

다중 LLM 교차 검증을 통한 선행 연구 검색 및 분석에서의 AI 환각 완화에 관한 연구

정한민

한국과학기술정보연구원, 과학기술연합대학원대학교

jhm@kisti.re.kr

Study on AI Hallucination Mitigation in Prior-Literature Search and Analysis Using Cross-validation across Multiple LLMs

Hanmin Jung

Korea Institute of Science and Technology Information, University of Science and Technology

요약

생성형 AI의 환각 현상은 R&D 및 학술 영역에서 심각한 왜곡을 초래할 수 있어 그 탐지와 완화가 중요한 이슈이다. 본 연구는 다중 LLM의 교차 검증 구조를 기반으로, 선행 연구 검색 및 분석 과정에서 환각을 효과적으로 발견하고 제거하는 다단계 프롬프트 실험을 설계하였다. 실험 결과, ChatGPT-4o와 Grok-3 DeepSearch/Think 간의 교차 검증을 통해 환각이 발생한 선행 연구를 성공적으로 식별하였으며, 두 LLM이 완전히 동의한 결과는 환각 없이 신뢰할 수 있는 정보를 제공하는 데 기여함을 확인하였다.

I. 서론

AI 환각(Hallucination) 현상은 생성형 AI가 사실과 무관하거나 근거 없는 정보를 마치 사실처럼 만들어내는 현상으로, R&D 및 학술 영역에서 잘못된 정보가 포함된 요약, 문헌 생성, 또는 인용으로 인해 후속 연구 왜곡, 정책 결정 오류로 인해 실질적 피해가 발생할 수 있기 때문에 중요하게 다루어야 한다. LLM을 활용한 자동화된 문헌 분석이나 요약 도구가 빠르게 확산함에 따라, 연구자가 AI가 생성한 내용을 비판 없이 수용할 경우, 환각이 발생한 정보가 검증되지 않은 채 학술 생태계에 퍼지는 '2차 환각 전파'의 위험성도 증가할 수밖에 없다. 따라서 AI 환각을 탐지하고 완화하는 기술은 학술의 정확성과 연구 윤리를 보장하는 핵심 인프라로서 필요성이 크다. 특히, 참고문헌의 구조는 일정한 포맷(저자, 제목, 연도 등)을 따르기 때문에, 모델은 그럴듯한 형식을 자동으로 생성할 수 있지만, 실제 존재 여부에 대한 근거가 부족하므로 논문이나 책 제목을 그럴듯하게 만들어내는 경향이 있다[1-2].

환각을 줄이기 위한 연구가 활발하게 진행되고 있는데, [3]은 LLaMA-v2의 텍스트 생성 과정에서 발생하는 환각 현상을 시각적인 힌트를 통해 완화하는 방법을, [4]는 Nltk와 Rouge_scorer를 사용하여 QA 데이터의 성능을 정량적으로 분석하는 방식으로 결과의 정확도를 올리는 방법을 제안한다. 그렇지만, 이러한 방법들은 하나의 LLM에만 의존함으로써 학습 과정에서 원천적으로 완벽하지 않은 상황에 대처가 쉽지 않으며, 모델의 편향성을 해결할 방법이 없다. 이에 본 연구는 다중 LLM에 기반하여 교차 검증(Cross-validation)하는 방식으로 빈번하게 환각이 발생하는 선행 연구 검색 이슈를 해결하고자 한다.

II. 본론

본 연구는 R&D 프로세스 내 '사전 단계' 중 '과거(선행) 연구 리뷰'를 효율적으로 수행하기 위한 목표 아래 다중 LLM의 교차 검증을 통해 환각을 최소화하는 방식으로 원하는 선행 연구를 정확하고 신속하게 검색과 분석하는 방안을 제시한다[5-6]. 실험은 "다중 LLM이 선행 연구 검색에 있어서 환각 현상을 줄여줄 수 있는가?" "다중 LLM이 선행 연구 내용에 대해 정확히 분석하는 데 도움을 줄 수 있는가?"라는 질문들에

대한 답을 얻기 위해 진행하였다. 본 실험을 위해, 4개의 프롬프트와 3개의 LLM(ChatGPT-4o, Grok-3 DeepSearch, Grok-3 Think)을 이용한다(표 1 참조). 첫 번째, 1번 프롬프트와 ChatGPT-4o를 이용하여 특정 조건에 맞는 선행 논문 10편을 검색하였다(표 2 참조). 두 번째, 2번 프롬프트와 Grok-3 DeepSearch를 이용하여 검색된 10편에 대해 교차 검증을 수행하여 환각이 발생했다고 의심되는 논문들을 제거하였다(표 2 참조). 세 번째, 3번 프롬프트와 ChatGPT-4o를 이용하여 검색된 10편에 대해 분석 과정을 수행하여 주요 기여점과 키워드 목록을 얻었다(표 3 참조). 네 번째, 4번 프롬프트와 Grok-3 Think를 이용하여 분석된 10편에 대해 교차 검증을 수행하여 환각 여부를 확인하였다.

표 1. 실험에 사용된 프롬프트와 LLM 목록

No.	Prompt	LLM
1	List 10 papers published in SCI(E) journals since 2023 that deal with the phenomenon of 'hallucinations in LLM', in order of citation index. List the author, paper title, source, year of publication, and citation index. If you disagree, please specify the reason along with the paper number.	ChatGPT-4o
2	The following papers are the top 10 papers published in SCI(E) journals since 2023 that deal with the 'hallucinations of LLM', listed in order of citation index as suggested by ChatGPT-4o. Please review these papers to verify whether the sources actually exist, whether the sources are accurate, and whether the results align with the specified conditions (such as topic and publication year). Then, select only those papers that you fully agree with. + [Results of 1]	Grok-3 DeepSearch
3	Describe the main contributions of each of the 10 papers in one paragraph and select the five most important keywords for each.	ChatGPT-4o
4	The following is ChatGPT-4o's summary of the main contributions of 10 papers in one paragraph each, along with the five most important keywords for each. Please indicate which ones you fully agree with, partially agree with, or disagree with, and explain your reasons. + [Results of 3]	Grok-3 Think

표 2. 1번과 2번 프롬프트로부터 얻어진 LLM 결과

No.	Title	Publication	Year	Citation Index	Agreement	Reason for Disagreement
1	Survey of Hallucination in Natural Language Generation	ACM Computing Surveys	2023	~310	Fully Agreed	-
2	Siren's Song in the AI Ocean: A Survey on Hallucination in LLMs ¹⁾	ArXiv	2023	~180	Partially Agreed	Not yet in journal

3	Detecting hallucinations in LLMs using semantic entropy	Nature	2024	~120	Fully Agreed	-
4	HILL: A Hallucination Identifier for Large Language Models	CHI	2024	~90	Fully Agreed	-
5	A Survey on LLM Hallucination via a Creativity Perspective	ArXiv	2024	~75	Partially Agreed	Not yet in journal
6	Hallucination is Inevitable: An Innate Limitation of LLMs	ArXiv	2024	~60	Partially Agreed	Not yet in journal
7	A Comprehensive Survey of Hallucination Mitigation Techniques in LLMs	Claimed IEEE/Elsevier	2024	~45	Disagreed	Not in journal, only on arXiv
8	Hallucination Detection and Mitigation: An Investigation	Claimed ACL/EACL	2024	~40	Partially Agreed	Not yet in conference or journal
9	On LLMs' Hallucination with Regard to Known Knowledge	NAACL	2024	~30	Fully Agreed	-
10	Quantifying Uncertainty in Answers and Enhancing Trustworthiness	ACL 2024	2024	~25	Fully Agreed	-

[1번 프롬프트로부터 얻어진 LLM 결과에 대한 Grok-3 DeepSearch 결과]

- **Fully Agreed Papers (5):** Papers 1, 3, 4, 9, and 10 meet all conditions, being published in high-impact venues since 2023 and focusing on LLM hallucinations.
- **Partially Agreed Papers (4):** Papers 2, 5, 6, and 8 are relevant and from 2023-2024, but they are currently only on arXiv, not yet in journals or top conferences.
- **Disagreed Paper (1):** Paper 7 does not meet the condition of being published in an SCI(E) journal, as it is only on arXiv despite the claim, leading to disagreement.

표 2에서 보듯이, Grok-3 DeepSearch에 의해 부분 동의(Partially Agreed)나 비동의(Disagreed) 된 논문 5편(2, 5, 6, 7, 8번) 중 2번 논문은 Grok-3 DeepSearch가 환각이 발생한 경우이며(빨간색), 7번 논문은 ChatGPT-4o가 환각이 발생한 경우이며(파란색), 5, 6, 8번 논문은 두 LLM 모두 환각이 발생한 경우(보라색)이다. 완전 동의(Fully Agreed) 된 논문 5편(1, 3, 4, 9, 10번)은 두 LLM 모두 환각이 발생하지 않은 경우이다. 표 3은 논문 10편을 대상으로 3번 프롬프트로부터 얻어진 결과 중, 두 LLM 모두 환각이 발생하지 않은 5편 논문의 주요 기여점과 키워드 목록을 보여준다. 4번 프롬프트를 이용하여 이 결과를 검증한 결과, Grok-3 Think 역시 모든 결과에 대해 완전 동의(Fully Agree) 의견을 제시하였다. 이는 문제 범위를 한정한 Retrieval-in-the-loop 구조가 사실 근거성을 높인다는 연구 결과를 뒷받침한다[7].

표 3. 3번 프롬프트로부터 얻어진 LLM 결과 중 일부

No.	Main Contributions	Keywords
1	Provides a foundational taxonomy of hallucination types (intrinsic vs. extrinsic), surveys evaluation metrics and mitigation methods across NLG tasks (summarization, QA, dialogue, MT, etc.), and highlights emerging challenges when scaling to transformer-based LLMs. Serves as a go-to reference for hallucination research.	intrinsic hallucination, extrinsic hallucination, evaluation metrics, mitigation strategies, faithfulness
3	Introduces semantic entropy as a robust signal for detecting hallucinations, demonstrating strong correlation between entropy spikes and factual inconsistency. Validated on GPT-series and open-source LLMs, it suggests real-time monitoring potential.	semantic entropy, detection, entropy spikes, factual inconsistency, real-time monitoring

4	Presents HILL, a user-centric tool that flags potential hallucinations via interface overlays, using model confidence and provenance signals. Empirical user studies show increased trust and reduced misinformation.	user interface, provenance logging, confidence score, human in loop, trust
9	Reveals that hallucinations occur even when knowledge is present in context—due to retrieval and attention misalignment. Proposes attention-aware retrieval refinement to improve grounding.	retrieval misalignment, attention bias, context grounding, knowledge retrieval, NAACL evaluation
10	Proposes probabilistic answer confidence scores applicable to any LLM, calibrates them on held-out factual queries, and shows using confidence thresholds significantly reduces hallucination rates.	uncertainty calibration, confidence scoring, probabilistic framework, calibration dataset, thresholding

III. 결론

본 연구는 단일 LLM에 의존하는 기존 접근의 한계를 넘어, 다중 LLM을 활용한 교차 검증 방식이 특히 선행 연구 검색 및 분석에서 환각을 줄이는 데 효과적임을 확인하였다. 이처럼 다중 LLM 기반 교차 검증은 자동화된 문헌 조사에서의 정확성 확보뿐만 아니라, 후속 연구와 정책 결정의 신뢰도를 높이는 데 실질적인 기여를 할 수 있다. 향후 추가적인 LLM과 대용량 데이터 셋 기반 실험을 통해 좀 더 신뢰할 만한 환각 검색 및 분석 방식을 제안할 예정이다.

ACKNOWLEDGMENT

This work was supported by UST Young Scientist+ Research Program 2024 through the University of Science and Technology(No. 2024YS12).

참 고 문 헌

- [1] Agrawal, A., Suzgun, M., Mackey, L., and Kalai, A. "Do Language Models Know When They're Hallucinating References?," In Proceedings of EACL 2024, pp. 912-928, Mar. 2024.
- [2] Park, D. and Lee, H. "Literature Review of AI Hallucination Research Since the Advent of ChatGPT: Focusing on Papers from arXiv," Journal of Information Policy 31(2), pp. 3-38, 2024.
- [3] Noh, T., Yeo, H., and Han, K. "Using Deep Learning-Based Visual Hints to Mitigate Hallucinations in Large Language Model," In Proceedings of KSC 2023, pp. 523-525, Dec. 2023.
- [4] Bae, S. and Lee, D. "Exploring Hallucination Types in Question-Answering Generation and Limitation of Text Evaluation Metrics," In Proceedings of KCC 2024, pp. 450-452, Jun. 2024.
- [5] Jung, H., Jang, Y., Choi, G., Kim, H., and Park, J. "Scenario Design and Prototype Development to Support R&D Process," In Proceedings of the KSCI 2023, pp. 131-132, Jan. 2023.
- [6] Jung, H. "The Generation of 'General Procedure' and 'Task-Oriented Procedure' from LLM Reasoning Processes," In Proceedings of the KSII 2025, Apr. 2025.
- [7] Shuster, K., Poff, S., Chen, M., Kiela, D., and Weston, J. "Retrieval Augmentation Reduces Hallucination in Conversation," In Proceedings of EMNLP 2021, pp. 3784-3803, Nov. 2021.

¹⁾ Computational Linguistics, July 07 2025.