

Shengyun Si

+4901745917025 | shengyun.si@tum.de/shengyuninmunich@gmail.com|

EDUCATION

Technical University of Munich

Master of Robotics, Cognition, Intelligence

Zhejiang University

Bachelor of Automotive Engineering

Munich, Germany

2021.10-present

Hangzhou, China

2017.09-2021.7

RESEARCH AND INDUSTRY INTEREST

My long-term research goal is to develop a system capable of grounding and reasoning over world knowledge to enable safe and meaningful interactions with humans. Motivated by this goal, I am mainly interested in **Language Reasoning**, **Large Language Models Safety**, and **Controllable Text Generation**.

PUBLICATION

Haowei Z.*, **Shengyun S.***, Yilun Z., Linyong N., Xiangru T., Arman C.

OpenT2T: An Open-Source Toolkit for Table-to-Text Generation. EMNLP 2024

 [paper]

Yilun Z.*, Haowei Z.*, **Shengyun S.***, Linyong N., Xiangru T., Arman C.

Investigating Table-to-Text Generation Capabilities of Large Language Models in Real-World Information Seeking Scenarios. EMNLP 2023

 [paper]

EXPERIENCE

Data Scientist Intern

Allianz SE

October 2023 - present

Munich, Germany

- Design and implement RAG-based document answering pipelines using **Haystack** and **LLM**, aiding finance experts in streamlining their workflow.
- Developed a backend API using **FastAPI** and **Haystack** to integrate prototypes into a production environment, facilitating interactions with the frontend and vector databases. Managed the deployment of necessary services using **Kubernetes**
- Fine-tuned the **sentenceTransformer** based ranker and retriever models in the pipeline using a collected ground truth dataset to achieve higher accuracy. After fine-tuning, the models' recall on the validation set improved from **72%** to **90%**.
- Developed a command-line tool using **Poetry** and **Click** libraries to simplify the workflow for non-technical team members, enabling them to independently manage and execute business logic.
- Collaborate with a multidisciplinary team to understand financial domain-specific challenges, ensuring the developed tools were aptly tailored to user needs.
- Authored technical documentation to assist non-technical colleagues in understanding and utilizing command-line tools for tasks such as one-click training and deployment.

Research Assistant

Yale University LILY Lab

March 2023 - Sep 2023

Munich, Germany(remote)

- Research on how **LLMs**, like **GPT3.5/GPT4/LLaMa** perform on **Table2Text** task. If LLMs show great performance in this task, we further research on how fine-tuned model benefit from LLMs' strong table-to-text abilities.
- Fine-tune models with different table-to-text datasets **LogicNLG/WTQ/Totto/FetaQA**. Evaluate the faithfulness of generation based on different metrics. Prompt **LLMs** to generate faithful statements and give feedback to previous fine-tuned models' generation with CoT.
- Demonstrate **LLMs** can generate statements with higher faithfulness compared with previous state-of-the-art fine-tuned models. We also demonstrate that **LLMs** using chain-of-thought prompting can generate high-fidelity natural language feedback for other table-to-text models' generations, provide insights for future work regarding the distillation of text generation capabilities from **LLMs** to smaller models.

- Propose **OpenT2T**, the first open-source toolkit for table-to-text generation, designed to reproduce existing table pre-training models for performance comparison and expedite the development of new models. We have implemented and compared 7 fine-tuned models as well as 44 large language models under zero- and few-shot settings on 9 table-to-text generation datasets, covering data insight generation, table summarization, and free-form table question answering
- Our work has been accepted by the **EMNLP2023**, **EMNLP2024**.
- <https://aclanthology.org/2023.emnlp-industry.17/>
- <https://aclanthology.org/2024.emnlp-demo.27/>

NLP Research Intern

June 2023 - Sep 2023

Siemens AG

Munich, Germany

- Utilizing the latest methods in natural language processing, review scholarly articles and propose recommendations for practical implementation.
- Fine-tune pre-trained models on industrial data, for example the Requirement-Model alignment data which is generated during the traces of industrial production. Use the fine-tuned model to serve for the upcoming raw data.

SELECTED AWARDS

- **2021:** Outstanding Graduates at Zhejiang University
- **2019:** Second-grade Scholarship at Zhejiang University
- **2018:** National Encouragement Scholarship at Zhejiang University

PROJECTS

Attack on Unfair ToS Clause Detection (Adverse Robustness via Prompts)

March 2023 - Jul 2023

- Research on if **Model-tuning Via prompts** have better robustness against adversarial attacks in legal data area.
- Fine-tune **Legal-Bert** with **Terms of Service** data and attack the model with **Textbugger** and **Textfooler** attack methods. Prompts-tune **Legal-Bert** with different templates and verbalizers and also do the same attack on the tuned model
- Use **PGD** adversarial training to defense the model and test the trained model's robustness on the same attack methodologies.
- Results show that **Model-tuning Via prompts** model prompted on legal data has great robustness than model with traditional fine-tuning approach.

Large-scale Protein Embedding Analysis with Hadoop and Spark

October 2022 - Jan 2023

- Deploying **Hadoop** and **HDFS** in the remote server first, and utilizing **Spark MLlib** to successfully cluster protein embeddings, resulting in an effective model that grouped similar proteins together based on their characteristics.
- Conducted extensive pre-processing of the protein embeddings to ensure compatibility with the Spark MLlib framework and performed parameter tuning and testing to optimize the model's performance.
- Gained valuable experience working with large-scale data processing and machine learning technologies, like **K-means** and its variations, and honed problem-solving and analytical skills.
- This experience has equipped me with a solid understanding of developing models for the analysis of complex biological data, as well as with the ability to troubleshoot issues and refine models for improved accuracy and efficiency.

TECHNICAL SKILLS

Languages: English(business-fluent), German(B2), Chinese(Native)

Frameworks: Tensorflow, Pytorch, NumPy, Scikit-learn, Matplotlib, Docker

Big data technologies: Hadoop, Spark, Dask, MLlib, Apache OpenNLP

Programming Language: Python, C++/C, Java, Scala