Jessica Tang

jessicao.tang@mail.utoronto.ca | jessicaxtang.github.io

Research areas: Interpretability, AI Safety, Large Language Models (LLMs), Natural Language Processing (NLP)

EDUCATION

University of Toronto, Toronto, Ontario **Sep 2022 – May 2026**

BASc in Engineering Science (Machine Intelligence)

- *Relevant Courses:* Machine Learning, Matrix Algebra & Optimization, Probability and Statistics, Vector Calculus, Computational Linguistics, Foundations of Computing, Control Theory

RESEARCH EXPERIENCE

Vector Institute for Artificial Intelligence, Toronto, Ontario **Sept 2025 – Present**

Senior Research Thesis

- *Supervisors:* Prof. Sheila McIlraith and Dr. Silviu Pitis
- *Paper:* **Editing Prompts to Optimize a Margin of Safety in Large Language Models.**
- Developed an attribution-driven method that defines a directional "margin of safety" between aligned and violating LLM behaviours and uses span-level prompt edits to reliably increase this safety margin without retraining the model.

Microsoft Research, Redmond, Washington **May 2025 – Present**

Undergraduate Research Intern (5% acceptance rate)

- *Mentors:* Dr. Sharad Agrawal and Dr. Shraddha Barke
- *Project*: **Tokengeist: Tracing Influential Tokens in Agentic LLM Systems.**
- Designing an attention-based tracing method for multi-turn agentic LLM systems that links influential inputs and tool fields to downstream decisions, with the goal of improving debugging, accountability, and provenance in LLM workflows.

KITE Research Institute, Toronto, Ontario **Apr 2024 – May 2025**

Machine Learning Research Intern

- *Supervisors:* Prof. Shehroz Khan and Dr. Ali Abedi
- *Project:* **Rehabilitation Exercise Quality Assessment and Feedback Generation Using Large Language Models.**
- Built an end-to-end LLM feedback system for rehabilitation exercise quality assessment, combining ST-GCN and VQ-VAE for spatiotemporal skeleton classification with explainability methods (Grad-CAM) to localize motion errors and support human clinicians' decisions.

Cognitive Neuroscience and Sensorimotor Integration Lab, Toronto, Ontario **May 2023 – May 2025**

Lab Manager, Computational Neuroscience Research Assistant

- *Supervisor:* Prof. Matthias Niemeier
- *Project:* **From Tasks to Topology: Dorsal and Ventral Streams Emerge in Optimized Neural Networks.**
- Modeled task-driven emergence of dorsal and ventral visual streams in CNNs, achieving >80% accuracy in biologically inspired tasks. Ran explainability studies (Neuron Shapley, activation maximization) to map neural information flow and graph connectivity. Engineered EEG–motion capture synchronization and experimental control interfaces in Python to scale data collection.

PUBLICATIONS & POSTERS

- Tokengeist: Tracing Influential Tokens in Agentic LLM Systems.
  J. Tang, S. Barke, S. Agarwal. *In preparation.*

- Editing Prompts to Optimize a Margin of Safety in Large Language Models.
  J. Tang, S. Pitis, S. McIlraith. *AAAI Machine Ethics Workshop, 2026.*

- From Tasks to Topology: Dorsal and Ventral Streams Emerge in Optimized Neural Networks.

T. Reza, E. Jordan, S. Luo, K. Patel, **J. Tang**, M. Niemeier. *Under review in Nature Neuroscience.*

- Rehabilitation Exercise Quality Assessment and Feedback Generation Using Large Language Models.
  J. Tang, A. Abedi, T. S. Colella, S. Khan. *IJCAI ARIAL Workshop, 2025. Published in Springer Nature, 2025.*

- Comparative analysis of optimization trends in dorsal and ventral stream using computational model (*Poster)*
  T. Reza, S. Luo, G. Singh, **J. Tang**, R. Jain, M. Niemeier. *Cognitive Neuroscience Society Conference, 2024.*

- Modular Can-Sized Satellite System with Active Attitude Control (*Outstanding Paper Award)*
  T. Cai, K. Howard, B. Zhou, **J. Tang**, V. He, D. Liu. *International CanSat Competition, 2022.*

- Deep Reinforcement Learning Controller for Indoor Farming *(Best Presenter Award)*
  Jessica Tang. *International Student Conference on Artificial Intelligence, 2021.*

AWARDS & HONOURS

Dean's Honours List, *University of Toronto*	2022–2025
$30,000 Award for Diversity and Innovation in Technology, *Royal Bank of Canada*	2024, 2025
$6000 Research Award, *Transform HF*	2024
$400 Best Project Overall, *Cohere*	2023
$2000 Dean's Merit Award, *University of Toronto*	2022
$1250 BC Achievement Scholarship	2022
$1250 District/Authority Scholarship in Technical and Trades Training	2022
$1000 Ingenious+ National Finalist and Regional Innovation Winner	2022
SGD 1000 Best Presenter, *International Student Conference On Artificial Intelligence*	2021
$560 AI4Impact Grant Award, *AI4ALL*	2021

LEADERSHIP, TEACHING

President and Founder, *IlluminAI* **Jul 2020 – Present**

- Founded an organization increasing AI ethics education; leading a team of 30, reached 1,000+ participants across 11 countries.
- Directed and hosted 16 events, featuring the **UofT AI Ethics Hackathon**, engaging 15+ disciplines.
- Raised $3,000+ in funding from AI4ALL, UofT, NCWIT, SAP, and Perplexity AI, with guests from MIT, IBM, Zoom, and UofT.

Creator and Editor, *YouTube Channel* **Aug 2023 – Present**

- Built an audience of **3,500 subscribers** and **120,000 views,** combining songwriting and storytelling to increase visibility for engineering, research, and women in STEM.

Course Instructor, *Wave Learning Festival* **June 2021 – Jul 2021**

- Designed and taught 'Introduction to Artificial Intelligence' to 70+ students, covering neural networks and AI ethics.

AI Scholar, Curriculum Developer, *AI4ALL* **Jul 2020 – Apr 2021**

- Selected among 32 students nationwide; developed a CNN for facial expression classification.
- Created AI curriculum later adopted by Bronx School of Science, featured by NCWIT.