Shihao LIANG

Date of birth: 28/08/1999

Residence: Zhongshan City, Guangdong Province

E-mail: shihaoliang0828@gmail.com | *Telephone:* 17627812352 | site

Education

B.Eng in Computer Science and Technology — *Tianjin University*
GPA: **3.49**/4.0 — *Sep. 2018 - Jun. 2022*
Average Score: **86.0**/100
Average Score for the third and the fourth academic year: **89.3**/100 and **88.1**/100

M.Eng in Computer Science — *HongKong University*
Incoming student — *Enroll in 2023 Fall*

Work experience

THUNLP Lab, Tsinghua University — *Jul. 2022 - Apr. 2023*
Research Assistant — *Beijing, China*

- **Question Answering:** Generate high-quality and large-scale QA data with an iterative bootstrapping framework. Assist in building the first Chinese Long-form question answering dataset with human web search behaviors.
- **Instruction Tuning:** Explore how different instruction formats affect instruction tuning and unify different formats into our proposed format.
- **Tool Learning:** Assist in BMTools project. Explore tool learning with foundation models and reinforce learning.

NLP Department, Baidu — *Dec. 2021 - Apr. 2022*
Research and Development Intern — *Beijing, China*

- **Deep Question-Answering:** Optimize the first search result of Baidu search engine. Given a query, use cross-encoder to re-rank paragraphs from topk websites. Improve the ranking model's robustness with unsupervised strong negatives.
- **Model Distillation:** Assist in the application of question-answering technology in search engine in industry. Distill Ernie2.0 into a 2-layer encoder model with minimal performance difference.
- **Paper Reproduction:** Reproduce papers with Paddlepaddle in computer vision and natural language processing.

Publications

Under review, first author	*QASnowball: An Iterative Bootstrapping Framework for High-Quality Question-Answering Data Generation*
Under review, co-author	*Interactive Web Search for Chinese Long-form Question Answering*
Under preparation, first author	*Unified Instruction Tuning*
Arxiv, co-author	*Tool Learning With Foundation Models*

Projects

BMTools	**An open-source repository** on Github that extends **language models using tools** and serves as a platform for the community to build and share tools. Develop the translation tools in the platform.
WebCPM	**The first Chinese long-form question answering dataset**, with retrieved information based on interactive web search. Collected 5500 question-answer pairs, together with the supporting facts and human behaviors, with web search behaviors recorded. QA Pipeline launched on Zhihu.
QASnowball	**Generate high-quality and large-scale QA data** continually with an iterative bootstrapping framework. Equip CPM-Live with question-answering ability by pre-training with the auto-generated data.

Language proficiencies

IELTS Academic	7.0
Cantonese	Native speaker

Awards and Honors

2019	Merit Student in Social Practice of TJU
2020	Scholarship for Academic Progress from TJU
2020	Advanced Individual in Academic Progress of TJU
2021	First Prize in Paper Reproduction Challenge of China Society of Image and Graphics