# Jiashu Xu 徐家澍

Cambridge, MA | 949-522-2936 | jxu1@g.harvard.edu | Website | Scholar | cnut1648 | jiashu-xu

## Education

**Harvard University** — Cambridge, USA
*M.S. in Computational Science and Engineering; cross-registered at MIT; GPA: 4.0/4.0* — *Fall 2022 – Spring 2024*

**University of Southern California** — Los Angeles, USA
*B.S. double major in Applied Math & Computer Science;* Summa Cum Laude*; GPA: 3.97/4.0* — *Fall 2020 – Spring 2022*

**University of California, Irvine** — Irvine, USA
*B.S. double major in Applied Math & Computer Science; GPA: 3.98/4.0* — *Fall 2018 – Spring 2020*

**Hong Kong University of Science and Technology** — Hong Kong, China
*UCEAP summer visiting student; study robotics; GPA: 4.0/4.0* — *Summer 2019*

## Awards

| | |
|---|---|
| CURVE Research Fellowship | $1250/semester Research Stipend |
| Jennifer Battat Scholarship | $3.5k for Mathematics Major |
| USC Transfer Merit Scholarship | Half-tuition Merit Scholarship for 2% Of Transfer Applicants |
| USC Academic Achievement Award | Double Major with 3.75+ GPA |
| USC & UCI Dean's List | All Semesters |

## Research Interests

My current research interests lie in **Reliable AI**. Particularly,

1. AI Safety. For example, defending against malicious exploitation of LLM vulnerabilities ([3], [4]), protecting open-sourced LLM ownership ([1]), and aligning LLMs with human preferences ([2]).
2. Training AI that excels in low-resource regimes, through indirect supervision ([7], [11]) or synthetic data generation ([5], [6], [8], [10]).
3. Explanation and model learning from explanation ([12] to [14]).

## Publications & Services

*=Equal Contribution

[1] **Training Large Language Models as Reward Models**
**Jiashu Xu**, Daniel Pressel, Prasoon Goyal, Luke Dai, Reza Ghanadan, Michael Johnston
*COLM*, 2024 (To be submitted, under Amazon Internal Review) paper

[2] **Instructional Fingerprinting of Large Language Models**
**Jiashu Xu**, Fei Wang*, Mingyu Derek Ma*, Pang Wei Koh, Chaowei Xiao, Muchao Chen
*NAACL*, 2024 (Under Review) paper

[3] **Instructions as Backdoors: Backdoor Vulnerabilities of Instruction Tuning for Large Language Models**
**Jiashu Xu**, Mingyu Derek Ma, Fei Wang, Chaowei Xiao, Muhao Chen
*NAACL*, 2024 (Under Review) paper

[4] **Test-time Backdoor Mitigation for Black-Box Large Language Models with Defensive Demonstrations**
Wenjie Mo, **Jiashu Xu**, Qin Liu, Jiongxiao Wang, Jun Yan, Chaowei Xiao, Muhao Chen
*NAACL*, 2024 (Under Review) paper

[5] **BEHAVIOR Vision Suite: Customizable Dataset Generation via Simulation**
Yunhao Ge*, Yihe Tang*, **Jiashu Xu***, Cem Gokmen*, Chengshu Li, Wensi Ai, Benjamin Jose Martinez, Arman Aydin, Mona Anvari, Ayush K Chakravarthy, Hong-Xing Yu, Josiah Wong, Sanjana Srivastava, Sharon Lee, Shengxin Zha, Laurent Itti, Yunzhu Li, Roberto Martín-Martín, Miao Liu, Pengchuan Zhang, Ruohan Zhang, Fei-Fei Li, Jiajun Wu
*CVPR*, 2024 (Under Review) paper

[6] **DreamDistribution: Prompt Distribution Learning for Text-to-Image Diffusion Models**
Brian Nlong Zhao, Yuhang Xiao*, **Jiashu Xu***, Xinyang Jiang, Yifan Yang, Dongsheng Li, Laurent Itti, Yunhao Ge, Vibhav Vineet
*CVPR*, 2024 (Under Review) code paper

[7] **Can NLI Provide Proper Indirect Supervision for Low-resource Biomedical Relation Extraction?**
**Jiashu Xu**, Mingyu Derek Ma, Muhao Chen
*ACL*, 2023 (**Oral**) code paper

[8] **Dall-e for detection: Language-driven context image synthesis for object detection**
Yunhao Ge*, **Jiashu Xu***, Brian Nlong Zhao, Neel Joshi, Laurent Itti, Vibhav Vineet
*arXiv*, 2022 code paper extension

[9] **X-Norm: Exchanging Normalization Parameters for Bimodal Fusion**
Yufeng Yin*, **Jiashu Xu***, Tianxin Zu, Mohammad Soleymani
*ICMI*, 2022 paper

[10] **Neural-Sim: Learning to Generate Training Data with NeRF**
Yunhao Ge, Harkirat Behl*, **Jiashu Xu***, Suriya Gunasekar, Neel Joshi, Yale Song, Xin Wang, Laurent Itti, Vibhav Vineet
*ECCV*, 2022 code paper

[11] **Unified Semantic Typing with Meaningful Label Inference**
James Y. Huang, Bangzheng Li*, **Jiashu Xu***, Muhao Chen
*NAACL*, 2022 code paper

[12] **Dissection Gesture Sequence during Nerve Sparing Predicts Erectile Function Recovery after Robot-Assisted Radical Prostatectomy**
Runzhuo Ma, **Jiashu Xu**, Ivan Rodriguez, Gina DeMeo, Aditya Desai, Loc Trinh, Jessica H. Nguyen, Anima Anandkumar, Jim C. Hu, Andrew J. Hung
*NPJ Digital Medicine*, 2022 paper

[13] **Dissection Assessment for Robotic Technique (DART) to Evaluate Nerve-Spare of Robot-Assisted Radical Prostatectomy**
Runzhuo Ma, Alvin Hui, **Jiashu Xu**, Aditya Desai, Michael Tzeng, Emily Cheng, Loc Trinh, Jessica H. Nguyen, Anima Anandkumar, Jim C. Hu, Andrew J. Hung
*American Urological Association Annual Conference (AUA)*, 2022 paper

[14] **SalKG: Learning From Knowledge Graph Explanations for Commonsense Reasoning**
Aaron Chan, **Jiashu Xu**, Boyuan Long, Soumya Sanyal, Tanishq Gupta, Xiang Ren
*NeurIPS*, 2021 code paper

**Reviewer Service:** ACL Rolling Review, ACL 2023, EMNLP 2023, CVPR 2022

## Research Experience

**NVIDIA Research** Santa Clara, USA
*(Incoming) Research Scientist Intern | Manager: Ming-Yu Liu* *Summer 2024*
- Plan to work on LLM.

**Amazon Science** New York, USA
*Applied Scientist Intern | Manager: Daniel Pressel, Michael Johnston* *Summer 2023*
- Collaborated closely with the LLM team on the reward modeling side.
- Finetuned LLMs directly as reward models such that models learn to align with human preferences implicitly. Further benefits included zero-shot generalization to unseen dimensions and domains, high-quality data filtering, rationale generation to explain decisions, and synthetic conversation curation for AI self-improvement (RLAIF) [1].

**USC LUKA Group** Los Angeles, USA
*Research Assistant | Advisor: Prof. Muhao Chen, Prof. Chaowei Xiao* *Fall 2021 – Present*
- Proposed a fingerprinting method to safeguard open-source LLM ownership via memorizing fingerprint instances. Such lightweight fingerprint persists large-scale user finetuning on arbitrary datasets, is robust to fingerprint guessing and various PEFT training methods, and supports multi-stage fingerprinting akin to MIT License [2].

- Investigated backdoor vulnerabilities of instruction-tuned LLMs that have high backdoor success with minimal malicious instructions, can generalize to multiple tasks, and cannot be cured by continual learning [3]. And proposed leveraging clean in-context demonstrations as effective test-time defense against various backdoor attacks [4].
- Proposed indirect supervision to borrow supervision signals from resource-rich tasks to enhance resource-limited tasks: cross-domain transfer general domain NLI knowledge to improve low-resource biomedical Relation Extraction [7]; cross-task transfer semantic typing knowledge to handle large label space inference [11].

**Harvard AI4LIFE Group** Cambridge, USA

*Research Assistant | Advisor: Prof. Himabindu Lakkaraju* *Spring 2023 – Present*

- Integrated dynamic knowledge graph, constructed with a language model agent dynamically, into inference to improve factuality (In Progress).
- Investigated mechanistic interpretability on LLaVA-like vision language models to investigate how models react to complex queries such as referring expressions and object counting (In Progress).

**Stanford SVL Group** Palo Alto, USA

*Research Assistant | Advisor: Prof. Jiajun Wu* *Fall 2023 – Present*

- Developed BEHAVIOR Vision Suite, a customizable dataset generator featuring photorealistic assets and physically plausible annotations. Demonstrated applications include holistic benchmarks for 2D and 3D vision models, robustness evaluation through parametric out-of-distribution evaluation (*e.g.* low lighting, extreme camera pose), and synthetic dataset generation to bolster performance in low-resource scenarios [5].

**Microsoft Research** Los Angeles, USA

*Student Collaborator | Manager: Prof. Laurent Itti, Vibhav Vineet* *Spring 2022 – Present*

- Proposed prompt distribution learning for text-to-image and text-to-3D diffusion models to lightweight control image quality and diversity [6].
- Utilized diffusion models and object cut-and-paste to create coherent synthetic training datasets for enhancing low-resource object detection and segmentation [8]. And proposed differentiable synthetic dataset generation with NeRF to improve out-of-distribution object detection of varying views [10].

## Mentoring

Wenjie Mo USC B.S. Fall 2023 – Present

Brian Nlong Zhao USC B.S. → M.S., Research Scientist Intern at Microsoft Research Asia Fall 2022 – Fall 2023

## Working & Teaching Experience

**Teaching Assistant at USC** Los Angeles, USA

*Role: Office Hours, Discussion Sections, Grading* *Spring – Fall 2021*

- CSCI 567: Graduate level Machine Learning with Prof. Haipeng Luo.
- MATH 499: Independent Research with Prof. Neelesh Tiruviluamala.

**Teach for Los Angeles** Los Angeles, USA

*Mentor* *Spring 2021*

- Tutored middle school students from LA K-12 community 1-on-1 on mathematics for two hours every week.
- Inspired students to reach full math potential in preparation for college and STEM careers.

**Math CEO** Irvine, USA

*Mentor* *Fall 2018 – Spring 2020*

- Coordinated meetings with Santa Ana middle school students and taught mathematical thinking.

**Johnson & Johnson** Shanghai, China

*Digital & Analytics Data Assistant* *Summer 2019*

- Tracked counterfeit or parallel products from various sales channels using NLP techniques including semantic role labeling and named entity recognition.
- Devised a context extractor based on Jieba tokenizer and Chinese word vectors.
- Presented at PCS 2019 medicine CIO summit about the NLP approach for tracking counterfeit products.

**Wind Information** Shanghai, China

*Quantitative Index Research Analyst* *Spring – Summer 2018*

- Collaborated with product managers to launch Wind's new product: Wind Equity Backtester and implemented multiple prototype algorithms with test codes using python-wind and Pytest.
- Code-reviewed index-related codes and queried Wind index database to resolve clients' complaints.