

YU FEI

yu.fei@uci.edu fywalter.github.io

RESEARCH INTERESTS

- Post-training and reinforcement learning for LLMs (policy optimization, preference learning, reward modeling)
- Decoding and sampling algorithms for diversity, controllability, and efficiency
- Reasoning and inference-time methods
- Alignment, evaluation, and interpretability of large language models

EDUCATION

University of California, Irvine

Ph.D. Student, Computer Science

09/2023 – Present

California, United States

- Advisor: Prof. Sameer Singh

ETH Zurich

M.Sc. in Computer Science, with distinction, GPA: 5.84/6, top 3%

09/2019 – 06/2022

Zurich, Switzerland

- Thesis advisor: Prof. Mrinmaya Sachan

Peking University

B.Sc. in Theoretical and Applied Mechanics, GPA: 3.71/4.0, top 10%

09/2015 – 07/2019

Beijing, China

B.Sc. in Computer Science, Double Degree

- Thesis advisor: Prof. Yizhou Wang

EMPLOYMENT

Amazon Rufus

Applied Scientist Intern

06/2025 – 09/2025

California, United States

- Developed a reflective prompt optimization framework that infers latent human preferences from limited data, enabling fine-grained instruction-following evaluation and targeted model customization

EPFL

Research Intern

07/2022 – 01/2023

Lausanne, Switzerland

- Advisor: Prof. Antoine Bosselut

- Conducted research on in-context learning and interpretability, leading to two publications at ACL 2023 and EMNLP 2023 ([3], [4]).

PUBLICATIONS & PREPRINTS

- [1] *Nudging: Inference-time Alignment of LLMs via Guided Decoding*
ACL 2025 Oral and panel (top 0.3%, 25/8360)
Yu Fei, Yasaman Razeghi, and Sameer Singh.

- [2] *MoCo: A One-Stop Shop for Model Collaboration Research*
Under Review at ICML 2026

Shangbin Feng, Yuyang Bai, Ziyuan Yang, Yike Wang, Zhaoxuan Tan, Jiajie Yan, Zhenyu Lei, Wenxuan Ding, Weijia Shi, Haojin Wang, Zhenting Qi, Yuru Jiang, Heng Wang, Chengsong Huang, Yu Fei, Jihan Yao, Yilun Du, Luke Zettlemoyer, Yejin Choi, Yulia Tsvetkov

- [3] *Towards a Mechanistic Interpretation of Multi-Step Reasoning Capabilities of Language Models*
EMNLP 2023
Yifan Hou, Jiaoda Li, **Yu Fei**, Alessandro Stolfo, Wangchunshu Zhou, Guangtao Zeng, Antoine Bosselut, Mrinmaya Sachan.
- [4] *Mitigating Label Biases for In-context Learning*
ACL 2023
Yu Fei, Yifan Hou, Zeming Chen, Antoine Bosselut.
- [5] *Beyond Prompting: Making Pre-trained Language Models Better Zero-shot Learners by Clustering Representations*
EMNLP 2022
Yu Fei, Zhao Meng, Ping Nie, Roger Wattenhofer, Mrinmaya Sachan.
- [6] *Align, Attend and Locate: Chest X-Ray Diagnosis via Contrast Induced Attention Network With Limited Supervision*
ICCV 2019
Jingyu Liu, Gangming Zhao, **Yu Fei**, Ming Zhang, Yizhou Wang, Yizhou Yu

SERVICES

Reviewer: Neurips 2024, ICLR 2025, AISTATS 2025 (**Outstanding Reviewer**), ARR 2025 Feb and May, Neurips 2025

HONORS & AWARDS

ICS Deans Fellowship , UC Irvine	2024, 2025
EDIC Fellowship , EPFL	2023
ETH Scholarship for International Students , ETH Zurich	2020, 2021
Second Prize of Zhou Peiyuan National Mechanics Competition	2017
Triple-A Outstanding Student with Xiexin Scholarship , Peking University	2016
Academic Innovation Award , Peking University	2016

SKILLS

Language: Native Mandarin, Fluent in English (TOEFL 113), Japanese (JLPT level N2-equivalent)

Programming: Python (PyTorch, HuggingFace Transformers, vLLM), and L^AT_EX