Ziyang Xiong

xziyang@berkeley.edu | 734-210-3298 | https://www.linkedin.com/in/ziyang-xiong/ |https://lemon-awa.github.io/

Education

University of California, Berkeley | M.S. in EECS — Aug 2025 - May 2026

Master of Engineering, Electrical Engineering & Computer Sciences

University of Michigan | B.S.E in Data Science | GPA: 3.95 / 4.0 — Aug 2023 - May 2025

Core courses: Foundation of LLM, Natural Language Processing, Data Mining, Database Management Systems, Computer Vision

Shanghai Jiaotong University | B.S.E in ECE | GPA: 3.65 / 4.0 — Aug 2021 - Aug 2025

Publications

[1]Map2Text: New Content Generation from Low-Dimensional Visualizations
KDD 2025, ***Accepted;*** *NAACL AISD,* ***Accepted***
Xingjian Zhang, **Ziyang Xiong**, Shixuan Liu, Yutong Xie, Tolga Ergen, Dongsub Shim, Hua Xu, Honglak Lee, Qiaozhu Mei

[2]Safeguard is a Double-edged Sword: Denial-of-service Attack on Large Language Models
ACM CCS-LAMPS, ***Accepted****;*
Qingzhao Zhang, **Ziyang Xiong**, Z. Morley Mao

[3]Making Small Language Models Efficient Reasoners: Intervention, Supervision, Reinforcement
ICML 2025 Workshop LCFM, ***Accepted,*** *AAAI 2025 In Review*
Xuechen Zhang, Zijian Huang, Chenshun Ni, **Ziyang Xiong**, Jiasi Chen, Samet Oymak

[4]MASSW: A New Dataset and Benchmark Tasks for AI-Assisted Scientific Workflows
NAACL 2024, ***Accepted***
Xingjian Zhang, Yutong Xie, …, **Ziyang Xiong**, et al.

Intern Experience

Byte dance | Model Algorithm Engineer — May 2025 – Aug 2025

- Led development of **Search Ads title rewriting models** using advanced fine-tuning techniques (**SFT, DPO, KTO**), and keyword extraction processes, achieving production deployment with measurable business impact: **CTR +0.467%、advv +1.963%、send +0.477%.**
- Built automated **bad case evaluation models** achieving **86.7% precision** and **89% recall** using **Qwen3-8B-CoT**, establishing robust quality control frameworks, and significantly improving model assessment efficiency.
- Developed **end-to-end** video-to-title multimodal models using **Qwen2.5-VL-7B**, trying to resolve information loss issues in multi-stage processing and enhancing selling point concentration through innovative two-stage Label to Title generation approach.

Research Experience

Denial-of-Service Attack on Large Language Models | Model Security Algorithm Engineer — Jun 2024 - Present

- Engineered white-box **adversarial attack** frameworks in **PyTorch**, achieving **97%** denial-of-service success rate on safeguard models (e.g., **LLaMA-Guard**) through gradient-based optimization with minimal 30-character prompts.
- Developed hybrid attack strategies combining adversarial prompts and **data poisoning techniques** for black-box models (e.g. **GPT**), implementing automated prompt generation pipelines to enhance attack transferability while maintaining low detection rates.
- Designed and optimized **prompt placement algorithms** (prefix/suffix/random insertion) with vocabulary constraints, systematically evaluating attack effectiveness across different model architectures and safety mechanisms.

The scientific discovery application with LLM | Model Application Algorithm Engineer — Dec 2023 – May 2025

- Introduced a novel Map2Text task for converting 2D coordinates into coherent text descriptions, proposing three candidate solutions: **fine-tuned LLMs**, **embedding inversion models**, and **RAG-based prompt engineering** to establish baseline performance.
- Designed a novel **evaluation metric** that decomposes generated and reference texts into atomic statements before comparison, achieving more accurate assessment of **logical consistency** and **semantic alignment** compared to traditional metrics (BLEU, ROUGE).
- Designed an automatically extracting and structuring scientific workflows to build a dataset named MASSW from research publications, which can enable large language models to analyze and summarize complex research methodologies with improved accuracy.

SOTA Lab | Research Assistant Intern. — May 2024 – May 2025

- Conducted comprehensive experiments on **Qwen** and **DeepSeek-Distill** models of varying sizes, comparing **Supervised Finetuning (SFT)** and **Reinforcement Learning (RL)** approaches for task alignment, revealing SFT's tendency to generate unnecessarily verbose reasoning.
- Proposed **RL-based optimization** algorithm with **length penalty** mechanism, achieving comparable accuracy while significantly reducing redundant reasoning steps, demonstrating efficiency improvement across different model architectures.
- Conducted comparative analysis between **fine-tuning** and **in-context learning** approaches on math tasks, then developed hybrid optimization strategy combining their strengths to achieve superior performance while maintaining computational efficiency.

Project Experience

Search Engine Infrastructure — Feb 2025 – May 2025

- Engineered high-performance web crawler with multithreading and efficient parsing mechanisms, implementing inverted index structure and data compression techniques to optimize storage and retrieval operations.
- Developed hybrid search ranking system combining **neural network** models with **PageRank** scores and **heuristic** signals, achieving efficient Boolean retrieval and phrase matching capabilities for complex queries.
- Designed and deployed scalable search infrastructure on **AWS** with **caching and query optimization** for high-throughput retrieval.

Prediction of depression risk — Aug 2022 – May 2023

- Data Collection and Preprocessing: Gathered and preprocessed data, including depression status, nearly 20 biochemical indicators, and about 10 socioeconomic factors, to build a comprehensive feature set for high-quality input into regression model training.
- Feature Selection and Modeling: Employed methods such as **oversampling, Lasso regularization, and random forests** to optimize model.
- Impact Analysis: Used **Bootstrap, SVM, LR, and KNN** to determine and analyze the impact of different factors on depression.