

# JADE (LEI) YU

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

## RESEARCH INTEREST

---

Large Language Models (LLMs), Reasoning Agents, Deep Reinforcement Learning,  
Post-training Alignment, World Models, Multimodal Reasoning

## WORK EXPERIENCE

---

<b>AI Research Scientist, Meta Superintelligence Labs</b> Reinforcement Learning and World Modeling for Reasoning Agents	<i>2025.05 - Present</i>
<b>Research internship, Google Research</b> Memory-Augmented LLM Agents	<i>2024.12 - 2025.02</i>
<b>Research internship, Meta Fundamental AI Research (FAIR)</b> LLM Post-training Alignment and Interpretability	<i>2024.06 - 2024.11</i>

## EDUCATION

---

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

## KEY SKILLS

---

<b>Large Language Model</b>	Distributed training (RL, SFT) Agentic systems (reasoning, web search, coding) Multimodal LLM planning and reasoning
<b>Programming</b>	Python, PyTorch, Slurm, Ray, Verl
<b>Machine Learning</b>	Deep Learning, Bayesian Modeling, Reinforcement Learning
<b>Mathematics</b>	Probability Theories, Statistical Learning Theories, Information Theory, Optimization

## SELECTED WORK

---

**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>

**Lei Yu\***, Ziwei Ji\*, Yeskendir Koishekenov, Yejin Bang, Anthony Hartshorn, Alan Schelten, Cheng Zhang, Pascale Fung, and Nicola Cancedda. Calibrating Verbal Uncertainty as a Linear Feature to Reduce Hallucinations. (To appear) in EMNLP 2025. <https://arxiv.org/abs/2503.14477>.

Yihuai Hong, **Lei Yu**, Shauli Ravfogel, Haiqin Yang, Mor Geva. (2024) Intrinsic Evaluation of Unlearning Using Parametric Knowledge Traces. (To appear) in EMNLP 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*.

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.

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*.

Haoran Xu, Jiacong Hu, Zhang Ke, **Lei Yu**, Yuxin Tang, Xinyuan Song, Yiqun Duan, Lynn Ai, Tianyu Shi. SEDM: Scalable Self-Evolving Distributed Memory for Agents. In *Workshop on Scaling Environments for Agents at NeurIPS 2025*

## PUBLICATIONS

---

Haoran Xu, Jiacong Hu, Zhang Ke, **Lei Yu**, Yuxin Tang, Xinyuan Song, Yiqun Duan, Lynn Ai, Tianyu Shi. SEDM: Scalable Self-Evolving Distributed Memory for Agents. In *Workshop on Scaling Environments for Agents at NeurIPS 2025*

Wannan Yang, Xinchu Qiu, **Lei Yu**, Yuchen Zhang, Aobo Yang, Narine Kokhlikyan, Nicola Cancedda, Diego Garcia-Olano. Hallucination Reduction with CASAL: Contrastive Activation Steering for Amortized Learning. In *Mechanistic Interpretability Workshop at NeurIPS 2025*

Ji, Ziwei\*, **Lei Yu**\*, Yeskendir Koishakenov, Yejin Bang, Anthony Hartshorn, Alan Schelten, Cheng Zhang, Pascale Fung, and Nicola Cancedda. Calibrating Verbal Uncertainty as a Linear Feature to Reduce Hallucinations. (To appear) in EMNLP 2025. <https://arxiv.org/abs/2503.14477>.

**Lei Yu**\*, Jingcheng Niu\*, Zining Zhu, Xi Chen, Gerald Penn. (2024) Dynamic Granularity in the Wild: Differentiable Sheaf Discovery with Joint Computation Graph Pruning. (To appear) in EMNLP 2025. <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. (To appear) in EMNLP 2025. <https://arxiv.org/pdf/2406.11614>.

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).

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*.

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, ICLR, NAACL

- 2024: ICLR, ACL, EMNLP
- 2023: ACL, EMNLP, NAACL
- 2022: CogSci, EMNLP