Yaowen Ye 叶耀文

Email | Homepage | LinkedIn

Education

- University of California, Berkeley 2025.08
  - Incoming CS PhD
- The University of Hong Kong 2021.09-2025.06
  - Major in Computer Science and Minor in Statistics
  - Machine Learning, Bayesian Inference, Stochastic Process, Data Structures & Algorithms, etc.
- University of California, Berkeley 2024.01-2024.08
  - Visiting via Berkerley Extension
  - LLM Foundation & Safety; Forecasting; Linear Modeling; Efficient Algorithms and Intractable Problems.

Research Experience

- Intern at Steinhardt Group at UC Berkeley (advisor: Prof. Jacob Steinhardt) 2024.01-2024.08
  - I studied language model post-training under systematically unreliable supervision and showed that the canonical RLHF method breaks down in this setting. I demonstrate with a novel algorithm that it is better to direct unreliable feedback towards improving the *data* rather than the *model* as in RLHF.
- Remote Intern at AIMING Lab at UNC-Chapel Hill (advisor: Prof. Huaxiu Yao) 2023.10-2023.12
  - I studied the trade-off between helpfulness and harmlessness in language model alignment, exploring methods that leverage in-context reasoning to mitigate over-safety issues.
- Intern at Cognitive Reasoning Lab at Peking University (advisor: Prof. Yixin Zhu) 2023.06-2023.10
  - I worked on computational models of human intuitive physical reasoning. I and colleague identified key limitations in current theories and developed a new framework that conceptualizes intuitive physics as a dual process involving both probabilistic simulations and heuristical reasoning.
- Intern at Data Intelligence Lab at HKU (advisor: Prof. Chao Huang) 2021.12-2023.06
  - I applied generative self-supervised learning techniques to address data noise and sparsity issues in graph neural network-based recommender systems, reducing bias for new users with short context.

Publication & Projects

- Iterative Label Refinement Matters More than Preference Optimization ICLR'25 (spotlight)
  Yaowen Ye, Cassidy Laidlaw, Jacob Steinhardt
- Iterative Label Refinement Matters More than Preference Optimization Arxiv
  Xiaoyuan Zhu, **Yaowen Ye**, Tianyi Qiu, H. Zhu, S. Tan, A. Mannan, J. Michala, R. A. Popa, W. Neiswanger
- A Simulation-Heuristics Dual-Process Model for Intuitive Physics
  Shiqian Li, Yuxi Ma, **Yaowen Ye**, Bo Dai, Yujia Peng, Chi Zhang, Yixin Zhu
- Graph Masked Autoencoder for Sequential Recommendation SIGIR'23
  Yaowen Ye, Chao Huang, and Lianghao Xia.
- Masked Graph Transformer for Recommendation SIGIR'23
  Chaoliu Li, Chao Huang, Lianghao Xia, Xubin Ren, **Yaowen Ye**, and Yong Xu.

Leadership & Teaching

- Co-founder of HKU AI4Good 2023-2025

- Co-founded a student interest group focused on AI safety, and alignment.
- Organized events including research seminars, paper reading groups, etc.

- Co-leader of HKU Astar (former HerKules Robomaster-AI) Team 2022-2023
  - Designed ML-based auto-aiming algorithms for the RoboMaster competition.
  - Designed localization algorithms based on point cloud reconstruction with stereo cameras.
- Teaching Assistant of COMP2113: Programming II 2024
  - Give tutorials on Linux usage, C++ and Bash programming.
  - Answer students' questions on assignments and exams.

Honors & Awards

Berkeley Global Access Scholarship for Visiting Students (US$ 2000) 2024
Undergraduate Research Fellowship Program at HKU (HK$ 40,000) 2024
Kenny Tung Scholarship (HK$ 10,000) 2024
YC Cheng Engineering Scholarship (HK$15,000) 2024
Silver Award of the 1st EEG & AI Competition in HKU (HK$ 2,000) 2023
Young Tsun Dart Scholarship (HK$ 15,000) 2023
Noel Chau Scholarship (HK$20,000) 2022
Ho Fook and Chan Kai Ming Prize (HK$ 5,000) 2022
Dean's Honours List 2021-22, 2022-23
Outstanding Student of the Year of the nationwide Science Talent Program 2019
Outstanding Student of the Tsinghua University Global Innovation Exchange Summer Camp 2019
First Prize of the National Olympiad in Informatics (Guangdong Province) 2018

Skills

- Programming Languages
  - Experienced with Python (along with PyTorch, TensorFlow, HuggingFace, etc.) and LaTeX.
  - Familiar with C++, Haskell, R, HTML, CSS, JavaScript, Java
- Human Study Deign: Qualtrics, CloudResearch Connect
- Graphics Design / Video Editting: PhotoShop, Final Cut Pro