

# Junhong Shen

junhongs@andrew.cmu.edu | Website | GitHub | Google Scholar

## Research Interests

---

Multi-Modal Reasoning, Generation, and Agent Applications

## Education

---

**Carnegie Mellon University**, *Ph.D. in Machine Learning* Sep 2021 – Present

- **Advisor:** Ameet Talwalkar, **GPA:** 4.0/4.0
- **Thesis:** Navigating Through Heterogeneous Data: Building AI Systems for Diverse Data Types, Domains, and Complexities
- **Thesis Committee:** Ameet Talwalkar (CMU/Datadog), Ruslan Salakhutdinov (CMU/Meta), Aviral Kumar (CMU/DeepMind), Ludwig Schmidt (Stanford/Anthropic), Alexander Toshev (Apple)
- **Available for Full-Time:** Oct 2025

**University of California, Los Angeles**, *B.S. in Mathematics of Computation* Sep 2017 – June 2021

- **Daus Prize:** top 5 undergraduate students in the mathematics department, **GPA:** 4.0/4.0

## Publications

---

### Preprint

**RECODE: Reasoning Through Code Generation for Visual Question Answering** [paper]

*Junhong Shen*, Mu Cai, Bo Hu, Ameet Talwalkar, David A. Ross, Cordelia Schmid, Alireza Fathi

**CodePDE: An Inference Framework for LLM-Driven PDE Solver Generation** [paper] [code]

Shanda Li, Tanya Marwah, *Junhong Shen*, Weiwei Sun, Andrej Risteski, Yiming Yang, Ameet Talwalkar

**Terminal-Bench: A Benchmark for AI Agents in Terminal Environments** [website] [code]

Mike Merrill, Alexander Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Yeon Shin, *Junhong Shen*, ...73 authors..., Andy Konwinski, Ludwig Schmidt

### Peer-Reviewed Articles

**Thinking vs. Doing: Agents that Reason by Scaling Test-Time Interaction** [paper] [code] [website]

*NeurIPS 2025*

*Junhong Shen\**, Hao Bai\*, Lunjun Zhang, Yifei Zhou, Amrith Setlur, Shengbang Tong, Diego Caples, Nan Jiang, Tong Zhang, Ameet Talwalkar, Aviral Kumar

**CAT: Content-Adaptive Image Tokenization** [paper]

*NeurIPS 2025*

*Junhong Shen*, Kushal Tirumala, Michihiro Yasunaga, Ishan Misra, Luke Zettlemoyer, Lili Yu\*, Chunting Zhou\*

**Mixture-of-Mamba: Enhancing Multi-Modal State-Space Models with Modality-Aware Sparsity** [paper] [code]

*ICLR 2025 Scalable Optimization for Efficient and Adaptive Foundation Models Workshop (Oral, top 8/96)*

Weixin Liang\*, *Junhong Shen\**, Genghan Zhang, Ning Dong, Luke Zettlemoyer, Lili Yu

**ScribeAgent: Towards Specialized Web Agents Using Production-Scale Workflow Data** [paper] [code] [blog]

*ICLR 2025 Foundation Models in the Wild Workshop*

*Junhong Shen*, Atishay Jain, Zedian Xiao, Ishan Amlekar, Mouad Hadji, Aaron Podolny, Ameet Talwalkar

**Specialized Foundation Models Struggle to Beat Supervised Baselines** [paper] [code]

*ICLR 2025*

Zongzhe Xu, Ritvik Gupta, Wenduo Cheng, Alexander Shen, **Junhong Shen**, Ameet Talwalkar, Mikhail Khodak

**Tag-LLM: Repurposing General-Purpose LLMs for Specialized Domains** [paper] [code]

*ICML 2024*

**Junhong Shen**, Neil Tenenholtz, James Brian Hall, David Alvarez-Melis, Nicolò Fusi

**UPS: Towards Foundation Models for PDE Solving via Cross-Modal Adaptation** [paper] [code]

*TMLR 2024 & ICML 2024 AI4Science Workshop (Spotlight)*

**Junhong Shen**, Tanya Marwah, Ameet Talwalkar

**Cross-Modal Fine-Tuning: Align then Refine** [paper] [code] [talk] [website]

*ICML 2023 (Oral, top 158/6538)*

**Junhong Shen**, Liam Li, Lucio M. Dery, Corey Staten, Mikhail Khodak, Graham Neubig, Ameet Talwalkar

**Efficient Architecture Search for Diverse Tasks** [paper] [code] [blog]

*NeurIPS 2022*

**Junhong Shen\***, Mikhail Khodak\*, Ameet Talwalkar

**NAS-Bench-360: Benchmarking Neural Architecture Search on Diverse Tasks** [paper] [website] [blog]

*NeurIPS 2022 Datasets and Benchmarks Track*

Renbo Tu\*, Nicholas Roberts\*, Mikhail Khodak, **Junhong Shen**, Frederic Sala, Ameet Talwalkar

**AutoML Decathlon: Diverse Tasks, Modern Methods, and Efficiency at Scale** [paper] [website]

*NeurIPS 2022 Competitions Track*

Nicholas Roberts, ... 24 authors ..., **Junhong Shen**, Evan Sparks

**Iterative Teacher-Aware Learning** [paper] [code]

*NeurIPS 2021*

Luyao Yuan, Dongruo Zhou, **Junhong Shen**, Jingdong Gao, Jeffrey Chen, Quanquan Gu, Ying Nian Wu, Song-Chun Zhu

**Theoretically Principled Deep RL Acceleration via Nearest Neighbor Function Approximation** [paper] [code]

*AAAI 2021*

**Junhong Shen**, Lin F. Yang

**Mathematical Reconstruction of Patient-Specific Vascular Networks Based on Clinical Images and Global Optimization** [paper] [code] [talk]

*IEEE Access*

**Junhong Shen**, Abdul Hannan Faruqi, Yifan Jiang, Nima Maftoon

**Emergence of Pragmatics from Referential Game between Theory of Mind Agents** [paper] [code]

*NeurIPS 2019 Emergent Communication Workshop*

Luyao Yuan, Zipeng Fu, Jingyue Shen, Lu Xu, **Junhong Shen**, Song-Chun Zhu

\* Equal Contribution

## Experience

**Student Researcher, Google DeepMind**, Mountain View, CA

May 2025 – Sep 2025

- Mentors: Alireza Fathi, Cordelia Schmid, David A. Ross
- Improving visual reasoning agents via code generation and image derendering.

**Research Intern, FAIR at Meta**, Seattle, WA

May 2024 – Dec 2024

- Mentors: Chunting Zhou, Lili Yu, Luke Zettlemoyer
- Worked on caption-based adaptive image tokenization. Paper accepted by NeurIPS 2025.

- Senior Machine Learning Researcher, Scribe AI/ML**, Pittsburgh, PA Feb 2024 – May 2024
- Post-training LLMs for web navigation. Developed ScribeAgent, the SOTA open-source web agent.
- Research Intern, Microsoft Research**, Cambridge, MA May 2023 – Aug 2023
- Mentors: David Alvarez-Melis, Nicolò Fusi
  - Aligning LLMs to specialized domains (e.g., low-resource languages, protein sequences, chemical formulas) via special tokens. Paper accepted by ICML 2024.
- Research Intern, Determined AI, Hewlett Packard Enterprise**, Pittsburgh, PA Jun 2022 – Dec 2022
- Mentor: Liam Li
  - Fine-tuning LLMs and ViTs for diverse modalities via tokenizer training and distribution alignment. Paper accepted by ICML 2023 as an oral presentation.
- Product Manager Intern, SenseTime Face ID Research**, Beijing, China Jun 2018 – Sep 2018
- Worked on 3D-structured-light Face ID; participated in 5 software version releases and testing.

## Honors & Awards

<b>CMU MLD Google PhD Fellowship Nomination</b> , one of 3 PhD students nominated	2025
<b>Wilson Center, Pathways to AI Policy Program</b> , fellow	2025
<b>J.P. Morgan AI Ph.D. Fellowship</b> , awardee (accepted)	2024
<b>Bloomberg Data Science Ph.D. Fellowship</b> , awardee (declined)	2024
<b>CMU MLD Two Sigma PhD Fellowship Nomination</b> , one of 2 students nominated	2023
<b>UCLA Daus Prize</b> , top-5 undergraduate students in mathematics	2021
<b>UCLA Dean's Honors List</b> , awardee	2017 – 21

## Talks

<b>Thinking vs. Doing: Agents that Reason by Scaling Test-Time Interaction</b> <i>New York Reinforcement Learning Workshop</i>	Sep 2025
<b>Production-Scale Workflow Data Empowers Specialized Web Agents</b> <i>Ai4 Research Summit</i>	Aug 2025
<b>Thinking vs. Doing: Agents that Reason by Scaling Test-Time Interaction</b> <i>Agentic AI Summit, Berkeley</i>	Aug 2025
<b>LLM Meets Web Browsing</b> <i>AIRe Lab @ CMU</i>	Apr 2025
<b>LLM Meets Web Browsing</b> <i>EFML Reading Group, Stanford</i>	Mar 2025
<b>Repurposing LLMs for Long-Tail ML Applications</b> <i>Ai4 Research Summit</i>	Aug 2024
<b>Machine Learning for Diverse Tasks</b> <i>Guest Lecture, ML with Large Datasets, CMU 10605</i>	Nov 2023
<b>Bridging LLMs and Long Tail ML Applications</b> <i>Catalyst Reading Group, CMU</i>	Nov 2023
<b>Cross-Modal Fine-Tuning</b> <i>AI4Science Talks</i>	Mar 2023
<b>DASH: How to Search Over Convolutions</b> <i>The AutoML Podcast</i>	Dec 2022

## Professional Service

---

**Co-organizer:** CMU Agent Workshop (2024 & 2025); AutoML Decathlon, NeurIPS 2022 Competition Track

**Committee Member:** CMU MSML Admissions Committee (Fall 2022); CMU MLD Open House Committee (Spring 2024/2025)

**Conference Reviewer:** NeurIPS (2022-2025), ICLR (2024/2025), ICML (2024/2025), AAAI (2025), CVPR (2025), ICCV (2025)

**Teaching Assistant:** Deep Learning Systems (CMU 10714), ML in Practice (CMU 10718), Linear Algebra (UCLA Math 115A)

## Skills

---

**Programming:** Python, C, C++, Bash, R (Proficient); MATLAB, Java, Arduino (Familiar)

**Tools:** Git, LaTeX, PyTorch, Tensorflow, Scikit-learn, OpenCV, OpenAI Gym, Google Cloud Platform, Docker, SolidWorks

## Research Experience

---

**SAGE Lab, CMU, Pittsburgh, PA** June 2021 – Present  
*Advisor: Ameet Talwalkar*

- Ph.D. research on developing effective and efficient ML/AutoML tools for solving diverse tasks in practice.

**Lin Yang's Group, UCLA, Los Angeles, CA** Jan 2020 – June 2021  
*Advisor: Lin F. Yang*

- Studied sample-efficient reinforcement learning; proposed an algorithm for estimating the value functions using nearest neighbor function approximator; provided theoretical justification on the sample complexity.

**Center for Vision, Cognition, Learning, and Autonomy, UCLA, Los Angeles, CA** Jan 2019 – June 2021  
*Advisors: Song-Chun Zhu, Ying Nian Wu*

- Studied how theory of mind (ToM) can be integrated into various ML settings to improve algorithm efficiency.
- **Project 1: Multi-Agent Deep Reinforcement Learning with ToM.** Proposed a ToM algorithm in a referential game setting where the teacher and the student model each other's action likelihood while learning Q-functions.
- **Project 2: Efficient Learners in Iterative Machine Teaching:** Integrated ToM into machine teaching; improved teaching efficiency by having the learners model the teacher's training sample selection strategy.
- **Project 3: Meta Machine Teaching:** Studied how meta-learning can be combined with machine teaching.

**Computational Metastasis Lab, Fields Institute, Toronto, Canada** Jul 2019 – Sep 2019  
*Advisor: Nima Maftoon (University of Waterloo)*

- Developed a vascular network reconstruction framework that uses the main vessel skeletons segmented from clinical images and global constructive optimization to generate patient-specific cerebral vascular models.