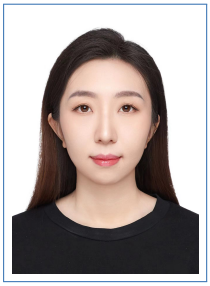
Weichen Li

Ph.D. Candidate in Machine Learning

Gottlieb-Daimler-Str. 67663
Kaiserslautern, Germany
weichen@cs.uni-kl.de
weichenli1223.github.io/weichenli/



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 research spans domain-specific agents in text-based environments to general-purpose methods for preference- and value-aligned decision-making.

Main Research Interests

- **Language-driven Reinforcement Learning**: Adapting stable RL algorithms for language-centric tasks. SAC can be effectively modified for text-based environments.
- **Ethical Reinforcement Learning Agents**: Aligning RL agents with moral guidelines using human or LLM-labeled scores. Constrained RL ensures adherence to ethical boundaries.
- **Human Preference Alignment in Reinforcement Learning**: Balancing safety, efficiency, and cost using diffusion-based planning, integrating human preferences at inference without retraining.

Education

2021–Present **Ph.D. Candidate in Computer Science**, *University of Kaiserslautern-Landau*, Germany, PhD thesis: Value-aligned Reinforcement Learning: From Language-based to Multi-objective Decision Making

2018–2021 **Master in Computational Linguistics and Computer Science**, *Ludwig Maximilian University of Munich (LMU)*, Germany

2015–2018 **Bachelor in Sociology and Computer Science**, *University of Bamberg*, Germany

Publications

- Weichen Li, Waleed Mustafa, Puyu Wang, Marius Klof, and Sophie Fellenz. Tora: Train once, realign anytime for offline multi-objective reinforcement learning. 2026. Under review at Association for the Advancement of Artificial Intelligence (**AAAI**) 2026
- 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
- Marcio Monteiro, Weichen Li, Puyu Wang, Marius Kloft, and Sophie Fellenz. Landmark-guided policy optimization for multi-objective language model selection. 2026. Under review at International Conference on Learning Representations (**ICLR**) 2026
- Waleed Mustafa, Naghmeh Ghanooni, Weichen Li, Andriy Balinskyy, Sophie Fellenz, and Marius Kloft. Non-vacuous generalization bounds for deterministic neural networks via parameter-space robustness. Under review at International Conference on Artificial Intelligence and Statistics (**AISTATS**) 2026

Student Research Supervision

I supervise Bachelor and Master theses in the following areas:

- Training LLMs for SMILES- and SMARTS-based molecular generation
- Reinforcement learning for thermodynamic group contribution methods
- RAG-based chatbot development for university library services