Weichen Li

Ph.D candidate

Gottlieb-Daimler-Str.67663 Kaiserslautern

weichenli1223.github.io/weichenli/

weichen@cs.uni-kl.de

Key Words

NLP, Machine Learning, Reinforcement Learning

Languages

Chinese, , English, German

Programming Languages

Python, Java

Deep Learning Framework

Pytorch

About me

I am a final-year Ph.D. student in the Machine Learning group at the University of Kaiserslautern-Landau, supervised by Professor Sophie Fellenz. My focus is on Reinforcement Learning. My research during PhD is from domain-specific agents in text-based environments to general-purpose methods for preference- and value-aligned decision-making. I thrive on exploring new challenges.

Research Interests

- **Language-driven Reinforcement Learning**: Adapting stable RL algorithms for language-centric tasks. Our experiments with text-based games show that SAC can be effectively modified for text-based environments with minimal adjustments.
- **Ethical RL Agents**: Aligning RL agents with moral guidelines using human or LLM-labeled scores. Constrained RL ensures agents maximize rewards while adhering to ethical boundaries.
- **Human Preference Alignment in RL**: Balancing competing objectives like safety, efficiency, and cost. Our diffusion-based planning framework integrates human preferences at inference, enabling flexible trade-offs without retraining.

Education

2021 - Present	University of Kaiserslautern-Landau,Germany Ph.D candidate in Computer Science, Machine Learning Group
2018 - 2021	Ludwig Maximilian University of Munich, Germany Master degree in Computational Linguistics and Computer science
2015 - 2018	University of Bamberg, Germany Bachelor degree in Sociology and Computer Science

Publications

- Weichen Li, Waleed Mustafa, Puyu Wang, Marius Klof, and Sophie Fellenz. Inference-time preference-aligned diffusion planning for safe offline reinforcement learning. In *Proceedings of the Third Workshop on Hybrid Human-Machine Learning and Decision Making (HHMLDM) at **ECML-PKDD***, 2025a. (Oral Presentation)
- Weichen Li, Waleed Mustafa, Rati Devidze, Marius Kloft, and Sophie Fellenz. Inference-time value alignment in offline reinforcement learning: Leveraging llms for reward and ethical guidance. In *workshop on WORDPLAY: WHEN LANGUAGE MEETS GAME at Empirical Methods in Natural Language Processing (**EMNLP**)*, 2025b
- Weichen Li, Rati Devidze, Waleed Mustafa, and Sophie Fellenz. Ethics in action: training reinforcement learning agents for moral decision-making in text-based adventure games. In *International Conference on Artificial Intelligence and Statistics (**AISTATS**)*, pages 1954–1962. PMLR, 2024
- Weichen Li, Rati Devidze, and Sophie Fellenz. Learning to play text-based adventure games with maximum entropy reinforcement learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (**ECML-PKDD**)*, pages 39–54. Springer, 2023
- Weichen Li, Patrick Abels, Zahra Ahmadi, Sophie Burkhardt, Benjamin Schiller, Iryna Gurevych, and Stefan Kramer. Topic-guided knowledge graph construction for argument mining. In *2021 IEEE International Conference on Big Knowledge (**ICBK**)*, pages 315–322. IEEE, 2021