LEI (JADE) YU

+1(437)297-4583 ◇ jadeleiyu@cs.toronto.edu ◇ https://jadeleiyu.github.io/

RESEARCH INTEREST

- **Safety and trustworthiness** of generative AI
- **Reasoning and planning** through **Large Language Models (LLM)**
- Reinforcement learning with human feedback (**RLHF**)
- Mechanistic interpretability and explainable AI

EDUCATION

University of Toronto, Toronto, Canada *January 2021 - Present*
Ph.D. in Computer Science (**expected graduation: January 2025**) GPA: 3.90/4.00
Research area: Natural language processing
Supervisor: Yang Xu

McGill University, Montreal, Canada *September 2016 - May 2019*
B.Sc. with Joint Major in Computer Science and Statistics GPA: 3.88/4.00
Graduated with 1st-class distinction.

PAPERS AND PUBLICATIONS

Lei Yu, Virginie Do, Karen Hambardzumyan, Nicola Cancedda. (2024) Robust LLM safeguarding via refusal adversarial training. https://arxiv.org/abs/2409.20089

Lei Yu, Meng Cao, Jackie CK Cheung, Yue Dong. (2024) Mechanistic Understanding and Mitigation of Language Model Non-Factual Hallucinations. In *Findings of EMNLP 2024*.

Meng Cao, Lei Shu, **Lei Yu**, Yun Zhu, Nevan Wichers, Yinxiao Liu, Lei Meng. (2024) Beyond Sparse Rewards: Enhancing Reinforcement Learning with Language Model Critique in Text Generation. In *EMNLP 2024*.

Lei Yu, Jingcheng Niu, Zining Zhu, Gerald Penn. (2024) Functional Faithfulness in the Wild: Circuit Discovery with Differentiable Computation Graph Pruning. https://arxiv.org/html/2407.03779v1.

Yihuai Hong, **Lei Yu**, Shauli Ravfogel, Haiqin Yang, Mor Geva. (2024) Intrinsic Evaluation of Unlearning Using Parametric Knowledge Traces. https://arxiv.org/pdf/2406.11614.

Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, **Lei Yu**, Alessandro Laio, Marco Baroni. (2024) Emergence of a High-Dimensional Abstraction Phase in Language Transformers. https://arxiv.org/pdf/2406.11614.

Meiling Tao, Liang Xuechen, Tianyu Shi, **Lei Yu**, Yiting Xie. (2024) RoleCraft-GLM: Advancing Personalized Role-Playing in Large Language Models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)* https://aclanthology.org/2024.personalize-1.1/.

Lei Yu. (2023) Systematic word meta-sense extension. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*. **Oral presentation.**

Lei Yu, Yang Xu. (2023) Word sense extension. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023).*

Lei Yu, Yang Xu. (2022) Infinite mixture chaining: Efficient temporal construction of word meaning. In *Proceedings of the 44th Annual Meeting of the Cognitive Science Society.* **Oral presentation.**

Lei Yu, Yang Xu. (2022) Probabilistic frame semantics for word class conversion. In *Computational Linguistics, Volume 48, Number 4*

Lei Yu, Yang Xu. (2021) Predicting emergent linguistic compositions through time: Syntactic frame extension via multimodal chaining. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021).* **Oral presentation.**

Lei Yu*, Chelsea Tanchip*, Aotao Xu, and Yang Xu. (2020) Inferring symmetry in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020.*

RESEARCH EXPERIENCE

Fundamental AI Research (FAIR), Meta June 2024 - November 2024
Research internship *Host: Dr. Nicola Cancedda*

· Improving LLM safety via mechanistic adversarial training
· Mitigating LLM hallucinations via representation engineering

Google Research September 2023 - January 2024
Student research collaborator *Project leader: Dr. Lei Shu*

· Intrinsic reward in reinforcement learning with human feedback (RLHF) via self-critique.

Tel Aviv University January 2024 - Present
Research assistant *Principal investigator: Prof. Mor Geva*

· Interpretability and mitigation of LLM hallucinations with irrelevant context.

NLP Lab, University of California, Riverside August 2023 - Present
Research collaborator *Principal investigator: Prof. Yue Dong*

· Mechanistic understanding of language model adversarial attacks.

Computational Linguistics Group, Universitat Pompeu Fabra September 2023 - Present
Research collaborator *Principal investigator: Prof. Marco Baroni*

· Intrinsic dimensionality and linguistic information in transformer language models.

Mila - Montreal Institute for Learning Algorithms May 2018 - December 2018
Undergraduate Research Assistant *Supervisor: Prof. Jackie Cheung*

· Automated abstractive text summarization through domain adaptation.

KEY SKILLS

Natural Language Processing	Text Generation, Language Model Pretraining, Efficient LLM Fine-tuning, Mechanistic Interpretability, Reinforcement Learning with Human Feedback,
Programming	Python (PyTorch, TensorFlow, JAX), Java, R, Matlab
Machine Learning	Deep Learning, Bayesian Modeling, Reinforcement Learning, Gradient-Based Meta-Learning
Mathematics	Probability Theories, Statistical Learning Theories, Differential Equations, Convex Optimization
Natural Languages	Mandarin Chinese (Native), English (Proficient), French (Intermediate)