JADE (LEI) YU

+(1)437-297-4583 ◇ jadeleiyu@meta.com ◇ https://jadeleiyu.github.io/

RESEARCH INTEREST

Large Language Models (LLM), AI Agents, AI Safety and Interpretability

WORK EXPERIENCE

Fundamental AI Research (FAIR), Meta *2025.05 - Present*
AI Research Scientist, AI Agent and World Modeling

Google Research *2024.12 - 2025.02*
Research internship, LLM Agents

Fundamental AI Research (FAIR), Meta *2024.06 - 2024.11*
Research internship, LLM Alignment

EDUCATION

University of Toronto, Toronto, Canada *2021.01 - 2025.01*
Ph.D. in Computer Science, Natural Language Processing

University of Toronto, Toronto, Canada *2019.09 - 2021.01*
M.Sc. in Computer Science

McGill University, Montreal, Canada *2016.09 - 2019.05*
B.Sc. in Computer Science and Statistics

PAPERS AND PUBLICATIONS

Lei Yu, Virginie Do, Karen Hambardzumyan, Nicola Cancedda. (2024) Robust LLM safeguarding via refusal adversarial training. (To appear) In ICLR 2025. https://arxiv.org/abs/2409.20089

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. (To appear) In ICLR 2025. https://arxiv.org/pdf/2406.11614.

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.

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.*

SERVICES

Reviewing

- 2025: NAACL, ACL (Area Chair)
- 2024: ICLR, ACL, EMNLP (Area Chair)
- 2023: ACL, EMNLP, NAACL
- 2022: CogSci, EMNLP

KEY SKILLS

Generative AI	Text Generation, LLM Pretraining, 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, Information Theory, Optimization
Natural Language	Mandarin Chinese (Native), Canadian English (Proficient), Metropolitan French (Intermediate)