Yaowen Ye 叶耀文

Email | LinkedIn | Homepage

EDUCATION

- The University of Hong Kong   2021.09-2025.06
  - Major in Computer Science and Minor in Statistics (4.09/4.3)
  - Machine Learning, Stochastic Process, Bayesian Inference, Data Structures & Algorithms, etc.
- University of California, Berkeley   2024.01-2024.08
  - Visiting via Berkerley Extension (4.0/4.0)
  - 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 showed that the canonical RLHF method for language model post-training breaks down under unreliable supervision, and I demonstrate with a novel algorithm that it is better to direct unreliable feedback towards improving the training *data* rather than the *model* as in RLHF.
- Intern at Yao's Lab at UNC-Chapel Hill (advisor: Prof. Huaxiu Yao)   2023.10-2023.12
  - I studied the trade-off between helpfulness and safety alignment in language models and explored 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 investigated the heuristics and mental simulations involved in human intuitive physical reasoning. We proposed a new framework that conceptualizes intuitive physics as a dual process. I also studied abductive reasoning and concept learning. [todo]
- Intern at Data Intelligence Lab at HKU (advisor: Prof. Chao Huang)   2021.12-2023.06
  - I used 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 (under review)
  Yaowen Ye, Cassidy Laidlaw, Jacob Steinhardt
- A Simulation-Heuristics Dual-Process Model for Intuitive Physics   Ongoing Project
  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 Association   2023-2025
  - Co-founded a student interest group focused on undergraduate AI research, 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, and 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)   2022
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 Deesign / Video Editting: PhotoShop, Final Cut Pro, Figma