Yu Ying Chiu (Kelly)

kellycyy@uw.edu | GScholar: Yu Ying Chiu | Linkedin: kellycyy | Github: @kellycyy

Education

University of Washington, MS in Computational Linguistics (NLP) Sep 2022 – Now

- **Courses:** Natural Language Processing, Language Models & Reasoning, Artificial Intelligence, GPA: 4.0/4.0

University of Hong Kong, BS in Decision Analytics (Stat./Comp. Sci.)& Psychology Sep 2017 – Jun 2022

- **Courses:** Statistical Machine Learning & Computational Social Psychology
- **Thesis:** Understanding Human Decision-Making in Stock Markets through Social Media Sentiment Analysis

Work Experience

ML Alignment & Theory Scholars (MATS) Program, MATS Scholar Jan 2025 – Now

- Design an evaluation framework on moral reasoning and human values of models to align models' behaviors, advised by Evan Hubinger (Anthropic) and Sydney Levine from Deepmind/NYU incoming prof.

Allen Institute for Artificial Intelligence, Research Collaborator May 2023 – Dec 2024

- Evaluated cultural knowledge of LLMs [6], advised by Dr. Bill Yuchen Lin and Prof. Yejin Choi.
- Built human-in-the-loop data collection platform [3] for collecting 2000^+ samples from 45^+ countries.

University of Washington Allen School of Computer Science, Research Assistant May 2023 – Dec 2024

- Designed evaluation framework and finetuned models for assessing LLM therapists behaviors during psychotherapy (e.g. reflecting upon client needs, normalizing expectations) [2], advised by Prof. Tim Althoff.
- Investigated the value preference of models in relation to psychological, sociological and philosophical theories, to provide insights on model alignment (e.g. effectiveness of AI Constitutions) [1], advised by Prof. Yejin Choi.

Publications

- [1] **Yu Ying Chiu**, Liwei Jiang, Yejin Choi. 2024. DailyDilemmas: Revealing Value Preferences of LLMs with Quandaries of Daily Life. *Accepted at ICLR 2025 as Spotlight.* [**Paper** | **Code** | **Data**]
- [2] **Yu Ying Chiu***, Ashish Sharma*, Inna Wanyin Lin, Tim Althoff. 2024. A Computational Framework for Behavioral Assessment of LLM Therapists. *Under Review at Nature Communications.* [**Paper** | **Code+Data**]
- [3] **Yu Ying Chiu**, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, Yejin Choi. 2024. CulturalBench: a Robust, Diverse and Challenging Benchmark on Measuring the (Lack of) Cultural Knowledge of LLMs. *Under Review at ACL.* [**Paper** | **Data** | **Leaderboard**]
- [4] Wenting Zhao, Tanya Goyal, **Yu Ying Chiu**, Liwei Jiang, Benjamin Newman, Abhilasha Ravichander, Khyathi Chandu, Ronan Le Bras, Claire Cardie, Yuntian Deng, Yejin Choi. 2024. WildHallucinations: Evaluating Long-form Factuality in LLMs with Real-World Entity Queries. *Under Review at ACL.* [**Paper**]
- [5] Zhilin Wang*, **Yu Ying Chiu***, Yu Cheng Chiu. 2023. Humanoid Agents: Platform for Simulating Human-like Generative Agents. *Published in EMNLP System Demo 2023.* [**Paper** | **Code** | **Demo**]
- [6] **Yu Ying Chiu**, Liwei Jiang, Maria Antoniak, Chan Young Park, Shuyue Stella Li, Mehar Bhatia, Sahithya Ravi, Yulia Tsvetkov, Vered Shwartz, Yejin Choi. 2024. CulturalTeaming: AI-Assisted Interactive Red-Teaming for Challenging LLMs'(Lack of) Multicultural Knowledge. *Preprint.* [**Paper**]
- [7] Abhinav Patil, Jaap Jumelet, **Yu Ying Chiu**, Andy Lapastora, Peter Shen, Lexie Wang, Clevis Willrich, Shane Steinert-Threlkeld. 2023. Filtered Corpus Training (FiCT) Shows that Language Models can Generalize from Indirect Evidence. *Accepted in TACL.* [**Paper**]

Service

STEM Outreach: Vice President at HKU Science Outreach Team (50 members, 300 beneficiaries) (2018), Organizing Committee at HKU Taster (2019) (200 participants), Volunteer mentor at Art x Science with Preface coding and Social Impact group (2022) (10 mentees)
Program Committee: International Conference on Learning Representations (ICLR) 2025