# Jiashu Xu

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

## Education

**Harvard University** — Cambridge, USA
*Master's in Computational Science and Engineering; GPA: 4.0/4.0* — *Fall 2022 – Spring 2024*

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

**University of California, Irvine** — Irvine, USA
*B.S. 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 study abroad, study robotics; GPA: 4.0/4.0* — *Summer 2019*

**Awards:** Center for Undergraduate Research in Viterbi Engineering Fellowship, Jennifer Battat Scholarship, USC Transfer Merit Scholarship, USC Academic Achievement Award, USC & UCI Dean's List (all semesters)

## Research Interest

My current research interest is in **reliable AI**. Particularly,

1. AI Security ([1] to [4])
2. Training AI that excels in low-resource regimes, through indirect supervision ([7], [12]) or synthetic data ([5], [6], [8], [9], [11])
3. Explanation and how can we learn from explanation ([13] to [15])

## Publication

[1] **Training Large Language Models as Reward Models**
**Jiashu Xu**, Daniel Pressel, Prasoon Goyal, Luke Dai, Michael Johnston
*COLM*, 2024 (Under Amazon Internal Review)

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

[3] **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)

[4] **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

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

[6] **BEHAVIOR Vision Suite: Customized Dataset Generation with Realistic 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, Li Fei-Fei, Jiajun Wu
*CVPR*, 2024 (Under Review)

[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

[9] **EXACT: Compositional Augmentation for Image-level Weakly-Supervised Instance Segmentation**
**Jiashu Xu***, Yunhao Ge*, Brian Nlong Zhao, Laurent Itti, Vibhav Vineet
*TMLR*, 2023 (Under Review)

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

[11] **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

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

[13] **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 Digit Medicine*, 2022 paper

[14] **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

[15] **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

## Work & Teaching Experience

**Amazon Alexa Science** New York, USA
*Applied Scientist* *Summer 2023*
- LLM research for science team.

**Teaching Assistant** Los Angeles, USA
*CSCI 567: Machine Learning with Prof. Haipeng Luo* *Fall 2021*
- Held Office Hours, monitored piazza to answer students' questions regarding math and code implementation and graded homework and projects.

**Teach for Los Angeles** Los Angeles, USA
*Mentor* *Spring 2021*
- Tutored middle school students from LA K-12 community 1-on-1 on mathematics 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 products or parallel products from various sales channels using NLP techniques including semantic role labeling and named entity recognition.
- Devised context extractor based on Jieba tokenizer and Chinese word vectors.
- Presented in PCS 2019 medicine CIO summit about 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, queried Wind index database to resolve clients' complaints.

## Skills

**Languages:** Python, C/C++, Java, Scala, MATLAB, R, {Java, Type}Script, SQL, LaTeX

**Frameworks:** PyTorch, TensorFlow, scikit-learn, Pandas, Spark, React.js, Spring, AWS, gradio, MariaDB, MongoDB