

Emily Xiao

emilyxxiao@gmail.com | [Website](#) | [Google Scholar](#) | [LinkedIn](#) | [Twitter](#)

EDUCATION

M.S. Carnegie Mellon University

Language Technologies Institute, School of Computer Science

Advisor: [Graham Neubig](#)

8/2024 – 12/2025

GPA: 3.9/4.3

B.A. University of California, Berkeley

Major: Computer Science

8/2017 – 5/2021

Major GPA: 3.7/4.0

PUBLICATIONS

Prompt-MII: Meta-Learning Instruction Induction for LLMs

Emily Xiao, Yixiao Zeng, Ada Chen, Chin-Jou Li, Amanda Bertsch, Graham Neubig

ICLR 2026

Efficient Many-Shot In-Context Learning with Dynamic Block-Sparse Attention

Emily Xiao, Chin-Jou Li, Yilin Zhang, Graham Neubig, Amanda Bertsch

ACL 2025 Main

DATE-LM: Benchmarking Data Attribution Evaluation for Large Language Models

Cathy Jiao*, Yijun Pan*, **Emily Xiao***, Daisy Sheng, Niket Jain, Hanzhang Zhao, Ishita Dasgupta, Jiaqi W. Ma, Chenyan Xiong

NeurIPS 2025

In-context learning with long-context models: An in-depth exploration

Amanda Bertsch, Maor Ivgi, **Emily Xiao**, Uri Alon, Jonathan Berant, Matthew R Gormley, Graham Neubig

NAACL 2025 Main [[SAC Award for Language Modeling](#)]

Automatically generating cause-and-effect questions from passages

Katherine Stasaski, Manav Rathod, Tony Tu, **Emily Xiao**, Marti A Hearst

EACL 2021, BEA Workshop

ACADEMIC RESEARCH EXPERIENCE

LLM Adaptation

9/2024 – 12/2025

CMU, Advised by [Graham Neubig](#)

- Analyzed behaviors of in-context learning vs finetuning for LLM task adaptation ([NAACL 2025](#)).
- Proposed efficient many-shot ICL using dynamic sparse attention; implemented with FlexAttention and transformers library modifications, achieving 2× speedup while recovering 95% accuracy ([ACL 2025](#)).
- Used GRPO to train an LLM for automatic prompt engineering; ran multi-node VeRL training using 3000+ datasets, achieving SOTA performance with 13× shorter prompt. ([ICLR 2026](#))

LLM Training Data Curation

9/2024 – 5/2025

CMU, Advised by [Chenyan Xiong](#) and [Jiaqi Ma](#)

- Explored pre-training and post-training data selection methods for LLMs.
- Implemented data-attribution methods efficiently using custom PyTorch hooks and LitGPT.
- Conducted 100+ training runs, and probing-based analysis using gradient and optimizer states.
- Designed efficient eval framework and comprehensive analysis of existing methods ([NeurIPS 2025 D&B](#))

Synthetic Data Generation

9/2020 – 1/2021

UC Berkeley, Advised by [Marti Hearst](#)

- Built synthetic data generation pipeline for finetuning BERT; designed autoevals ([BEA@EACL 2021](#))

INDUSTRY EXPERIENCE

Research Resident, Bridgewater Associates

1/2026 - Present

AIA Labs

- Developing post-training methods for LLM-based forecasting.
- Building financial foundation models.
- Co-authored [blogpost](#) with Thinking Machines Lab

Machine Learning Engineer, ByteDance

9/2021 – 8/2023

Query Auto-Completion

- Proposed a new training data and modeling method that mitigates position bias and click-baiting; Launched globally with +2% prediction accuracy, and gave talk at TikTok Search.
- Designed predictive pre-caching, making short-prefix responses 2× faster.
- Optimized the end-to-end multi-stage retrieval and ranking pipeline.

Related Search

- Sole developer of the modeling and recommendation pipeline; drove 3× growth in search volume.

Software Engineer Intern, Instagram

5/2020 – 8/2020

Instagram Reels – Feed Ranking (founding team)

- Developed a personalized short-video retrieval strategy; feature-engineered CTR ranking model.

Founding Engineer, SuiteSocial

9/2019 – 3/2020

Startup building a brand-influencer matching platform

Won 2nd Place at UC LAUNCH Accelerator Demo Day

- Built supervised model for brand-influencer affinity ranking.

ADDITIONAL

Programming: Python, C++, SQL

Tools: PyTorch, TensorFlow, Huggingface, vLLM, LitGPT, Hadoop, Spark, Kafka, CUDA C++

Languages: English, Chinese, Japanese, Spanish

CS Coursework: Data Structures; Probability and Statistics; CS Theory; Computer Architecture; Information Devices and Systems; Computer Security; UI/UX Design; Database Systems

ML Coursework: Machine Learning; NLP; Optimization Models; Advanced NLP; Deep Learning Systems; LLM Systems; Trustworthy AI Theory and Practice; Inference Algorithms for Language Modeling