# Kawin Ethayarajh

✉ kawin.ethayarajh@chicagobooth.edu  ethayarajh  kawine.github.io

## Positions

**University of Chicago, Booth School of Business**
Assistant Professor, Applied AI Group — 07/2025 –
Kathryn and Grant Swick Faculty Scholar — 07/2025 –

**Princeton University**
Postdoctoral Research Associate, Princeton Language & Intelligence (PLI) — 2024 – 2025

**Contextual AI**
Visiting Researcher (hosted by Douwe Kiela) — 2023

**Allen Institute for Artificial Intelligence**
Research Scientist Intern (hosted by Yejin Choi and Swabha Swayamdipta) — 2021

**Google Research**
SWE Intern (hosted by the AdsAI Team) — 2019
SWE Intern (hosted by the Research & Machine Intelligence Team) — 2018

## Education

**Stanford University**
Ph.D., Computer Science — 2019 – 2024
*Committee*: Dan Jurafsky (advisor), Percy Liang, Diyi Yang
*Thesis*: Behavior-Bound Machine Learning

**University of Toronto**
M.Sc., Computer Science — 2017 – 2019
*Advisor*: Graeme Hirst

**University of Toronto**, Victoria College
B.Sc. Hons., Computer Science — 2013 – 2017

## Highlights

- SHP, the first large-scale open-source dataset of human preferences over text. SHP was the only dataset not made by OpenAI/Anthropic/Meta used for post-training Llama 2; post-training LLMs with human preferences inferred from social media data has since become mainstream.
- KTO, a post-training method that has seen wide adoption due to its practicality (offline, supports unpaired class-imbalanced feedback). KTO was part of a broader discovery that PPO, DPO, and other alignment objectives belong to a class of losses with deep connections to behavioral economics.

## Awards

ICML Spotlight (Top 3.5% of accepted) — 2024

ICML Outstanding Paper (Top 10 of 1233 accepted) — 2022

Facebook (Meta) PhD Fellowship: $84,000 USD — 2021
*1 of 2 recipients in the field of natural language processing.*

NSERC Postgraduate Scholarship - Doctoral: $63,000 CAD — 2019

NSERC Canada Graduate Scholarship - Doctoral: $105,000 CAD (declined) — 2019

Best Paper – Repl4NLP, ACL 2018 — 2018

Rhodes Scholarship Finalist — 2017

University of Toronto Fellowship: $11,200 CAD — 2017

John H. Moss Scholarship: $16,650 CAD — 2017
*Given to the top graduating student, for academics and leadership.*

Chancellor Northrop Frye Gold Medal — 2017
*For the graduating student with the highest academic standing at Victoria College.*

Bank of Montreal National Scholarship: $75,000 CAD — 2013
*Merit–based university scholarship granted to 8 Canadians.*

## Invited Talks

**Behavior-Bound Machine Learning**

| | |
|---|---|
| Northwestern University | 10/2024 |

**Human-Aware Losses**

| | |
|---|---|
| Georgia Tech | 10/2024 |
| Princeton PLI | 10/2024 |
| Amazon Research | 09/2024 |
| Shah Lab (Stanford) | 02/2024 |
| Moveworks | 02/2024 |
| SAP | 04/2024 |

**Machine Learning under Real-World Incentives**

| | |
|---|---|
| UChicago Booth | 03/2024 |
| MIT Media Lab | 03/2024 |
| University of Southern California Natural Language Seminar | 11/2023 |

**From *In Vitro* to *In Vivo* AI Evaluation**

| | |
|---|---|
| Stanford CS224U Guest Lecture | 05/2023 |
| University of Washington CS Colloquium | 04/2023 |
| IBM Research (Zurich) | 05/2022 |

**Understanding Dataset Difficulty with V-Usable Information**

| | |
|---|---|
| RIKEN Center for Advanced Intelligence Project (Japan) | 09/2022 |
| Stanford NLP | 08/2022 |
| ICML | 07/2022 |

## Teaching

**Stanford CS224U: Natural Language Understanding**, Teaching Assistant — Spring 2023

*Theoretical understanding and practical application of NLP systems. Specific topics include information retrieval, transformers, domain adaptation, and evaluation.*

**Stanford CS221: Artificial Intelligence**, Teaching Assistant — Fall 2022

*Foundational principles and practice implementing various AI systems. Specific topics include machine learning, search, Markov decision processes, game playing, constraint satisfaction, graphical models, and logic.*

## Service

**Workshop on Behavioral Machine Learning**, Founding Co-Organizer — 2024

*Co-founded a workshop for research at the intersection of machine learning and human behavior with colleagues from Harvard, UC Berkeley, and Cambridge, which was accepted to NeurIPS 2024.*

**Review of Undergraduate Computer Science**, Founding Editor-in-Chief — 2015 – 2017

*Started non-archival publication dedicated to CS undergrad research, helping students write and edit their first scientific articles and publishing research from from UToronto, Cornell, and MIT.*

**Governing Council of the University of Toronto**, Board Member — 2015 – 2016

*Appointed to a board of the university's highest governing body to shape student affairs, where I debated and voted on several key issues, including student privacy and data collection.*

## Publications


1. Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., and Kiela, D. (2024). Model alignment as prospect theoretic optimization. In *Forty-first International Conference on Machine Learning* **(spotlight).**

2. Vivek, R., Ethayarajh, K., Yang, D., and Kiela, D. (2024). Anchor points: Benchmarking models with much fewer examples. In Graham, Y. and Purver, M., editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1576–1601, St. Julian's, Malta. Association for Computational Linguistics

3. Ethayarajh, K., Choi, Y., and Swayamdipta, S. (2022). Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR. **Outstanding Paper.**

4. Ethayarajh, K. and Jurafsky, D. (2022). The authenticity gap in human evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6070

5. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*

6. Hewitt, J., Ethayarajh, K., Liang, P., and Manning, C. D. (2021). Conditional probing: measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1626–1639.

7. Ethayarajh, K. and Jurafsky, D. (2021). Attention flows are shapley value explanations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 49–54.

8. Ma, Z., Ethayarajh, K., Thrush, T., Jain, S., Wu, L., Jia, R., Potts, C., Williams, A., and Kiela, D. (2021). Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking. *Advances in Neural Information Processing Systems*, 34.

9. Ethayarajh, K. and Jurafsky, D. (2020). Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics.

10. Ethayarajh, K. (2019). How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

11. Ethayarajh, K., Duvenaud, D., and Hirst, G. (2019b). Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.

12. Ethayarajh, K., Duvenaud, D., and Hirst, G. (2019a). Towards understanding linear word analogies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3253–3262, Florence, Italy. Association for Computational Linguistics.

13. Ethayarajh, K. (2018). Unsupervised random walk sentence embeddings: A strong but simple baseline. In *Proceedings of The Third Workshop on Representation Learning for NLP @ ACL*, pages 91–100, Melbourne, Australia. Association for Computational Linguistics. **Best Paper.**