

JADE (LEI) YU

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

RESEARCH INTEREST

Large Language Model (LLM), AI Agent, Reinforcement Learning, AI Safety and Interpretability

WORK EXPERIENCE

Meta Superintelligence Labs AI Research Scientist, Multimodal Agent and LLM Reasoning	<i>2025.05 - Present</i>
Google Research Research internship, LLM Agent	<i>2024.12 - 2025.02</i>
Meta Fundamental AI Research (FAIR) Research internship, LLM Alignment	<i>2024.06 - 2024.11</i>

EDUCATION

University of Toronto, Toronto, Canada Ph.D. in Computer Science Research area: Natural Language Processing and Computational Cognitive Science Supervisor: Yang Xu	<i>2021.01 - 2025.01</i>
University of Toronto, Toronto, Canada M.Sc. in Computer Science	<i>2019.09 - 2021.01</i>
McGill University, Montreal, Canada B.Sc. in Computer Science and Statistics	<i>2016.09 - 2019.05</i>

PAPERS AND PUBLICATIONS

Shayegani, Erfan, G. M. Shahariar, Sara Abdali, **Lei Yu**, Nael Abu-Ghazaleh, and Yue Dong. Misaligned Roles, Misplaced Images: Structural Input Perturbations Expose Multimodal Alignment Blind Spots. arXiv preprint arXiv:2504.03735 (2025).

Ji, Ziwei, **Lei Yu**, Yeskendir Koishkenov, Yejin Bang, Anthony Hartshorn, Alan Schelten, Cheng Zhang, Pascale Fung, and Nicola Cancedda. Calibrating Verbal Uncertainty as a Linear Feature to Reduce Hallucinations. arXiv preprint arXiv:2503.14477 (2025).

Hong, Yihuai, Dian Zhou, Meng Cao, **Lei Yu**, and Zhijing Jin. The reasoning-memorization interplay in language models is mediated by a single direction. In ACL 2025. arXiv:2503.23084.

Lee, Jin Hwa, Thomas Jiralerspong, **Lei Yu**, Yoshua Bengio, and Emily Cheng. Geometric Signatures of Compositionality Across a Language Model’s Lifetime. In ACL 2025. **SAC Highlight Award (Top 2.5%)**. arXiv:2410.01444.

Lei Yu, Virginie Do, Karen Hambardzumyan, Nicola Cancedda. (2024) Robust LLM safeguarding via refusal adversarial training. 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. 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

Area Chair

- 2025: ACL, EMNLP, NAACL
- 2024: ACL, EMNLP

Reviewer

- 2025: ACL, EMNLP, Neurips, ICML, NAACL
- 2024: ICLR, ACL, EMNLP
- 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)