

Andrew Lee

Contact Information	E-mail: andrewlee@g.harvard.edu Website: https://ajyl.github.io Google Scholar: Link	
Current Position	Post-doctoral Fellow - Harvard University Hosts: Martin Wattenberg, Fernanda Viegas	September 2024 - Present
Research	Interpretability, Feature Geometry, Representations, Alignment, Language Models	
Education	University of Michigan Ph.D. in Computer Science - Advisor: Rada Mihalcea	2020 - 2024
	University of Michigan Master's in Computer Science	2015
	Northwestern University Bachelor of Science in Computer Science	2013
Fellowships	OpenAI Superalignment Grant Awarded to 50 out of 2,700 applicants.	2024
Awards	Outstanding Paper Award COLM (4 / 418 accepted papers) Spotlight Presentation Mechanistic Interpretability Workshop Best Poster Award Pre-ACL Workshop	2025
	Spotlight Presentation NeurIPS (Top 2% of submissions) Oral Presentation ICML (Top 1.5% of submissions) Honorable Mention - Best Paper BlackboxNLP Workshop	2024
	University of Michigan CSE Diversity, Equity, and Inclusion Service Award University of Michigan CSE Research Honors AI Lab Nominee	2023
Publications: Conference Proceedings (*: Equal Contribution)	[15] Aaron Mueller, Andrew Lee , Shruti Joshi, Ekdeep Singh Lubana, Dhanya Sridhar, Patrik Reizinger. From Isolation to Entanglement: When Do Interpretability Methods Identify and Disentangle Known Concepts? <i>ACL Main</i> 2026	
	[14] Thomas Fel, Binxu Wang, Michael A Lepori, Matthew Kowal, Andrew Lee , Randall Balestriero, Sonia Joseph, Ekdeep S Lubana, Talia Konkle, Demba Ba, Martin Wattenberg. Into the Rabbit Hull: From Task-Relevant Concepts in DINO to Minkowski Geometry. <i>ICLR</i> . 2025	
	[13] Yushi Yang, Filip Sondej, Harry Mayne, Andrew Lee , Adam Mahdi. How Does DPO Reduce Toxicity? A Mechanistic Neuron-Level Analysis. <i>Empirical Methods in Natural Language Processing (EMNLP)</i> . 2025	
	[12] Andrew Lee , Melanie Weber, Fernanda Viegas, Martin Wattenberg. Shared global and local geometry of language model embeddings. <i>Conference on Language Modeling (COLM)</i> , 2025. Outstanding paper award (4 / 418 accepted papers) .	
	[11] Siyang Liu, Bianca Brie, Wenda Li, Laura Biester, Andrew Lee , James Pennebaker, Rada Mihalcea. Eeyore: Realistic Depression Simulation via Supervised and Preference Optimization. <i>Findings of Association for Computational Linguistics (ACL)</i> , 2025.	

[10] Core Francisco Park*, **Andrew Lee***, Ekdeep Singh Lubana*, Yongyi Yang*, Maya Okawa, Kento Nishi, Martin Wattenberg, Hidenori Tanaka. ICLR: In-Context Learning of Representations. *International Conference on Learning Representations (ICLR)*, 2025.

[9] Core Francisco Park*, Maya Okawa*, **Andrew Lee**, Ekdeep Singh Lubana*, and Hidenori Tanaka*. Emergence of Hidden Capabilities: Exploring Learning Dynamics in Concept Space. *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Spotlight presentation (Top 2% of submissions).

[8] **Andrew Lee**, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K. Kummerfeld, Rada Mihalcea. A Mechanistic Understanding of Alignment Algorithms: A Case Study on DPO and Toxicity. *The International Conference on Machine Learning (ICML)*, 2024.

Oral presentation (Top 1.5% of submissions).

[7] **Andrew Lee**, Jonathan K. Kummerfeld, Larry An, Rada Mihalcea. A Comparative Multidimensional Analysis of Empathetic Systems. *European Chapter of the Association for Computational Linguistics (EACL)*, 2024.

[6] Oana Ignat, Zhijing Jin, Artem Abzaliev, Laura Biester, Santiago Castro, Naihao Deng, Xinyi Gao, Aylin Gunal, Jacky He, Ashkan Kazemi, Muhammad Khalifa, Namho Koh, **Andrew Lee**, Siyang Liu, Do June Min, Shinka Mori, Joan Nwatu, Veronica Perez-Rosas, Siqi Shen, Zekun Wang, Winston Wu, Rada Mihalcea. Has It All Been Solved? Open NLP Research Questions Not Solved by Large Language Models. In *The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 2024.

[5] Shinka Mori, Oana Ignat, **Andrew Lee**, Rada Mihalcea. Towards Algorithmic Fidelity: Mental Health Representation across Demographics in Synthetic vs. Human-generated Data. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 2024.

[4] **Andrew Lee**, Jonathan K. Kummerfeld, Larry An, Rada Mihalcea. Empathy Identification Systems are not Accurately Accounting for Context. *European Chapter of the Association for Computational Linguistics (EACL)*, 2023.

[3] **Andrew Lee**, Jonathan K. Kummerfeld, Larry An, Rada Mihalcea. Micromodels for Efficient, Explainable, and Reusable Systems: A Case Study on Mental Health. *Findings of Empirical Methods in Natural Language Processing (Findings of EMNLP)*, 2021.

[2] Stefan Larson, Anish Mahendran, Joseph J. Peper, Chris Clarke, **Andrew Lee**, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael Laurenzano, Lingjia Tang and Jason Mars. An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction. *Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[1] Stefan Larson, Anish Mahendran, **Andrew Lee**, Jonathan K. Kummerfeld, Parker Hill, Michael Laurenzano, Johann Hauswald, Lingjia Tang, Jason Mars. Outlier Detection for Improved Data Quality and Diversity in Dialog Systems. *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

**Publications:
Workshops**

[5] Prakhar Gupta, Henry Conklin, Sarah-Jane Leslie, **Andrew Lee**. Better World Models Can Lead to Better Post-Training Performance. *Mechanistic Interpretability Workshop @ NeurIPS*. 2025. **Spotlight presentation**

[4] Core Francisco Park*, **Andrew Lee***, Ekdeep Singh Lubana, Kento Nishi, Maya Okawa, Hidenori Tanaka. Structured In-Context Task Representations. *Workshop on Symmetry and Geometry in Neural Representations @ NeurIPS*. 2024

[3] Core Francisco Park, Maya Okawa, **Andrew Lee**, Ekdeep Singh Lubana, Hidenori Tanaka. Hidden Learning Dynamics of Capability before Behavior in Diffusion Models. *High-dimensional Learning Dynamics 2024 @ ICML*. 2024.

[2] Neel Nanda*, **Andrew Lee***, Martin Wattenberg. Emergent Linear Representations in World Models of Self-Supervised Sequence Models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP)*, 2023. **Honorable Mention for Best-Paper.**

[1] **Andrew Lee**, Zhenguo Chen, Kevin Leach, Jonathan K. Kummerfeld. Augmenting Task-Oriented Dialogue Systems with Relation Extraction. In *Proceedings of the 10th Dialog System Technology Challenges (AAAI Workshop)*, 2022.

**Publications:
Preprints**

[7] **Andrew Lee**, Yonatan Belinkov, Fernanda Viegas, Martin Wattenberg. Decomposing Query-Key Feature Interactions Using Contrastive Covariances. 2026

[6] Lihao Sun, Lewen Yan, Xiaoya Lu, **Andrew Lee**, Jie Zhang, Jing Shao. Valence–Arousal Subspace in LLMs: Circular Emotion Geometry and Multi-Behavioral Control. 2026

[5] Yushi Yang, Shreyansh Padarha, **Andrew Lee**, Adam Mahdi. Agentic Reinforcement Learning for Search is Unsafe. *Preprint*. 2025

[4] Xiaoyan Bai, Itamar Pres, Yuntian Deng, Chenhao Tan, Stuart Shieber, Fernanda Viegas, Martin Wattenberg, **Andrew Lee**. Why Can't Transformers Learn Multiplication? Reverse-Engineering Reveals Long-Range Dependency Pitfalls. *Preprint*. 2025

[3] **Andrew Lee**, Lihao Sun, Chris Wendler, Fernanda Viegas, Martin Wattenberg. The Geometry of Self-Verification in a Task-Specific Reasoning Model. *Preprint*. 2025.

[2] Jing Xu, **Andrew Lee**, Sainbayar Sukhbaatar, Jason Weston. Some things are more CRINGE than others: Preference Optimization with the Pairwise Cringe Loss. *Preprint*, 2023.

[1] **Andrew Lee**, David Wu, Emily Dinan, Michael Lewis. Improving Chess Commentaries by Combining Language Models with Symbolic Reasoning Engines. *Preprint*, 2022.

**Past
Employments**

Meta AI Research New York, NY
Research Intern - RAM (Reasoning, Attention, Memory) Team May 2023 - October 2023
Advisors: Jing Xu, Sainbayar Sukhbaatar, Jason Weston

Meta AI Research New York, NY
Research Intern - Diplomacy Team May 2022 - December 2022
Advisors: Emily Dinan, Mike Lewis

Microsoft Research Redmond, WA
Research Intern - KTX Team May 2021 - August 2021
Advisor: Silviu-Petru Cucerzan

Clinic, Inc. Ann Arbor, MI
Core AI R&D - Senior Software Engineer, Team Lead June 2019 - August 2020
Core AI R&D - Software Engineer June 2017 - June 2019

Ford Motor Company Dearborn, MI
Software Engineer March 2016 - June 2017

- Invited Talks**
- Decomposing Query-Key Feature Interactions Using Contrastive Covariances** 2026
PRISM Journal Club (Harvard)
Boston University
ML Collective Deep Learning: Classics and Trends
 - Understanding Representations to Understand Phenomena** 2025
Yonsei University
 - Shared Global and Local Geometry of Language Model Embeddings** 2025
PRISM Journal Club (Harvard)
University of Michigan NLP Reading Group
Google DeepMind
New England Mechanistic Interpretability Workshop
 - Reverse-Engineering Language Models to Understand Alignment, Reasoning** 2025
MIT Language & Intelligence Lab
Northeastern University
Princeton University
Oxford University
University of Chicago
Microsoft Research
 - The Geometry of Self-Verification in a Task-Specific Reasoning Model** 2025
ML Collective Deep Learning: Classics and Trends
Eleuther AI Reading Group
 - A Mechanistic Understanding of Alignment Algorithms** 2024
University of Texas - Austin: Social Applications and Impact of NLP
University of Cambridge

Patents	<p>Systems and methods for slot relation extraction for machine learning task-oriented dialogue systems. Andrew Lee, Zhenguo Chen, Jonathan K. Kummerfeld. <i>US Patent 11,734,519. 2023.</i></p> <p>Systems and methods for constructing an artificially diverse corpus of training data samples for training a contextually-biased model for a machine learning-based dialogue system. Andrew Lee, Stefan Larson, Chris Clarke, Kevin Leach, Jonathan K. Kummerfeld, Parker Hill, Johann Hauswald, Michael Laurenzano, Lingjia Tang, Jason Mars. <i>US Patent 10,796,104. 2020.</i></p> <p>Systems and methods for automatically configuring training data for training machine learning models of a machine learning-based dialogue system including seeding training samples or curating a corpus of training data based on instances of training data identified as anomalous. Stefan Larson, Anish Mahendran, Andrew Lee, Jonathan K. Kummerfeld, Parker Hill, Michael Laurenzano, Johann Hauswald, Lingjia Tang, Jason Mars. <i>US Patent 10,679,150. 2020.</i></p>
Teaching	<p>Information Retrieval and Web Search - University of Michigan 2022 Teaching Assistant</p> <p>Introduction to Computer Security - University of Michigan 2015 Teaching Assistant</p>
Professional Services	<p>Mechanistic Interpretability Workshop (@ ICML) 2026 - Program Chair</p> <p>Mechanistic Interpretability Workshop (@NeurIPS) 2025 - Program Chair</p> <p>CLPsych Workshop on Computational Linguistics and Clinical Psychology (@NAACL) 2022 - Organizing Committee</p> <p>CBO International Symposium on Code Generation and Optimization Program Committee 2019</p>