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 — September 2024 - Present
Hosts: Martin Wattenberg, Fernanda Viegas

Research

Interpretability, Representations, Alignment, Language Models, Natural Language Processing

Education

University of Michigan — 2020 - 2024
Ph.D. in Computer Science - Advisor: Rada Mihalcea

University of Michigan — 2015
Master's in Computer Science

Northwestern University — 2013
Bachelor of Science in Computer Science

Fellowships

OpenAI Superalignment Grant — 2024
Awarded to 50 out of 2,700 applicants.

Publications
(*: Equal Contribution)

[14] 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.

[13] 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).

[12] **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).

[11] **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.

[10] 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.

[9] 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.

[8] 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.**

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

[6] **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.

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

[4] **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.

[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.

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

Clinc, 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

Oxford University 2025
Understanding Language Models by Reverse-Engineering Toy Models

University of Chicago 2025
Understanding Language Models by Reverse-Engineering Toy Models

Microsoft Research 2025
Reverse-Engineering Language Models to Explain Their Behavior

University of Texas - Austin: Social Applications and Impact of NLP 2024
A Mechanistic Understanding of Alignment Algorithms

University of Cambridge 2024
A Mechanistic Understanding of Alignment Algorithms

Awards

University of Michigan CSE Research Honors AI Lab Nominee 2023

Patents

Systems and methods for slot relation extraction for machine learning task-oriented dialogue systems.
Andrew Lee, Zhenguo Chen, Jonathan K. Kummerfeld.
US Patent 11,734,519. 2023.

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.
US Patent 10,796,104. 2020.

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.
US Patent 10,679,150. 2020.

Teaching

Information Retrieval and Web Search - University of Michigan 2022
Teaching Assistant

Introduction to Computer Security - University of Michigan 2015
Teaching Assistant

Professional Services

ACL Area Chair 2025
CLPsych - Workshop on Computational Linguistics and Clinical Psychology (@NAACL)
2022 - Organizing Committee
2021 - Volunteer
CBO International Symposium on Code Generation and Optimization Program Committee 2019