# Thomas Jiralerspong

📱 +1 514 625 9308 | @ thomasjiralerspong@gmail.com | LinkedIn | GitHub | Website | Google Scholar

**Citizenships:** Canadian and Italian (European Union)

## Education

**Université de Montréal/Mila**
*PhD – Computer Science – Co-supervised by* ***Prof. Yoshua Bengio*** *&* ***Prof. Doina Precup*** *Sep 2023 – April 2027 (Expected)*
- **Awards:** Vanier Scholarship (150 000$), FRQNT Scholarship (40 000$) (Rank #1), NSERC Scholarship (17 500$)

**Massachusetts Institute of Technology (MIT)**
*Brains, Minds & Machines Summer Course* *2024*

**McGill University**
*B.Sc. – Honours Computer Science –* ***GPA:4.00/4.00*** *– Supervised by* ***Prof. Blake Richards*** *Sep 2020 – May 2023*

## Industry Research Experience

**Anthropic** PyTorch, Python
*Research Fellow - San Francisco, United States* *Mar 2025 – Present*
- **Project:** Mechanistic interpretability of emergent misalignment models

**Occam AI** PyTorch, Python
*Research Scientist - New York City, United States* *Jun 2024 – Present*
- **Projects:** Optimization of interactions between network of LLM agents, SQL query generation using LLMs

**Waabi** PyTorch, Python
*Deep Learning Research Intern - Toronto, Canada* *Jun 2023 – Sep 2023*
- **Project:** Realistic and controllable traffic simulation using a transformer based variational autoencoder

**Vector Institute for A.I.** PyTorch, Python
*Machine Learning Research Intern – Toronto, Canada* *Sep 2022 – Dec 2022*
- **Project:** Model-based reinforcement learning for HVAC control

## Software Development Experience

**Amazon Web Services (AWS) – S3 Team** Python, JavaScript
*Software Development Engineer Intern – Vancouver, Canada* *May 2022 – Jul 2022*

**Expedia Group** JavaScript, TypeScript, React
*Software Development Engineer Intern – Montreal, Canada* *Jun 2019 – Aug 2019*

## Selected Publications

***A Complexity-based Theory of Compositionality*** (Accepted)
E. Elmoznino*, **T. Jiralerspong***, Y. Bengio, G. Lajoie. ***ICML 2025***

***Geometric Signatures of Compositionality Across a Language Model's Lifetime*** (Accepted)
J. Lee*, **T. Jiralerspong***, L. Yu, Y. Bengio, E. Cheng. ***ACL 2025***

***Efficient Causal Graph Discovery Using Large Language Models***
**T. Jiralerspong***, X. Chen*, Y. More, V. Shah, Y. Bengio ***ICLR Workshop 2024***

***Forecaster: Towards Temporally Abstract Tree-Search Planning from Pixels***
**T. Jiralerspong***, F. Kondrup*, D. Precup, K. Khetarpal. ***NeurIPS Workshop 2023***

***Towards Safe Mechanical Ventilation Treatment Using Deep Offline Reinforcement Learning***
F. Kondrup*, **T. Jiralerspong***, E. Lau*, N. de Lara, J. Shkrob, M.D. Tran, D. Precup, S. Basu. ***AAAI 2023***

*Equal Contribution

## Awards & Achievements

Chosen as one of the 200 most promising young researchers in math & CS by the **Heidelberg Laureate Forum** 2023
Winner of **Project X 2021** (25 000$) 2021