

# Tianjian Li

Baltimore, MD | [tl104@jhu.edu](mailto:tl104@jhu.edu) | [Personal Website](#) | [Github](#) | [Google Scholar](#)

## Research Interests

---

My research lies at the intersection of Natural Language Processing and Machine Learning. I focus on **data engineering** for language models (LMs), aiming to develop **online and adaptive data methods** that adjust dynamically to the model's training progress — including [online filtering of low-quality data](#), [adaptive data balancing across sources](#), and [enhancing data diversity in online Reinforcement Learning](#).

## Education

---

**Johns Hopkins University**, Ph.D. in Computer Science Aug 2024 – Present

- Advisor: Daniel Khashabi
- Research Focus: Natural Language Processing

**Johns Hopkins University**, M.S. in Computer Science Aug 2022 - May 2024

- Advisor(s): Kenton Murray, Philipp Koehn, Daniel Khashabi
- Research Focus: Natural Language Processing

**New York University**, B.A. Joint Major in Mathematics and Computer Science Aug 2017 - May 2021

## Industry Experience

---

**Research Scientist Intern**, Meta Fundamental AI Research (FAIR) – Bellevue, WA May 2025 – Dec 2025

- Advisor(s): Tianlu Wang, Jack Lanchantin, Jason Weston

**Machine Learning Engineer**, Baidu Inc. – Beijing, China Aug 2021 – Jan 2022

## First-Authored Publications

---

**Jointly Reinforcing Diversity and Quality in Language Model Generations** [\[Paper\]](#) [\[Code\]](#)

Tianjian Li, Yiming Zhang, Ping Yu, Swarnadeep Saha, Daniel Khashabi, Jason Weston, Jack Lanchantin, Tianlu Wang.

Preprint 2025.

**SIMPLEMIX: Frustratingly Simple Mixing of Off- and On-policy Data in Language Model Preference Learning** [\[Paper\]](#)

Tianjian Li, Daniel Khashabi.

ICML 2025.

**Upsample or Upweight: Balanced Training on Heavily Unbalanced Datasets** [\[Paper\]](#) [\[Video\]](#)

Tianjian Li, Weiting Tan, Haoran Xu, Kenton Murray, Daniel Khashabi.

NAACL 2025.

**Error Norm Truncation: Robust Training in the Presence of Data Noise in Text Generation Models** [\[Paper\]](#) [\[Code\]](#) [\[Video\]](#)

Tianjian Li, Haoran Xu, Philipp Koehn, Daniel Khashabi, Kenton Murray.

ICLR 2024. **Spotlight Presentation** (367/7304  $\approx$  5%).

**Why Does Zero-shot Cross-lingual Transfer Fail? An Explanation and a Solution** [\[Paper\]](#)

Tianjian Li, Kenton Murray.

ACL 2023 Findings.

## Non-First Authored Publications

---

**The Flaw of Averages: Quantifying Uniformity of Performance on Benchmarks** [\[Paper\]](#)

Arda Uzunoglu, Tianjian Li, Daniel Khashabi. Preprint 2025.

The Translation Barrier Hypothesis: Multilingual Generation with Large Language Models Suffers from Implicit Translation Failure

[\[Paper\]](#)

Niyati Bafna, Tianjian Li, Kenton Murray, David R. Mortensen, David Yarowsky, Hale Sirin, Daniel Khashabi.

Preprint 2025.

Benchmarking Language Model Creativity: A Case Study on Code Generation

[\[Paper\]](#)[\[Code\]](#)[\[Video\]](#)

Yining Lu, Dixuan Wang, Tianjian Li, Dongwei Jiang, Sanjeev Khudanpur, Meng Jiang, Daniel Khashabi.

NAACL 2025.

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

[\[Paper\]](#)[\[Code\]](#)[\[Video\]](#)

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

NAACL 2025. [Oral Presentation.](#)

## Technologies

---

**Languages:** Python, C++ , Java, Shell Script.

**Frameworks and Tools:** Linux, PyTorch, HuggingFace, Inference Engines (vLLM, SgLang), RL frameworks (verl, slime), Distributed Training (FSDP, DeepSpeed), vim, Latex.

## Awards

---

National Olympiad in Informatics Provinces (NOIP) First Prize.

2015

## Service

---

**Reviewer for —**

- NLP Conferences: ACL (2023, 2024, 2025), EMNLP (2023, 2024, 2025), NAACL (2025), EACL (2024, 2025), COLM (2024, 2025).
- ML Conferences: ICLR (2025, 2026), NeurIPS (2025)

**Organizer:** Mid-Atlantic Student Colloquium on Speech, Language and Learning (2024)

## Teaching

---

Head Teaching Assistant, JHU CS 601.471/671 NLP: Self-supervised Models

[\[Video 1\]](#)[\[Video 2\]](#)

Spring 2025

Course Assistant, JHU CS 601.471/671 NLP: Self-supervised Models

Spring 2024