

## RESEARCH EXPERIENCE

### Contextual AI, Member of Technical Staff

Palo Alto, CA, USA

Research Engineer, Douwe Kiela & Amanpreet Singh

2023-07 – 2024-11

Developed model alignment methods, synthetic data pipelines, and ideas for process-level feedback optimization.

*Topics: retrieval, RLHF, Kahneman-Tversky Optimization, instruction finetuning, large language models*

### Meta, Fundamental AI Research (FAIR Labs)

New York, NY, USA

Research Scientist Intern with Dr. Karen Ullrich & Dr. Matthew Muckley

2022-09 – 2023-05

Researched ideas at the intersection of generative modeling, neural compression, and representation learning.

*Topics: generative models, compression, information theory, representation learning, autoencoding*

### Stanford University, Stanford AI Laboratory

Palo Alto, CA, USA

Visiting Research Scholar with Prof. Stefano Ermon & Chelsea Finn

2021-06 – 2021-11

Introduced self-referential operators for fractal data encoding, efficient compression, and controllable generation.

*Topics: score-based generative models, diffusion processes, latent variable models, implicit representation learning*

### Google DeepMind,

Mountain View, CA, USA

Research Scientist Intern with Igor Mordatch & Durk Kingma & David Dohan

2021-10 – 2022-08

(1) Improve Decision Transformer models to extrapolate in creative and general ways towards embodied game play and online decision-making. (2) Develop spectral diffusion models leveraging resolution agnostic architectures and signal adaptive scheduling. (3) Formalize language models as probabilistic programs via *Cascades* framework.

*Topics: diffusion models, Transformers, large language models, reinforcement learning, robotics, decision-making*

### Vector Institute & University of Toronto

Toronto, ON, Canada

Undergraduate Researcher with Prof. David Duvenaud

2020-01 – 2021-01

Derive variance-reducing gradient estimator and improve Neural ODE robustness through Bayesian inference w/ SDEs.

*Topics: stochastic differential equations, Bayesian neural networks, variational inference*

## SELECT PUBLICATIONS

### PEER-REVIEWED

- [8] Karel DOosterlinck, **Winnie Xu**, Chris Chris Develder, Thomas Demeester, Amanpreet Singh, Christopher Potts, Douwe Kiela, and Shikib Mehri, “Anchored preference optimization and contrastive revisions: Addressing underspecification in alignment,” Safe Generative AI Workshop at NeurIPS 2024, 2024.
- [7] Kawin Ethayarajh, **Winnie Xu**, Dan Jurafsky, and Douwe Kiela, “KTO: Model alignment as prospect theoretic optimization,” International Conference on Machine Learning [Spotlight Award], 2024.
- [6] **Winnie Xu**, Nikita Vassilyev, Douwe Kiela, and Shikib Mehri, “Learning from natural language preferences,” In progress, 2024.
- [5] **Winnie Xu**, Matthew Muckley, Yann Dubois, and Karen Ullrich, “Revisiting associative compression: I can’t believe it’s not better,” *International Conference on Machine Learning Neural Compression Workshop*, 2023.
- [4] David Dohan\*, **Winnie Xu\***, Aitor Lewkowycz, Jacob Austin, David Bieber, Raphael Gontijo Lopes, Yuhuai Wu, Henryk Michalewski, Rif A. Saurous, Jascha Sohl-dickstein, Kevin Murphy, and Charles Sutton, “Language model cascades,” *Beyond Bayes: Paths Towards Universal Reasoning Systems, International Conference on Machine Learning [Contributed Talk]*, 2022.
- [3] †Kuang-Hui Lee\*, Ofir Nachum\*, Mengjiao Yang, Lisa Lee, **Winnie Xu**, Daniel Freeman, Sergio Guadarrama, Ian Fischer, Eric Jang, Henryk Michalewski, and Igor Mordatch\*, “Multi-game decision transformers,” *Neural Information Processing Systems [Oral]*, 2022.
- [2] Michael Poli\*, **Winnie Xu\***, Stefano Massaroli, Chenlin Meng, and Stefano Ermon, “Self-similarity priors: Neural collages as differentiable fractal representations,” *Neural Information Processing Systems*, 2022.
- [1] **Winnie Xu**, Ricky T.Q. Chen, Xuechen Li, and David Duvenaud, “Infinitely deep bayesian neural networks with stochastic differential equations,” *International Conference on Artificial Intelligence and Statistics*, 2022.

\*co-first authorship, †ordering by seniority

## PROFESSIONAL EXPERIENCE

---

**Cohere, Large Language Models** Toronto, ON, Canada  
Machine Learning Researcher with Nick Frosst and Aidan Gomez 2021-01 – 2021-06  
Apply deep learning algorithms to improve training cost and personalization of billion parameter language models.  
*Topics: GPT, attention, distillation, distributed cloud training, TPUs*

**Nvidia, Simulations & Robotics** Toronto, ON, Canada  
Deep Learning Research Intern with Gavriel State and Prof. Animesh Garg 2020-08 – 2020-12  
Build performant GPU-accelerated environments towards time / resource efficient reinforcement learning for robotics.  
*Topics: Omniverse, IsaacGym, robotics simulation*

**Google, Tensorflow** Mountain View, CA, USA  
Research Engineering Intern with Dr. Tomer Kaftan 2020-05 – 2020-08  
Actualize state of the art pre-/post-hoc pruning methods for easy experimentation and efficient hardware computation.  
*Topics: lottery tickets, dynamic sparsity, Tensorflow Model Optimization Toolkit (contributor)*

## EDUCATION

---

**University of Toronto** 2017 – 2020, 2021 – 2022  
Honours Bachelors of Science in *Computer Science, Statistics, Mathematics* High Distinction  
Graduate coursework: Natural Language Processing (CSC401), Probabilistic Reasoning and Uncertainty (CSC412), Deep Learning (CSC413), Stochastic Processes (STA447), Computer Vision (CSC420)  
Natural/Social Sciences (2017-2019): Evolutionary/Molecular Genetics (BIO120/130), Physical/Organic Chemistry (CHM135/135), Calculus (MAT135/136/235), Political Sciences (MUN101), Global Affairs (MUN102)

## TEACHING

---

**CSC258: Intro. to Computer Systems**, University of Toronto Fall 2020  
Teaching Assistant with Prof. Steve Engels. Head of content development (labs/assignments). Ran office hours.

## AWARDS

---

**Research Fellowship**, Constellation / UC Berkeley 2024  
Awarded to researchers working on AI safety and alignment in collaboration with various organizations focused on risk mitigation, capabilities research, and evaluations.

**Finalist Award, Outstanding Undergraduate Researcher**, Computing Research Association (CRA) 2022  
Awarded to top undergraduate computer science researchers in North America. Finalist awarded to Top 20 overall.

**Scholar Award**, Neural Information Processing Systems (NeurIPS) 2022  
Awarded to fund in-person conference attendance for select first-author student presenters.

**Cloud TPU Research Award**, Google Research 2022  
Awarded to fund independent researchers in AI with access to Google's Cloud TPU compute resources.

**Undergraduate Student Research Award**, NSERC [*declined*] 2020  
Awarded to fund a summer research internship in Canada. Declined due to dual employment in industry internship.

**Dean's List Scholar**, University of Toronto 2018, 2019, 2021  
Awarded on the basis of grade point average (cGPA).

**Trinity College Academic Scholarship**, University of Toronto 2019  
Awarded on the basis of academic standing.