

# Tianjian Li

Updated October 8, 2024

Email: tli104@jhu.edu Website: [tianjianl.github.io](https://tianjianl.github.io)

## Education

### Johns Hopkins University

PhD in Computer Science

Advisor: Daniel Khashabi

Baltimore, MD

2024 – Present

### Johns Hopkins University

MSE in Computer Science

Advisors: Kenton Murray, Daniel Khashabi, Philipp Koehn

Baltimore, MD

2022 – 2024

### New York University

BA in Computer Science and Mathematics

New York, NY

2017 – 2021

## Publications

### ***Upsample or Upweight?* Balanced Training on Heavily Imbalanced Datasets**

**Tianjian Li**, Haoran Xu, Weiting Tan, Dongwei Jiang, Kenton Murray, Daniel Khashabi

*Under Review.* [Link](#)

### ***Verifiable by Design: Aligning Language Models to Quote from Pre-Training Data***

Jingyu Zhang, Marc Marone, **Tianjian Li**, Benjamin Van Durme, Daniel Khashabi

*Under Review.* [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

#### **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](#))

- Locating errors in training data. ([ICLR' 24](#))

2022 – 2024

### Tsinghua University & Zhipu.AI

Research Intern

Advisor: Jie Tang

- Data curation and pre-training of large multilingual language models.

- Multilingual Language Model evaluation.

Spring 2022

## Industry experience

### Baidu Inc. Baidu Maps

Machine Learning Engineer (Intern)

- Optimization of estimated trip time with graph neural networks.

Beijing, China

Fall 2021

## Skills

**Programming:** Python, C/C++, Java

**Frameworks:** PyTorch, Huggingface (Accelerate), Fairseq, DeepSpeed, Jax

**Tools:** Docker, git, Hadoop streaming, Spark, Vim, LaTeX

## Service

**Reviewer:** ACL (2023, 2024), EMNLP (2023, 2024), EACL (2024), COLM (2024), ICLR (2025)