

Summary

- Industry** End-to-end experience in designing and deploying 0.01B–10B+ models in production
Expertise in bridging theoretical analysis with practical problem-solving in industrial settings
3 years at Google Core ML & 9 months at MIT-IBM Watson AI Lab
- Research** 21 publications (9 as first-author, 7 as corresponding-author), 2 patents, and 400+ citations
- Expertise** Machine Learning (large-small model collaboration, personalization, learning to rank, LLM, recommendation systems), Trustworthy AI (differential privacy, robustness, explainability), Computational Social Choice, and Theoretical Computer Science

Education

- 1/2018–5/2023 **Ph.D. in Computer Science**, Rensselaer Polytechnic Institute (RPI) NY, USA
Thesis: *Group Decision Makings from Partial Preferences*
- 8/2015–5/2018 **M.Eng. in Material Engineering**, Rensselaer Polytechnic Institute NY, USA
Received *Presidential Graduate Research Fellowship*
- 8/2010–5/2014 **B.S. in Mathematics and Physics**, Tsinghua University Beijing, China
Minor in Computer Technology, and in *Academic Talent Program*

Experience (all full-time)

Google, Core Machine Learning (Recommendation ML Group) CA, USA
Skills: Python, C/C++, TensorFlow, TFData, Keras, TPU Training & Inference

- ML Engineer** Project 1: SplitFormer, A Low-Latency, Splittable Transformer Architecture
7/2023–1/2026
- Theoretically showed that existing methods (e.g., LinFormer, FlashAttention) cannot reduce inference latency by more than 25%, far below the target for production use.
 - Proposed SplitFormer, which uses an offline Transformer to compress long-term user activity history into summaries and uses an online Transformer to combine the summaries with short-term history for complete user representations; proved a theoretical latency reduction of 90%+.
 - Led end-to-end design, training, and deployment of SplitFormer in the *Google Shopping* recommendation system.
 - Achieved 3× model size and 5% accuracy improvement while cutting online inference compute cost by 30%, driving an estimated multi-million-dollar annual revenue impact for *Google Shopping*.

Project 2: User Behavior Summarization via Tokenization

- Used VQ-VAE or RQ-VAE (depending on data) to discretize user behavior summary embeddings into tokens; aligned tokens can be directly used as prefix prompts for LLMs.
- Built an end-to-end pipeline covering training, storage, inference, and low-latency serving.
- Reduced storage cost by ~60% and training cost by ~25% vs. raw embedding.

Project 3: Large Encoder Model for User Behavior Prediction & Summarization

- Designed and productionized its *time embedding* module – from design to system integration – improving prediction accuracy by 17% and summarization accuracy by 4% (relative).
- Contributed to the full model lifecycle, including architecture design, data processing, full-behavior pre-training, downstream task fine-tuning, and production deployment.

- Research Intern** Intern Project: Unbiased Learning to Rank with a More Accurate Position Bias Estimator
5/2022–8/2022
- Found that the classic position bias model's independence assumption fails when a single data entry contains multiple clicks; proposed a new model with a probabilistic correlation coefficient and proved it is theoretically unbiased.
 - Designed an EM-like training framework to optimize the new model, which has no explicit loss function; achieved ~6% higher accuracy than the previous SOTA.

Visiting Scholar MIT-IBM Watson AI Lab (5/2019–8/2019 & 9/2018–1/2019) MA, USA
Project: Certifiably Robust Interpretation via Rényi Differential Privacy

- Proposed and implemented the first algorithm with provable robustness against ℓ_∞ -norm attacks (Skills: PyTorch, MATLAB).
- Outperformed SOTA by 12% in robustness with higher accuracy and no extra computational cost.
- Results published as a top-tier paper and two U.S. patents.

Selected Research Outputs (* denotes corresponding author)

- AIJ & AAI **Certifiably Robust Interpretation via Rényi Differential Privacy**
Ao Liu*, Xiaoyu Chen, Sijia Liu, Lirong Xia, and Chuang Gan
- TMLR **Smoothed Differential Privacy**
Ao Liu*, Yu-Xiang Wang, and Lirong Xia
- AAAI (oral) **Near-Neighbor Methods in Random Preference Completion**
Ao Liu*, Qiong Wu, Zhenming Liu, and Lirong Xia
- AAAI (oral) **Learning Plackett-Luce Mixture from Partial Preferences**
Ao Liu*, Zhibing Zhao, Chao Liao, Pinyan Lu, and Lirong Xia
- AAAI (spotlight) **The Semi-Random Likelihood of Doctrinal Paradoxes**
Ao Liu* and Lirong Xia
- UAI (oral) **Accelerating Voting by Quantum Computation**
Ao Liu*, Qishen Han, Lirong Xia, and Nengkun Yu
- ETRA (oral) **Differential Privacy for Eye-Tracking Data**
Ao Liu*, Lirong Xia, Andrew Duchowski, Reynold Bailey, Kenneth Holmqvist, and Eakta Jain
- UAI (oral) **How Private Are Commonly-Used Voting Rules?**
Ao Liu, Yun Lu*, Lirong Xia, and Vassilis Zikas
- Polymer Physics **Simulation of Pulse Responses of Lithium Salt-Doped Poly-Ethyleneoxide**
Ao Liu, F. Zeng*, Y. Hu, S. Lu, W. Dong, X. Li, C. Chang, and D. Guo
- U.S. Patent **Certifiably Robust Interpretation**
Ao Liu, Sijia Liu, Bo Wu, Lirong Xia, Qi Cheng Li, and Chuang Gan
- U.S. Patent **Interpretation Maps with Guaranteed Robustness**
Ao Liu, Sijia Liu, Abhishek Bhandwalder, Chuang Gan, Lirong Xia, and Qi Cheng Li
- AAAI (oral) **Differentially Private Condorcet Voting**
Zhechen Li, Ao Liu, Lirong Xia, Yongzhi Cao*, and Hanpin Wang
- IJCAI (oral) **Learning Mixtures of Random Utility Models from Incomplete Preferences**
Zhibing Zhao*, Ao Liu, and Lirong Xia

► See all 21 publications and 2 patents at [[Personal Website](#)] [[Google Scholar](#)]

Academic Impact

- Reviewer Journal: Information Sciences, ACM ToIS, TMLR, DMLR, IPA, and Sankhya B
Conference: NeurIPS, ICML, ICLR, AAI, IJCAI, UAI, and AISTATS
- Invited Talk Institute for Theoretical Computer Science, Shanghai University of Finance & Economics
Topic: *Preference Learning from General Partial Orders*
- Scholarships **Presidential Graduate Research Fellowship** [[Certificate](#)]
RPI-IBM AI Horizon Scholarship (fund 3 years' tuition and stipend)