Ming Li

Personal Page | Google Scholar | Github | minglii@umd.edu

Education

University of Maryland — Maryland, US
Ph.D. in Computer Science — *Aug. 2023 – present*

Texas A&M University — Texas, US
M.S. in Computer Science — *Sep. 2021 – May 2023*

Xi'an Jiaotong University — Xi'an, China
B.S. in Computer Science — *Aug. 2016 – June 2020*

Research & Internship experience

Research Assistant — Aug. 2023 – present
University of Maryland — *Maryland, US*
- Supervisor: Prof. Tianyi Zhou
- Focus: Instruction-tuning on Large Language models

Algorithm Engineer (Intern) — May 2023 – Aug. 2023
Ping An Technology (Shenzhen) Co., Ltd. — *Shenzhen, China*
- Data selection for instruction-tuning on LLMs
- Black-Box Large Language Models for Retrieval Question Answering

Research Assistant — Sep. 2021 – May 2023
Texas A&M University — *Texas, US*
- Supervisor: Prof. Ruihong Huang
- Focus: General Discourse Parsing in Natural Language Processing

Research Assistant (Intern) — Jun. 2019 – Jun. 2021
Shenzhen Institutes of Advanced Technology, Chinese Academy of Science — *Shenzhen, China*
- Supervisor: Prof. Yu Qiao
- Focus: Scene Text Recognition and Text Detection

Algorithm Engineer (Intern) — Jan. 2021 – Apr. 2021
Shenzhen Fitlab Co. Ltd — *Shenzhen, China*
- Deep Learning based dumbbell detection and weight recognition
- Application on Deep Learning based pose estimation

Research Student — Sep. 2017 – May 2018
Xi'an Jiaotong University — *Xi'an, China*
- Supervisor: Prof. Hongzhe Xu
- Focus: Knowledge Graph, Information Extraction and Natural Language Processing

Publications

[1] **Ming Li**, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, Tianyi Zhou, Superfiltering: Weak-to-Strong Data Filtering for Fast Instruction-Tuning. *arXiv preprint arXiv:2402.00530.*

[2] **Ming Li**, Lichang Chen, Jiuhai Chen, Shwai He, Heng Huang, Jiuxiang Gu, Tianyi Zhou, Reflection-Tuning: Data Recycling Improves LLM Instruction-Tuning. *arXiv preprint arXiv:2310.11716, Accepted by NIPS 2023 Workshop.*

[3] **Ming Li**, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, Jing Xiao, From Quantity to Quality: Boosting LLM Performance with Self-Guided Data Selection for Instruction Tuning. *arXiv preprint arXiv:2308.12032.*

[4] Haoyan Yang, Zhitao Li, Yong Zhang, Jianzong Wang, Ning Cheng, **Ming Li**, Jing Xiao, PRCA: Fitting Black-Box Large Language Models for Retrieval Question Answering via Pluggable Reward-Driven Contextual Adapter. *Accepted by EMNLP 2023.*

[5] **Ming Li**, Ruihong Huang, Less is More: A Lightweight and Robust Neural Architecture for Discourse Parsing. *arXiv preprint arXiv:2210.09537.*

[6] **Ming Li**, Ruihong Huang, RST-style Discourse Parsing Guided by Document-level Content Structures. *arXiv preprint arXiv:2309.04141.*

[7] **Ming Li**, Ruihong Huang, Semi-supervised News Discourse Profiling with Contrastive Learning.*arXiv preprint arXiv:2309.11692.*

[8] **Ming Li**, Bin Fu, Zhengfu Zhang, Yu Qiao, Character-Aware Sampling and Rectification for Scene Text Recognition. *Accepted by IEEE Transactions on Multimedia.*

[9] **Ming Li**, Bin Fu, Han Chen, Junjun He, Yu Qiao, Dual Relation Network for Scene Text Recognition. *Accepted by IEEE Transactions on Multimedia.*

[10] Qitong Wang, Bin Fu, **Ming Li**, Junjun He, Yu Qiao, Region-aware Arbitrary-shaped Text Detection with Progressive Fusion *Accepted by IEEE Transactions on Multimedia.*

Research Projects

Selective Reflection-Tuning [Project Repo] Aug. 2023 – present

University of Maryland *Maryland, US*

- Proposed the Reflection-Tuning and Selective Reflection-Tuning, a data recycle method for instruction tuning
- Win rate of 83% on Alpaca Eval Leaderboard, best 7B model with only a little recycled instruction data

Cherry data selection for instruction-tuning on LLM [Project Repo] May 2023 – Aug. 2023

University of Maryland *Maryland, US*

- Used approximately 5% or 10% of the data to have comparable performances to the models trained on full data, which is experimented on the Alpaca and WizardLM datasets.
- The selection of cherry data is entirely self-guided and does not need ANY extra outside models, ranging from BERT to chatGPT.

How Chain-of-Thaught affects the instruction-tuning on LLM Apr. 2023 – June 2023

University of Maryland *Maryland, US*

- Implemented Chain-of-Thaught during the instruction-tuning of LLM
- Exprimentd on how paraphrasing of COT affects LLM's performance on following COT.

Semi-Supervised Learning on News Discourse Profiling Sep. 2022 – Jan. 2023

Texas A&M University *Texas, US*

- Researched towards semi-supervised methods for discourse-level tasks, especially News Discourse Profiling.
- Designed Knowledge Distillation and Contrastive Learning based methods and achieved state-of-the-art performance in News Discourse Profiling task

Natural Language Processing on Rhetorical Structure Theory Parsing Jun. 2022 – Feb. 2023

Texas A&M University *Texas, US*

- Proposed to construct the rhetorical structure with the high-level event-related representation of each sentence
- The proposed method achieved state-of-the-art performance with only few layers introduced

Natural Language Processing on News Discourse Profiling Jan. 2022 – July 2022

Texas A&M University *Texas, US*

- Analyzed the structure of news articles and categorized every sentence based on its role in the article.
- Proposed a simple yet effective model that achieves promising performance in several discourse parsing tasks with lower parameters and processing time.

Computer Vision on Scene Text Recognition and Detection Jun. 2019 – Jun. 2021

Shenzhen Institutes of Advanced Technology, Chinese Academy of Science *Shenzhen, China*

- A paper is accepted which focuses on recognizing curved texts in natural scene
- A paper is accepted where local visual and long-range contextual information are utilized simultaneously to get a better recognition performance
- A paper is accepted where effective multi-scale contextual features are utilized for locating text instances

Technical Skills

Programming Languages: Python, C/C++, Java, MATLAB, SQL // Pytorch, TensorFlow
Languages: Chinese (Native), English (TOEFL: 100; GRE: 322)