Qingyi Si (佀庆一)

☎ (+86)131-6168-5288 | 1997.12.01 | ✉ siqingyi@iie.ac.cn | ⌂ https://phoebussi.github.io/ | https://github.com/PhoebusSi | Beijing, China | **Research Interests**: vision & language, LLMs and OOD generalization. | **Google Scholar**: https://scholar.google.com/citations?user=OhNtMsYAAAAJhl=zh-CN

Education

University of Chinese Academy of Sciences (UCAS), Institute of Information Engineering *Beijing*
PhD in Cyberspace Security and **MS** in Computer Applied Technology, advised by Prof. Zheng Lin and Weiping Wang. Sep. 2019 - Jun. 2024

Beijing Language and Culture University (BLCU) *Beijing*
BS in Computer Science and Technology Sep. 2015 - Jun. 2019

Publications (Co-*) First author of 10 papers

[1] **Qingyi Si**, Yuchen Mo, Zheng Lin, et al. "Combo of Thinking and Observing for Outside-Knowledge VQA". ***ACL'23***

[2] **Qingyi Si**, Yuanxin Liu et al. "Compressing And Debiasing Vision-Language Pre-Trained Models for VQA". ***EMNLP'23***

[3] **Qingyi Si**, Tong Wang et al. "An Empirical Study of Instruction-tuning Large Language Models in Chinese". ***EMNLP'23***

[4] **Qingyi Si**, Fandong Meng, et al. "Language Prior Is Not the Only Shortcut: A Benchmark for Shortcut Learning in VQA". ***EMNLP'22***
*ps: This paper received **high praise** from **Damien Teney**.*

[5] **Qingyi Si**, Fandong Meng, et al. "Towards Robust VQA: Making the Most of Biased Samples via Contrastive Learning". ***EMNLP'22***

[6] **Qingyi Si**, Zheng Lin, et al. "Check It Again: Progressive Visual Question Answering via Visual Entailment". ***ACL'21***

[7] **Qingyi Si**, Yuanxin Liu, et al. "Learning Class-Transductive Intent Representations for Zero-shot Intent Detection". ***IJCAI'21***

[8] Zhe Wen*, **Qingyi Si***, Zheng Lin, et al. "Schema Item Matters in Knowledge Base Question Answering". *IJCNN'23*

Under Review Papers

[9] Huishan Ji*, **Qingyi Si***, Zheng Lin, et al. "Towards Many-to-one Visual Question Answering". ***AAAI'24***

[10] Jie Xu*, **Qingyi Si***, Hanbo Zhang*, et al. "Towards Unified Interactive Visual Grounding in The Wild". ***ICRA'24***

Co-author Papers

[11] Peize Li, **Qingyi Si**, et al. "Cross-modality Multiple Relations Learning for Knowledge-Based Visual Question Answering". ***TOMM'23***

[12] Duo Zheng, Fandong Meng, **Qingyi Si**, et al. "Visual Dialog for Spotting the Differences between Pairs of Similar Images". ***MM'22***

[13] Yeqi Sun, **Qingyi Si**, et al. "Outside-knowledge Visual Question Answering for Visual Impaired People". *MedAI'23*

[14] Ran Li, **Qingyi Si**, et al. "A Multi-channel Neural Network for Imbalanced Emotion Recognition". *ICTAI'19*

[15] Chunhua Liu, Yan Zhao, **Qingyi Si**, et al. "Multi-perspective fusion network for commonsense reading comprehension".*CCL'18*

Research Projects

Alpaca-CoT (***https://github.com/PhoebusSi/alpaca-CoT/***) ***2.1k Star***

An Instruction-tuning Platform with Unified Interface of Instruction Collection, Parameter-efficient Methods, and Large Language Models. *(I was invited by Machine Heart (jiqizhixin) to give a talk about this project.)* ***2.9k Downloads last month***

We unified the interfaces of instruction-tuning data, multiple LLMs and parameter-efficient methods (e.g., lora, p-tuning) together for easy use. Besides, we are the first to extend CoT data into LLaMA to improve its reasoning ability. We constructed the largest collection of instruction tuning datasets currently available. On this basis, we conduct a thorough empirical study of instruction tuning in Chinese.

A Multimodal Large Language Model ***In Progress***

We empower LLMs with multi-modality understanding, image generation, and visual grounding capabilities.

Awards and Honors

[1] **ZhuLiYueHua Scholarship for Excellent Doctoral Student (Top 1%, RMB ¥ 5,000).** *Bureau of Personnel, CAS. 2023.*

[2] **National Scholarship for Doctoral students (Top 2%, RMB ¥ 30,000).** *Ministry of Education of P.R. China. 2021.*

[3] **Merit Student, University of Chinese Academy of Sciences.** *UCAS. 2020, 2021.*

[4] **Excellent Undergraduate Graduation Project (Thesis) (Top 0.7%).** *Education Commission of Beijing. 2019.*

[5] **Regular Institutions of Higher Education Outstanding Graduate.** *Education Commission of Beijing. 2019.*

[6] **National Scholarship for Undergraduates (Top 2%, RMB ¥ 8,000).** *Ministry of Education of P.R. China. 2017, 2018.*

Working Experience

Microsoft Research Asia (MSRA) *Beijing*

Research Intern in NLC group. Research in the mixture of experts (MoE) implementation for large NMT models. November.2021 - May.2022

WeChat, Tencent *Beijing*

Research Intern in the Pattern Recognition Center (PRC), WeChat AI. Research in the OOD robustness of VQA. March.2021 - November.2021

Research Review

Language & Vision

OOD robustness of VQA

[1] *ACL'21*: **Overcoming the OOD problem in VQA.** Specifically, we propose a select-and-rerank (SAR) progressive framework formalizing VQA as Visual Entailment, which can make full use of the interactive information of image, question and candidate answers. This paper constructs a new state-of-the-art accuracy with large margins on VQA-CP.

[2] *EMNLP'22*: **Overcoming the OOD problem while maintaining in-distribution performance.** Specifically, instead of undermining the importance of the biased samples like the previous works, our method makes the most of them based on a contrstive learning method. This paper is the first to break the trade-off between OOD performance and IID performance of VQA models.

[3] *EMNLP'22*: **Building a more reliable and comprehensive testbed for OOD robustness in VQA.** We verify that the vast majority of existing debiasing methods are dataset-specific rather than truly robust. We construct different OOD test sets, based on a range of VQA elements, existing debiasing methods fail to simultaneously generalize to.

[4] *EMNLP'23*: **Searching sparse and robust subnetworks from visual-language pretrained models.** We systematically study the design of a debiasing and pruning pipeline. The obtained sparse and robust subnetworks clearly outperform the SoTAs with fewer parameters.

Open-domain VQA

[1] *ACL'23*: **Extending the multimodal encoder into LLMs.** Specifically, we mimic human behavior, ”thinking while observing”, i.e., benefiting from the vast knowledge in LLMs while making the most of the visual features for better image understanding. The proposed multimodal LLM constructs a new state-of-the-art accuracy with large margins on outside-knowledge VQA.

[2] *AAAI'24*: **Towards one-to-many VQA.** Specifically, we collect 12 VQA tasks to empirically explore how to balance the three ability of real open-domain VQA models: knowledge capacity, visual attribute recognition and scene comprehension.

Large Language Models

Instruction-tuning

[1] *Alpaca-CoT*: **An instruction-tuning platform** with unified interface of instruction collection, parameter-efficient methods, and LLMs.

[2] *EMNLP'23*: **Empirical study on instruction tuning open LLMs in Chinese.** Specifically, we systematically explore the impact of a range of LLM bases, parameter-efficient methods, instruction data types, which are the three most important elements for instruction-tuning. Besides, we also conduct experiment to study the impact of other factors, e.g., chain-of-thought data, language of prompt, and human-value alignment.

Multimodal LLMs (MLLM)

[1] *ICRA'24*: **A multimodal LLM unifying visual dialog and visual grounding.** We propose an interactive visual-language Transformer (MLLM) that can hold a natural and informative dialog with human users to disambiguate their expression and identify the target object, which can also generalize to real-world products, e.g., household robot.

[2] *ACL'23*: **A multimodal LLM bringing in more comprehensive types of knowledge.** Specifically, we mimic human behavior, ”thinking while observing”, i.e., benefiting from the vast knowledge in LLMs while making the most of the visual features for better image understanding. The proposed multimodal LLM constructs a new state-of-the-art accuracy with large margins on outside-knowledge VQA.

[3] *In process*: **A multimodal LLM can handle both image and text input and image and text generation simultaneously.**

Self-Assessment

- Outgoing, energetic and a nice personality.
- Creative, self-motivated and a passion for problem-solving.
- Strong team work, but also ability to work independently.