Aaron Chan

CONTACT INFORMATION

Website: aarzchan.com
Email: aarzchan@gmail.com

RESEARCH INTERESTS

artificial intelligence (AI), machine learning (ML), natural language processing (NLP), generative AI (GenAI), large language models (LLMs), AI safety, AI alignment, trustworthy AI, model explainability

EDUCATION

University of Southern California, Los Angeles, CA

Doctor of Philosophy (PhD), Computer Science Aug 2017 - Dec 2022

- Dissertation: "Generating and Utilizing Machine Explanations for Trustworthy NLP"
- Adviser: Prof. Xiang Ren
- Committee: Prof. Xiang Ren (chair), Prof. Robin Jia, Prof. Jesse Thomason, Prof. Bistra Dilkina, Prof. Morteza Dehghani

University of Pennsylvania, Philadelphia, PA

Master of Science in Engineering (MSE), Robotics Aug 2015 - May 2017

- Advisers: Prof. Kostas Daniilidis, Prof. Jianbo Shi

University of Maryland, College Park, College Park, MD

Bachelor of Science (BS), Electrical Engineering Aug 2011 - May 2015

- Advisers: Prof. Rama Chellappa, Prof. David Jacobs

EXPERIENCE

Meta, Menlo Park, CA (Remote)

Research Scientist, GenAI Oct 2024 - Present

- Meta AI Safety Team
- Developed LLM post-training algorithms for finetuning or prompting Llama models to detect/classify Meta AI content that violates Meta's safety policies.
- Built automated pipelines for data collection, generation, and processing, across multiple Meta product surfaces.
- Developed a general algorithm for using LLM prompting to efficiently estimate the prevalence of arbitrary safety issues in Meta AI production traffic.
- Achieved major gains in safety classification (precision, recall, F1, FPR) and prevalence estimation (prevalence error, precision) performance, yielding downstream improvements in Meta AI topline metrics (VR, FRR).

Research Scientist, MRS Dec 2022 - Oct 2024

- Modern Recommendation Systems (MRS) AI Team
- Developed ML models to improve Reels ranking on Facebook and Instagram, with a focus on long-term user value optimization.
- Achieved significant gains in Reels topline metrics like DAU, sessions, and watch time.

Student Researcher, AI Integrity Jan 2022 - Apr 2022

- Managers: Maziar Sanjabi, Hamed Firooz
- Developed FRAME, a framework for evaluating rationale-label consistency metrics for free-text rationales [9].

Research Intern, AI Integrity Sep 2021 - Jan 2022

- Managers: Maziar Sanjabi, Hamed Firooz
- Developed UNIREX, a unified learning framework for jointly optimizing language model rationale extractors with respect to faithfulness, plausibility, and task performance [7].
- Achieved 5x speedup, 6.7% topline metric gains, and $0.5M annual cost savings for Meta's Bullying & Harassment detection system.

University of Southern California, Los Angeles, CA

Graduate Research Assistant Oct 2020 - Dec 2022

- Intelligence and Knowledge Discovery (INK) Lab
- Adviser: Prof. Xiang Ren

- Conducted fundamental research in model explainability [7, 9, 11], explanation-based learning [6, 8, 10, 11, 12, 13], and commonsense reasoning [4, 5, 6] for NLP.

Graduate Teaching Assistant Jan 2022 - May 2022

- CSCI 566 – Deep Learning and its Applications
- Instructor: Prof. Xiang Ren

Graduate Teaching Assistant Sep 2020 - Dec 2020

- CSCI 100xg – Explorations in Computing
- Instructor: Prof. Saty Raghavachary

Google, Mountain View, CA

Hardware Engineering Intern May 2017 - Aug 2017

- Android Camera Team
- Manager: Ying Chen Lou
- Worked on designing a saliency detection algorithm to improve camera autofocus on the Google Pixel phone.

GRASP Lab, University of Pennsylvania, Philadelphia, PA

Graduate Research Assistant Feb 2017 - May 2017

- Adviser: Prof. Jianbo Shi
- Constructed a first-person video dataset of one-on-one basketball games to train a model for egocentric trajectory prediction from a single image [3].

Graduate Research Assistant May 2016 - Oct 2016

- Adviser: Prof. Kostas Daniilidis
- Helped develop an algorithm to robustly estimate 6-DoF object pose from a single RGB image of the object [2].

Publications

[15] **ResPrompt: Residual Connection Prompting Advances Multi-Step Reasoning in Large Language Models**
S. Jiang, Z. Shakeri, A. Chan, M. Sanjabi, H. Firooz, Y. Xia, B. Akyildiz, Y. Sun, J. Li, Q. Wang, A. Celikyilmaz
NAACL 2024

[14] **Tailoring Self-Rationalizers with Multi-Reward Distillation**
S. Ramnath, B. Joshi, S. Hallinan, X. Lu, L. Li, A. Chan, J. Hessel, Y. Choi, X. Ren
ICLR 2024

- SeT LLM Workshop at ICLR 2024

[13] **KNIFE: Distilling Reasoning Knowledge From Free-Text Rationales**
A. Chan*, Z. Zeng*, W. Lake, B. Joshi, H. Chen, X. Ren
Technical Report - 2023

- TrustML-(un)Limited Workshop at ICLR 2023

[12] **XMD: An End-to-End Framework for Interactive Explanation-Based Debugging of NLP Models**
D. Lee*, A. Kadakia*, B. Joshi, A. Chan, Z. Liu, K. Narahari, T. Shibuya, R. Mitani, T. Sekiya, J. Pujara, X. Ren
ACL 2023 - Demo Track

[11] **Are Machine Rationales (Not) Useful to Humans? Measuring and Improving Human Utility of Free-Text Rationales**
B. Joshi*, Z. Liu*, S. Ramnath, A. Chan, Z. Tong, Q. Wang, Y. Choi, X. Ren
ACL 2023 (**Oral**)

- TRAIT Workshop at CHI 2023

[10] **PINTO: Faithful Language Reasoning Using Prompt-Generated Rationales**
P. Wang, A. Chan, F. Ilievski, M. Chen, X. Ren
ICLR 2023

- TL4NLP Workshop at NeurIPS 2022
- TSRML Workshop at NeurIPS 2022

[9] **FRAME: Evaluating Rationale-Label Consistency Metrics for Free-Text Rationales**
A. Chan, S. Nie, L. Tan, X. Peng, H. Firooz, M. Sanjabi, X. Ren
Technical Report - 2022
- BlackboxNLP Workshop at EMNLP 2022

[8] **ER-Test: Evaluating Explanation Regularization Methods for NLP Models**
B. Joshi*, A. Chan*, Z. Liu*, S. Nie, M. Sanjabi, H. Firooz, X. Ren
Findings of EMNLP 2022
- TrustNLP Workshop at NAACL 2022

[7] **UNIREX: A Unified Learning Framework for Language Model Rationale Extraction**
A. Chan, M. Sanjabi, L. Mathias, L. Tan, S. Nie, X. Peng, X. Ren, H. Firooz
ICML 2022 (**Spotlight**)
- SRML Workshop at ICLR 2022
- BigScience Workshop at ACL 2022

[6] **SalKG: Learning From Knowledge Graph Explanations for Commonsense Reasoning**
A. Chan, J. Xu, B. Long, S. Sanyal, T. Gupta, X. Ren
NeurIPS 2021
- XAI Workshop at ICML 2021

[5] **Learning Contextualized Knowledge Structures for Commonsense Reasoning**
J. Yan, M. Raman, A. Chan, T. Zhang, R. Rossi, H. Zhao, S. Kim, N. Lipka, X. Ren
Findings of ACL 2021
- KR2ML Workshop at NeurIPS 2020

[4] **Learning to Deceive Knowledge Graph Augmented Models via Targeted Perturbation**
M. Raman, A. Chan*, S. Agarwal*, P. Wang, H. Wang, S. Kim, R. Rossi, H. Zhao, N. Lipka, X. Ren
ICLR 2021
- KR2ML Workshop at NeurIPS 2020 (**Best Paper Award Finalist**)

[3] **Egocentric Basketball Motion Planning from a Single First-Person Image**
G. Bertasius, A. Chan, J. Shi
CVPR 2018
- MIT Sloan Sports Analytics Conference (SSAC) 2018

[2] **6-DoF Object Pose from Semantic Keypoints**
G. Pavlakos, X. Zhou, A. Chan, K. Derpanis, K. Daniilidis
ICRA 2017

[1] **Scalable Vision System for Mouse Homecage Ethology**
G. Salem, J. Krynitsky, B. Kirkland, E. Lin, A. Chan, S. Anfinrud, S. Anderson, M. Garmendia-Cedillos, R. Belayachi, J. Alonso-Cruz, J. Yu, A. Iano-Fletcher, G. Dold, T. Talbot, A. Kravitz, J. Mitchell, G. Wu, J. Dennis, M. Hayes, K. Branson, T. Pohida
ACIVS 2016

* Equal contribution.

Awards

Amazon Research Award – Alexa Fairness in AI (PI: Prof. Xiang Ren) 2022

Best Paper Award Finalist, KR2ML Workshop at NeurIPS 2020

Mentoring

Research Interns at Meta
- Song Jiang (2023-2024), PhD Student at UCLA [15]

Research Assistants at USC
- Sahana Ramnath (2022-2023), PhD Student at USC [11, 14]

- Zhiyuan Zeng (2022-2023), Undergraduate Student at Tsinghua University [13]
- Zhewei Tong (2022-2023), Undergraduate Student at Tsinghua University [11]
- Ziyi Liu (2022-2023), Master's Student at USC [8, 11, 12, 13]
- Brihi Joshi (2022-2023), PhD Student at USC [8, 11, 12, 13]
- Wyatt Lake (2021-2023), High School Student at Harvard-Westlake School [13]
- Siba Smarak Panigrahi (2021), Undergraduate Student at IIT Kharagpur
- Tanishq Gupta (2021), Undergraduate Student at IIT Delhi [6]
- Boyuan Long (2021), Undergraduate Student at USC [6]
- Jiashu Xu (2021), Undergraduate Student at USC [6]
- Siddhant Agarwal (2020-2021), Undergraduate Student at IIT Delhi [4]
- Mrigank Raman (2020-2021), Undergraduate Student at IIT Delhi [4, 5]

Skills

Programming Languages: Python, LaTeX

ML Libraries: PyTorch, Lightning, Captum, Scikit-learn

NLP Libraries: Hugging Face Transformers, Hugging Face Datasets

Data Analysis Libraries: NumPy, Pandas, Matplotlib, Seaborn

Other Tools: VSCode, GitHub, Neptune, Hydra, Slurm

[Last updated: Feb 20, 2025]