LEI (JADE) YU

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

RESEARCH INTEREST

- Natural language processing
- Interpretability and safety of large language models
- Aligning generative AI with human values

EDUCATION

University of Toronto, Toronto, Canada *September 2019 - Present*
Ph.D. in Computer Science (**expected graduation date: December 2024**) 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.

PUBLICATIONS

Lei Yu, Jingcheng Niu, Zining Zhu. (2024) Functionally faithful neural circuit discovery via differentiable masking. Submitted to ICML 2024.

Lei Yu, Meng Cao, Jackie CK Cheung, Yue Dong. (2024) Mechanisms of non-factual hallucinations in language models. https://arxiv.org/pdf/2403.18167.

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. https://arxiv.org/abs/2401.07382.

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, Meta London March 2024 -
Research scientist intern, Large Language Models *Host: Dr. Nicola Cancedda*

· Interpretable LLM compression via spectral filtering

Google Research, Mountain View 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*

· Interpretation 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)