Yu Ying Chiu (Kelly)

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

Personal Website: https://kellycyy.github.io/

Education

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

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

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

Work Experience

New York University, Research Associate Jun 2025 – Now

- Built procedural moral reasoning benchmark in collaboration with Scale AI and 50+ PhD-level annotators, advised by Prof. Mitchell Gordon (MIT/OpenAI) and Prof. Sydney Levine (NYU/Google DeepMind).

ML Alignment & Theory Scholars (MATS) Program, Research Fellow Jan 2025 – Mar 2025

- Designed an evaluation framework on moral reasoning and human values of models to align models' behaviors, advised by Evan Hubinger (Anthropic).

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

- Evaluated cultural knowledge of LLMs and built human-in-the-loop data collection platform to curate data from 40+ countries, advised by Dr. Bill Yuchen Lin and Prof. Yejin Choi.

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

- Investigated the value preference of models in relation to psychological, sociological and philosophical theories, to provide insights on model alignment (e.g. effectiveness of Constitutional AI), advised by Prof. Yejin Choi.
- Designed evaluation framework and finetuned models for assessing LLM as psychotherapists demonstrating empathy (e.g. reflecting upon client needs, normalizing expectations), advised by Prof. Tim Althoff.

Highlighted Publications

- [6] **Yu Ying Chiu**, Michael S. Lee, Rachel Calcott, Brandon Handoko, Paul de Font-Reaulx, Paula Rodriguez, Chen Bo Calvin Zhang, Ziwen Han, Udari Madhushani Sehwag, Yash Maurya, Christina Q Knight, Harry R. Lloyd, Florence Bacus, Mantas Mazeika, Bing Liu, Yejin Choi, Mitchell L Gordon, Sydney Levine. 2025. MoReBench: Evaluating Procedural and Pluralistic Moral Reasoning in Language Models, More than Outcomes. *Under Review at ICLR 2026.* [**Project Website | Paper | Code | Data**]
- [5] **Yu Ying Chiu**, Zhilin Wang, Sharan Maiya, Yejin Choi, Kyle Fish, Sydney Levine, Evan Hubinger. 2025. Will AI Tell Lies to Save Sick Children? Litmus-Testing AI Values Prioritization with AIRiskDilemmas. *Under Review at ICLR 2026.* [**Paper | Code | Data**]
- [4] **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. 2025. CulturalBench: a Robust, Diverse and Challenging Benchmark on Measuring the (Lack of) Cultural Knowledge of LLMs. *ACL 2025* [**Paper | Data**]
- [3] **Yu Ying Chiu**, Liwei Jiang, Yejin Choi. 2025. DailyDilemmas: Revealing Value Preferences of LLMs with Quandaries of Daily Life. *ICLR 2025 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**]
- [1] Zhilin Wang*, **Yu Ying Chiu***, Yu Cheng Chiu. 2023. Humanoid Agents: Platform for Simulating Human-like Generative Agents. *EMNLP System Demo 2023.* [**Paper | Code | Demo**]

Service

Program Committee: International Conference on Learning Representations (ICLR): 2025-26, ACL ARR: 2025
Volunteer Work: EMNLP 2023 Volunteer; Art x Science Social Impact Group 2022 Volunteer Mentor; University of Hong Kong Science Outreach Society 2018 Vice President.