ZIWEI JI

Email: zjiad@connect.ust.hk ⋄ Website: ziweiji.github.io ⋄ Phone: +86-15927277932

EDUCATION

The Hong Kong University of Science and Technology Hong Kong SAR
Ph.D. Candidate in Electronic and Computer Engineering *Sept. 2019 - Present (Expected in Spring 2025)*
Supervisor: Pascale Fung; Research Topic: Hallucination in NLG, NLP

Huazhong University of Science and Technology Wuhan, China
B.Sc. in Electronic Science and Technology *Sept. 2015 - Jun. 2019*
GPA: 3.97/4.0, Top 1%, Graduate with Honors

SELECTED AWARDS

Area Chair Award (Language Modeling and Analysis) at IJCNLP-AACL	2023
Silver medal (Top 2%) in Kaggle Competition: Stable Diffusion - Image to Prompts	2023
National Scholarship (3 times, Top 0.2%), Ministry of Education of P.R.China	2016, 2017, 2018
Second Prize of Hubei Province in National Undergraduate Mathematics Competition	2017
Outstanding Overseas Exchange Undergraduate in UC Berkeley Hi-Tech Program	2017
Outstanding Scientific Research Achievement Award for University Students in Hubei Province	2018
Outstanding Undergraduate in Terms of Academic Performance (Top 1%)	2016

WORK EXPERIENCE

Shanghai Artificial Intelligence Laboratory Shanghai, China
Applied Scientist Intern *Jul. 2023 - Jan. 2024*

- Build an analytical hallucination annotation dataset in Large Language Models. Employing the dataset, we train hallucination annotators based on InternLM-7B/20B. Accepted in ACL 2024.
- Scale analytical hallucination annotation progressively and improve the accuracy of the annotator with an iterative self-training framework. And apply the annotator for hallucination mitigation. Submitted to NeurIPS 2024.
- Mentor: Wenwei Zhang (OpenMMLab)

SELECTED PROJECT

Building Analytical Annotation of Hallucinations in Large Language Models *Jul. 2023 - Jan. 2024*

- We establish a Chinese-English dataset offering Analytical Annotation of Hallucinations in LLMs. The dataset consists of $\sim$12k sentence-level annotations covering over 700 topics.
- Due to the fine granularity, we quantitatively confirm that the hallucinations progressively accumulate in the answer.
- Employing the dataset, we train hallucination annotators based on InternLM-7B and InternLM-20B, obtaining performance competitive with GPT-4.

Scaling Analytical Hallucination Annotation of Large Language Models *Jan. 2024 - May. 2024*

- Current hallucination detection and mitigation datasets are limited in domains and sizes and struggle to scale due to prohibitive labor costs and insufficient reliability of existing annotators.
- We introduce an iterative self-training framework that simultaneously and progressively scales up the hallucination annotation dataset and improves the accuracy of the hallucination annotator.
- Based on the Expectation Maximization (EM) algorithm, in each iteration, the framework first annotates data scaled in multi-dimension with a self-consistency strategy. Then a more accurate annotator is trained on the data.
- Our final dataset consists of $\sim$822k annotations. Our final annotator with only 7B parameters surpasses GPT-4 and obtains new SOTA on HaluEval and HalluQA by zero-shot inference. The annotator also helps to mitigate the hallucination, with the NLI metric increasing from 25% to 37% on HaluEval.

Mitigating Hallucination in Large Language Models via Self-Reflection *Feb. 2023 - Jun. 2023*

- LLMs are still prone to generate hallucinations, *i.e.* plausible-sounding but unfaithful or nonsensical information, in generative and knowledge-intensive tasks including QA.
- We analyze the hallucination phenomenon in medical generative QA systems using widely adopted LLMs (Vicuna, Alpaca-LoRA, ChatGPT, MedAlpaca, Robin-medical) and datasets (PubMedQA, MedQuAD, MEDIQA2019, LiveMedQA2017, MASH-QA) Our investigation centers on the identification and comprehension of common problematic answers, with a specific emphasis on hallucination.
- To tackle this challenge, we present an interactive self-reflection methodology that incorporates knowledge acquisition and answer generation. Our approach steadily enhances the factuality, consistency, and entailment of the generated answers. Consequently, we harness the interactivity and multitasking ability of LLMs and produce progressively more precise and accurate answers.
- Experimental results on both automatic and human evaluation demonstrate the superiority of our approach in hallucination reduction compared to baselines. Published in EMNLP 2023 Findings.

Reducing Hallucination in Open-domain Dialogues with Knowledge Graph Grounding *Aug. 2022 - Jan. 2023*

- Dialogue systems are still prone to produce hallucinated responses not supported by the input source. We adopt the dataset OpenDialKG containing 15k open-domain KG-grounded dialogues to explore the problem.
- To handle the heterogeneity between external knowledge and dialogue context and generate more faithful responses, we propose RHO with 1) Local Knowledge Grounding combining textual embeddings with the corresponding KG embeddings. 2) Global Knowledge Grounding via the attention mechanism for multi-hop reasoning abilities. 3) A response re-ranking technique based on walks over KG sub-graphs for better conversational reasoning.
- Experimental results show that our approach significantly outperforms SOTA on both automatic and human evaluation by a large margin, especially in hallucination reduction (17.54% in FeQA). Published in ACL 2023 Findings.

AI Film *Feb. 2021 - Feb. 2022*

- In order to offer a customized film tool and inspire professional filmmakers, we present an automatic, real-time film-producing system cooperating with the Central Academy of Fine Arts.
- We adopt a hierarchical structure, which first generates the plot, then the script and its visual presentation: 1) Design a genre-controllable and plot-guided film script generation system. 2) Collect a video database from social media and retrieve video clips based on the scripts. 3) Develop a user interface for demonstration.
- The experiment results show that our approach outperforms the baselines on both automatic and human evaluations, especially in genre control.
- Exhibited at Pingyao International Film Festival and Xu Bing's Language Art Exhibition. Published in AACL Demo.

SELECTED PUBLICATIONS

ANAH: Analytical Annotation of Hallucinations in Large Language Models ACL 2024
Ziwei Ji, Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen

Towards Mitigating Hallucination in Large Language Models via Self-Reflection EMNLP 2023 Findings
Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Pascale Fung, et al.

RHO(ρ): Reducing Hallucination in Open-domain Dialogues with Knowledge Grounding ACL 2023 Findings
Ziwei Ji, Zihan Liu, Nayeon Lee, Tiezheng Yu, Pascale Fung, et al.

Survey of Hallucination in Natural Language Generation ACM Computing Surveys 2022
Ziwei Ji, Nayeon Lee, Rita Frieske, Pascale Fung, et al. Get 2100+ citations

VScript: Controllable Script Generation with Visual Presentation AACL Demo 2022
Ziwei Ji, Yan Xu, I-Tsun Cheng, Pascale Fung, et al.

Plausible May Not Be Faithful: Probing Object Hallucination in Vision-Language Pre-training EACL 2023
Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, Pascale Fung

Diverse and Faithful Knowledge-Grounded Dialogue Generation via Sequential Posterior Inference ICML 2023
Yan Xu, Deqian Kong, Dehong Xu, Ziwei Ji, Bo Pang, Pascale Fung, Yingnian Wu

SKILLS AND OTHERS

Sub-Tasks	Have experience in Question Answering, Dialogue Generation, Image Captioning, NER, Storytelling, Question Generation, Fake News Detection
Academic Service	Reviewer in EMNLP and ACL
Programming Language	Python, C, Java, JavaScript, MATLAB
Skills	Pytorch, TensorFlow, Slurm, DeepSpeed, Linux, Git, SVN
Languages	Chinese (Mother Tongue), English (Full-Proficiency, IELTS 7)