Yifei He | Curriculum Vitae

yifeihe3@illinois.edu • yifei-he.github.io

My research interest is centered around trustworthy machine learning. I currently work on: i) Improving the multi-task capabilities of foundation models (e.g, multilingual LLMs, model merging). ii) Enhancing the efficiency of LLMs (e.g., semi-supervised learning, MoE models). Previously, I have worked on multi-objective optimization, domain adaptation/generalization and multimodal learning.

Education

University of Illinois Urbana-Champaign (UIUC) — **Urbana, IL, USA**
Ph.D. in Computer Science — *May 2023 - Present*
M.S. in Computer Science — *Aug 2021 - May 2023*
Advisor: Prof. Han Zhao

University of Michigan (UM) — **Ann Arbor, MI, USA**
B.S.E. in Data Science, minor in Mathematics — *Aug 2019 - Apr 2021*
Summa Cum Laude

Shanghai Jiao Tong University (SJTU) — **Shanghai, China**
B.S.E. in Electrical and Computer Engineering — *Sept 2017 - Aug 2021*

Industry Experience

Microsoft — **Redmond, WA, USA**
Research Intern, GenAI — *Aug 2024 - present*
Manager: Dr. Yang Liu

- Worked on distillation of Mixture-of-Experts (MoE) models.

Microsoft — **Redmond, WA, USA**
Applied Scientist Intern, Turing — *May 2024 - Aug 2024*
Manager: Dr. Alon Benhaim

- Worked on scaling law for multilingual language models.

Amazon — **Seattle, WA, USA**
Applied Scientist Intern, Search Science and AI — *May 2023 - Aug 2023*
Manager: Dr. Alejandro Mottini

- Improved large-scale multi-task tuning of foundation models.
- Developed a vision-language retrieval foundation model with instruction tuning.

Publications (* denotes equal contribution)

[1] **Scaling Laws for Multilingual Language Models.**
Yifei He, Alon Benhaim, Barun Patra, Praneetha Vaddamanu, Sanchit Ahuja, Parul Chopra, Vishrav Chaudhary, Han Zhao, Xia Song.
Under review.

[2] **Towards Understanding the Fragility of Multilingual LLMs against Fine-Tuning Attacks.**
Samuele Poppi, Zheng-Xin Yong, **Yifei He**, Bobbie Chern, Han Zhao, Aobo Yang, Jianfeng Chi.
Under review.

[3] **Localize-and-Stitch: Efficient Model Merging via Sparse Task Arithmetic.**
Yifei He, Yuzheng Hu, Yong Lin, Tong Zhang, Han Zhao.
In *Transactions of Machine Learning Research*. **(TMLR)**

[4] **Semi-Supervised Reward Modeling via Iterative Self-Training.**
Yifei He*, Haoxiang Wang*, Ziyan Jiang, Alexandros Papangelis, Han Zhao.
In *Findings of the Association for Computational Linguistics: EMNLP 2024*. **(EMNLP 2024 Findings)**

[5] **Robust Multi-Task Learning with Excess Risks.**
Yifei He, Shiji Zhou, Guojun Zhang, Hyokun Yun, Yi Xu, Belinda Zeng, Trishul Chilimbi, Han Zhao.
In *Proceeding of the 41st International Conference on Machine Learning*. **(ICML 2024)**

[6] **Gradual Domain Adaptation: Theory and Algorithms.**
Yifei He*, Haoxiang Wang*, Bo Li, Han Zhao.
In *Journal of Machine Learning Research*. **(JMLR)**

[7] **Efficient Modality Selection in Multimodal Learning.**
Yifei He*, Runxiang Cheng*, Gargi Balasubramaniam*, Yao-Hung Hubert Tsai, Han Zhao.
In *Journal of Machine Learning Research*. **(JMLR)**
(Extended version of publication [8].)

[8] **Greedy Modality Selection via Approximate Submodular Maximization.**
Runxiang Cheng*, Gargi Balasubramaniam*, **Yifei He***, Yao-Hung Hubert Tsai, Han Zhao.
In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence*. **(UAI 2022)**

[9] **Conformer-RL: A Deep Reinforcement Learning Library for Conformer Generation.**
Runxuan Jiang, Tarun Gogineni, Joshua Kammeraad, **Yifei He**, Ambuj Tewari, Paul Zimmerman.
In *Journal of Computational Chemistry*. **(JCC)**

[10] **A Hierarchical Approach to Multi-Event Survival Analysis.**
Donna Tjandra, **Yifei He**, Jenna Wiens.
In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*. **(AAAI 2021)**

Professional Service

Reviewer: UAI, NeurIPS, ICLR, AISTATS, ICML, TMLR

Teaching Experience

Teaching assistant at UIUC

- CS 357 Numerical Methods I — 2022 Fall, 2022 Spring
- CS 441 Applied Machine Learning — 2021 Fall

Teaching assistant at UM

- EECS 445 Intro to Machine Learning — 2020 Fall

Skills

Programming: Python, Java, C++, Matlab, R, LaTeX, Mathematica

Framework: PyTorch, DeepSpeed, TensorFlow, Keras, Gym