Mengzhou Xia

CONTACT INFORMATION

35 Olden Street
Princeton University
Princeton, NJ 08540

E-mail: mengzhou@princeton.edu
Twitter: @xiamengzhou
Home: Webpage

EDUCATION

Princeton University, Princeton — 2020 - 2025 (expected)
Ph.D. candidate in Computer Science
Advisor: Danqi Chen

Carnegie Mellon University, Pittsburgh — 2018 - 2020
Masters in Computational Data Science at LTI, School of Computer Science
Advisor: Graham Neubig

Fudan University, Shanghai, China — 2014 - 2018
Bachelor of Engineering, Software Engineering (Data Science & Technology)

EMPLOYMENT

Princeton University, Princeton, NJ — Aug 2020 - Present
Graduate research assistant (Advisor: Danqi Chen)

Meta AI Research, Menlo Park, CA — May 2022 - Sept 2022
Research intern (Mentor: Veselin Stoyanov)

Meta AI Research, Remote — May 2021 - Jan 2022
Research intern (Mentor: Veselin Stoyanov)

Microsoft Research, Redmond, WA — May 2020 - Aug 2020
Research intern (Mentor: Guoqing Zheng)

Carnegie Mellon University, Pittsburgh, PA — May 2019 - Aug 2019
Graduate research assistant (Advisor: Graham Neubig)

Tencent AI Lab, Shenzhen, China — Mar 2018 - Aug 2018
Research intern (Mentor: Lemao Liu)

Google, Shanghai, China — Jul 2017 - Sept 2017
Software engineering intern

Google, Shanghai, China — Jul 2016 - Sept 2016
Engineering practicum intern

AWARDS

MIT EECS Rising Star	2024
Princeton SEAS Award for Excellence (4 awarded in Computer Science)	2024
Apple Scholars in AIML PhD Fellowship (21 awarded in 2024)	2024-2025
Best paper award at the DPFM Workshop at ICLR 2024	2024
Qualcomm Innovation Fellowship Finalist	2023
Bloomberg Data Science Fellowship (4 awarded in 2022)	2022-2023
Princeton Graduate Student Teaching Assistant Award	2021-2022
The Hisashi and Masae Kobayashi *67 Fellowship	2020-2021
Outstanding Graduates, Fudan University	2018
Shanghai City Scholarship	2015

Publications

Google Scholar https://scholar.google.com/citations?user=fNexhaUAAAAJ&hl=en&oi=ao
Semantic Scholar https://www.semanticscholar.org/author/Mengzhou-Xia/67284811

* denotes equal contribution

Preprints

[1] Hongjin Su*, Howard Yen*, **Mengzhou Xia***, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S. Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan Ö. Arik, Danqi Chen, Tao Yu. "BRIGHT: A Realistic and Challenging Benchmark for Reasoning-Intensive Retrieval." arXiv 2407.12883. *Submitted to ICLR 2025.* [PDF]

Peer-Reviewed Publications

[26] Meng Yu*, **Mengzhou Xia***, Danqi Chen. "SimPO: Simple Preferences Optimization with a Reference-Free Reward." NeurIPS 2024. [PDF]
> 700 Github Stars
Best <10B model on Chatbot Arena
> 100K monthly downloads and 400K accumulated downloads on Hugging Face

[25] Zirui Wang, **Mengzhou Xia**, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, Danqi Chen. "CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs." NeurIPS 2024 Datasets and Benchmarks Track 2024. [PDF]
Spotlight Presentation at the EVAL-FoMo Workshop at ECCV 2024

[24] Anirudh Ajith, **Mengzhou Xia**, Alexis Chevalier, Tanya Goyal, Danqi Chen, Tianyu Gao. "LitSearch: A Retrieval Benchmark for Scientific Literature Search." EMNLP 2024. [PDF]

[23] Luxi He*, **Mengzhou Xia***, Peter Henderson. "What is in Your Safe Data: Identifying Benign Data that Breaks Safety." COLM 2024; DPFM Workshop@ICLR 2024. [PDF]
Best Paper Award at the DPFM Workshop at ICLR 2024

[22] Zexuan Zhong **Mengzhou Xia**, Danqi Chen, Mike Lewis. "Lory: Fully Differentiable Mixture-of-Experts for Autoregressive Language Model Pre-training." COLM 2024. [PDF]

[21] **Mengzhou Xia***, Sadhika Malladi*, Suchin Gururangan, Sanjeev Arora, Danqi Chen. "LESS: Selecting Influential Data for Targeted Instruction Tuning." ICML 2024. [PDF]

[20] Abhishek Panigrahi, Sadhika Malladi, **Mengzhou Xia**, Sanjeev Arora. "Trainable Transformer in Transformer." ICML 2024. [PDF]

[19] Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, **Mengzhou Xia**, Prateek Mittal, Mengdi Wang, Peter Henderson. "Assessing the brittleness of safety alignment via pruning and low-rank modifications." ICML 2024. [PDF]

[18] Alexis Chevalier, Jiayi Geng, Alexander Wettig, Howard Chen, Sebastian Mizera, Simon Machado, Arturo Rodriguez Fanlo, Simon Frieder, Zirui Wang, Akshara Prabhakar, Jiachen T. Wang, Xindi Wu, **Mengzhou Xia**, Wenhan Xia, Jiatong Yu, Ellie Thieu, Max Aragon,

Zhiyong Ren, Junjie Zhu, Toni Annala, Sanjeev Arora, Danqi Chen. "Language Models as Science Tutors." ICML 2024. [PDF]

[17] **Mengzhou Xia**, Tianyu Gao, Zhiyuan Zeng, Danqi Chen. "Sheared LLaMA: Accelerating Language Model Pre-training via Structured Pruning." ICLR 2024. [PDF]
> 500 Github Stars
Best 1.3B model at the time of release (2023-10)
> 100K monthly downloads and 800K accumulated downloads for the series

[16] Weijia Shi*, Anirudh Ajith*, **Mengzhou Xia**, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, Luke Zettlemoyer. "Detecting Pretraining Data from Large Language Models." ICLR 2024. [PDF]

[15] Yangsibo Huang, Samyak Gupta, **Mengzhou Xia**, Kai Li, Danqi Chen. "Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation." ICLR 2024. [PDF]
Spotlight Presentation

[14] Anirudh Ajith, Chris Pan, **Mengzhou Xia**, Ameet Deshpande, Karthik Narasimhan. "InstructEval: Systematic Evaluation of Instruction Selection Methods." NAACL 2024 Findings. [PDF]

[13] **Mengzhou Xia**, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, Veselin Stoyanov. "Training Trajectories of Large Language Models Across Scales." ACL 2023. [PDF]

[12] **Mengzhou Xia**, Mikel Artetxe, Jingfei Du, Danqi Chen, Veselin Stoyanov. "Prompting ELECTRA: Few-Shot Learning with Discriminative Pre-Trained Models." EMNLP 2022 short. [PDF]

[11] Mozes van de Kar, **Mengzhou Xia**, Danqi Chen, Mikel Artetxe. "Don't Prompt, Search! Mining-based Zero-Shot Learning with Language Models." EMNLP 2022 short. [PDF]

[10] Jacqueline He, **Mengzhou Xia**, Christiane Fellbaum, Danqi Chen. "MABEL: Contrastive Gender Bias Mitigation using Entailment Pairs." EMNLP 2022. [PDF]

[9] **Mengzhou Xia**, Zexuan Zhong, Danqi Chen. "Structured Pruning Learns Compact and Accurate Models." ACL 2022. [PDF]

[8] Howard Chen, **Mengzhou Xia**, Danqi Chen. "Non-Parametric Few-Shot Learning for Word Sense Disambiguation." NAACL 2021 short. [PDF]

[7] **Mengzhou Xia**, Guoqing Zheng, Subhabrata Mukherjee, Milad Shokouhi, Graham Neubig, Ahmed Hassan Awadallah. "MetaXL: Meta representation transformation for low-resource cross-lingual learning." NAACL 2021. [PDF]

[6] **Mengzhou Xia**, Antonios Anastasopoulos, Ruochen Xu, Yiming Yang, Graham Neubig. "Predicting Performance for Natural Language Processing Tasks." ACL 2020. [PDF]

[5] **Mengzhou Xia**, Anjalie Field, Yulia Tsvetkov. "Demoting Racial Bias in Hate Speech Detection." SocialNLP Workshop@ACL 2020. [PDF]

[4] **Mengzhou Xia**, Xiang Kong, Antonios Anastasopoulos, Graham Neubig. "Generalized Data Augmentation for Low-Resource Translation." ACL 2019. [PDF]

[3] Junjie Hu, **Mengzhou Xia**, Graham Neubig, Jaime Carbonell. "Domain Adaptation of Neural Machine Translation by Lexicon Induction." ACL 2019. [PDF]

[2] Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, **Mengzhou Xia**, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, Graham Neubig. "Choosing Transfer Languages for Cross-Lingual Learning." ACL 2019. [PDF]
Oral Presentation

[1] **Mengzhou Xia**, Guoping Huang, Lemao Liu, Shuming Shi. "Graph based Translation Memory for Neural Machine Translation." AAAI 2019. [PDF]

Invited Talks

Columbia NLP Seminar 11/2024
Title: *Aligning Language Models with LESS Data and a Simple (SimPO) Objective*

Stanford NLP Seminar 11/2024
Title: *Aligning Language Models with LESS Data and a Simple (SimPO) Objective*

MIT Embodied Intelligence Seminar 10/2024
Title: *Aligning Language Models with LESS Data and a Simple (SimPO) Objective*

Google Research (Host: Yangsibo Huang) 08/2024
Title: *CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs*

Microsoft Research Asia (Host: Li Dong) 06/2024
Title: *Training and Aligning Language Models: Algorithmic Advances in Objectives and Data Curation*

Apple Machine Learning Research (Host: Fartash Faghri) 04/2024
Title: *Exploring the Pareto-Frontier of Performance and Efficiency of Large Language Models*

Google Research (Host: Chiyuan Zhang) 03/2024
Title: *Data- and Parameter-Efficient Adaptation of Large Language Models*

Tencent AI (Host: Huayang Li) 11/2023
Title: *Sheared-LLaMA: Accelerating Language Model Pre-training via Structured Pruning*

ML Collective (Host: Rosanne Liu) 02/2023
Title: *Training Trajectories of Large Language Models Across Scales*

Apple Natural Language Understanding Workshop 02/2023
Title: *Towards Building Efficient Language Models*

Google Research (Host: Tao Lei) 10/2022
Title: *Structured Pruning Learns Compact and Accurate Models*

Professional Services

Workshop Co-Organizer

- 2nd Workshop on Adapting and Aligning Foundation Models, ICLR 2025 (*in submission*)
- 2nd Workshop on High-dimensional Learning Dynamics (HiLD): The Emergence of Structure and Reasoning, ICML 2024

Area Chair

- EMNLP 2023 (Track: Efficiency Methods for NLP)

Reviewer/Program Committee

- AAAI (2020, 2021)
- EMNLP (2020, 2021, 2022, 2023)
- ACL (2021, 2022, 2023)
- NAACL (2021)
- ICLR (2022, 2023, 2024)
- ICLR ME-FoMo Workshop (2024)
- ICML (2022, 2023)
- NeurIPS (2022, 2023)
- NeurIPS Workshop on Instruction Tuning and Instruction Following (2023)
- COLM (2024)

Mentorship

- Adithya Bhaskar (2024-), Princeton PhD.
- Luxi He (2024), Princeton PhD.
  - Best Paper Award at DPFM Workshop@ICLR 2024
  - COLM 2024
- Howard Yen (2024), Princeton BS → PhD at Princeton.
  - ICLR 2025 in submission
- Jiayi Geng (2023-), Princeton MS.
- Colin Wang (2023-), Princeton MS.
  - NeurIPS Datasets and Benchmarks Track 2024
  - Spotlight Presentation at ECCV EVAL-FoMo Workshop 2024
- Anirudh Ajith (2022-2024), Princeton MS → Research Scientist at AI2.
  - EMNLP 2024
  - NAACL 2024 Findings
  - ICLR 2024
- Jane Pan (2022-2023), Princeton MS → PhD at NYU.
- Zhiyuan Zeng (2023), Tsinghua BS → PhD at UW.
- Chris Pan (2022-2023), Princeton MS → Quantitative Researcher at HRT.
  - NAACL 2024 Findings
- Jacqueline He (2021-2022), Princeton BS → PhD at UW.
  - EMNLP 2022

Teaching

Guest lecturer (Instructor: Sanjeev Arora, Danqi Chen) — Fall 2024
COS 597R: Deep Dive into Large Language Models, Princeton University
Title: *The Power of Small Language Models*

Guest lecturer (Instructor: Stella Yu) — Fall 2024
EECS 542, Advanced Topics in Computer Vision, University of Michigan
Title: *Evaluations of Multimodal Language Models*

Guest lecturer (Instructor: Kai Li) — Spring 2024
COS 598D: Systems and Machine Learning, Princeton University
Title: *Pre-trained Large Language Models*

Lecturer — Fall 2023
Computing Center Seminar, Princeton University
Title: *Distributed Training with Model Parallel Techniques*

Teaching assistant (Instructor: Karthik Narasimhan) — Spring 2022
COS 484: Natural Language Processing, Princeton University

Teaching assistant (Instructor: Sanjeev Arora, Danqi Chen) — Fall 2021
COS 324: Introduction to Machine Learning, Princeton University

Teaching assistant (Instructor: Graham Neubig) — Fall 2019
11-731: Machine Translation and Sequence-to-sequence Models, CMU