

Kawin Ethayarajh

✉ kawin@stanford.edu

🐦 [ethayarajh](https://twitter.com/ethayarajh)

🔗 kawine.github.io

EDUCATION

Stanford University

Ph.D., Computer Science
Advisor: Dan Jurafsky

2019 – 2024

University of Toronto

M.Sc., Computer Science
Advisor: Graeme Hirst

2017 – 2019

University of Toronto, Victoria College

B.Sc. Hons., Computer Science

2013 – 2017

HIGHLIGHTS

I created the second largest dataset of human preferences over text, Stanford Human Preferences (SHP), with 4.8 million preferences. SHP is used by AWS for reranking generations and one of the few datasets used by Meta's Llama-2 for alignment.

I trained and released the largest suite of human feedback-aligned LLMs, called Archangel: 70+ models spanning the Pythia and Llama families, aligned with methods like PPO, KTO, and DPO. This was part of a broader project on Human-Aware Losses (HALOs).

I co-created Dynaboard, an end-to-end evaluation-as-a-service platform used to host well-known challenges like DataPerf, Dynabench, and BabyLM.

AWARDS

ICML Outstanding Paper

2022

Facebook (Meta) PhD Fellowship: \$84,000 USD

2021

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.

NSERC Undergraduate Student Research Award: \$4,500 CAD

2015

Awarded by NSERC (Canadian NSF) to undergraduate researchers.

Bank of Montreal National Scholarship: \$75,000 CAD

2013

Merit-based university scholarship granted to 8 Canadians.

POSITIONS

Contextual AI, Visiting Researcher

Summer 2023

Hosts: Douwe Kiela

Project: Aligning large language models with human preferences.

Allen Institute for Artificial Intelligence, Research Scientist Intern

Summer 2021

Hosts: Yejin Choi and Swabha Swayamdipta

Project: Understanding dataset difficulty with information theory.

Google, SWE Intern

Summer 2019

Hosts: AdsAI Team

Project: Embed hypergraphs at scale using autoencoders.

Google , SWE Intern	Summer 2018
<i>Hosts</i> : Research & Machine Intelligence Team	
<i>Project</i> : Zero-shot relation extraction using pre-trained QA models.	
University of Toronto , Graduate Research Assistant	2017 – 2019
<i>Hosts</i> : Graeme Hirst and David Duvenaud	
<i>Project</i> : Theoretical analysis of word embeddings and sentence embeddings.	
University of Toronto , Undergraduate Research Assistant	2015 – 2016
<i>Hosts</i> : Michalis Famelis and Marsha Chechik	
<i>Project</i> : Transferability across domain-specific languages in software engineering.	

REPRESENTATIVE * denotes equal contribution (see full list on Google Scholar)
PUBLICATIONS

Model Alignment as Prospect Theoretic Optimization.
Kawin Ethayarajh, Winnie Xu, Dan Jurafsky, and Douwe Kiela.
under review.

The Authenticity Gap in Human Evaluation.
Kawin Ethayarajh and Dan Jurafsky.
Empirical Methods in Natural Language Processing (EMNLP), 2022.

Understanding Dataset Difficulty with \mathcal{V} -Usable Information.
Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta.
International Conference on Machine Learning (ICML), 2022. 🏆 **Outstanding Paper.**

Dynaboard: An Evaluation-As-A-Service Platform.
 Zhiyi Ma*, Kawin Ethayarajh*, Tristan Thrush*, Somya Jain, Ledell Wu, Robin Jia, Christopher Potts, Adina Williams, and Douwe Kiela.
Neural Information Processing Systems (NeurIPS), 2021.

Utility is in the Eye of the User: A Critique of NLP Leaderboards.
Kawin Ethayarajh and Dan Jurafsky.
Empirical Methods in Natural Language Processing (EMNLP), 2020.

INVITED TALKS	Machine Learning with Human Fault-Tolerance	
	University of Southern California Natural Language Seminar	11/2023
	From <i>In Vitro</i> to <i>In Vivo</i> AI Evaluation	
	Stanford CS224U Guest Lecture	05/2023
	University of Washington CS Colloquium	04/2023
	IBM Research (Zurich)	05/2022
	Understanding Dataset Difficulty with \mathcal{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
	<i>Theoretical understanding and practical application of NLP systems. Specific topics include information retrieval, transformers, domain adaptation, and evaluation.</i>	
	Stanford CS221: Artificial Intelligence , Teaching Assistant	Fall 2022
	<i>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.</i>	

SERVICE

Socially Responsible Language Modelling Research (SoLaR), Founding Co-Organizer 2023
Co-founded a workshop for the responsible use of language models with colleagues from UCL, Cambridge, ETH, and MILA, which was accepted to NeurIPS 2023 and received over 150 submissions.

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.