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, China
Ph.D. Candidate in ECE Department (NLP, Supervisor: Pascale Fung) — Sept. 2019 - Present

Huazhong University of Science and Technology — Wuhan, China
B.Sc. in Electronic Science and Technology (GPA: 3.97/4.0, Top 1%, Honored) — Sept. 2015 - Jun. 2019

SELECTED AWARDS

National Scholarship (3 times, Top 0.2%), Ministry of Education of P.R.China — 2016, 2017, 2018
Area Chair Award (Language Modeling and Analysis) at IJCNLP-AACL — 2023
Silver medal (Top 2%) in Kaggle Competition: Stable Diffusion - Image to Prompts — 2023
Second Prize of Hubei Province in National Undergraduate Mathematics Competition — 2017
Merit Student at Huazhong University of Science and Technology (3 times, Top 3%) — 2016, 2017, 2018
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

Applied Scientist Intern — Jul. 2023 - Present
Shanghai AI Lab — Shanghai, China

- Evaluate and mitigate hallucination in large language models.
- Mentor: Wenwei Zhang

PROJECT

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

- Large language models (LLMs) have shown promise for generative and knowledge-intensive tasks including question-answering (QA) tasks. However, the practical deployment still faces hallucination, plausible-sounding but unfaithful or nonsensical information.
- Analyse the phenomenon of hallucination 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. Through this feedback process, 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.

Reducing Hallucination in Open-domain Dialogues — Aug. 2022 - Jan. 2023

- Dialogue systems can leverage large pre-trained language models and knowledge to generate fluent and informative responses. However, these models are still prone to produce hallucinated responses not supported by the input source. We adopt the dataset OpenDialKG containing 15k open-domain dialogues grounded on the knowledge graph (KG).

- 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).

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.

Probing Object Hallucination in Vision-Language Pre-training Sept. 2022 - Jan. 2023

- Large-scale vision-language pre-trained (VLP) models are prone to hallucinate non-existent visual objects when generating text based on visual information.
- We systematically study the object hallucination problem: 1) Examine recent SOTA VLP models, showing that they still hallucinate frequently, and models achieving better scores on standard metrics (e.g., CIDEr) could be more unfaithful. 2) Investigate how different types of image encoding in VLP influence hallucination, including region-based, grid-based, and patch-based. We find that patch-based features perform the best and smaller patch resolution yields a non-trivial reduction in object hallucination. 3) Decouple various VLP objectives and demonstrate that token-level image-text alignment and controlled generation are crucial to reducing hallucinations. Based on that, we propose a simple yet effective VLP loss named ObjMLM to further mitigate object hallucination.
- Experimental results show that our approach reduces object hallucination by up to 17.4% when tested on COCO Caption for in-domain and NoCaps for out-of-domain evaluation.

A Multitask, Multilingual, Multimodal Evaluation of ChatGPT Jan. 2023 - Feb. 2023

- We evaluate the multitask, multilingual and multimodal aspects of ChatGPT using 23 datasets covering 8 different common NLP application tasks.
- We find that ChatGPT outperforms LLMs with zero-shot learning on most tasks and even outperforms fine-tuned models on some tasks. It is better at understanding non-Latin script languages than generating them. It is able to generate multimodal content from textual prompts, via an intermediate code generation step.
- ChatGPT is 63.41% accurate on average in 10 different reasoning categories under logical reasoning, non-textual reasoning, and commonsense reasoning, hence making it an unreliable reasoner. It suffers from hallucination problems like other LLMs and it generates more extrinsic hallucinations from its parametric memory. The interactive feature of ChatGPT enables human collaboration with the underlying LLM to improve its performance.

SELECTED PUBLICATIONS

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 954 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

Model Generalization on COVID-19 Fake News Detection AAAI CONSTRAINT 2021
Ziwei Ji[1], Yejin Bang[1], Etsuko Ishii[1], Samuel Cahyawijaya[1], Pascale Fung

SKILLS AND OTHERS

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

[1]Equal Contribution