

# Yixiao Song

✉ yixiaosong@umass.edu • 🌐 yixiao-song.github.io

5<sup>th</sup>-year linguistics Ph.D and CS master's student at University of Massachusetts Amherst. Experienced in interpretability and evaluation of NLP models and theoretical semantics.

## Education

### Ph.D Student in Linguistics

(Advised by Prof. Rajesh Bhatt & Prof. Mohit Iyyer)

University of Massachusetts Amherst

3.93/4.0

2019-current

### M.S. in Computer Science

University of Massachusetts Amherst

2023-current

### M.A. in Germanic Linguistics

University of Konstanz

1.1/6.0 (excellent)

2016-2018

### B.A. in German

University of Shanghai for Science and Technology

3.71/4.0

2011-2015

## Publications

### kNN-LM Does Not Improve Open-ended Text Generation

arXiv 2023

Shufan Wang, Yixiao Song, Andrew Drozdov, Aparna Garimella, Varun Manjunatha, and Mohit Iyyer

- Revealed that retrieval-augmented language models do not improve open-ended generation quality

### A Critical Evaluation of Evaluations for Long-form Question Answering

ACL 2023

Fangyuan Xu\*, Yixiao Song\*, Mohit Iyyer, and Eunsol Choi (\*Equal contribution)

- Comprehensively evaluated text generation metrics on long-form open-ended question answering generation
- First expert-annotated long-form question answering dataset

### Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense

arXiv 2023

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer

- Introduced a paraphrase generation model DIPPER which leverages context and offers diversity control
- Stress-tested and successfully evaded major AI-generated text detectors (e.g., watermarking, GPTZero)
- Proposed a simple but effective defense that relies on retrieving semantically-similar generations

### SLING: Sino Linguistic Evaluation of Large Language Models

EMNLP 2022

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer

- A benchmark with 38K minimal sentence pairs in Mandarin Chinese
- Tested 18 publicly available pretrained monolingual and multi-lingual language models
- Showed that the average accuracy for LMs is far below human performance (69.7% vs. 97.1%)
- Revealed the strengths and weaknesses of large language models

### DEMETR: Diagnosing Evaluation Metrics for Translation

EMNLP 2022

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer

- A diagnostic dataset with 31K English sentences (translated from 10 source languages)
- Evaluated the sensitivity of MT evaluation metrics to 35 different linguistic perturbations
- Found that learned metrics perform substantially better than string-based ones
- Revealed the strengths and weaknesses of learned metrics

## Works and Presentations

### Mandarin Chinese Alternative Questions are not Disjoined Polar Questions

Yixiao Song

2021

- First qualification paper supervised by Prof. Rajesh Bhatt and Prof. Seth Cable

### Early Cue Effects of Chinese Relative Clause Comprehension in Pre-trained Language Model

Yixiao Song

2021

- LING692C Cognitive Modeling, individual final project

### A Comparative Study of German and Chinese Alternative Questions

Yixiao Song

2018

- Poster presentation at Semantics and Philosophy in Europe
- Talk at the 17<sup>th</sup> China International Conference on Contemporary Linguistics

## Internships

---

- **Quillbot, Learneo, Inc.**  
*Platform for grammar correction, translation and text rewriting* June - September 2023  
Research Engineer Intern
- **Cultural Institute of the Federal Republic of Germany (Goethe Institut)**  
*Non-profit German cultural association, promoting the German language study abroad* August 2014 - February 2016  
Translator, interpreter, project assistant
- **Friedrich Ebert Stiftung Shanghai Office**  
*Non-profit German foundation funded by the Government of the Federal Republic of Germany* April-July 2014  
Translator, project assistant

## Positions of Responsibility

---

- **Instructor** of LING201 at UMass Amherst: How Language Works—Introduction to Linguistic Theory (R2)
- **Research Assistant** advised by Prof. Mohit Iyyer (Summer 2022): Human evaluation of model performance of long-form question answering.
- **Teaching Assistant** for courses at master level (University of Konstanz) and (under)graduate level (UMass Amherst)
- **Proceeding Editor** of NELS50 and SULA11
- **Student Research Assistant** in Deutsche Forschungsgemeinschaft Project—Questions at the Interfaces (P3 Alternative Questions)

## Skills

---

- **Natural Languages:** Shanghai Wu, Mandarin Chinese, English, German
- **Programming:** Python (PyTorch, Hugging Face), R, Linux, Perl
- **Others:**  $\text{\LaTeX}$