Wenzheng Zhang

Updated January 13, 2025

Email: wz283@cs.rutgers.edu **Homepage**: https://wenzhengzhang.github.io/

Research interests

Natural Language Processing, Large Language Models, Information Retrieval, Retrieval Augmented Generation (code at GitHub)

Education

Rutgers University — New Brunswick, NJ, US
PhD in Computer Science — Sept 2021 – Present
MS in Computer Science — Jan 2020 – June 2021
Advisor: Prof. Karl Stratos

University of Science and Technology of China — Hefei, China
BS in Applied Physics — Sept 2015 – June 2019

Honors and Awards

Outstanding Publications Award (Rutgers University) 2021

ICLR Spotlight Paper 2022

Publications

Wenzheng Zhang, Sam Wiseman, Karl Stratos. Seq2seq is All You Need for Coreference Resolution *In EMNLP 2023*

Wenzheng Zhang, Chenyan Xiong, Karl Stratos, Arnold Overwijk. Improving Multitask Retrieval by Promoting Task Specialization *In TACL 2023*

Wenzheng Zhang, Wenyue Hua, Karl Stratos. EntQA: Entity Linking as Question Answering. *In Proceedings of ICLR 2022 Spotlight*

Wenzheng Zhang, Karl Stratos. Understanding Hard Negatives in Noise Contrastive Estimation *In Proceedings of NAACL 2021*

Internships

Microsoft — Jun 2022 – Sep 2022
Applied Scientist Intern
Hosts: Prof. Chenyan Xiong and Arnold Overwijk

- Designed a novel multitask learning framework to enhance task specialization in multitask information retrieval.
- Leveraged optimized pretrained models, compatible prompting, and a novel adaptive learning method to ensure parameter specialization for individual tasks.
- Achieved state-of-the-art performance, surpassing task-specific retrievers on the KILT benchmark. Results published in **TACL 2023**.

Meta FAIR May 2024 – Dec 2024

Research Scientist Intern.

Hosts: Mingda Chen, Victoria Lin and Scott Yih

- Designed a novel model architecture for a Retrieval-Augmented Generation (RAG) system based on the Llama-3 model.
- Utilized the attention mechanism of the large language model (LLM) for retrieval and enhanced it to improve retrieval performance.
- Unified the retriever and generator components within a single LLM for RAG system.
- Explored various key-value state compression techniques to reduce disk usage and accelerate inference.
- Achieved strong performance on knowledge-intensive tasks, including Natural Questions (NQ).

Teaching experience

TA, Natural Language Processing(CS533), Rutgers University Spring 2023

TA, Natural Language Processing(CS533), Rutgers University Spring 2022

TA, Machine Learning (CS461), Rutgers University Fall 2021

Service

Reviewer, NAACL 2021.

Reviewer, EMNLP 2023.

Skills

Programming Languages

Python, Java, R, C

Toolkits

Pytorch, Git, Bash, Latex

Relevant Course Work

Natural Language Processing, Machine Learning, Computer Vision, Probability and Statistics, Artificial Intelligence, Algorithms