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 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 Scientific Research Achievement Award for University Students in Hubei Province	2018
Outstanding Undergraduate in Terms of Academic Performance (Top 1%)	2016

WORK EXPERIENCE

Meta FAIR Paris, France
Research Scientist Intern *Nov. 2024 - Apr. 2025*

- Discover linear "verbal uncertainty" feature in Large Language Model (LLM) representation space and calibrate via inference-time intervention to reduce hallucinations by 32%. (Submitted to ACL 2025)
- Build a dynamic hallucination benchmark to evaluate LLMs on factual questions, both short and long-form, and evaluate their ability to abstain from answering questions about non-existent entities. (Submitted to ACL 2025)
- Build a video benchmark for high-level world modeling and long-horizon procedural planning. Given initial and final states, the task is to distinguish the properly ordered action sequence in different contexts. (Submitted to ICCV 2025)
- Build a Vision Language World Model that predicts procedural video planning given the goal.
- Mentor: Pascale Fung, Nicola Cancedda

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

- Build an analytical hallucination annotation dataset in LLMs. Employing the dataset, we train hallucination annotators based on InternLM-7B/20B. (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. (NeurIPS 2024)
- Mentor: Wenwei Zhang

SELECTED PUBLICATIONS

ANAH: Analytical Annotation of Hallucinations in Large Language Models ACL 2024
Ziwei Ji[1], Yuzhe Gu[1], Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen

ANAH-v2: Scaling Analytical Hallucination Annotation of Large Language Models NeurIPS 2024
Yuzhe Gu[1], Ziwei Ji[1], Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen

[1] Equal Contribution

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 3500+ citations

LLM Internal States Reveal Hallucination Risk Faced With a Query EMNLP 2024 Blackbox
Ziwei Ji, Delong Chen, Etsuko Ishii, Pascale Fung et al.

Calibrating Verbal Uncertainty as a Linear Feature to Reduce Hallucinations ACL 2025 (Under Review)
Ziwei Ji, Lei Yu, Yeskendir Koishekenov, Nicola Cancedda et al.

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

SELECTED PROJECT

Building and Scaling Analytical Annotation of Hallucinations in LLMs *Jul. 2023 - May. 2024*

- Detecting hallucinations in LLMs is increasingly challenging due to their fluent and convincing responses. Current datasets often label entire responses as hallucinations without explanations or references, hindering trigger identification and mitigation. Additionally, existing datasets are limited in domain and size, struggling to scale due to high labor costs and unreliable automatic annotators.
- We establish ANAH, a Chinese-English dataset offering sentence-level Analytical Annotation of Hallucinations in LLMs. Using ANAH, we train hallucination annotators based on InternLM-7B and InternLM-20B.
- We introduce an iterative self-training framework that progressively scales up the ANAH dataset and improves annotator accuracy. 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, ANAH-v2, expands from ~12k to ~822k annotations. Our 7B parameter annotator surpasses GPT-4, achieving new SOTA on HaluEval and HalluQA by zero-shot inference. It also mitigates hallucinations, increasing the NLI metric from 25% to 37% on HaluEval.
- Published in ACL 2024 and NeurIPS 2024.

Mitigating Hallucination in LLMs via Self-Reflection *Feb. 2023 - Jun. 2023*

- LLMs are prone to generating hallucinations, *i.e.*, plausible-sounding but unfaithful or nonsensical information, in generative and knowledge-intensive tasks like QA.
- We analyze hallucinations in medical generative QA systems using LLMs (Vicuna, Alpaca-LoRA, ChatGPT, MedAlpaca, Robin-medical) and datasets (PubMedQA, MedQuAD, MEDIQA2019, LiveMedQA2017, MASH-QA), focusing on identifying and understanding common problematic answers.
- We present an interactive self-reflection methodology incorporating knowledge acquisition and answer generation, enhancing the factuality, consistency, and entailment of generated answers.
- Experimental results 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 often produce hallucinated responses not supported by the input source. We used the OpenDialKG dataset, containing 15k open-domain KG-grounded dialogues, to explore this problem.
- We propose RHO, which includes: 1) Local Knowledge Grounding combining textual embeddings with KG embeddings, 2) Global Knowledge Grounding via attention mechanisms for multi-hop reasoning, and 3) A response re-ranking technique based on walks over KG sub-graphs for better conversational reasoning.
- Experimental results show our approach significantly outperforms SOTA in both automatic and human evaluations, especially in hallucination reduction (17.54% in FeQA). Published in ACL 2023 Findings.

AI Film *Feb. 2021 - Feb. 2022*

- To offer a customized film tool and inspire professional filmmakers, we developed an automatic, real-time film-producing system in collaboration with the Central Academy of Fine Arts.
- We adopt a hierarchical structure to generate the plot, script, and visual presentation: 1) A genre-controllable and plot-guided film script generation system, 2) A video database from social media for script-based video retrieval, and 3) A user interface for demonstration.
- Experimental results show our approach outperforms baselines in both automatic and human evaluations, particularly in genre control.
- Exhibited at Pingyao International Film Festival, Xu Bing's Language Art Exhibition, and published in AACL 2022.

SKILLS AND OTHERS

Sub-Tasks	Have experience in Textual and Visual Question Answering, Dialogue Generation, Image Captioning, Video Procedural Planning, Named Entity Recognition, 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)