

THOMAS JIRALERSPONG

☎ +1 514 625 9308 | @ thomasjiralerspong@gmail.com | [LinkedIn](#) | [GitHub](#) | [Website](#) | [Google Scholar](#) | [Games](#)

Citizenships: Canadian and Italian (European Union)

EDUCATION

Université de Montréal/Mila

PhD – Computer Science – Supervised by [Yoshua Bengio](#), [Guillaume Lajoie](#) & [Doina Precup](#) 2023 – Present

- **Awards:** Vanier Scholarship (150 000\$), FRQNT Scholarship (40 000\$) (Rank #1), NSERC Scholarship (17 500\$)

EXPERIENCE

Astra Fellowship - Mentored by [Dan Mossing](#) (Anthropic)

Research Fellow

PyTorch, Python

Apr 2026 – Aug 2026

- **Project:** Characterizing the mapping from context to answer activations in LLMs

Astra Fellowship - Google DeepMind Stream

Research Fellow

PyTorch, Python

Jan 2026 – Mar 2026

- **Project:** Benchmark for control intervention awareness in LLMs

Anthropic - Mentored by [Trenton Bricken](#)

Research Fellow

PyTorch, Python

May 2025 – Oct 2025

- **Project:** Mechanistic interpretability to discover behavioral differences between models
- Published as [official Anthropic research release](#) (500K+ views)

Occam AI

Research Scientist - New York City, United States

PyTorch, Python

Jun 2024 – Apr 2025

- **Projects:** Optimization of interactions between networks of LLM agents, SQL query generation using LLMs

MENTORSHIP

SPAR (Supervised Program for Alignment Research) – Research Mentor, 3 AI safety projects

Aug 2026 – Present

AlgoVerse AI Research – Principal Investigator, overseeing 20 AI safety research teams

Jun 2026 – Present

MARS V - Meridian Cambridge – Research Mentor, 7 AI safety researchers (incl. PhD/MSc)

Jun 2026 – Present

Mila/McGill University – Research Mentor, 2 PhD students, 1 research intern & 1 Master's student

2023 – Present

SELECTED PUBLICATIONS

[When Does Interleaving Prevent Emergent Misalignment?](#)

S. Mammeri, R. N. Kwon, C. Sun, S. Sahoo, C. Gagné, [T. Jiralerspong](#)

Mech Interp Workshop at ICML 2026

[Hidden-State Similarity Predicts Re-Elicitation After Inoculation Prompting](#)

R. N. Kwon, S. Mammeri, S. Sahoo, C. Gagné, [T. Jiralerspong](#)

Mech Interp Workshop at ICML 2026

[Measuring Control Intervention Awareness Across Frontier LLMs](#)

J. Schaeffer* [T. Jiralerspong*](#), A. Panfilov, R. Zimmermann

Best Paper Award, Oral Presentation

CAO Workshop at ICLR 2026

[Noticing the Watcher: LLM Agents Can Infer CoT Monitoring from Blocking Feedback](#)

[T. Jiralerspong*](#), F. Kondrup, Y. Bengio

AIWILD Workshop at ICLR 2026

[Cross-Architecture Model Diffing With Crosscoders: Unsupervised Discovery of Differences Between LLMs](#)

[T. Jiralerspong](#), T. Bricken

Official Anthropic Release

[Learning What Matters: Steering Diffusion via Spectrally Anisotropic Forward Noise](#)

L. Scimeca*, [T. Jiralerspong*](#), B. Earnshaw, J. Hartford, Y. Bengio.

ICML 2026 (Submitted)

[A Complexity-based Theory of Compositionality](#)

E. Elmoznino*, [T. Jiralerspong*](#), Y. Bengio, G. Lajoie.

ICML 2025

[Geometric Signatures of Compositionality Across a Language Model's Lifetime](#)

J. Lee*, [T. Jiralerspong*](#), L. Yu, Y. Bengio, E. Cheng.

Oral Presentation, Top 1% of papers

ACL 2025

*Equal Contribution

AWARDS & ACHIEVEMENTS

Manifund research grants – US\$473,332 for [mentored AI safety](#) and [alignment pretraining](#) research.

2026

Y Combinator (W26) – Selected from 20,000+ applicants (<1% acceptance rate); declined to focus on PhD.

2025

Chosen as one of the 200 most promising young researchers in math & CS by the [Heidelberg Laureate Forum](#)

2023

Winner of **Project X 2021** (ML Research Competition - 25 000\$)

2021