Zichong LI *(D.o.B.: 03/30/2002)*

+1-4049526617 | zli911@gatech.edu | https://zichongli5.github.io

EDUCATION

Georgia Institute of Technology, *Atlanta, GA, USA* August 2023 – June 2027
Ph.D. in Machine Learning, *H. Milton Stewart School of Industrial and Systems Engineering*

University of Science and Technology of China (USTC), *Hefei, Anhui, China* September 2020 - June 2023
M.S. in Data Science, *School of Data Science* GPA: 3.98/4.3 Ranking: 1/56
Relevant Coursework: Machine Learning and Knowledge Discovery, Deep Learning, Reinforcement Learning, Digital image Processing, Fundamentals of Data Science, Natural Language Understanding, Optimization Algorithm, Social Computing

University of Science and Technology of China (USTC), *Hefei, Anhui, China* September 2016 - June 2020
B.S. in Mathematics and Applied Mathematics/Probability Statistics, *School of the Gifted Young.* Major GPA: 3.91/4.3
Relevant Coursework: Mathematical Statistics, Advanced Probability Theory, Regression Analysis, Multivariate Analysis, Mathematical Analysis, Combinatorics, Applied Stochastic Processes, Time Series Analysis, Functional Analysis

PUBLICATIONS

- **SlimMoE: Structured Compression of Large MoE Models via Expert Slimming and Distillation**
  Zichong Li, Chen Liang, Zixuan Zhang, Ilgee Hong, Young Jin Kim, Weizhu Chen and Tuo Zhao
  Submitted to the 2nd Conference on Language Modeling (COLM), 2025
- **LLMs Can Generate a Better Answer by Aggregating Their Own Responses**
  Zichong Li, Xinyu Feng, Yuheng Cai, Zixuan Zhang, Tianyi Liu, Chen Liang, Weizhu Chen, Haoyu Wang and Tuo Zhao
  arXiv preprint arXiv:2503.04104, 2025
- **Adaptive Preference Scaling for Reinforcement Learning with Human Feedback**
  Ilgee Hong*, **Zichong Li***, Alexander Bukharin, Yixiao Li, Haoming Jiang, Tianbao Yang and Tuo Zhao
  The Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS), 2024
- **Robust Reinforcement Learning from Corrupted Human Feedback**
  Alexander Bukharin, Ilgee Hong, Haoming Jiang, **Zichong Li**, Qingru Zhang, Zixuan Zhang and Tuo Zhao
  The Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS), 2024
- **Beyond Point Prediction: Score Matching-based Pseudolikelihood Estimation of Neural Marked Spatio-Temporal Point Process**
  Zichong Li, Qunzhi Xu, Zhenghao Xu, Yajun Mei, Tuo Zhao and Hongyuan Zha
  International Conference on Machine Learning (ICML), 2024
- **SMURF-THP: Score Matching-based UnceRtainty quantiFication for Transformer Hawkes Process**
  Zichong Li, Yanbo Xu, Simiao Zuo, Haoming Jiang, Chao Zhang, Tuo Zhao and Hongyuan Zha
  International Conference on Machine Learning (ICML), 2023
- **Efficient Deep Ensemble Inference via Query Difficulty-dependent task Scheduling**
  Zichong Li, Lan Zhang, Mu Yuan, Miaohui Song and Qi Song
  International Conference on Data Engineering (ICDE), 2023
- **CoTel: Ontology-Neural Co-Enhanced Text Labeling**
  Miaohui Song, Lan Zhang, Mu Yuan, **Zichong Li**, Qi Song, Yijun Liu and Guidong Zheng
  The Web Conf (WWW), 2023
- **Transformer Hawkes Process**
  Simiao Zuo, Haoming Jiang, **Zichong Li**, Tuo Zhao and Hongyuan Zha
  International Conference on Machine Learning (ICML), 2020
- **PRIMAL: A Linear Programming-based Sparse Learning Library in R and Python**
  Qianli Shen*, **Zichong Li***, Yujia Xie and Tuo Zhao

WORK EXPERIENCE

Research Intern, **Microsoft Research**, Redmond, WA, USA **May 2024 – present**
Research Assistant, **Nanshan Bureau of Statistics**, Shenzhen, Guangdong, China **August 2018 - September 2018**

RESEARCH EXPERIENCE (Selected)

Microsoft Research, USA
Advisor: Chen Liang
Project: Structured Compression of Large MoE Models via Expert Slimming and Distillation

- Proposed a multi-stage prune-and-distill approach for reducing the size of MoE model while preserving performance.
- Reduced Phi 3.5 MoE to less than 20% of the original size using <10% of pretraining data and developed Phi-mini-MoE and Phi-tiny-MoE, achieving superior performance compared to open-sourced models with similar parameters.
- Paper submitted to COLM 2025.

Foundations of Learning Systems for Alchemy, *Georgia Institute of Technology*, USA
Advisor: Professor Tuo Zhao
Project: Adaptive Preference Scaling for Reinforcement Learning with Human Feedback.

- Proposed an adaptive preference loss for reward learning in RLHF to address the uncertainty in preference data.
- Incorporated an adaptive scaling parameter for each pair of preference, increasing the flexibility of the reward.
- Paper accepted to NeurIPS 2024.

Project: Score Matching-based Uncertainty Quantification for Point Process.

- Proposed training the model using score matching technique to circumvent computation of the intractable integral.
- Designed a sampling algorithm based on Langevin Dynamics to generate event samples for uncertainty quantification.
- Paper accepted to ICML (International Conference on Machine Learning) 2023.

Project: Transformer Hawkes Process

- Proposed leveraging a Transformer model to capture long-term dependencies in the temporal point process.
- Paper accepted to ICML (International Conference on Machine Learning) 2020.

Lab for Intelligent Networking and Knowledge Engineering, *USTC*, China
Advisor: Professor Lan Zhang
Project: Efficient Deep Ensemble Inference via Query Difficulty-dependent Task Scheduling

- Proposed an adaptive scheduler for deep ensemble inference pipeline to reduce its query deadline miss rate.
- Created an algorithm that dynamically scheduled model inference tasks for queries according to workload and query difficulty, effectively reduced deadline miss rate by 5x and increased overall accuracy by 30.8%.
- Paper accepted to ICDE (International Conference on Data Engineering) 2023.

AWARDS

- First-level Freshman Scholarship, awarded by USTC **September 2016**
- Endeavour Scholarship, awarded by USTC **October 2017**
- Outstanding Student (Top 10 in the special class), awarded by USTC **October 2019**

VOLUNTEER WORK

Volunteer, **Rural Poverty Alleviation**, Shanwei, Guangdong, China, **July 2017 - August 2017**
Dance Performer, **University of Science and Technology of China**, Hefei, Anhui, China **September 2021**

SKILLS

- **Programming Language**: Python, C, R
- **Other Software**: Photoshop, Mathematica, LaTeX, MATLAB
- **Mathematics:** Complex Analysis, Differential Equations, Probability Theory, Stochastic Processes