Google Scholar

Research Interests

Systems and Networking for Machine Learning

Education

University of Chicago — Chicago, IL
Ph.D. in Computer Science — 2021–2026
GPA: 3.95/4.0
Advisors: Prof. Shan Lu and Prof. Junchen Jiang

University of Wisconsin – Madison — Madison, WI
Bachelor of Science in Computer Science (Honors) & Minor in Math — 2017–2021
GPA: 4.0/4.0
Advisor: Prof. Shivaram Venkataraman

Employment

The University of Texas at Austin — Austin, TX
Incoming Tenure-track Assistant Professor, Department of Computer Science — Fall 2027

Publications

- **Yuhan Liu**, Yuyang Huang, Jiayi Yao, Shaoting Feng, Zhuohan Gu, Kuntai Du, Hanchen Li, Yihua Cheng, Junchen Jiang, Shan Lu, Madan Musuvathi, Esha Choukse
  DroidSpeak: KV Cache Sharing Across Fine-tuned Model Variants.
  NSDI '26
- **Yuhan Liu**, Hanchen Li, Yihua Cheng, Siddhant Ray, Yuyang Huang, Qizheng Zhang, Kuntai Du, Jiayi Yao, Shan Lu, Michael Maire, Henry Hoffmann, Ari Holtzman, Ganesh Ananthanarayanan, Junchen Jiang
  CacheGen: KV Cache Compression and Streaming for Fast Large Language Model Serving.
  SIGCOMM '24
- **Yuhan Liu**, Chengcheng Wan, Kuntai Du, Henry Hoffmann, Junchen Jiang, Shan Lu, Michael Maire
  Automatic and Efficient Customization of Neural Networks for ML Applications.
  OSDI '24
- Jiayi Yao, Hanchen Li, **Yuhan Liu**, Siddhant Ray, Yihua Cheng, Qizheng Zhang, Kuntai Du, Shan Lu, Junchen Jiang
  CacheBlend: Fast Large Language Model Serving with Cached Knowledge Fusion.
  EuroSys'25 Best Paper Award
- Yihua Cheng, Ziyi Zhang, Hanchen Li, Anton Arapin, Yue Zhang, Qizheng Zhang, **Yuhan Liu**, Kuntai Du, Xu Zhang, Francis Y. Yan, Amrita Mazumdar, Nick Feamster, Junchen Jiang

GRACE: Loss-Resilient Real-Time Video through Neural Codecs.
NSDI '24

- Kuntai Du, **Yuhan Liu**, Yitian Hao, Qizheng Zhang, Haodong Wang, Yuyang Huang, Ganesh Ananthanarayanan, Junchen Jiang
  OneAdapt: Fast Adaptation for Deep Learning Applications via Backpropagation
  SoCC '23
- Chengcheng Wan, **Yuhan Liu**, Kuntai Du, Henry Hoffmann, Junchen Jiang, Michael Maire, Shan Lu.
  Run-Time Prevention of Software Integration Failures of Machine Learning APIs
  OOPSLA '23
- Chengcheng Wan, Shicheng Liu, Sophie Xie, **Yuhan Liu**, Henry Hoffmann, Michael Maire, Shan Lu
  Keeper: Automated Testing and Fixing of Machine Learning Software
  TOSEM '24
- Ryan Jacobs*, Mingren Shen*, **Yuhan Liu***, Wei Hao, Xiaoshan Li, Ruoyu He, Jacob RC Greaves, Donglin Wang, Zeming Xie, Zitong Huang, Chao Wang, Kevin G Field, Dane Morgan
  Performance and limitations of deep learning semantic segmentation of multiple defects in transmission electron micrographs
  Cell Reports Physical Science '22

Preprints

- **Yuhan Liu***, Yihua Cheng*, Jiayi Yao*, Yuwei An, Xiaokun Chen, Shaoting Feng, Yuyang Huang, Samuel Shen, Kuntai Du, Junchen Jiang
  LMCache: An Efficient KV Cache Layer for Enterprise-Scale LLM Inference
- Yuyang Huang, **Yuhan Liu**, Haryadi S. Gunawi, Beibin Li, Changho Hwang
  Alchemist: Towards the Design of Efficient Online Continual Learning System
- Xingyu Xiang, Raj Joshi, **Yuhan Liu**, Jiayi Yao, Chenxingyu Zhao, Junchen Jiang, Yang Zhou, Eddie Kohler, Minlan Yu
  ShadowServe: Interference-Free KV Cache Fetching for Distributed Prefix Caching
- **Yuhan Liu**, Saurabh Agarwal and Shivaram Venkataraman.
  AutoFreeze: Automatically Freezing Model Blocks to Accelerate Fine-tuning
- Arjun Balasubramanian, Adarsh Kumar, **Yuhan Liu**, Han Cao, Shivaram Venkataraman, Aditya Akella
  Accelerating deep learning inference via learned caches

Honors & Awards

- EECS Rising Star, *MIT* 2025
- Best Paper Award, *Eurosys* 2025
- UU Fellowship, *University of Chicago* 2023
- Neubauer Graduate Scholarship, *University of Chicago* 2021
- Outstanding Undergraduate Researcher Awards, Honorable Mention, *Computing Research Association* 2021
- Trewartha Honors Senior Thesis Award, *University of Wisconsin – Madison* 2020

Open-Source Projects

- **LMCache** (GitHub, 12k stars)
  State-of-the-art open-source KV Cache Layer for efficient, cross-query reuse and offloading in large language model inference.
- **vLLM Production Stack** (GitHub, 2.5k stars)
  vLLM's reference system for Kubernetes-native, cluster-wide model deployment with community-driven performance optimization.

Patents

- Efficient LLM Agent communication with KV cache sharing, *Microsoft* ..... March, 2025
  Filed with Microsoft. Based on my internship at Microsoft Research and the paper DroidSpeak.

Workshops

- Chen Wang, Xunzhuo Liu, **Yuhan Liu**, Yue Zhu, Xiangxi Mo, Junchen Jiang, Huamin Chen
  When to Reason: Semantic Router for vLLM
  ML for Systems at NeurIPS'25
- Shaoting Feng, Hanchen Li, Kuntai Du, Zhuohan Gu, **Yuhan Liu**, Jiayi Yao, Siddhant Ray, Samuel Shen, Yihua Cheng, Ganesh Ananthanarayanan, Junchen Jiang
  AdaptCache: KV Cache Native Storage Hierarchy for Low-Delay and High-Quality Language Model Serving
  BigMem at SOSP'25
- Hanchen Li, **Yuhan Liu**, Yihua Cheng, Siddhant Ray, Kuntai Du, Junchen Jiang.
  Chatterbox: Robust Transport for LLM Token Streaming under Unstable Network
  NAIC at SIGCOMM'24

Posters

- Hanchen Li*, **Yuhan Liu***, Yihua Cheng, Kuntai Du, Junchen Jiang.
  Towards More Economical Context-Augmented LLM Generation by Reusing Stored KV Cache
  NSDI '24
- **Yuhan Liu**, Saurabh Agarwal and Shivaram Venkataraman.
  AutoFreeze: Automatically Freezing Model Blocks to Accelerate Fine-tuning
  Women in Machine Learning Workshop 2020 co-located with the NeurIPS conference.

Talks

LMCache: An Efficient KV Cache Layer for Enterprise-Scale LLM Inference

- University of Pennsylvania ..... November 2024
- Duke University ..... November 2024
- University of Göttingen ..... January 2025
- University of California San Diego ..... March 2025
- University of Maryland ..... March 2025
- Carnegie Mellon University ..... April 2025
- Rutgers University ..... May 2025

- Amazon ..... August 2025
- Northeastern University ..... October 2025
- ByteDance ..... November 2025
- KubeCon ..... November 2025
- MLSys Young Professionals Symposium ..... May 2026
- Google System Research Group ..... June 2026

Rethinking AI Systems through Efficient Model Communication

- Virginia Tech ..... February 2026
- National University of Singapore ..... February 2026
- Rutgers University ..... March 2026
- Rice University ..... March 2026
- University of California, Irvine ..... March 2026
- University of Michigan ..... March 2026
- ETH Zurich ..... March 2026
- Columbia University ..... March 2026
- Chinese University of Hong Kong ..... March 2026
- University of Maryland, College Park ..... April 2026

Work Experience

- **Microsoft Research, Redmond**
  Research Internship June 2024 – Sept. 2024
  Mentors: Esha Choukse, Madan Musuvathi, Shan Lu

- **Google, Shanghai**
  Software Engineering Internship Sep. 2019 – Dec. 2019
  Mentors: Lu Sun, Fan Lin

Teaching

- Graduate Networking, *University of Chicago* ..... Fall, 2024
- Introduction to Computer Systems, *University of Chicago* ..... Winter, 2022
- Introduction to Database systems, *University of Wisconsin – Madison* ..... Fall, 2020

Mentoring

- Hanchen Li, *BS at University of Chicago to PhD at UC Berkeley* ..... June, 2023 – August, 2025
- Zhuohan Gu, *BS at University of Chicago to PhD at MIT* ..... January, 2024 – August, 2025
- Shaoting Feng, *MS at University of Chicago* ..... September, 2024 – Present
- Ashton Tang, *BS at University of Chicago* ..... June, 2024 – September, 2024
- Jiayi Yao, *BS at CUHK (Shenzhen) to PhD at University of Chicago* ..... June, 2023 – June, 2024
- Yitian Hao, *MS at University of Chicago to Amazon AWS* ..... September, 2022 – August, 2024

Service

- Organizer of SIGCOMM '25 tutorial: Networking for Stateful LLM Inference

- Co-Chair of Graduate Women in Computer Science (GWiCS) at University of Chicago
- Reviewer for NeurIPS 2022, ICML 2022
- Silver reviewer for ICML 2026

References

- **Junchen Jiang**
  Associate Professor, Department of Computer Science, University of Chicago
  junchenj@uchicago.edu

- **Shan Lu**
  Professor, Department of Computer Science, University of Chicago
  shanlu@uchicago.edu

- **Esha Choukse**
  Principal Researcher, Microsoft
  esha.choukse@microsoft.com

- **Ganesh Ananthanarayanan**
  Senior Principal Researcher, Microsoft Research
  ga@microsoft.com