Tianjian Li

Updated February 13, 2025

Email: tli104@jhu.edu **Personal Website**: tianjianl.github.io **Google Scholar**: Link

Education

Johns Hopkins University — Baltimore, MD
PhD in Computer Science — 2024 – Present
Advisor: Daniel Khashabi

Johns Hopkins University — Baltimore, MD
MSE in Computer Science — 2022 – 2024
Advisor(s): Kenton Murray, Daniel Khashabi, Philipp Koehn

New York University — New York, NY
BA in Computer Science and Mathematics — 2017 – 2021

Publications

***Upsample or Upweight?* Balanced Training on Heavily Imbalanced Datasets**
Tianjian Li, Haoran Xu, Weiting Tan, Kenton Murray, Daniel Khashabi
North American Chapter of the Association for Computational Linguistics (NAACL). 2025. Link

Benchmarking Language Model Creativity: A Case Study on Code Generation
Yining Lu, Dixuan Wang, **Tianjian Li**, Dongwei Jiang, Daniel Khashabi
North American Chapter of the Association for Computational Linguistics (NAACL). 2025. Link

***Verifiable by Design:* Aligning Language Models to Quote from Pre-Training Data**
Jingyu Zhang, Marc Marone, **Tianjian Li**, Benjamin Van Durme, Daniel Khashabi
North American Chapter of the Association for Computational Linguistics (NAACL). 2025. Link

Error Norm Truncation: Robust Training in the Presence of Data Noise for Text Generation Models
Tianjian Li, Haoran Xu, Philipp Koehn, Daniel Khashabi, Kenton Murray
International Conference on Learning Representations (ICLR), 2024. Spotlight (Top 5%). Link

Why Does Zero-Shot Cross-Lingual Generation Fail? An Explanation and a Solution
Tianjian Li, Kenton Murray
Association of Computational Linguistics (ACL) - Findings, 2023. Link

Research experience

Johns Hopkins University — 2022 – 2024
Center for Language and Speech Processing (CLSP)
Research Assistant. Advisors: Kenton Murray, Daniel Khashabi, Philipp Koehn
- Data re-weighting for heavily imbalanced datasets. (Under Review)
- Detecting token-level errors in training data. (ICLR' 24)

Tsinghua University — Spring 2022
Research Intern. Advisor: Jie Tang
- Data curation and pre-training of large multilingual language models.
- Multilingual Language Model evaluation.

Industry experience

Meta. Fundamental AI Research (FAIR) — Seattle, WA
Research Scientist Intern. Manager: Tianlu Wang — Summer 2025

Baidu Inc. Baidu Maps — Beijing, China
Machine Learning Engineer (Intern) — Fall 2021

Skills

Programming: Python, C/C++, Java, Shell script
Frameworks: Pytorch (Distributed Training), Huggingface, Fairseq, Jax, vLLM, Ray

Service

Reviewer: ACL (2023, 2024), EMNLP (2023, 2024), NAACL (2025), EACL (2024), COLM (2024), ICLR (2025)
Organizer: Mid-Atlantic Student Colloquium on Speech, Language and Learning (2024)

Teaching

Teaching Assistant Baltimore, MD
CS 601.471/671 NLP: Self-supervised Models Spring 2025

Course Assistant Baltimore, MD
CS 601.471/671 NLP: Self-supervised Models Spring 2024