Review Paper
Informatization Policy Vol. 31, No.2 (2024), pp.003-038
https://doi.org/10.22693/NIAIP.2024.31.2.003

챗GPT 등장 이후 인공지능 환각 연구의 문헌 검토: 아카이브(arXiv)의 논문을 중심으로*

박 대 민** · 이 한 종***

요 약

환각은 대형언어모형이나 대형 멀티모달 모형의 활용을 막는 큰 장벽이다. 본 연구에서는 최신 환각 연구 동향을 살펴보기 위해 챗 GPT 등장 이후인 2022년 12월부터 2024년 1월까지 아카이브(arXiv)에서 초록에 '환각'이 포함된 컴퓨터과학 분야 논문 654건을 수집해 빈도분석, 지식연결망 분석, 문헌 검토를 수행했다. 이를 통해 분야별 주요 저자, 주요 키워드, 주요 분야, 분야 간 관계를 분석했다. 분석 결과 '계산 및 언어'와 '인공지능', '컴퓨터비전 및 패턴인식', '기계학습' 분야의 연구가 활발했다. 이어 4개 주요 분야 연구 동향을 주요 저자를 중심으로 데이터 측면, 환각 탐지 측면, 환각 완화 측면으로 나눠 살펴보았다. 주요 연구 동향으로는 지도식 미세조정(SFT)과 인간 피드백 기반 강화학습(RLHF)을 통한 환각 완화, 생각의 체인(CoT) 등 추론 강화, 자동화와 인간 개입의 병행, 멀티모달 AI의 환각 완화에 대한 관심 증가 등을 들 수 있다. 본 연구는 환각 연구 최신 동향을 파악함으로써 공학계는 물론 인문사회계 후속 연구의 토대가 될 것으로 기대한다.

주제어 : 인공지능, 환각, 환각 탐지, 환각 완화, 인간 피드백 기반 강화학습, 생각의 체인, 멀티모달 인공지능

Literature Review of AI Hallucination Research Since the Advent of ChatGPT: Focusing on Papers from arXiv*

Park, Dae-Min** · Lee, Han-Jong***

Abstract

Hallucination is a significant barrier to the utilization of large-scale language models or multimodal models. In this study, we collected 654 computer science papers with "hallucination" in the abstract from arXiv from December 2022 to January 2024 following the advent of Chat GPT and conducted frequency analysis, knowledge network analysis, and literature review to explore the latest trends in hallucination research. The results showed that research in the fields of "Computation and Language," "Artificial Intelligence," "Computer Vision and Pattern Recognition," and "Machine Learning" were active. We then analyzed the research trends in the four major fields by focusing on the main authors and dividing them into data, hallucination detection, and hallucination mitigation. The main research trends included hallucination mitigation through supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF), inference enhancement via "chain of thought" (CoT), and growing interest in hallucination mitigation within the domain of multimodal AI. This study provides insights into the latest developments in hallucination research through a technology-oriented literature review. This study is expected to help subsequent research in both engineering and humanities and social sciences fields by understanding the latest trends in hallucination research.

Keywords : artificial intelligence, hallucination, hallucination detection, hallucination mitigation, Reinforcement Learning from Human Feedback(RLHF), Chain of Thought(CoT), multimodal artificial intelligence

Received Apr 30, 2024; Revised May 18, 2024; Accepted May 20, 2024

* This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No. 2022R1A5A7083908).

** First Author & Corresponding Autor , Assistant professor, School of Media & Communication, Sunmoon University (dmpark@sunmoon.ac.kr, https://orcid.org/0000-0002-0789-9851)

*** Ph.D. student, Graduate School of Communication, Seoul National University (hanjonglee@snu.ac.kr https://orcid.org/0009-0005-4101-0496)

I. 문제제기

2022년 11월 오픈AI(Open AI)의 챗GPT(ChatGPT) 출시 이후 모든 사회 분야에 걸쳐 인공지능 전환(Artificial Intelligence Transformation, AIX)이 일어날 것이라는 전망이 힘을 받고 있다. 디지털 전환, 모바일 전환에 이어 인공지능(Artificial Intelligence, AI)이 사회 전반을 변화시킨다는 것이다.

그러나 이에 대한 우려도 적지 않다. AI를 프로젝트나 의사결정에 활용하려는 시도도 적지 않지만, 이에 적합한 조직 구조나 거버넌스가 마련되어 있는지는 여전히 의문이다(Lee & Nam, 2022; Eun & Hwang, 2020). AI 기술의 발전으로 저작권 등 법적 문제, AI 관련 윤리 문제, 일자리 감소 등 인류의 미래에 대한 문제 등 다양한 우려도 제기되고 있다(Song & Lee, 2024). 무엇보다 AI의 신뢰성은 개별 도메인(Domain)와 최종 사용자(end-user) 수준에서도 보장되어야 한다(Park, 2023a).

환각(Hallucination) 문제는 도메인별 AI 활용을 주저하게 만드는 요소 중 하나다. 특히 언론과 법과 같이 사회적 영향력이 큰 분야에서는 높은 수준의 신뢰성이 요구된다(Park, 2023a). 이러한 분야에서 AI의 신뢰성을 높이기 위해서는 환각의 원인을 파악하고 이를 탐지(Hallucination Detection) 및 완화(Hallucination Mitigation)하는 연구가 필수적이다.

환각 완화는 인간처럼 텍스트를 생성하는 대규모 언어 모델(Large Language Model, LLM) 개발을 주도하는 빅테크 기업에서도 가장 중요하게 생각하는 문제이다. 예컨대 오픈AI는 이러한 문제를 해결하기 위해 지도식 미세조정(Supervised Fine Tuning, SFT)과 인간 피드백 기반 강화학습(Reinforcement Learning from Human Feedback, RLHF)를 도입한 사례를 소개한다(Ouyang, et al., 2022). 이밖에 검색과 생성을 결합한 검색 증강 생성(Retrieval-Augmented Generation, RAG)이나 생각의 연쇄(Chain of Thought, CoT)와 같은 추론 단계의 환각 완화 기법도 주목을 받고 있다.

그러나 오픈AI의 최신 멀티모달(Multimodal) AI인 GPT-4o조차도 환각을 완전히 해소하지는 못한 것으로 알려져 있다. 예컨대 GPT-4o에게 중국어로 "最新高清无码(최신 비검열 고화질 컨텐츠)"에 대해 설명하라고 요청했을 때, 이와 전혀 관계없는 "밀러-울만 꿈 분석법"을 설명하는 사례가 보고되었다.[1] 이는 결함 토큰(glitch token)에 의한 문제로 알려졌다. 오픈AI가 GPT-4o를 토큰화 단계부터 재설계한 것으로 알려졌음에도 불구하고 결함 토큰에 의한 환각이 발생한 셈이다. 그만큼 환각 완화는 여전히 해결해야 할 난제로 남아있다.

환각 완화 기법은 모델 유형, 환각 유형, 환각 완화 기법의 적용 단계 등에 따라 매우 다양하다. 새로운 환각 유형도 보고되고 있으며 환각 완화 기법도 추가되고 있다. 특히 과거에도 생성 AI의 환각 완화 문제가 다뤄져 왔으나 LLM 기반의 챗GPT 등장 이후 학계는 물론 업계와 일반인들까지 AI 환각에 대한 관심이 크게 높아진 상태다.

이 연구는 챗GPT 등장 이후 본격화된 환각 완화 관련 최신 연구를 리뷰한다. 이를 통해 연구자들이 환각 완화와 관련된 세부 연구 주제를 파악하고 관련 연구를 세분화하여 단계별로 체계적으로 수행하는데 도움을 주고자 한다. 특히 AIX가 사회 전반에 확산된다고 본다면, 환각 문제는 공학계만 다루어야 할 문제는 아니다. 정책 당국은 물론 법률가나 언론인, 의료인, 예술가 등 다양한 사회 영역(Domain)의 전문가부터 일반 사용자까지 환각과 그 완화에 대한 이해를 돕는데 기여하고자 한다.

이를 위한 세부 연구 문제는 다음과 같다. 첫째, 챗GPT 등장 이후 환각 완화 관련 연구의 주요 분야, 주요 연구자, 주요 연구 주제는 무엇인가? 둘째, 기존 리뷰 논문을 통해 보았을 때 환각 완화 연구는 어떻게 분류할 수 있는가? 셋째, 주요 분야의 주요 연구자가 작성한 논

1) https://chatgpt.com/share/9f434a84-b316-4114-a185-c8c7ebc1b496?oai-dm=1

문을 통해 보았을 때 환각 완화 최신 연구는 어떻게 진행되고 있는가?

분석 대상은 챗GPT 등장 이후인 2022년 12월부터 2024년 1월까지 오픈 엑세스(Open Access) 사전 출판(Preprint) 논문 데이터베이스(Database, DB) 아카이브(arXiv)에서 환각 관련 논문 654건이다. 분석 방법은 빈도분석과 지식연결망 분석(Knowledge Network Analysis), 질적 내용분석을 활용했다. 빈도분석과 지식연결망 분석은 순위화에 활용했다. 즉 아카이브에서 제시하는 세부 분야별로 주목받는 분야는 무엇이고, 분야별 주요 저자는 누구이며, 어떤 키워드가 중점적으로 다뤄지는지 살펴본다. 지식연결망 분석은 지식연결망의 특성에 따른 연구 활성화 정도, 연구 분야 간 관계 등을 파악하는데에도 도움을 준다. 이어 아카이브에 게재된 리뷰 논문 28개를 검토하여 분석틀을 마련했다. 이에 따라 연구 경향은 리뷰 논문 검토 결과를 바탕으로 크게 데이터 접근과 모델 접근으로, 모델 접근은 환각 탐지와 환각 완화로 나눈 뒤, 세부적인 내용을 살펴보았다. 다음으로 빈도분석 및 지식연결망 분석의 순위화 결과를 토대로 집중 분석할 4개 분야 주요 저자의 논문 108개를 선별해 그 내용을 분석했다.

이 논문은 생성 AI의 환각 완화 관련 방대한 연구를 세세하게 검토하는 대신, 개발자와 비개발자 독자를 포함하여 환각의 기제와 완화 기술의 동향을 포괄적으로 살펴보는 것을 목적으로 한다. 이를 통해 공학 분야는 물론 루프 속 인간(Human in the Loop)을 구체적인 방안을 모색하는 인문사회 연구자들에게도 도움이 되길 기대한다.

Ⅱ. 이론적 논의

1. 생성 AI의 정의, 유형, 구조, 과업, 문제점

본격적인 논의에 앞서 생성 AI에 대한 기본적인 개념을 간략하게 설명하고자 한다. AI는 크게 판별 모델(Discriminative Model)과 생성 모델(Generative Model)로 나눠진다(Foster, 2022). 분류(Classification)에 사용되는 판별 모델과 달리 생성 모델은 텍스트나 음성, 이미지 등을 만드는데 사용된다. 쉽게 말해 영상에서 판별 모델이 개와 고양이를 찾아주는 것이라면, 생성 모델은 개와 고양이를 그려주는 모델이라고 할 수 있다. 또한 통상 지도학습(Supervised Learning)으로 학습하는 판별 모델과 달리 생성 모델은 주로 비지도학습(Unsupervised Learning)으로 구축된다. 생성 AI의 주요 아키텍처로는 트랜스포머(Transformer), 잠재 확산 모델(Latent Diffusion Model), 오토 인코더(Autoencoder), 생성적 적대 신경망(Generative Neural Network, GAN) 등이 있다.

현재 가장 각광 받는 생성 AI는 언어 모델(Language Model, LM), 특히 챗GPT와 같은 LLM이다. 텍스트 형태의 프롬프트를 입력하면, 텍스트를 생성한다. LLM의 매개변수는 많게는 수조 개에 달한다. 이보다 매개변수 수가 적은 sLM(small Language Model)이나 sLLM(smaller Large Language Model)도 있다.

멀티모달 AI도 주목받고 있다. 멀티모달 AI 중 가장 일반적인 것이 대형 멀티모달 모델(Large Multimodal Model, LMM), 또는 대규모 시각 언어 모델(Large Vision Language Model, LVLM)이다. 이들은 텍스트 프롬프트를 입력하면 사진을 생성하는 T2I(Text to Image) 기능을 구현한다. 달리3(DALLE-3), 미드저니(Midjourney), 스테이블 디퓨전(Stable Diffusion) 등이 대표적이다. 멀티모달 AI는 텍스트 외에 이미지나 동영상, 음성 등을 입력으로 받을 수도 있다. 출력으로도 이미지 외에 텍스트나 오디오, 동영상, 3D를 생성할 수도 있다. 일례로 2024년 5월 공개된 GPT-4o는 텍스트와 이미지는 물론 음성까지 입력으로 받고 생성한다.

생성 AI는 통상 인코딩(Encoding)-디코딩(Decoding) 구조로 이루어져 있다. 인코딩 과정에서는 텍스트나 이미지 등 입력 값의 다양한 특징(Feature)을 벡터 형태로 벡터 공간(Vector Space)에 임베딩(Embedding)한다. 디코딩 과정에서는 벡터 공간에 임베딩된 특징을 출력으로 복원한다(Foster, 2022).

LLM은 기본적으로 트랜스포머가 사용된다. 트랜스포머는 텍스트에서 인코딩과 디코딩에 중요한 부분을 병렬로 찾아내는 멀티헤드 어텐션(Multi-Head Attention) 구조를 갖는다(Vaswani, et al., 2017). 어텐션이란 문장 전체 입력이 아닌 출력에 중요한 입력을 찾아내는 법을 학습해 기계번역이나 생성의 효율을 높이는 방식이다. 멀티헤드 어텐션은 여러 셀프 어텐션(Self Attention)을 동시에 수행한 결과를 합치는 방식이다.

LVLM에서는 잠재 확산 모델(Latent Diffusion Model)이 널리 활용된다. 우선 확산은 인코딩 단계에서 구체적인 영상에 표준 가우스 잡음을 추가하는 노이징(Noising) 과정과 디코딩 단계에서 노이징으로 완전히 평균적인 잡음이 된 영상을 질서 정연한 생성 영상으로 복원하는 디노이징(Denoising) 과정으로 이루어진다. 잠재 확산 모델은 이 확산 과정을 저수준의 잠재 공간(Latent Space)에서 수행한다. 멀티모달 AI는 텍스트와 이미지를 모두 임베딩하고 텍스트 임베딩과 이미지 임베딩을 매핑한다(Rombach, et al., 2022).

생성 AI는 sLM이라고 하더라도 학습할 매개변수가 굉장히 많다. 때문에 학습의 효율성을 높이기 위해 전이학습(Transfer Learning)을 적극 활용한다. 생성 AI는 세 종류의 학습을 수행한다(Huang, et al., 2023a). 첫째, 사전학습(pre-Training)이다. 대규모 학습을 통해 범용으로 사용할 수 있는 기반 모델(Foundation Model)을 만드는 과정이다. 둘째, 미세조정(Fine Tuning)이다. 사전학습 모델에서 약간의 학습을 추가해 구체적인 목표 과업의 성능을 더욱 높이는 방식이다. 지도학습 방식의 SFT이 널리 사용된다(Dong, et al., 2023a). 미세조정은 사전학습 모델의 지식, 즉 매개변수의 가중치를 약간 업데이트하는 방식과 최종층에 가까운 은닉층을 추가하는 방식이 있다(Sarker, et al., 2018, 2019). 셋째, 강화학습(Reinforcement Learning)이다. RLHF이 주로 활용된다. 미세조정 후 보상 모형(Reward Model)을 통해 인간의 선호를 모방하게 한 뒤, 이를 근거리 정책 최적화(Proximal Policy Optimization, PPO)와 같은 최적화 알고리듬을 이용한 강화학습을 진행한다(Ouyang, et al., 2022).

다양한 생성 AI는 일반인공지능(Artificial General Intelligence, AGI)에 가까운 AI로 높은 기대를 받고 있다. 그러나 생성 AI는 생성 결과만 놓고 볼 때도 해결해야 할 다양한 문제가 남아있다(Zhang, et al., 2023a). 첫째, 모호성(Ambiguity)이다. 프랑스의 수도를 묻는 질문에 "유럽 국가의 수도"라고 답변하는 식이다. 둘째, 불완전성(Incompleteness)이다. 차량 타이어 교체 방법을 부분적으로만 알려주는 식이다. 셋째, 편향(Bias)이다. 전형적인 초등학교 교사를 알려달라고 할 때, 여교사라고 답변하는 예를 들 수 있다. 넷째, 정보 부족(Under Informativeness)이다. 질문에 대해 최신 정보는 모르겠다고 답변하는 경우이다. 다섯째, 이 논문에서 살펴보고자 하는 환각이다.

2. 환각의 정의, 관련 과업, 연구 주제

환각은 원래 정신의학 용어다. 의학적 환각은 감각기관의 자극이 없었는데도 자극이 있는 것처럼 잘못 지각하는 증상을 뜻한다(Montagnese, et al., 2021). 외부 자극은 있었지만 잘못 지각하는 착각(Illusion)과는 다르다. 이와 유사하게 생성 AI의 환각은 AI가 근거가 없이 믿을 수 없거나(Unfaithful) 말이 안 되는(Non-Sensical) 정보를 생성하는 현상을 뜻한다(Ji, et al., 2023a).

환각 완화(Hallucination Mitigation)는 LLM 이전의 자연어 생성에서도 사실 불일치(Factual Inconsistency) 개선 등으로 다루어졌다. 환각 완화는 환각 감소(Hallucination Reduction), 환각 교정(Hallucination Correction), 탈환각(Dehalluciation) 등의 용어로도 다뤄진다(구다훈 등, 2022; 오동석, 2024; Jha, et al. 2023; Rehman, et al., 2023; Ji, et al., 2023a). 이 논문에서는 환각 완화라는 용어를 사용한다.

챗GPT가 공개된 이후 환각의 논의는 LLM 중심으로 진행됐다. LLM에서 환각을 일으키는 주요 과업으로는 기계 번역(Machine Translation), 질의응답(Question

and Answer), 대화 시스템(Dialogue System), 요약(Summarization), 지식 그래프(Knowledge Graph) 생성 등이 있다. LVLM과 관련된 과업으로는 시각적 질의응답(Visual Question Answering), 이미지 캡셔닝(Image Captioning), 보고서 생성(Report Generation) 등이 대표적이다(Ye, et al., 2023). 최근에는 사진이나 영상 생성, 음성 생성, 3D 모델 생성 등 다양한 멀티모달 AI의 환각 문제도 주목받기 시작했다(Wang, 2024).

생성 AI의 환각 문제를 이해하기 어려운 이유는 데이터-학습-추론으로 이어지는 AI의 생애주기별로 설명할 수 있다. 우선 학습데이터가 방대하다. 특히 판별 모델의 지도학습 방식과 달리 생성 AI의 비지도학습에서는 학습데이터 구축 시 인간의 주석(Annotation) 작업이 최소화된다. 이 때문에 웹 등 다양한 출처의 정보에 포함된, 오래되거나 잘못되고 편향된 정보를 충분히 제거하지 못할 수 있다. 또한 생성 AI의 종류와 서비스 방식이 다양해지면서 환각에 대응하기 위한 일반적인 평가 방법이나 모델 개발 등에 어려움을 겪을 수 있다. 뿐만 아니라 생성 AI가 너무 그럴듯한 콘텐츠를 생성하기 때문에 생성물에 환각이 포함됐는지를 인지하는 것조차 어려울 수 있다(Zhang, et al., 2023a).

기존 리뷰 논문(Review Paper)를 정리해보면 생성 AI의 환각 연구의 주제는 다음과 같다. 첫째, 환각 자체를 유형화하고 그 원인을 살펴본다. 둘째, 사실이나 환각 여부를 평가할 벤치마크 데이터세트를 구축하고 성능 평가 지표를 마련한다. 셋째, 환각 탐지나 환각 완화 방법을 고안한다(Huang, et al., 2023a; Ji, et al, 2023a; Wang, 2024; Ye, et al., 2023; Zhang, et al., 2023a). 다음 절에서는 환각의 유형과 원인을 살펴본다. 환각 관련 데이터와 환각 탐지, 환각 완화는 리뷰 논문 내용 분석 결과에서 살펴본다.

3. 환각 유형과 원인

생성 AI의 환각은 환각의 성질에 따라서 사실성(Factuality) 환각과 충실성(Faithfulness) 환각으로 나눌 수도 있다(Huang, et al., 2023a). 사실성은 출력된 생성물이 실제 사실과 일치하는지를 뜻한다. 충실성은 입력 데이터와 출력된 생성물 간, 또는 생성물 간의 일관성을 의미한다.

환각은 환각 발생이 어디인지에 따라 내재적 환각(Intrinsic Hallucination)과 외재적 환각(Extrinsic Hallucination)으로 나눌 수 있다(Ji, et al. ,2023a). 내재적 환각은 모델 결함에 의한 환각이다. 학습데이터나 입력 데이터는 맞는데 이에 상충하는 출력을 생성하는 것이다. 이 경우 데이터에서 사실 확인을 할 증거를 찾을 수 있다. 외재적 환각은 데이터 결함에 의한 환각이다. 모델이 제대로 작동한다면 그럴 듯한 출력을 생성할 것이다. 이 경우 데이터만으로는 출력을 증명하지도 반박하지도 못할 수 있다.

환각의 특성에 따라 입력 상충(Input-Conflicting), 맥락 상충(Context-Conflicting), 사실 상충(Fact-Conflicting)으로도 나눌 수도 있다(Zhang, et al., 2023a). 입력 상충은 프롬프트 입력과 다른 내용을 출력하는 것이다. 프롬프트 입력은 크게 과업 지시와 과업 입력으로 구성된다. 입력 상충은 지시와 다른 과업을 실행하거나, 입력과 모순된 응답을 내놓는 경우를 뜻한다. 예컨대 요약을 하라고 했는데 설명을 하거나, A라는 사람을 입력했는데 B라고 출력하는 경우가 해당된다. 보통 요약이나 기계번역에서 자주 일어난다. 맥락 상충은 장기 기억 문제와 맥락 인식 문제로 발생한다. 생성 AI는 장기 기억 유지가 쉽지 않기 때문에 긴 응답이나 연속적인 질의응답 시 일관성 유지 못하는 경우가 있다. 맥락 인식 자체가 AI에게 어려운 과제이기도 하다. 사실 상충은 실제 세계와 다른 사실을 생성하는 것을 의미한다. 이는 생성 AI가 실제 세계에서 직접 감각해 학습하기보다는 기호를 통해 통계적 확률을 학습하기 때문에 나타난다. 입력 상충과 맥락 상충은 모델의 결함에 의한 내재적 환각으로, 사실 상충은 데이터의 결함에 의한 외재적 환각으로도 볼 수 있다. 또한 사실성 환각은 외재적 환각 및 사실 상충과, 충실성 환각은 내재적 환각과 입력 상충 및 맥락 상충과 관련성이 높다.

환각의 원인은 크게 데이터 측면과 모델 측면에서 살펴볼 수 있다(Huang, et al., 2023a; Ji, et al. ,2023a; Zhang, et al., 2023a). 데이터 측면의 환각은 오정보(Misinformation), 지식 경계(Knowledge Boundary) 초과, 데이터 활용 부족(Inferior Data Utilization) 등으로 나타난다. 휴리스틱 데이터 수집(Heuristic Data Collection) 때문에 데이터가 편향될 경우도 있다. 실수 또는 고의로 데이터에 틀린 정보나 사회적 편견을 담은 잘못된 정보가 담겨 있을 수도 있다. 잘못된 지식이나 견해차, 또는 오래된 정보와 새로운 정보의 차이로 데이터 내에 모순된 정보가 포함된 경우가 있다. 잘못된 데이터가 중복돼 암기될 수도 있다. 충분한 데이터를 수집하지 못해서, 또는 최신 데이터를 수집하지 못해서 생기는 환각도 생길 수 있다. 영역 지식(Domain Knowledge) 내지 전문 지식이 부족한 경우도 있다. 지식 지름길(Knowledge Shortcut)을 따라가다가 데이터를 제대로 활용하지 못할 수 있다. 데이터가 적은 긴 꼬리 지식(Long Tailed Knowledge)에 이를 때도 환각이 발생할 수 있다. 추론이 필요한 약간 복잡한 질문에 응답하기 어려울 수도 있다. 예컨대 반전의 저주(Reversal Curse)를 들 수 있다(Berglund, et al., 2023). 이는 LLM이 "A는 B이다"를 알아도 "B는 A인가"에 잘못 답하는 현상을 말한다.

모델 측면의 환각 원인은 모델 학습 과정 내지 인코딩 시 오류와 모델에 의한 추론 과정 내지 디코딩 단계에서의 오류로 나눌 수 있다. 모델 학습 과정에서 환각은 사전학습, SFT, RLHF의 모든 단계에서 유발될 수 있다.

사전학습 단계의 환각은 아키텍처의 결함(architecture flaw)에 의한 것이다. 예컨대 인코딩 단계에서 모델이 단방향 순서만 고려하거나 양방향이더라도 비교적 짧은 문장만 고려해 학습됐다면, 장문에서 맥락을 제대로 이해하지 못하고 환각을 일으킬 수 있다. 어텐션(Attention)의 오류로 불완전한 표현 학습(Imperfect Representation Learning)을 할 수도 있다. 다음 단어의 예측 시 무질서도(Entropy)가 높거나 정답과 오답 분포가 유사해 잘못된 선택을 하거나 잘못된 선택 후에 이를 일관되게 밀어붙이는 눈덩이 환각(Snowballing Hallucination)도 나타난다(Zhang, et al., 2023b). 학습 과정에서 제대로 된 전략을 학습하지 않고 따라 하기만 하는 행동 복제(Behavior Cloning) 또는 확률적 앵무새(Stochastic Parrot) 문제도 있을 수 있다(Chiesurin, et al., 2023).

SFT나 RLHF 단계에서도 다양한 지식 충돌(Knowledge Conflict), 모델 순위화와 인간 선호도 간 정렬의 불일치(Misalignment) 등도 생길 수 있다(Zhang, et al., 2023a; Huang, et al., 2023a). 예컨대 DB와 달리 AI의 지식은 매개변수로 저장된다. 이러한 매개변수적 지식(Parametric Knowledge)과 SFT, RLHF, 실제 사용 등의 단계에 입력되는 추가된 지식 간의 차이가 있을 수 있다(Schulman, 2023). 사용자가 입력하는 정보에 대한 무조건적인 동조화(Sycophancy)나 최대한 답하려는 긍정 편향도 문제가 된다(Cotra, 2021).

추론 과정의 오류는 디코딩의 결함으로 나타날 수 있다. 우선 확률 함정(Likelihood Trap)이 있다(Zhang, et al., 2020). 다음 단어를 추론할 때 무작위성에 기초한 확률에 의존하기 때문에 나타나는 문제다. 기계 번역이나 요약에서 유창함을 위해 원천 텍스트보다 생성 텍스트에 의존하게 되는 과신(over-Confidence) 문제도 있다(Chen, et al., 2022; Miao, et al., 2021). 지식을 제대로 활용하지 못하는 지식 회상(Knowledge Recall) 실패 문제도 있다(Zhong, et al., 2023b). 어텐션이 제대로 작동하지 않아 생기는 문제도 있다(Shi, et al. 2023b). 어텐션이 잘못되면 망각 위험에 따른 환각 가능성도 커진다(Chen, et al., 2023a). 최종층 출력에 흔히 사용되는 소프트맥스의 처리 한계에 따른 병목(Softmax Bottleneck) 현상도 환각을 낳을 수 있다(yang, et al., 2018a). 알지 못하는데 아는 것처럼 답변하는 긍정 편향(Positive Bias), 반대로 너무 조심한 나머지 아는데도 모른다고 답변하는 과도한 보수성(over-Conservative Phenomenon) 문제도 있다(Zhang, et al., 2023a).

Ⅲ. 연구방법

1. 데이터 수집

검토 대상 논문은 아카이브에서 수집했다. 아카이브는 미국 코넬 대학교에서 운영하는 무료 사전 출판 논문을 제공하는 DB로 컴퓨터공학을 비롯해 경제학, 전기공학 및 시스템 과학, 수학, 물리학, 계량생물학, 통계학, 계량금융학 분야의 논문이 공개돼 있다. 동료 평가를 받기 전이지만 최신 연구를 빠르게 접할 수 있다는 것이 장점이다. 특히 구글(Google), 메타(Meta) 등 저널 논문 실적에 구애받지 않는 대형 IT 기업의 최신 연구 성과도 발표된다. 아카이브는 연구 분야도 세분하고 있다. 특히 컴퓨터과학은 계산 및 언어(cs.CL), 인공지능(cs.AI), 컴퓨터비전 및 패턴인식(cs.CV), 기계학습(cs.LG), 정보검색(cs.IR) 등 40개 분야로 나눠진다.

검색 조건은 다음과 같다. 검색어는 'hallucination'이다. 검색 기간은 2022년 11월 30일부터 2024년 1월 31일까지다. 2022년 11월 30일은 GPT3.5가 대중에 공개된 날이다. 검색 영역(Field)은 초록(Abstract)으로 했다. 환각을 주제로 한 연구는 초록에서 환각을 언급하지 않을 수 없기 때문이다. 검색 분야(Subject)는 컴퓨터과학(Computer Science, cs)으로 한정했다. 분야를 한정함으로써 의학적 환각 등 타 분야에서 AI와 무관하게 논의된 환각 연구를 대부분 배제할 수 있다. 아카이브에서 한 논문은 2순위까지 분류된다. 2순위까지의 분류 중 하나라도 컴퓨터과학으로 분류되는 경우 분석 대상에 포함시켰다.

메타데이터와 논문 PDF 파일은 크롤링으로 수집했다. 메타데이터 수집은 크롤링 도구인 스크랩스톰(ScrapeStorm)[2]을 사용했다. 수집 항목은 논문 ID(ID), 논문 링크(Link), 제목(Title), 저자(Authors), 초록(Abstract), 분야(Subject), 최종본 제출일(Submitted), 최초 게재 연월(Originally Announced), 버전 1 제출일(v1 Submitted), 비고(Comments) 등이다. 논문 PDF 파일 다운로드는 파이썬(Python)의 뷰티풀스프(BeautifulSoup) 라이브러리 등을 활용해 다운로드 링크를 파싱(Parsing)해 수집했다. 아카이브에서는 논문의 키워드가 제시되지 않는다. 따라서 키워드는 파이썬의 PKE(Python Keyphrase Extraction toolkit)[3]를 활용해 초록에서 상위 키워드 5개를 추출했다. 파일 전처리에는 MS 오피스365의 엑셀(Excel)과 넷마이너(Netminer)를 사용했다. 연결망 분석과 시각화는 UCINET과 넷드로우(NetDraw)를 활용했다.

2. 분석방법

1) 기술통계, 빈도분석, 워드클라우드

우선 기술통계를 통해 검색 기간 발행된 논문 총수와 월별 논문 수, 분야별 논문 수, 저자별 논문 수를 제시한다. 월별 논문 수를 통해서는 환각 관련 연구의 관심도 추이를 확인할 수 있다. 분야별 논문 수로는 컴퓨터과학 중에서도 특히 어떤 분야에서 환각 관련 연구가 많이 다뤄졌는지 확인한다. 저자별 논문 수를 통해서는 발표한 논문 수 기준으로 주요 연구자를 파악할 수 있다. 논문을 많이 작성한 주요 저자는 분야별로도 살펴보았다. 구글 스칼라(Google Scholar), 소속 기관의 홈페이지, 개인 홈페이지 등을 토대로 주요 연구자들의 소속과 연구 관심 분야도 간략하게 소개했다. 이를 통해 AI 연구를 주도하는 국가, 기관 등도 파악할 수 있다.

키워드는 빈도분석을 거쳐 전체 논문에서 주요 키워드로 4회 이상 제시된 상위 30위권(동 순위 포함 35개) 단어를 워드 클라우드로 시각화했다. 이를 통해 환각 연구의 전체적인 주요 주제를 가늠할 수 있다.

2) https://scrapestorm.com/
3) https://github.com/boudinfl/pke.git

2) 지식연결망 분석

이 연구에서는 지식연결망의 일종인 공저자 연결망(co-Authorship Network), 키워드 연결망(co-Keyword Network), 연구 분야 연결망(Subject Network)을 분석한다(Hong, et al., 2019).

공저자 연결망은 연구자를 노드(Node)로, 논문 공저 여부를 엣지(Edge)로 하는 1원 연결망(1 Mode Network)이다. 키워드 연결망은 키워드를 노드로, 같은 논문 초록에 등장했는지 여부를 엣지로 하는 1원 연결망이다. 분야 연결망은 분야를 노드로, 엣지를 공동 저자 또는 공동 키워드 중첩 여부로 부여한다.

각 연결망에서는 연결정도 중앙성(Degree Centrality)과 연결 강도(Tie Centrality), 구성집단(Component) 등을 필요에 따라 분석한다. 공저자 연결망에서 연결정도 중앙성은 저자별 공저자 수를 의미한다. 공저자가 많은 저자는 그만큼 다양한 저자와 연구 협업을 주도한다고 볼 수 있다. 논문 편수도 많다면 해당 연구 분야에서 연구를 주도하는 활발한 연구자로 간주할 수 있다. 연결 강도는 두 연구자 간의 공저 건수를 의미한다. 연결 강도가 큰 연구자들은 공동 주제를 연구하는 팀으로 볼 수 있다. 구성집단은 일종의 연구회로, 관련 연구를 하는 연구자들을 묶은 것이다. 각 구성집단의 크기(Size)는 노드 수, 즉 해당 집단에 속한 저자 수이다. 한 구성집단이 크다는 것은 그 연구회와 관련된 연구자가 많다는 것을 의미하며, 보통 해당 주제에 대한 세분화된 연구가 진행되고 있음을 의미한다. 구성집단의 수가 많다는 것은 여러 연구회가 존재한다고 볼 수 있지만, 최대 구성집단의 크기가 작아질 수 있다. 이는 연구가 파편화되어 진행 중임을 보여준다.

키워드 연결망에서 연결정도 중앙성은 각 키워드의 연관 키워드 수를 뜻한다. 한 키워드의 연관 키워드가 많다는 것은 해당 키워드가 다양한 세부 주제로 논의된 상위 의제라고 이해할 수 있다. 연결 강도는 두 키워드가 같은 논문에 키워드로 제시된 수를 나타낸다. 연결 강도가 큰 두 키워드는 그만큼 관련성이 높음을 의미한다. 키워드 연결망의 각 구성집단은 하위 연구 주제를 묶은 상위 의제를 나타낸다. 구성집단의 크기가 크다면 해당 의제가 세분화되어 깊이 있게 연구되고 있다고 볼 수 있다. 반대로 구성집단이 많은 것은 연구가 파편화되어 진행되고 있음을 의미한다. 유기적으로 연구가 깊이 있게 진행된 경우 거대 구성집단 1개와 시작 단계의 주변화된 작은 구성집단들이 공존하는 형태를 갖춘다. 키워드 연결망은 특정 키워드의 중심어 연결망(Keyword-Centric Network) 분석도 수행한다. 본 연구에서는 검색어인 “hallucination”이 포함된 키워드들의 연관어를 살펴본다. 이를 통해 환각과 직결된 연구 주제를 살펴볼 수 있다.

분야 연결망에서는 연결 강도를 살펴본다. 이 연결망에서 연결 강도는 저자 또는 키워드 기준으로 연구 분야 간 얼마나 중첩돼 있으며 긴밀하게 연관되는지를 보여준다.

3) 문헌 검토

주요 논문에 대한 문헌 검토는 두 단계로 이뤄진다. 첫째, 분석 대상에 포함된 리뷰 논문들을 살펴본다. 일부 리뷰 논문의 내용은 앞서 이론적 논의에서도 자세히 다뤘다. 분석에서는 벤치마크 데이터, 환각 탐지, 환각 완화 기법을 중심으로 살펴본다. 이어 빈도분석과 지식연결망 분석의 결과를 토대로 주요 4개 분야의 주요 연구자를 대상으로 논문의 핵심 내용을 정리한다. 분석은 리뷰 논문 분석 결과를 바탕으로 데이터, 환각 탐지, 환각 완화를 분석 유목으로 하여 유목별 세부 접근과 주요 내용을 요약 제시한다. 내용분석 결과는 분야별로 표로 정리한다.

4. 분석 결과

1) 기술통계 및 빈도분석

(1) 총 논문 수, 월별 논문 수

환각 관련 논문 수는 대체로 꾸준히 증가하고 있다. 〈그림 1〉은 2022년 12월부터 2024년 1월까지 월 발행 논문 수 추이를 보여주고 있다. 총 논문 수는 654건

〈그림 1〉 환각 관련 논문 수 (월별)
〈Fig. 1〉 Monthly Count of Papers Related to Hallucination.

이다. 2022년 12월 10건이던 논문 수는 2023년 10월 111건으로 월 100편을 넘겼다. 2023년 연말에 다소 주춤했지만 2024년 1월 다시 증가세로 돌아섰다. 참고로 연구 분야별 추세는 대동소이했다.

(2) 연구 분야별 논문 수

분야별 논문 수는 〈표 1〉과 같다. 계산 및 언어(cs.CL) 분야의 논문 수가 431건(41.32%)으로 가장 많았다. 인공지능(cs.AI) 분야와 컴퓨터비전 및 패턴인식(cs.CV) 분야도 각각 283건(27.85%)과 145건(14.27%)에 달한다. 이어 기계 학습(cs.LG) 53건(5.22%), 정보검색(cs.IR) 34건(3.35%), 소프트웨어공학(cs.SE) 17건(1.67%), 인간-컴퓨터 상호작용(cs.HC) 11건(1.08%) 등의 순으로 논문 수가 많다.

2순위 내에서 컴퓨터과학 분야 1개 외에 비컴퓨터과학 분야에 포함된 연구는 15편이었다. 해당 분야는 오디오 및 음성 처리(eess.AS) 3편, 포트폴리오 관리(q-fin.PM), 의학물리학(physics.med-ph), 재료과학(cond-mat.mtrl-sci) 각 2편, 뉴런과 인지(q-bio.NC), (math.OC), 최적화 및 제어(eess.SP), 지구물리학(physics.geo-ph), 통계방법론(stat.ME), 천체물리학의 계측 및 방법(astro-ph.IM) 각 1편 등이다. 이들 연구는 각 영역에서 딥러닝을 적용하는 것에 초점을 뒀다.

(3) 논문 수 기준 주요 저자

전체 저자 수는 3119명이다. 주요 저자로는 중국계 연구자들이 많았다. 가장 많이 쓴 저자는 8편의 논문을 쓴 헝 지(Heng Ji)다. 미국 일리노이 대학교(University of Illinois at Urbana-Champaign)의 컴퓨터과학 교수로 주요 연구 분야는 자연어처리이다. 이어 쉬밍 시(Shuming Shi), 민 장(Min Zhang), 패스칼 펑(Pascale Fung), 페이 황(Fei Huang) 등 중국계 연구자가 6편의

〈표 1〉 분야별 논문 수와 비중
〈Table 1〉 Number and Proportion of Papers by Categories.

Sub categories	Abbreviation	Number*	Proportion**
Computation and Language	cs.CL	431	41.32
Artificial Intelligence	cs.AI	283	27.85
Computer Vision and Pattern Recognition	cs.CV	145	14.27
Machine Learning	cs.LG	53	5.22
Information Retrieval	cs.IR	34	3.35
Software Engineering	cs.SE	17	1.67
Human-Computer Interaction	cs.HC	11	1.08
Robotics	cs.RO	9	0.89
Computers and Society	cs.CY	7	0.69
Sound	cs.SD	7	0.69
Graphics	cs.GR	6	0.59
Cryptography and Security	cs.CR	6	0.59
Multimedia	cs.MM	4	0.39
Logic in Computer Science	cs.LO	1	0.1
Multi-agent Systems	cs.MA	1	0.1
Data Structures and Algorithms	cs.DS	1	0.1
Emerging Technologies	cs.ET	1	0.1
Etc.	NA	15	NA

* Including duplicate classifications.
** The proportion of papers in the given category out of the total number of papers (654).

논문을 작성했다. 이들은 텐센트 AI 연구소(Tencent AI Lab), 알리바바 다모 아카데미(Alibaba DAMO Academy) 등 중국 기업과 쑤저우대학교(Soochow University), 홍콩과학기술대학교(Hong Kong University of Science & Technology) 등 중국 또는 홍콩 대학 소속 연구자들이다. 통계적 기계 번역(Statistical Machine Translation), 대화형 AI(conversational AI) 등 자연어처리 관련 연구를 내놓았다. 중국계 이외 연구자로는 모힛 반살(Mohit Bansal)이 6편의 논문을 작성해 가장 많았다. 미국 노스캐롤라이나 대학교(The University of North Carolina at Chapel Hill)의 컴퓨터과학과 교수로, 자연어처리와 컴퓨터비전, 멀티모달 AI 등을 연구했다. 논문 편수 기준 10위권 저자(총 14명)는 〈표 2〉와 같다.

분야별 주요 저자와 논문 수는 〈표 3〉과 같다. 주요 저자를 간단히 소개하면, 먼저 cs.CL 분야에는 헝 지, 민 장, 패스칼 펑 등 전체 논문 수가 많은 저자들이 상

〈표 2〉 논문 편수 기준 주요 저자
〈Table 2〉 Primary Authors Based on the Number of Papers.

Number of Papers	Author
8	Heng Ji
6	Shuming Shi, Min Zhang, Pascale Fung, Mohit Bansal, Fei Huang
5	Ziwei Ji, Xiang Li, Yang Liu, Liangming Pan, Ji-Rong Wen, Huaxiu Yao, Dan Roth, André F. T. Martins

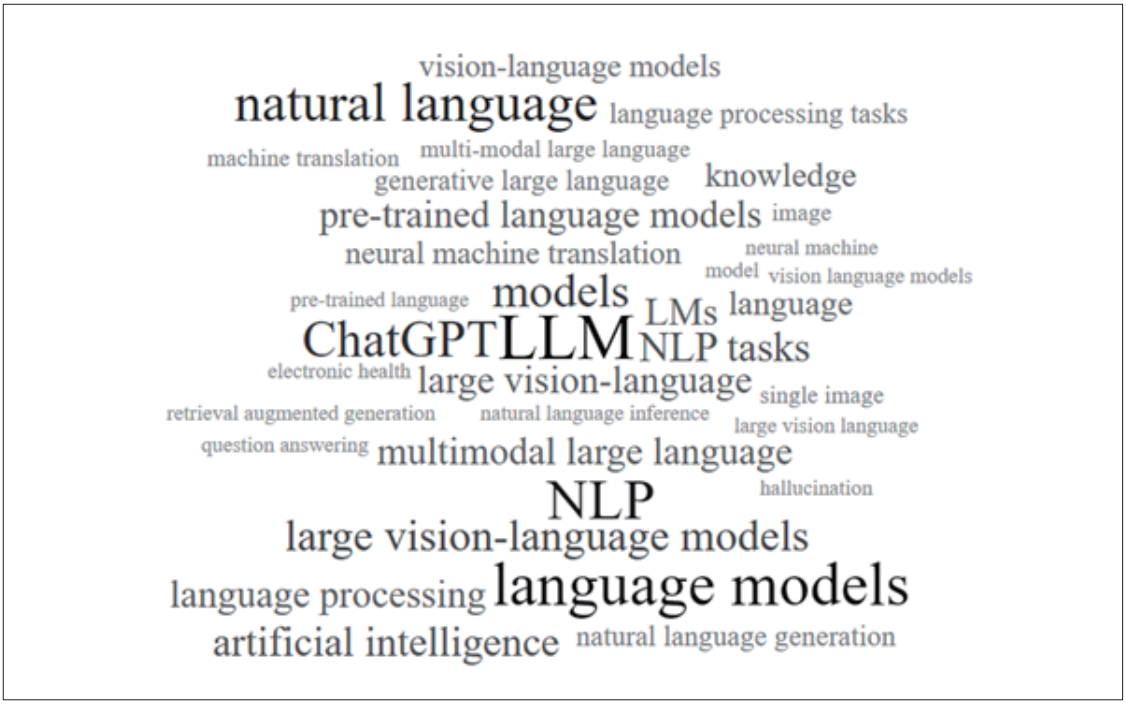
〈표 3〉 세부 분야별 주요 저자
〈Table 3〉 Primary Authors by Sub categories.

Sub categories (Abb.)	Author	Number of Papers
cs.CL	Heng Ji, Min Zhang, Pascale Fung, Shuming Shi, Dan Roth	8
cs.AI	Liangming Pan, Mohit Bansal, Yue Zhang, Dan Roth, Amitava Das	4
cs.CV	Ping Luo, Yu Qiao, Conghui He, Jiaqi Wang, Kaipeng Zhang	4
cs.LG	Yiyang Zhou, Huaxiu Yao, Yue Guo, André F. T. Martins, Nuno M. Guerreiro	3
cs.IR	Negar Arabzadeh, Yongfeng Zhang, Siqing Huo, Charles L. A. Clarke, Zihan Wang	3
cs.SE	Shin Yoo, Felix Juefei-Xu, Mateusz Dolata, Angela Fan, Zhuolin Xu 등 83명	2
cs.HC	Ari Kouts, Gregor Donabauer, Jinjun Xiong, Zehavi Horowitz-Kugler, Davinia Hernandez-Leo 등 72명	1
cs.RO	Chen Feng, Yiming Li, Fei Xia, Jianning Deng, Leila Takayama 등 54명	2
cs.CY	Kamalpreet Kaur, Ned Cooper, Ajith Abraham, Zachary Kilhoffer, Margaret Eby 등 42명	1
cs.SD	Rita Frieske, Hankun Wang, Kai Yu, Zhongjie Yu, Andrew T. Campbell 등 30명	1
cs.GR	Abhimitra Meka, Eli Shechtman, Jiebo Luo, Yuqian Zhou, Christian Theobalt 등 32명	1
cs.MM	Matthieu Cord, Arvind Krishna Sridhar, Erik Visser, Yinyi Guo, Zhou Wang 등 13명	1
cs.CR	Mannudeep Kalra, Francesco Marchiori, Jordan Vice, Tianyu Chen, Ajmal Mian 등 28명	1
cs.LO	Susmit Jha, Nathaniel D. Bastian, Patrick Lincoln, Sumit Kumar Jha, Alvaro Velasquez 등 7명	1
cs.MA	Zili Wang, Jinlin Wang, Jürgen Schmidhuber, Zijuan Lin, Chenglin Wu 등 15명	1
cs.DS	Felipe Cucker, Alexander Bastounis, Anders C. Hansen 포함 3명	1

위권을 차지했다. cs.AI 분야의 주요 저자로는 모힛 반살, 량밍 판(Liangming Pan), 웨 장 등이 있다. 량밍 판은 캘리포니아대학 산타바바라 캠퍼스(University of California, Santa Barbara) 박사후 연구원으로, 자연어처리와 지식 그래프를 연구했다. 컴퓨터비전 및 패턴인식(cs.CV) 분야의 주요 저자는 핑 뤄(Ping Luo), 위 챠오(Yu Qiao), 충후이 허(Conghui He) 등이다. 핑 뤄(Ping Luo)는 홍콩대학교(The University of Hong Kong) 소속 부교수이다. 위 챠오는 상하이 AI 연구소(Shanghai AI Laboratory)의 교수로, 대형 멀티모달 모델 등을 연구한다. 기계학습(cs.LG) 분야의 저자로는 화슈 야오(Huaxiu Yao), 이양 쥬(Yiyang Zhou), 웨 궈(Yue Guo) 등이 있다. 화슈 야오는 노스캐롤라이나 대학교 소속 조교수로 파운데이션 모델(Foundation Model), AI 안전(AI Safety) 등을 연구한다. 정보검색(cs.IR) 분야의 주요 저자는 융펑 장(Yongfeng Zhang), 네가르 아라브자데흐(Negar Arabzadeh), 씨칭 훠(Siqing Huo) 등이 있다. 융펑 장은 미국 럿거스 대학교(Rutgers University)의 조교수로, 주 연구 분야는 정보 검색 및 추천 시스템(Recommendation Systems) 등이다.

(4) 빈도 기준 주요 키워드

검토한 전체 주제어 수는 2206개였다. 〈그림 2〉는 전체 분야에서 4회 이상 언급된 키워드 35개를 워드클라우드로 시각화한 것이다. 복수형이나 약어 등 의미상 중복되는 키워드는 병합하고 약어로 통일했다. “LLM” 등 대형 언어 모델을 지칭하는 키워드가 503개의 논문에 등장해 가장 많았다. 이밖에 “natural language processing” 등 자연어처리 관련 키워드, “ChatGPT”, “large vision-language models”과 같은 멀티모달 AI



〈그림 2〉 빈도 기준 주요 키워드 워드클라우드

〈Fig. 2〉 Wordcloud of Major Keywords Based on Frequency.

관련 키워드가 자주 언급됐다.

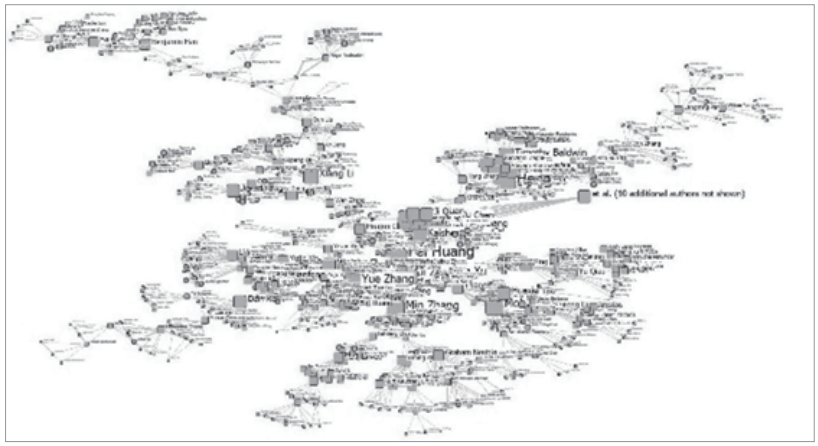
2) 지식연결망 분석

(1) 공저자 연결망 분석

공저자 연결망의 전체 연결망과 주 구성집단 연결망의 시각화는 〈그림 3〉에 제시했다. 공저자 연결망의 노드 수는 저자 수와 동일한 3119개이다. 노드가 2개 이상으로 이루어진 구성집단의 개수는 371개이다. 환각 연구를 다양한 연구 집단이 수행하고 있음을 알 수 있다. 가장 큰 구성집단의 크기는 818로 꽤 크다. 즉 818명의 저자가 직간접적으로 연결돼 있다. 분석 기간이 짧음에도 불구하고 연구자들 간 연계가 잘 되어 있다고 할 수 있다.

연결정도 중앙성 값이 큰 상위권 주요 저자는 〈표 5〉와 같다. 공저자가 가장 많은 저자는 헝 지(Heng Ji)로 46명과 공저했다. 헝 지(Heng Ji)는 논문 수도 많은 활

전체 연결망



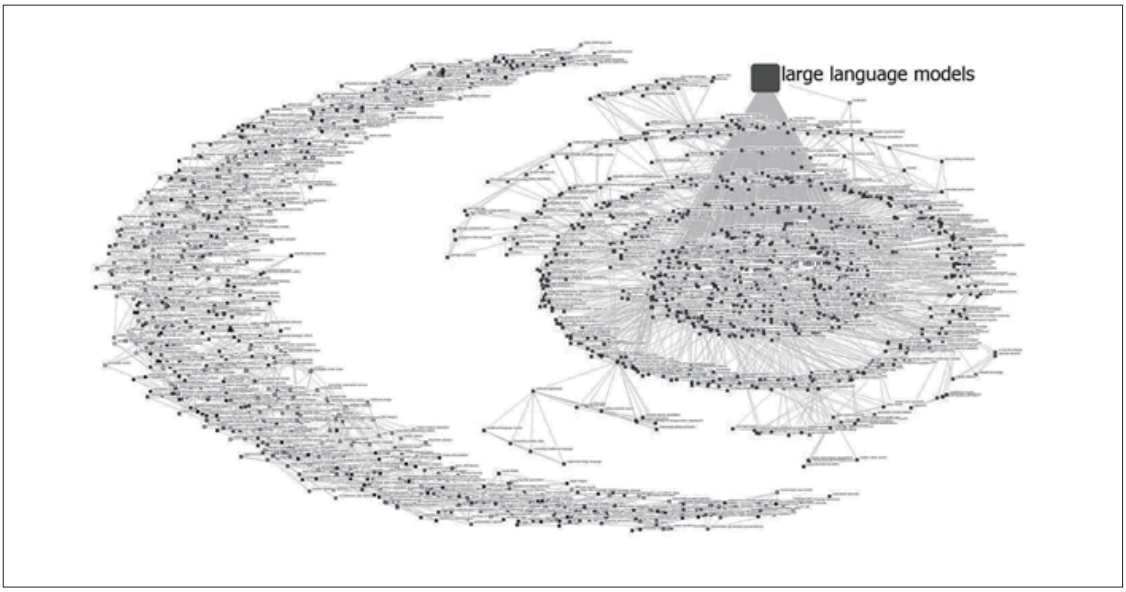
주 구성집단

〈그림 3〉 공저자 연결망 시각화

〈Fig. 3〉 Visualization of Co-Authorship Networks.

〈표 5〉 공저자 연결망 분석 결과: 연결정도 중앙성 상위 20위권 저자
〈Table 5〉 Result of Co-Authorship Network Analysis: Top 20 Authors by Degree Centrality.

Degree Centrality	Author
46	Heng Ji
45	Fei Huang
34	Mohit Bansal, Min Zhang
33	Hao Peng
32	Ming Yan, Haiyang Xu, Ji Zhang, Yue Zhang
29	Shuming Shi, Kaisheng Zeng, Xiang Li
28	Qinghao Ye, Timothy Baldwin
27	Weizhu Chen, Guohai Xu
25	Jifan Yu, David Thulke 등 총 52명



〈그림 4〉 전체 키워드 연결망
〈Fig. 4〉 The Whole Keyword Network.

발한 연구자이다. 전체적으로 중국인 연구자가 주요 연구자의 다수를 차지했다. 한편 공저자가 25명인 연구자가 과도하게 많은데, 이는 한 논문을 26명 이상 작성한 경우가 2편 있기 때문이다. 실제로는 해당 논문의 저자는 26명보다도 많지만, 아카이브에서는 메타데이터 상으로 26명까지 수집한다. 이러한 저자의 연결정도 중앙성은 사실상 예외 값이다.

(2) 키워드 연결망 분석

키워드 연결망의 노드 수는 2206개이다. 키워드 연결망의 전체 연결망 시각화는 〈그림 4〉와 같다.

구성집단의 개수는 171개다. 주 구성집단은 1257개의 노드로 구성돼 상당히 크다. 두 번째로 큰 구성집단의 크기가 49로 차이도 크다. 공저자 연결망보다 노드 수가 적은데도 주 구성집단의 크기는 더 크다. 이는 환각 연구가 주제 측면에서는 좀 더 유기적으로 연구되고 있음을 시사한다. 논문별로 5개의 키워드를 추출했기 때문에 가장 작은 구성집단 크기는 5이다. 이러한 구성집단에 속한 노드 수는 805개로 많다. 이는 주변화돼 있긴 하지만 새로운 연구 주제도 활발하게 발굴되고 있음을 의미한다.

키워드 연결망 분석을 통해 상위 20위권 키워드를 추리면 〈표 6〉과 같다. 연결정도 중앙성이 가장 큰 키워드는 "large language models"로 연관어 수는 575개였다. 빈도분석 결과와 마찬가지로 전체적으로 LLM과 NLP 관련 키워드가 많다. 전체적 분포 역시 압도적으로 연관어가 많은 중심(hub) 키워드와 두터운 꼬리(fat tailed)에 속한 키워드가 공존하는 척도 없는 연결망의 분포를 보인다. 이 역시 환각 연구가 주제 측면에서 깊이 있고 다양하게 진행됐음을 시사한다(Park, 2023b).

"hallucination"이란 단어가 포함된 용어의 연관어를 중심어 연결망으로 시각화한 결과는 〈그림 5〉와 같다. 중심어 연결망을 구성집단별로 살펴보면, 우선 LLM을 중심으로 하는 구성집단의 키워드가 가장 많다. 사

〈표 6〉 키워드 연결망 분석 결과: 연결정도 중앙성 상위 20위권 키워드

〈Table 6〉 Results of Keyword Network Analysis: Top 20 Keywords by Degree Centrality.

Degree Centrality (Number of Related Words)	Keywords
575	large language models
315	large language
291	language models
158	LLMs
62	ChatGPT
58	natural language processing
53	natural language
49	models
36	artificial intelligence
32	large vision-language models
31	large language model
27	pre-trained language models
26	NLP tasks
26	knowledge
22	neural machine translation, large vision-language, multimodal large language
20	LLM, single image, image, LMS

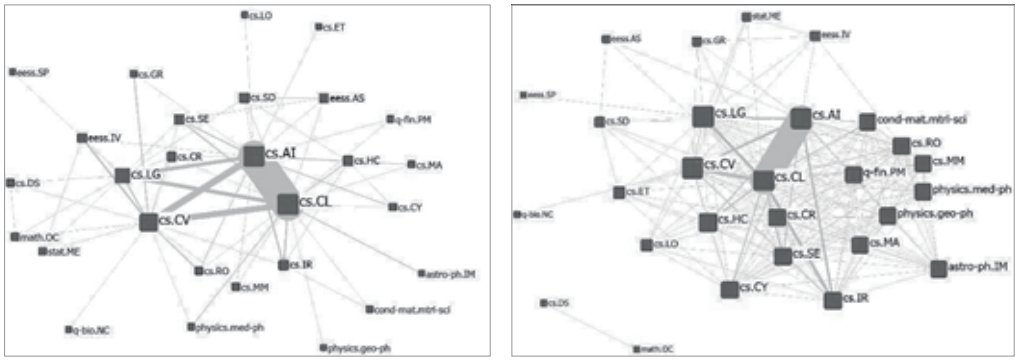
〈그림 5〉 "hallucination" 중심어로 하는 키워드 연결망
〈Fig. 5〉 Keyword Network Centered on "Hallucination."

실 환각 감소(Reducing Factual Hallucination), 외부 지식(External Knowledge)이나 실제 외부 감각(Real External Sensory) 등이 주요 키워드로 제시돼 있다. 다른 구성집단의 주요 키워드로는 비디오 기반 텍스트(Video-Grounded Text), 요약(Summarization), 중국어 질의응답 (Chinese Hallucination Question-Answering) 등 환각이 일어나는 분야 관련 키워드, 객체 환각(Object Hallucination) 등 환각 유형에 대한 키워드가 제시됐다.

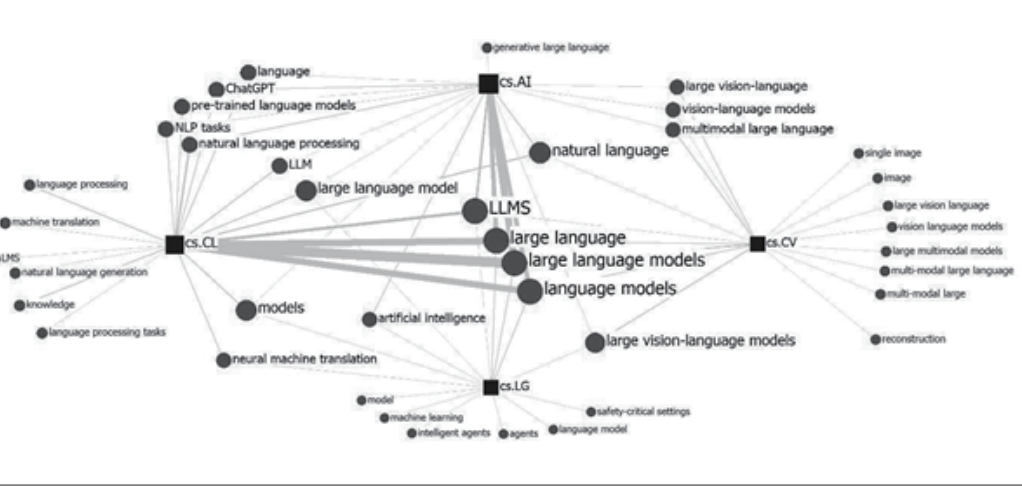
(3) 연구 분야 연결망 분석

〈그림 6〉은 전체 연구 분야 연결망을 저자와 키워드 기준으로 시각화한 것, 〈그림 7〉은 연구 분야-키워드 연결망을 시각화한 것이다. 연구 분야 연결망으로 볼 때 환각 연구의 중심 분야는 계산 및 언어(cs.CL)와 인공지능(cs.AI)이었다. 저자 기준 연구 분야 연결망의 연결강도로 보면 인공지능 분야는 다른 분야와 연구자가 2,336명이 겹친다. 또한 키워드 기준 연결망의 연결 강도로 보면 계산 및 언어 분야는 다른 분야와 80,659개 키워드가 겹친다. 두 분야는 1,623명의 저자와 54,928개의 키워드가 겹쳐 상호 연관성도 가장 높았다. 컴퓨터 비전 및 패턴인식(cs.CV), 기계학습(cs.LG)도 주요 연구 분야로 cs.CL, cs.AI와의 연관성도 높았다.

다음으로 4개 주요 연구 분야와 분야별 빈출 상위 20개 안팎의 주요 키워드를 연구 분야-키워드 연결망으로 살펴보자. 우선 LLMs이 4개 주요 연구 분야에서 모두 중요하게 다뤄진다. 계산 및 언어와 인공지능은 챗GPT가 공통 주제어로 나왔다. LVLM은 인공지능, 컴퓨터비전, 기계학습의 공통 주제어였다. 신경망 기계 번역은 계산 및 언어와 기계학습 분야의 공통 주제어였다. 분야별 고유 주제어로는 계산 및 언어는 자연어생성, 컴퓨터비전은 멀티모달, 기계학습은 에이전트와 안전이 눈에 띈다.



〈그림 6〉 연구 분야 연결망
〈Fig. 6〉 Research Field Network.



〈그림 7〉 연구 분야-키워드 연결망
〈Fig. 7〉 Research Categories Keyword Network.

3) 리뷰 논문 내용분석

(1) 리뷰 논문 분류

LLM이 폭발적인 관심을 받게 된 계기가 챗GPT 출시 이후라는 점을 감안할 때, 2년 남짓한 짧은 기간임에도 불구하고 꽤 많은 수의 리뷰 논문이 포함돼 있다. 먼저 전체 조망 성격의 리뷰 논문를 살펴보자. 초기에는 LLM 중심으로 환각 정의, 분류에 초점을 두었다. 이후 벤치마크, 탐지, 완화 관련 연구를 조사했다(Yue, et al., 2023; Ye, et al., 2023; Huang, et al., 2023a). 최근에는 LLM뿐만 아니라 멀티모달 AI의 환각 문제를 살펴본 논문도 나왔다(Wang, 2024).

세부적으로는 우선 환각 탐지와 환각 완화에 초점을 둔 여러 리뷰 논문이 있다. 환각 탐지와 환각 완화(Luo, et al., 2024), 또는 환각 완화 전반을 살펴본 연구가 있다(Tonmoy, et al., 2024). 세부 환각 완화 기법을 살펴본 논문으로는 지식 기반 LLM 증강(Andriopoulos & Pouwelse, 2023), 지식 그래프 활용(Agrawal, et al., 2023), 프롬프트 엔지니어링 활용(Chen, et al., 2023c), 정렬(Wang, et al., 2023a), 자체 수정(Liangming, et al., 2023) 등이 있다. 환각 관련 세분화된 주제를 다룬 리뷰 논문도 적지 않다. 우선 대형 사전학습 모델인 파운데이션 모델의 환각을 다룬 연구가 있다(Rawte, et al., 2023b). 특정 과업별 환각 문제를 다룬 연구로는 기계 번역의 환각(Nuno, et al., 2022), 의료 분야의 LLM 환각(Yun, et al. 2023), 멀티모달 LLM의 망각 문제(Zhai, et al., 2023), 요약의 환각(Chrysostomou, et al., 2023) 등이 있다. 이 밖에 GPT-3 (Zong & Krishnamachari, 2022), 이미지 캡셔닝 (Ghandi, et al., 2022), QA(Xavier Daull, 2023), 과학논문 자동 리뷰(Kasanishi, et al., 2023), 소프트어 엔지니어링(Fan, et al., 2023), 멀티모달 에이전트(Durante, et al, 2024) 등 부차적이지만 환각을 다룬 리뷰 논문이 있다. 환각 대응을 위한 사회적, 정책적 방안을 모색한 논문도 있다(Augenstein, et al., 2023).

(2) 벤치마크와 성능 평가에 대한 리뷰

환각 관련 벤치마크 데이터세트는 사실성 또는 충실성을 평가한다. 사실성 평가 벤치마크로는 TruthfulQA (Lin, et al., 2022), FACTOR(Muhlgay, et al., 2023), HaluQA(Cheng, et al., 2023b), FreshQA(Vu, et al., 2023), REALTIMEQA(Kasai, et al., 2022), Med-HALT(Umapathi, et al., 2023) 등이, 충실성 평가 벤치마크로는 SelfCheckGPT-Wikibio(Miao, et al., 2023), HaluEval(Li, et al., 2023a), BAMBOO (Dong, et al., 2023b), PHD(Yang, et al., 2023), ScreenEval(Lattimer, et al., 2023), RealHall(Friel & Sanyal, 2023), LSum(Feng, et al., 2023a), SAC3(Zhang, et al., 2023c) 등이 있다. FELM(Chen, et al., 2023b)은 둘 다 평가한다. 대부분 영어 질의응답으로 되어 있지만 ChineseFactEval, HaluQA와 같이 중국어나 Med-HALT와 같이 다국어로 된 경우도 있다.

환각 방지를 위한 핵심 과업은 사실적 질의 생성과 환각 탐지로 나눌 수 있다(Zhang, et al., 2023a; Huang, et al., 2023a). 이에 따라 각 벤치마크는 사실 생성이나 환각 탐지 과업에 특화돼 있다. 사실 생성 관련 벤치마크는 생성된 응답이나 선택의 사실성을 평가한다. TruthfulQA, FACTOR, REALTIMEQA, HalluQA, FreshQA 등이 있다(Lin, et al., 2021; Min, et al., 2023). 환각 탐지 관련 벤치마크는 생성된 응답에 환각이 포함됐는지를 판단한다. 관련 벤치마크로는 HaluEval, BAMBOO, FELM, SelfCheckGPT-Wikibio, ChineseFactEval, PHD, ScreenEval, RealHall, LSum, SAC3가 있다.

벤치마크 데이터세트는 두 가지 방식으로 구축될 수 있다. 첫째, 인간이 수작업으로 주석을 달거나 큐레이션(Curation) 하는 방식이다. TruthfulQA, REALTIMEQA, ChineseFactEval, HaluQA, FreshQA 등이 해당된다. 둘째, 데이터세트를 자동 생성하는 방식이다. Med-HALT, FACTOR, FELM, SelfCheckGPT-Wikibio, HaluEval, BAMBOO, PHD, ScreenEval, RealHall, LSum, SAC3이 해당된다. 챗GPT와 같은 고성능 LLM을 사용해 환각 데이터세트 등을 만드는 식이다. FACTOR처럼 자동 생성된 데이터를 수작업으로 확인하는 방식도 있다.

성능 평가 방법은 인간 평가(Human Evaluation)와 자동화된 평가로 나뉜다. 자동화된 평가는 다시 모델 기반 평가(Model based Evaluation)와 규칙 기반 평가(Rule based Evaluation)로 나눌 수 있다. 인간 평가는 생성된 응답에 인간이 “지지됨, 지지되지 않음, 관련 없음”(supported, not-supported, irrelevant) 등 3개 중 하나를 라벨링하는 방식을 들 수 있다(Min, et al., 2023). 모델 기반 평가는 정답 데이터를 학습한 LLM으로 다른 LLM 응답을 판정하는 식이다. HaluQA가 이러한 방식이다. 정답 세트와 생성 응

답 간의 사실 일치 또는 불일치를 확인하는 정렬함수(Alignment Function)나 손실함수(Loss Function)를 활용하기도 한다(구다훈 등, 2023; Zha, et al., 2023). 규칙 기반 평가는 정답 데이터세트와의 일치 여부에 따른 정확도를 계산하는 식이다(Bang, et al., 2023). 성능 평가 지표로는 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1점수(F1 Score) 외에 우도(Likelihood), AUROC(Area Under ROC Curve) 등이 활용된다.

(3) 환각 탐지에 대한 리뷰

환각 탐지는 환각 유형에 따라 사실성 환각 탐지와 충실성 환각 탐지로 나눌 수 있다(Huang, et al., 2023a). 탐지 기준을 인간 평가나 검색 결과 등 외부 정답에 두느냐, 내부 상태(Internal State)로부터 추정하는지로 나눌 수도 있다. 탐지 주체가 인간이나 웹처럼 모델 밖에 있는지, 모델 자체나 다른 모델인가에 따라 외부 평가와 자체 평가(Self-Evaluation)로도 구분할 수 있다.

사실성 환각 탐지 기법은 외부 사실 검색(Retrieve External Facts) 방식과 불확실성 추정(Uncertainty Estimation) 방식이 있다. 외부 사실 검색은 생성 결과를 외부 사실 출처와 비교하는 것이다. 활용되는 사실 출처는 비정형 코퍼스로부터 정형 DB, 위키피디아 등 신뢰할만한 사이트, 구글과 같은 검색엔진 등 연구자마다 다양하다. 사실 검증 도구나 구글 스칼라 API와 같은 전문 DB를 활용하는 경우도 있다(Zhang, et al., 2023a).

불확실성 지표(Uncertainty Metrics)를 이용하는 불확실성 추정은 내부 상태 기반 추정과 행동(Behavior) 기반 추정으로 나눌 수 있다(Huang, et al., 2023a; Ye, et al., 2023). 불확실성이란 AI가 하나의 토큰을 생성했을 때, 다음 토큰을 예측하는 것이 얼마나 어려운지 그 정도를 나타낸다. 내부 상태 기반 방식은 토큰 생성 확률 분포와 같은 모델 내부 상태를 알 수 있는 경우에만 활용된다. 로짓 기반 추정은 생성할 다음 토큰을 예측할 때 그 확률이 무질서도의 임계값을 넘어간다면 환각을 일으킬 가능성이 크다고 보는 방식이다(Luo, et al., 2023; Yao, et al., 2023a).

그러나 모델 개발사도 아니고 모델도 오픈소스가 아니어서 모델을 API 호출을 통해서만 이용할 경우 모델 내부 상태를 알기는 어렵다. 이 경우 행동 기반 방식을 사용한다. 행동 기반 방식은 프롬프트를 이용한다. 우선 LLM에게 사실성이나 충실성을 직접 수치로 표현해달라고 물어볼 수 있다(Xiong, et al., 2023). 다른 방법으로는 일관성 기반 추정이 있다. 직접적 또는 개방형 질의를 반복해 얻은 모델의 여러 응답 간 사실적 자기 일관성을 파악하는 방식이다(Manakul, et al., 2023). 생성 AI가 서로 질의응답을 주고받는 다중 토론(Multi-Debate)을 통해 사실성 환각 탐지를 하는 방식도 제안됐다(Cohen, et al., 2023). 다만 불확실성 추정은 환각 가능성이 큰 부분을 탐지할 수는 있지만, 사실 자체를 입증하기는 어렵다.

충실성 환각 탐지는 사실 기반 지표(Fact based Metrics), 분류기 기반 지표(Classifier based Metrics), QA 기반 지표(QA based Metrics), 불확실성 추정, 프롬프트 기반 지표(Prompt based Metrics) 등을 활용한다. 사실 기반 지표 방식은 출처와 응답 간 N-그램(N-gram), 수치, 인명이나 기관 장소와 같은 개체명(Named Entities), 개체 간 관계를 나타내는 지식 그래프 등의 중첩 정도를 따진다. 분류기 기반 지표 방식은 출처와 응답 간 연관성을 자연어추론(Natural Inference)을 통해 살펴본다. QA 기반 지표 방식은 QA 시스템을 통해 소스와 응답 간 일관성을 파악한다. 불확실성 추정은 앞서 살펴본 무질서도 외에도 로그 확률(Log Probability)을 활용할 수 있다. 프롬프트 기반 방식은 프롬프트를 이용한다.

(4) 환각 완화에 대한 리뷰

환각 완화는 환각을 줄여 AI가 충실성과 사실성을 갖춘 결과물을 내놓도록 만드는 과업이다(Huang, et al., 2023a). 환각 완화는 보완 대상에 따라 크게 데이터 중

심 접근과 모델 중심 접근으로 나눌 수 있다. 개선 방식에 따라서는 프롬프트 엔지니어링과 모델 개발로 나눌 수 있다(Tonmoy, et al., 2024). 모델 개발은 어떤 단계를 보완하는지에 따라 학습 전 보완, 학습 단계의 인코딩 보완, 추론 단계의 디코딩 보완으로 나눌 수 있다(Huang, et al., 2023a). 학습 단계의 완화는 사전학습, 미세조정, 강화학습에서 이뤄질 수 있다(Zhang, et al., 2023a). 다만 개별 환각 완화 기법은 여러 유형에 걸쳐 있다. 환각 완화 대상 모델의 모달리티에 따라 텍스트만 다루는 LLM의 환각 완화와 오디오나 3D 관련 생성 AI 및 멀티모달 AI 등 텍스트 이외의 환각 완화로 나눠 살펴볼 수도 있다(Wang, 2024).

데이터 중심 접근은 고품질 학습데이터 구축, 구축한 데이터의 정제, 학습데이터 추가의 방식으로 진행될 수 있다. 사전학습 단계에서 수작업을 통해 고품질 학습데이터를 구축하는 것은 환각 완화의 기본이다(Radford, et al., 2019). 그러나 학습데이터 규모가 커짐에 따라 책, 논문, 뉴스, 보고서 등 영역 특화된 교과서적인 DB를 활용하는 것이 효과적일 수 있다(Tourvron, et al., 2023). 편향 제거(Debias)는 구축된 데이터에서 의미론적 중복을 포함한 중복 편향이나 편견을 담은 말뭉치 데이터 등을 활용해 사회적 편견을 제거하는 것이다(Abbas, et al., 2023; Ferrara, 2023; Ladhak, et al., 2023).

지식 경계 완화(Knowledge Boundary Mitigation)는 데이터 중심 접근과 모델 중심 접근이 혼재됐다. 미세조정 단계에서 데이터를 활용해 모델의 지식을 수정하는 지식 편집(Knowledge Editing)과 추론 단계에서 외부 검색을 활용하는 RAG가 있다.

지식 편집은 문제 있는 매개변수를 직접 수정하는 방식(locate-then-edit methods)과 특정 과업에 특화된 하이퍼네트워크(Hypernetwork)[1]를 이용해 매개변수의 가중치를 업데이트하는 메타 학습 방식(meta-learning method)이 있다(Ha, et al, 2016; Mitchell, et al., 2022a; Meng, et al., 2022).

RAG는 생성 단계에서 입력과 함께 검색 결과를 조건으로 부여하여 출력을 생성한다(Lewis, et al., 2020b). RAG는 응답의 정확성과 최신성을 개선한다. 검색은 생성 전이나 생성 중에 할 수도 있고, 생성 후에 할 수도 있다(Tonmoy, et al., 2024). 우선 생성 전 일회성 검색(One-time Retrieval)이 제안됐다(Ram, et al., 2023). 다단계 추론이나 장문의 답변 시에는 반복 검색(Iterative Retrieval)이 필요하다. 추론 과정을 단계별로 나눈 뒤, 단계별로 외부 검색을 통합하는 CoT 기법이 대표적이다(Wei, et al., 2023). CoT를 활용해 지식 회상 실패 문제를 개선하기도 한다(Zhong, et al. 2023b). 선행 질문과 검색 결과를 바탕으로 후행 질문을 스스로, 또는 사용자에게 질의함으로써 보완하는 방법도 있다(Press, et al., 2022; Zhang, et al., 2023d). 사후 검색(Post-hoc Retrieval)은 생성된 응답을 최종 검색해 확인하고 근거가 부족하거나 무질서도가 높아 환각이 있을 가능성이 있는 단어를 수정기(fixer)로 사후 수정하는 방식이다(Zhao, et al., 2023a; Gao, et al., 2023; Rawte, et al., 2023a). RAG는 일종의 프롬프트 엔지니어링으로 볼 수 있다. 프롬프트 엔지니어링을 통한 환각 완화는 구독 기반 API를 이용한 LLM에서 특히 유효하다. 생성된 응답에 대한 적절한 피드백 프롬프트를 제공하면 사실적이고 충실한 응답을 얻어낼 수 있다(Si, et al., 2022). MixAlign은 사람에게 설명을 요청하는 질의를 요청함으로써 입력, 사실, 응답 간의 지식 정렬(Knowledge Alignment)를 달성한다. 좋은 응답을 얻을 수 있는 범용 프롬프트를 활용하는 방안도 있다. 구조화된 프롬프트를 인간이 입력하기도 한다. UPRISE(Universal Prompt Retrieval for Improving zero-Shot Evaluation)처럼 내재적으로 작동하는 소프트 프롬프트(Soft Prompt)를 미세조정해 활용하기도 한다(Cheng, et al., 2023). 피드백을 SLM 등 별도 모델이나 모델 스스로 주는 방식도 있다(Liu, et al., 2023a; Dhuliawala, et al., 2023). CoVe(Chain-of-Verification), CoNLI(Chain of Natural Language Inference)는 단계별 검증과 수정을 하는 CoT를 응용 개선한 것이다(Dhuliawala, et al., 2023;

Lei, et al., 2023).

모델 개발 단계의 환각 완화로는 우선 학습 전 단계에서 아키텍처 개선하는 방식이 있다. 생성 AI의 약점인 양방향 맥락 파악, 장문 인코딩 및 디코딩, 정확한 어텐션 등을 개선하는 것이다(Li, et al., 2023b; Liu, et al., 2023b; Beltagy, et al., 2020; Huang, et al., 2023a). 학습 과정에서 입력 내지 참조(Reference)와 출력 간의 개체명이나 수치 등을 비교하는 손실함수를 이용해 최적화하는 방식도 있다(구다훈 등, 2023). 사전학습 단계와 SFT 단계의 모델 개선은 통상 학습데이터 개선과 병행된다. 다만 SFT 단계의 모델 개선에 필요한 학습데이터의 양은 사전학습 모델의 학습데이터 양보다 훨씬 적다(Zhou, et al., 2023a). SFT의 예로는 사실 또는 환각 데이터를 입력해 성능을 개선하는 지식 주입(Knowledge Injection)이나 HAR(Hallucination Augmented Recitations) 등이 있다(Elaraby, et al., 2023; Köksal, et al., 2023). RLHF 단계의 환각 완화는 인간 피드백을 바탕으로 모델의 3H(Helpful, Honest, Harmless)를 강화한다 (Bai, et al., 2022). 동조화나 긍정 편향 문제를 완화하기 위해 지식 경계를 벗어나는 질문에 대해 모른다고 답변하는 정직성 지향 SFT(Honesty-oriented SFT)도 제안됐다(Sun, et al., 2023a).

추론 단계의 환각 완화는 디코딩 전략을 개선하는 방식으로 이뤄진다. 예컨대 ITI(Inference-Time Intervention, ITI)와 같이 사실성을 강화하는 디코딩 전략을 선택한다(Lee, et al., 2022b; Li, et al., 2023c). 충실성을 강화하는 기법도 있다. 예컨대 문맥 인식 디코딩(Context Aware Decoding, CAD)처럼 문맥 일관성을 강화한다(Shi, et al., 2023b). 가짜 상관관계를 유발하는 데이터를 삭제해 지식 지름길 문제함으로써 논리적 일관성을 높이기도 한다(Kang & Choi, 2023). 단계별 과정을 검증하는 CoVe 기법은 사실성 강화와 추론 과정을 개선하는 방식이다(Dhuliawala, et al., 2023). 정렬 시 발생하는 오류 개선에 초점을 두는 접근도 있다(Saunders, et al., 2022).

4) 아카이브 논문 리뷰: 분야별 주요 저자 중심으로

이 절에서는 cs.CL, cs.AI, cs.CV, cs.LG 등 4개 주요 분야의 빈도 및 연결정도 중앙성 기준 주요 저자의 연구를 정리한다. cs.CL과 cs.AI의 경우 1순위와 2순위에 함께 분류된 경우가 많아 같이 분석한다. cs.CL나 cs.AI가 cs.CV나 cs.LG와 2순위까지 함께 분류된 연구는 cs.CV나 cs.LG의 표에서 소개했다. cs.CL, cs.AI의 연구는 〈표 7〉, cs.CV는 〈표 8〉, cs.LG는 〈표 9〉와 같다. 각 연구는 우선 데이터, 탐지, 완화 등 3개 과업으로 분류했다. 이어 세부적인 접근을 나누고 내용을 요약했다.

전체적인 연구 추이를 보면, 환각 완화에 대한 논문이 많았다. 먼저 데이터는 사전학습의 대규모 학습데이터 구축보다는 성능 평가 등을 위한 벤치마크 데이터를 구축하는 연구가 많았다. 해당 데이터는 주로 사실성이나 충실성 성능 평가, 또는 환각 탐지의 학습에 활용된다. 이는 환각 탐지나 환각 완화와 같은 과업과 깊은 관련이 있기는 하다. 환각 완화에서는 RAG, 미세조정, RLHF와 관련된 내용이 많았다. 메타가 리마3 개발시 RLHF의 중요성을 강조했듯이, 현업에서는 RLHF에 주목하고 있지만, 연구에서는 자동화된 에이전트를 통한 환각 완화와 같이 자동화 기법에 관심이 높았다. LVLM의 환각에 대한 관심도 높아지는 추세로 보인다.

5. 요약 및 제언

분석 결과를 보면 비교적 짧은 기간임에도 많은 연구자들 사이에서 다양한 주제로 연구됐음을 알 수 있다. 지식연결망의 분포를 봐도 멱함수 분포와 거대 구성집단의 출현은 유기적 연구가 진행됐다는 점을 알려준다.

주목할 만한 연구 동향과 그에 대한 제언은 다음과 같다. 첫째, 빈도분석 및 지식연결망 분석 결과 중국계 연구자들이 활발하게 활동하고 있었다. AI 분야에서 중국 또는 중국계 연구자의 활약은 이미 알려진 바이기는 하다. 다만 이러한 점은 특히 동료 평가나 언어의 장벽이 낮은 아카이브에서 더 두드러질 수 있다. 또한 국내

〈표 7〉 계산 및 언어, 인공지능 분야 주요 연구

〈Table 7〉 Key Research in the Fields of Computing, Language, and Artificial Intelligence.

Category	Approach	Articles	Summary
Data	Benchmarks	(Cheng, 2024)	Building a Idk ("I don't know") dataset
		(Li, et al., 2023d)	EureQA benchmark for the evaluation of hallucination by knowledge shortcuts
		(Amayuelas, et al., 2023)	Known-Unknown Question (KUQ) dataset for uncertainty classification
		(Chen, et al., 2023d)	FactCHD, a hallucination detection benchmark incorporating fact-based evidence chains
	Automated data generation	(Bang, et al., 2023)	Evaluating multitask, multilingual, multimodal performance of ChatGPT with 23 datasets
	Adversarial evaluation	(Yu, et al., 2023)	AutoDebug, an LLM-based framework to automatically generate hallucination-inducing adversarial evaluation data
	Metrics	(Du, et al., 2023)	Assoicantion analysis combining each risk factor hallucination level with the cause of the hallucination
Detection	Autonomous detection	(Li, et al., 2024)	Autonomous detection of hallucinations
Mitigation	Knowledge alignment	(Zhang, et al., 2023d)	MixAlign to improve knowledge alignment based on human-user and knowledge DB
	Knowledge graph	(Ji, et al., 2022)	RHO(ρ) utilizing knowledge graphs that contain representations of objects and relational predicates
	RAG	(Peng, et al., 2023)	LLM-Augmenter, where LLM generates responses referencing external knowledge in the DB on a task-by-task basis
		(Kang, et al., 2023)	EVER for real-time, step-by-step verification and correction
	Model development	(Saha, et al., 2022)	A model to generate text from semi-structured data
		(Wang, et al., 2024a)	TechGPT-2.0, an improved model for building knowledge graphs
		(Li, et al., 2023e)	Development of Context-Bias Whisper (CB-Whisper), a new ASR system based on OpenAI's Whisper model
	Encoding improvements	(Zhang, et al., 2023e)	Attention structure that emphasizes informative tokens, untrustworthy tokens, token types and frequency
	Functions	(Feng, et al., 2023)	A DEcoupling method to disentangle the Comprehension and EmbellishmeNT (DECENT) to detect LLM's comprehension and writing skills
	Decoding	(Zhang, et al., 2023f)	Induce-then-Contrast Decoding (ICD) to penalize LLM's hallucinatory induction followed by contrast decoding
	Inference	(Chen, et al., 2023e)	Abstract Reasoning Induction (ARI) framework to improve temporal knowledge reasoning
	Fine-tuning	(Yao, et al., 2023b)	Unlearning through deletion of hallucination-inducing negative data

〈표 7〉애서 계속

Category	Approach	Articles	Summary
Mitigation	Fine-tuning	(Tian, et al., 2023)	Fine-tuning based on automatically generated fact preference rankings
		(Yang, et al., 2023c)	Fine-tuning of LLMs by providing plug-in supervised knowledge
		(Gupta, et al., 2023)	Fine-tuning to avoid multiple input perturbations
	Sorting	(Wan, et al., 2023)	HistAlign complements cache alignment in memory
	Prompts	(Wang, et al., 2023b)	Solo Performance Prompting (SPP) to improve response by setting up multiple personas on a single LLM
		(Ji, et al., 2023b)	Improves response performance through human interaction in medical QA systems
	Engineering	(Sun, et al., 2023b)	Corex, an autonomous agent that pioneers multi-model collaboration including discussion, review, and discovery modes
	Agents	(Nathani, et al., 2023)	Leveraging multi-perspective feedback combining multiple modules, including LM and external tools, to iteratively improve hallucinations in the reasoning chain
		(Liu, et al., 2023)	Apply SFT, RLHF to avoid answering unanswerable questions through adversarial question-answering benchmarks
		(Gou, et al., 2023)	CRITIC framework to improve model for utilizing external tools and human feedback
	RLHF	(Kang & Liu., 2023)	Validation of LLM competencies and hallucination mitigation techniques in the financial domain
		(Fung. et al., 2022)	NormSage, a model for automatically discovering culture-specific norms in documents, with self-validation to mitigate hallucinations

〈표 8〉 컴퓨터 비전 및 패턴 인식 분야 주요 연구

〈Table 8〉 Key Research in the Field of Computer Vision and Pattern Recognition

Category	Approach	Articles	Summary
Data	Benchmarks	(Li, et al., 2023f)	POPE: object hallucination assessment benchmark
		(Shi, et al., 2023)	ChEF: a framework for synthesizing different assessments
		(Lovenia, et al., 2023)	NOPE(Negative Object Presence Evaluation): object hallucination evaluation benchmark
		(Cho, et al., 2023)	VDSG-1k: QG/A improvement benchmark for evaluating multimodal AI generation
		(Xu, et al., 2023)	LVLM Evaluation Hub (LVLM-eHub)
		(Wang, et al., 2024b)	Mementos: evaluate the ability to generate image sequences
		(Wang, et al., 2023c)	AMBER: Multidimensional benchmark for simultaneously evaluating the existence, properties, and relationships of objects

〈표 8〉애서 계속

Category	Approach	Articles	Summary
Detection	OOD detection	(Dai, et al., 2023)	Improve out-of-distribution (OOD) detection
Mitigation	Alignment	(Chen, et al., 2023f)	DRESS: an LVLM that introduces sorting using natural language feedback (NLF)
	Decoding improvements	(Huang, et al., 2023b)	OPERA: a decoding strategy with improved partial overtrust propensity
	Prompts	(Wang, et al., 2023d)	VIGC: visual instruction generation and correction framework
	Engineering	(Zhao, et al., 2023b)	Learning to choose facts between hallucinations and facts using hallucination-aware direct preference optimization (HA-DPO)
	Loss function	(Dai, et al., 2022)	ObjMLM: improving object hallucination in large-scale vision-language pre-trained models
	Model	(Jiang, et al., 2023)	Contrastive learning using hallucinated text as negative examples
	Fine tuning	(Liu, et al., 2023d)	Text shearing to learn by suggesting multiple captions for images

〈표 9〉 기계학습 분야 주요 연구
〈Table 9〉 Key Research in the Field of Machine Learning

Category	Approach	Articles	Summary
Detection	Knowledge conflict detection	(Guerreiro, et al., 2022)	Fully unsupervised plugin detector for detecting knowledge conflicts between ground truth and models in NMT
Mitigation	Data	(Qiu, et al., 2022)	Improving supervised learning with data hallucination teaching (DHT) to intelligently generate input data
	Addition	(Zhou, et al., 2023b)	LURE(LVLM Hallucination Revisor) to improve object hallucinations in LVLM
	Model	(Cui, et al., 2023)	Bingo (Bias and Interference Challenges in Visual Language Models) benchmark for evaluating bias and interference in LVLMs
	Benchmarks	(Wang, et al., 2023e)	LLM-based hallucination evaluation framework for LVLMs HaELM (Hallucination Evaluation based on Large Language Models)

에서는 중국계 연구자의 성과가 미국 빅테크 기업에 비해 주변화된 경향이 있다고도 볼 수 있다. 다만 아카이브에 소개된 중국계 연구자의 연구는 미국 빅테크 기업과 비교할 때 완전히 선구적인 연구는 아닌 측면도 있다. 중국계 연구자의 특성은 다른 논문 DB나 빅테크 기업과의 비교 연구를 통해 좀 더 정확히 이해할 수 있을 것이다.

둘째, 빈도분석과 지식연결망 분석 결과, 분석 DB가

아카이브로 한정돼 있고 분석 기간도 짧은 편이었지만 논문 수도 많고, 공동 연구도 활성화된 편이었으며, 연구 주제도 다양한 편이었다. 이는 환각 관련 연구가 그만큼 높은 관심 속에 깊이 있게 연구되고 있음을 시사한다. 이러한 경향은 분석 기간 및 컨퍼런스 발표 논문이나 동료 리뷰 논문을 포함하는 논문 DB로 분석 대상을 확대한다면 더욱 분명하게 드러날 것이다.

셋째, 내용분석을 통해 살펴본 결과 학습데이터, 벤치마크 데이터, 환각 탐지, 환각 완화가 핵심 과업이 되고 있다. 물론 환각 완화를 위해서는 사전학습 단계에서나 미세조정 단계에서나 데이터 구축을 병행하고, 환각 탐지도 필요하지만, 최종 목표 과업은 환각 완화이다.

넷째, 파운데이션 모델 개발이 어렵고 구독 기반 API를 통해서만 접근 가능한 블랙박스형 LLM이 다수인 상황에서 사전학습 모델의 미세조정과 강화학습 단계의 환각 완화 적용이 주목받는다. 이는 라마3(Llama 3)와 같은 오픈소스 LLM이 공개되면 오히려 더 관심을 끌 수 있다. 미세조정과 강화학습을 통해 도메인에 특화된 SLM을 구축하는 것이 효율적이기 때문이다.

다섯째, 학습 단계의 인코딩을 개선하는 시도를 넘어서 추론 단계의 디코딩을 개선하려는 노력이 강화되고 있다. 추론 단계를 정교하게 나누고 단계별로 외부 지식 기반 검증과 자체 평가와 에이전트 간의 경쟁을 통해 최종 생성물의 환각을 최소화한다. RAG와 결합한 CoT가 고도화되는 추세다.

여섯째, 인간 개입 접근과 자동화 접근이 경쟁적으로 발전하고 있다. 궁극적으로는 높은 수준의 자동화를 기반으로 더 정교한 SFT와 RLHF가 수행될 것으로 예상된다.

일곱째, LLM의 환각 완화를 넘어서 LVLM의 환각 완화에 대한 관심이 늘고 있다. 즉 텍스트 외에 이미지, 동영상, 음성, 3D 등 다양한 모달리티 및 멀티모달의 환각 문제가 제기될 것이다.

여덟째, 이러한 상황에서 도메인별 전문성 높고 업데이트되는 DB의 중요성이 더욱 커지고 있다. 이러한 데이터는 책 DB, 학술 DB, 언론사 DB 등으로 제공된다. 파운데이션 모델을 만드는 개발사는 이러한 고품질 원천데이터를 생산하는 창작자를 지원하는 차원에서도 창작자의 저작권을 적극 보호할 필요가 있다. 창작자 역시 적극적으로 루프 속 인간으로서 AI 개발에 참여할 필요가 있다. 이를 통해 사회적 편견 제거 등 사회적 측면에서도 신뢰성이 제거된 고성능 AI를 개발하고, 이를 활용하여 자신의 창작성을 개선하는 선순환을 모색하길 기대한다. 특히 LVLM 개발에 필요한 고품질 이미지나 동영상 데이터는 더 많이 필요함에도 불구하고, 텍스트보다 더 적게 생산돼 있다. 따라서 LVLM의 환각 완화에서 데이터 문제는 더 중요하게 될 것이다.

본 연구에서 자세하게 다루지 않은 환각 관련 논쟁도 검토할 필요가 있다. 예컨대 환각은 좋은 것인가? 어느 정도 용인하면 출력의 다양성을 높일 수 있다(Zhang, et al., 2020). 이는 AI의 창작성 논쟁으로도 확장될 수 있다(Park, 2024). 텍스트와 이미지, 동영상은 물론 3D와 음성, 행동 등을 포괄하는 멀티모달 환경에서 환각의 유형을 다양화할 필요도 있다(Wang, 2024). 모호성, 불완전성, 편향, 정보 부족 등 환각 외의 문제도 해결해야 한다(Zhang, et al., 2023a). 딥페이크(Deepfake)나 보안(Security) 문제 등도 해결해야 한다(Barrett, et al., 2023; Khalid, et al., 2021). LVLM에서 일관성 등의 문제 또한 상업 영화 감독이나 웹툰 작가 등 최종 사용자의 활용을 주저하게 만든다(Moses, 2024).

본 연구에서는 일반적 리뷰 논문에 더해 지식연결망 분석 등 지식사회학적 관점을 도입했다. 본 연구에서 연구 집단 분석은 자세히 하지는 않았지만, 이를 통해 AI 연구의 주요 저자와 저자 간 관계, 연구 집단 등 학계 상황을 조망할 수 있는 단초를 제시했다. 이를 통해 이 연구에서 검토한 연구의 흐름을 고려하여 연구 과제를 제시하는 한편, 각 흐름을 대표하는 해외의 신진 연구 집단과의 협력을 모색하는 데에도 참고가 될 것으로 기대한다.

본 연구는 기술적인 부분을 세세하게 다루지는 못했다는 한계가 있다. 특히 후속 연구에서는 의료나 금융, 언론 등 영역별 생성 AI 활용 시, 또는 기계 번역이나 요약, 캡셔닝, 코딩 등 특정 과업 수행 시 발생하는 고유한

환각 문제를 좀 더 자세히 살펴볼 필요가 있다. 또한 이 연구에서 살펴본 2024년 1월까지의 연구 이후, 논문 출간까지 몇 개월 동안에도 관련 연구가 지속적으로 빠르게 늘어나고 있어 관련 연구 동향 지속적인 업데이트가 필요하다.

■ References

Abbas, A., Tirumala, K., Simig, D., Ganguli, S. & Morcos, A.S. (2023). *SemDeDup: Data-efficient learning at web-scale through semantic deduplication.* arXiv preprint. https://arxiv.org/abs/2303.09540

Agrawal, G., Kumarage, T., Alghami, Z. & Liu, H. (2023). *Can Knowledge Graphs Reduce Hallucinations in LLMs? : A Survey.* arXiv preprint. https://arxiv.org/abs/2311.07914

Amayuelas, A., Pan, L., Chen, W. & Wang, W. (2023). *Knowledge of Knowledge: Exploring Known-Unknowns Uncertainty with Large Language Models.* arXiv preprint. https://arxiv.org/abs/2305.13712

Andriopoulos, K. & Pouwelse, J. (2023). *Augmenting LLMs with Knowledge: A survey on hallucination prevention.* arXiv preprint. https://arxiv.org/abs/2309.16459

Augenstein, I., Baldwin, T., Cha, M., Chakraborty, T., Ciampaglia, G. L., Corney, D., DiResta, R., Ferrara, E., Hale, S., Halevy, A., Hovy, E., Ji, H., Menczer, F., Miguez, R., Nakov, P., Scheufele, D., Sharma, S. & Zagni, G. (2023). *Factuality Challenges in the Era of Large Language Models.* arXiv preprint. https://arxiv.org/abs/2310.05189.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T.B., Clark, J., McCandlish, S., Olah, C., Mann, B. & Kaplan, J. (2022). *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback.* arXiv preprint. https://arxiv.org/abs/2204.05862

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y. & Fung, P. (2023). *A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity.* arXiv preprint. https://arxiv.org/abs/2302.04023

Barrett, C., Boyd, B., Bursztein, E., Carlini, N., Chen, B., Choi, J., Chowdhury, A., Christodorescu, M., Datta, A., Feizi, S., Fisher, K., Hashimoto, T., Hendrycks, D., Jha, S., Kang, D., Kerschbaum, F., Mitchell, E., Mitchell, J., Ramzan, Z., Shams, K., Song, D., Taly, A. & Yang, D. (2023). "Identifying and mitigating the security risks of generative ai." *Foundations and Trends® in Privacy and Security, 6*(1), 1-52.

Beltagy, I., Peters, M.E. & Cohan, A. (2020). *Longformer: The Long-Document Transformer.* arXiv preprint. https://arxiv.org/abs/2004.05150

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T. & Evans, O. (2023). *The Reversal Curse: LLMs trained on" A is B" fail to learn" B is A".* arXiv preprint. https://arxiv.org/abs/2309.12288

Chen, X., Li, M., Gao, X. & Zhang, X. (2022). "Towards improving faithfulness in abstractive summarization." *Advances in Neural Information Processing Systems, 35*, 24516-24528.

Chen, Y., Liu, Y., Meng, F., Chen, Y., Xu, J. & Zhou, J. (2023a). *Improving Translation Faithfulness of Large Language Models via Augmenting Instructions.* arXiv preprint. https://arxiv.org/abs/2308.12674

Chen, S., Zhao, Y., Zhang, J., Chern, E., Gao, S., Liu, P. & He, J. (2023b). *FELM: Benchmarking Factuality Evaluation of Large Language Models.* arXiv preprint. https://arxiv.org/abs/2310.00741

Chen, B., Zhang, Z., Langrené, N. & Zhu, S. (2023c). *Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review.* arXiv preprint. https://arxiv.org/abs/2310.14735

Chen, X., Song, D., Gui, H., Wang, C., Zhang, N., Yong, J., Huang, F., Lv, C., Zhang, D. & Chen, H. (2023d). *FactCHD: Benchmarking Fact-Conflicting Hallucination Detection.* arXiv preprint. https://arxiv.org/abs/2310.12086

Chen, Z., Li, D., Zhao, X., Hu, B. & Zhang, M. (2023e). *Temporal Knowledge Question Answering via Abstract Reasoning Induction.* arXiv preprint. https://arxiv.org/abs/2311.09149

Chen, Y., Sikka, K., Cogswell, M., Ji, H. & Divakaran, A. (2023f). *DRESS: Instructing Large Vision-Language Models to Align and Interact with Humans via Natural Language Feedback.* arXiv preprint. https://arxiv.org/abs/2311.10081

Cheng, Q., Sun, T., Zhang, W., Wang, S., Liu, X., Zhang, M., He, J., Huang, M., Yin, Z., Chen, K. & Qiu, X. (2023a). *Evaluating Hallucinations in Chinese Large Language Models.* arXiv preprint. https://arxiv.org/abs/2310.03368

Cheng, D., Huang, S., Bi, J., Zhan, Y., Liu, J., Wang, Y., Sun, H., Wei, F., Deng, D. & Zhang, Q. (2023b). *UPRISE: Universal Prompt Retrieval for Improving Zero-Shot Evaluation.* arXiv preprint. https://arxiv.org/abs/2303.08518

Cheng, Q., Sun, T., Liu, X., Zhang, W., Yin, Z., Li, S., Li, L., He, Z., Chen, K. & Qiu, X. (2024). *Can AI Assistants Know What They Don't Know?.* arXiv preprint. https://arxiv.org/abs/2401.13275

Chiesurin, S., Dimakopoulos, D., Cabezudo, M. A. S., Eshghi, A., Papaioannou, I., Rieser, V. & Konstas, I. (2023). *The dangers of trusting stochastic parrots: Faithfulness and trust in open-domain conversational question answering.* arXiv preprint. https://arxiv.org/abs/2305.16519

Cho, J., Hu, Y., Garg, R., Anderson, P., Krishna, R., Baldridge, J., Bansal, M., Pont-Tuset, J. & Wang, S. (2023). *Davidsonian Scene Graph: Improving Reliability in Fine-grained Evaluation for Text-to-Image Generation.* arXiv preprint. https://arxiv.org/abs/2310.18235

Chrysostomou, G., Zhao, Z., Williams, M. & Aletras, N. (2023). *Investigating Hallucinations in Pruned Large Language Models for Abstractive Summarization.* arXiv preprint. https://arxiv.org/abs/2311.09335

Cohen, R., Hamri, M., Geva, M. & Globerson, A. (2023). *LM vs LM: Detecting Factual Errors via Cross Examination.* arXiv preprint. https://arxiv.org/abs/2305.13281

Cotra, A. (2021). "Why AI Alignment Could Be Hard with Modern Deep Learning." https://www.cold-takes.com/why-ai-alignment-could-be-hard-with-modern-deep-learning/. (Retrieved on April 27, 2024).

Cui, C., Zhou, Y., Yang, X., Wu, S., Zhang, L., Zou, J. & Yao, H. (2023). *Holistic Analysis of Hallucination in GPT-4V(ision): Bias and Interference Challenges.* arXiv preprint. https://arxiv.org/abs/2311.03287

Dai, W., Liu, Z., Ji, Z., Su, D. & Fung, P. (2022). *Plausible May Not Be Faithful: Probing Object Hallucination in Vision-Language Pre-training.* arXiv preprint. https://arxiv.org/abs/2210.07688

Dai, Y., Lang, H., Zeng, K., Huang, F. & Li, Y. (2023). *Exploring Large Language Models for Multi-Modal Out-of-Distribution Detection.* arXiv preprint. https://arxiv.org/abs/2310.08027

Daull, X., Bellot, P., Bruno, E., Martin, V. & Murisasco, E. (2023). *Complex QA and Language Models Hybrid Architectures, Survey.* arXiv preprint. https://arxiv.org/abs/2302.09051

Deng, H., Ding, L., Liu, X., Zhang, M., Tao, D. & Zhang, M. (2022). Improving Simultaneous Machine Translation with Monolingual Data. arXiv preprint. https://arxiv.org/abs/2212.01188.

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A. & Weston, J. (2023). *Chain-of-Verification Reduces Hallucination in Large*

Language Models. arXiv preprint. https://arxiv.org/abs/2309.11495

Ding, Y., Wang, Z., Ahmad, W. U., Ramanathan, M. K., Nallapati, R., Bhatia, P., Roth, D. & Xiang, B. (2022). CoCoMIC: *Code Completion By Jointly Modeling In-file and Cross-file Context.* arXiv preprint. https://arxiv.org/abs/2212.10007

Dong, G., Yuan, H., Lu, K., Li, C., Xue, M., Liu, D., Wang, W., Yuan, Z., Zhou, C. & Zhou, J. (2023a). *How Abilities in Large Language Models are Affected by Supervised Fine-tuning Data Composition.* arXiv preprint. https://arxiv.org/abs/2310.05492

Dong, Z., Tang, T., Li, J., Zhao, W. X. & Wen, J. R. (2023b). BAMBOO: *A Comprehensive Benchmark for Evaluating Long Text Modeling Capacities of Large Language Models.* arXiv preprint. https://arxiv.org/abs/2309.13345.

Du, L., Wang, Y., Xing, X., Ya, Y., Li, X., Jiang, X. & Fang, X. (2023). *Quantifying and Attributing the Hallucination of Large Language Models via Association Analysis.* arXiv preprint. https://arxiv.org/abs/2309.05217.

Durante, Z., Huang, Q., Wake, N., Gong, R., Park, J. S., Sarkar, B., Taori, R., Noda, Y., Terzopoulos, D., Choi, Y., Ikeuchi, K., Vo, H., Fei-Fei, L. & Gao, J. (2022). *Agent AI: Surveying the Horizons of Multimodal Interaction.* arXiv preprint. https://arxiv.org/abs/2401.03568

Elaraby, M. S., Lu, M., Dunn, J., Zhang, X., Wang, Y. & Liu, S. (2023). *Halo: Estimation and Reduction of Hallucinations in Open-Source Weak Large Language Models.* arXiv preprint. https://arxiv.org/abs/2308.11764

Eun, J., & Hwang, S. (2020). "An Exploratory Study on Policy Decision Making with Artificial Intelligence: Applying Problem Structuring Typology on Success and Failure Cases." *Informatization Policy, 27*(4), 47-66.

{은종환 · 황성수 (2020). 인공지능을 활용한 정책의사결정에 관한 탐색적 연구: 문제구조화 유형으로 살펴 본 성공과 실패 사례 분석. 〈정보화정책〉, 27권 4호, 47-66.}

Fadeeva, E., Vashurin, R., Tsvigun, A., Vazhentsev, A., Petrakov, S., Fedyanin, K., Vasilev, D., Goncharova, E., Panchenko, A., Panov, M., Baldwin, T. & Shelmanov, A. (2023). *LM-Polygraph: Uncertainty Estimation for Language Models.* arXiv preprint. https://arxiv.org/abs/2311.07383

Fan, A., Gokkaya, B., Harman, M., Lyubarskiy, M., Sengupta, S., Yoo, S. & Zhang, J. M. (2023). *Large Language Models for Software Engineering: Survey and Open Problems.* arXiv preprint. https://arxiv.org/abs/2310.03533

Farinhas, A., de Souza, J. G. C. & Martins, A. F. T. (2023). *An Empirical Study of Translation Hypothesis Ensembling with Large Language Models.* arXiv preprint. https://arxiv.org/abs/2310.11430

Fei, H., Liu, Q., Zhang, M., Zhang, M. & Chua, T. S. (2023). *Scene Graph as Pivoting: Inference-time Image-free Unsupervised Multimodal Machine Translation with Visual Scene Hallucination.* arXiv preprint. https://arxiv.org/abs/2305.12256

Feng, H., Fan, Y., Liu, X., Lin, T. E., Yao, Z., Wu, Y., Huang, F., Li, Y. & Ma, Q. (2023). *Improving Factual Consistency of Text Summarization by Adversarially Decoupling Comprehension and Embellishment Abilities of LLMs.* arXiv preprint. https://arxiv.org/abs/2310.19347

Ferrara, E. (2023). "Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies." *Sci, 6*(1), 3.

Foster, D. (2022). *Generative deep learning.* O'Reilly Media, Inc..

Fung, Y. R., Chakraborty, T., Guo, H., Rambow, O., Muresan, S. & Ji, H. (2022). *NormSAGE: Multi-Lingual Multi-Cultural Norm Discovery from Conversations On-the-Fly.* arXiv preprint. https://arxiv.org/abs/2210.08604

Friel, R. & Sanyal, A. (2023). *Chainpoll: A High Efficacy Method for LLM Hallucination Detection.* arXiv preprint. https://arxiv.org/abs/2310.18344

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M. & Wang, H. (2023). *Retrieval-Augmented Generation for Large*

Language Models: A Survey. arXiv preprint. https://arxiv.org/abs/2312.10997

Ghandi, T., Pourreza, H. & Mahyar, H. (2023). Deep Learning Approaches on Image Captioning: A Review. arXiv preprint. https://arxiv.org/abs/2201.12944.

Gou, Z., Shao, Z., Gong, Y., Shen, Y., Yang, Y., Duan, N. & Chen, W. (2023). *CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing.* arXiv preprint. https://arxiv.org/abs/2305.11738

Guan, J., Dodge, J., Wadden, D., Huang, M. & Peng, H. (2023). *Language Models Hallucinate, but May Excel at Fact Verification.* arXiv preprint. https://arxiv.org/abs/2310.14564

Guerreiro, N. M., Colombo, P., Piantanida, P. & Martins, A. F. T. (2022). *Optimal Transport for Unsupervised Hallucination Detection in Neural Machine Translation.* arXiv preprint. https://arxiv.org/abs/2212.09631

Gu, D., On, B. & Jeong, D. (2022). "Relevance and Redundancy-based Loss Function of KoBART Model for Improvement of the Factual Inconsistency Problem in Abstractive Summarization." *The Journal of Korean Institute of Information Technology, 20*(12), 25-36.

{구다훈 · 온병원 · 정동원 (2022). 생성요약의 사실 불일치 문제 개선을 위한 관련성과 중복성을 고려한 손실 함수 기반의 KoBART 모델. 〈한국정보기술학회논문지〉, 20권 12호, 25-36.}

Gupta, V., Pandya, P., Kataria, T., Gupta, V. & Roth, D. (2023). *Multi-Set Inoculation: Assessing Model Robustness Across Multiple Challenge Sets.* arXiv preprint. https://arxiv.org/abs/2311.08662

Ha, D., Dai, A. & Le, Q. V. (2016). Hypernetworks. arXiv preprint. https://arxiv.org/abs/1609.09106

He, Z., Liang, T., Jiao, W., Zhang, Z., Yang, Y., Wang, R., Tu, Z., Shi, S. & Wang, X. (2023). Exploring *Human-Like Translation Strategy with Large Language Models.* arXiv preprint. https://arxiv.org/abs/2305.04118

Hua, W., Xu, S., Ge, Y. & Zhang, Y. (2023). *How to Index Item IDs for Recommendation Foundation Models.* arXiv preprint. https://arxiv.org/abs/2305.06569

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B. & Liu, T. (2023a). *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions.* arXiv preprint. https://arxiv.org/abs/2311.05232

Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W. & Yu, N. (2023b). OPERA: *Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation.* arXiv preprint. https://arxiv.org/abs/2311.17911

Jafari, M., Sadeghi, D., Shoeibi, A., Alinejad-Rokny, H., Beheshti, A., García, D. L., Chen, Z., Acharya, U. R. & Gorriz, J. M. (2023). *Empowering Precision Medicine: AI-Driven Schizophrenia Diagnosis via EEG Signals: A Comprehensive Review from 2002-2023.* arXiv preprint. https://arxiv.org/abs/2309.12202

Ji, Z., Liu, Z., Lee, N., Yu, T., Wilie, B., Zeng, M. & Fung, P. (2022). *RHO (ρ): Reducing Hallucination in Open-domain Dialogues with Knowledge Grounding.* arXiv preprint. https://arxiv.org/abs/2212.01588

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A. & Fung, P. (2023a). "Survey of Hallucination in Natural Language Generation." *ACM Computing Surveys, 55*(12), 1-38.

Ji, Z., Yu, T., Xu, Y., Lee, N., Ishii, E. & Fung, P. (2023b). *Towards Mitigating Hallucination in Large Language Models via Self-Reflection.* arXiv preprint. https://arxiv.org/abs/2310.06271

Jiang, C., Xu, H., Dong, M., Chen, J., Ye, W., Yan, M., Ye, Q., Zhang, J., Huang, F. & Zhang, S. (2023). *Hallucination Augmented Contrastive Learning for Multimodal Large Language Model.* arXiv preprint. https://arxiv.org/abs/2312.06968

Jiao, W., Wang, W., Huang, J., Wang, X., Shi, S. & Tu,

Z. (2023). *Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine.* arXiv preprint. https://arxiv.org/abs/2301.08745

Jha, S., Jha, S. K., Lincoln, P., Bastian, N. D., Velasquez, A. & Neema, S. (2023). *Dehallucinating Large Language Models Using Formal Methods Guided Iterative Prompting.* Paper presented at 2023 IEEE International Conference on Assured Autonomy (ICAA), June 6-8.

Kamalloo, E., Dziri, N., Clarke, C. L. A. & Rafiei, D. (2023). *Evaluating Open-Domain Question Answering in the Era of Large Language Models.* arXiv preprint. https://arxiv.org/abs/2305.06984

Kanda, N., Yoshioka, T. & Liu, Y. (2023). *Factual Consistency Oriented Speech Recognition.* arXiv preprint. https://arxiv.org/abs/2302.12369

Kang, C. & Choi, J. (2023). *Impact of Co-occurrence on Factual Knowledge of Large Language Models.* arXiv preprint. https://arxiv.org/abs/2310.08256

Kang, H. & Liu, X. Y. (2023). *Deficiency of Large Language Models in Finance: An Empirical Examination of Hallucination.* arXiv preprint. https://arxiv.org/abs/2311.15548

Kang, H., Ni, J. & Yao, H. (2023). *Ever: Mitigating Hallucination in Large Language Models through Real-Time Verification and Rectification.* arXiv preprint. https://arxiv.org/abs/2311.09114

Kasai, J., Sakaguchi, K., Takahashi, Y., Le Bras, R., Asai, A., Yu, X.V., Radev, D.R., Smith, N.A., Choi, Y. & Inui, K. (2022). *RealTime QA: What's the Answer Right Now?* arXiv preprint. https://arxiv.org/abs/2207.13332

Kasanishi, T., Isonuma, M., Mori, J. & Sakata, I. (2023). *SciReviewGen: A Large-scale Dataset for Automatic Literature Review Generation.* arXiv preprint. https://arxiv.org/abs/2305.15186

Khalid, H., Tariq, S., Kim, M. & Woo, S. (2021). *FakeAVCeleb: A novel audio-video multimodal deepfake dataset.* arXiv preprint. https://arxiv.org/abs/2108.05080

Köksal, A., Aksitov, R. & Chang, C. (2023). *Hallucination Augmented Recitations for Language Models.* arXiv preprint. https://arxiv.org/abs/2311.07424

Ladhak, F., Durmus, E., Suzgun, M., Zhang, T., Jurafsky, D., McKeown, K. & Hashimoto, T. (2023). *When Do Pre-Training Biases Propagate to Downstream Tasks? A Case Study in Text Summarization.* Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 3206–3219.

Lattimer, B. M., Chen, P., Zhang, X. & Yang, Y. (2023). Fast and *Accurate Factual Inconsistency Detection Over Long Documents.* arXiv preprint. https://arxiv.org/abs/2310.13189

Lee, N., Ping, W., Xu, P., Patwary, M., Shoeybi, M. & Catanzaro, B. (2022). *Factuality Enhanced Language Models for Open-Ended Text Generation.* arXiv preprint. https://arxiv.org/abs/2206.04624

Lee, Z. & Nam, H. (2022). "A Literature Review Study in the Field of Artificial Intelligence (AI) Aplications, AI-Related Management, and AI Application Risk." *Informatization Policy, 29*(2), 3-36.

{이준기 · 남효경 (2022). 인공지능의 활용, 프로젝트 관리 그리고 활용 리스크에 대한 문헌 연구. 〈정보화정책〉, 29권 2호, 3-36.}

Lei, D., Li, Y., Hu, M., Wang, M., Yun, V., Ching, E. & Kamal, E. (2023). *Chain of Natural Language Inference for Reducing Large Language Model Ungrounded Hallucinations.* arXiv preprint. https://arxiv.org/abs/2310.03951

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S. & Kiela, D. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." *Advances in Neural Information Processing Systems, 33*, 9459-9474.

Li, J., Cheng, X., Zhao, W. X., Nie, J. Y. & Wen, J. R. (2023a). *HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language*

Models. arXiv preprint. https://arxiv.org/abs/2305.11747

Li, Z., Zhang, S., Zhao, H., Yang, Y. & Yang, D. (2023b). *BatGPT: A Bidirectional Autoregessive Talker from Generative Pre-trained Transformer.* arXiv preprint. https://arxiv.org/abs/2307.00360

Li, K., Patel, O., Vi'egas, F., Pfister, H. & Wattenberg, M. (2023c). *Inference-Time Intervention: Eliciting Truthful Answers from a Language Model.* arXiv preprint. https://arxiv.org/abs/2306.03341

Li, B., Zhou, B., Wang, F., Fu, X., Roth, D. & Chen, M. (2023d). *Deceiving Semantic Shortcuts on Reasoning Chains: How Far Can Models Go without Hallucination?.* arXiv preprint. https://arxiv.org/abs/2311.09702

Li, Y., Li, Y., Zhang, M., Su, C., Ren, M., Qiao, X., Zhao, X., Piao, M., Yu, J., Lv, X., Ma, M., Zhao, Y. & Yang, H. (2023e). *A Multitask Training Approach to Enhance Whisper with Contextual Biasing and Open-Vocabulary Keyword Spotting.* arXiv preprint. https://arxiv.org/abs/2309.09552

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X. & Wen, J. R. (2023f). *Evaluating Object Hallucination in Large Vision-Language Models.* arXiv preprint. https://arxiv.org/abs/2305.10355

Li, J., Chen, J., Ren, R., Cheng, X., Zhao, W. X., Nie, J. Y. & Wen, J. R. (2024). *The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models.* arXiv preprint. https://arxiv.org/abs/2401.03205

Lin, S.C., Hilton, J. & Evans, O. (2022). *TruthfulQA: Measuring How Models Mimic Human Falsehoods. Annual Meeting of the Association for Computational Linguistics.* Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 3214–3252.

Liu, W., Li, G., Zhang, K., Du, B., Chen, Q., Hu, X., Xu, H., Chen, J. & Wu, J. (2023a). *Mind's Mirror: Distilling Self-Evaluation Capability and Comprehensive Thinking from Large Language Models.* arXiv preprint. https://arxiv.org/abs/2311.09214

Liu, B., Ash, J.T., Goel, S., Krishnamurthy, A. & Zhang, C. (2023b). *Exposing Attention Glitches with Flip-Flop Language Modeling.* arXiv preprint. https://arxiv.org/abs/2306.00946

Liu, G., Wang, X., Yuan, L., Chen, Y. & Peng, H. (2023c). *Prudent Silence or Foolish Babble? Examining Large Language Models' Responses to the Unknown.* arXiv preprint. https://arxiv.org/abs/2311.09731

Liu, Y., Wang, K., Shao, W., Luo, P., Qiao, Y., Shou, M. Z., Zhang, K. & You, Y. (2023d). *MLLMs-Augmented Visual-Language Representation Learning.* arXiv preprint. https://arxiv.org/abs/2311.18765

Lovenia, H., Dai, W., Cahyawijaya, S., Ji, Z. & Fung, P. (2023). *Negative Object Presence Evaluation (NOPE) to Measure Object Hallucination in Vision-Language Models.* arXiv preprint. https://arxiv.org/abs/2310.05338

Luo, J., Xiao, C. & Ma, F. (2023). *Zero-Resource Hallucination Prevention for Large Language Models.* arXiv preprint. https://arxiv.org/abs/2309.02654

Luo, J., Li, T., Wu, D., Jenkin, M., Liu, S. & Dudek, G. (2024). *Hallucination Detection and Hallucination Mitigation: An Investigation.* arXiv preprint. https://arxiv.org/abs/2401.08358

Ma, W., Liu, S., Wang, W., Hu, Q., Liu, Y., Zhang, C., Nie, L. & Liu, Y. (2023). *ChatGPT: Understanding Code Syntax and Semantics.* arXiv preprint. https://arxiv.org/abs/2305.12138

Manakul, P., Liusie, A. & Gales, M.J. (2023). *SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models.* arXiv preprint. https://arxiv.org/abs/2303.08896

Mei, K. & Zhang, Y. (2023). *LightLM: A Lightweight Deep and Narrow Language Model for Generative Recommendation.* arXiv preprint. https://arxiv.org/abs/2310.17488.

Meng, K., Bau, D., Andonian, A. & Belinkov, Y. (2022).

"Locating and Editing Factual Associations in GPT." *Advances in Neural Information Processing Systems, 35*, 17359-17372.

Miao, M., Meng, F., Liu, Y., Zhou, X. H. & Zhou, J. (2021). *Prevent the Language Model from Being Overconfident in Neural Machine Translation.* arXiv preprint. https://arxiv.org/abs/2105.11098

Miao, N., Teh, Y.W. & Rainforth, T. (2023). *SelfCheck: Using LLMs to Zero-Shot Check Their Own Step-by-Step Reasoning.* arXiv preprint. https://arxiv.org/abs/2308.00436

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W., Koh, P., Iyyer, M., Zettlemoyer, L. & Hajishirzi, H. (2023). *FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation.* arXiv preprint. https://arxiv.org/abs/2305.14251

Mitchell, E., Lin, C., Bosselut, A., Finn, C. & Manning, C.D. (2022, April 25-29). *Fast Model Editing at Scale* [Poster Presentation]. The Tenth International Conference on Learning Representations, ICLR 2022 Virtual Event. https://openreview.net/forum?id=0DcZxeWfOPt

Mohammadshahi, A., Vamvas, J. & Sennrich, R. (2023). *Investigating Multi-Pivot Ensembling with Massively Multilingual Machine Translation Models.* arXiv preprint. https://arxiv.org/abs/2311.07439

Montagnese, M., Leptourgos, P., Fernyhough, C., Waters, F., Larøi, F., Jardri, R., McCarthy-Jones, S., Thomas, N., Dudley, R., Taylor, J.-P., Collerton, D. & Urwyler, P. (2021). "A review of multimodal hallucinations: categorization, assessment, theoretical perspectives, and clinical recommendations." *Schizophrenia Bulletin, 47*(1), 237-248.

Moses, L. (2024). "OpenAI's Sora's Best Features and Biggest Limitations." *Business Insider,* April 19.

Muhlgay, D., Ram, O., Magar, I., Levine, Y., Ratner, N., Belinkov, Y., Abend, O., Leyton-Brown, K., Shashua, A. & Shoham, Y. (2023). *Generating Benchmarks for Factuality Evaluation of Language Models.* arXiv preprint. https://arxiv.org/abs/2307.06908

Nathani, D., Wang, D., Pan, L. & Wang, W. Y. (2023). *MAF: Multi-Aspect Feedback for Improving Reasoning in Large Language Models.* arXiv preprint. https://arxiv.org/abs/2310.12426

Oh, D. (2024). "Research Trends for Dehallucination of Natural Language Generation Model." *Communications of the Korean Institute of Information Scientists and Engineers, 42*(1), 15-20.

{오동석 (2024). 자연어 생성 모델의 탈환각을 위한 연구 동향. 〈정보과학회지〉, 42권 1호, 15-20.}

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J. & Lowe, R. (2022). "Training Language Models to Follow Instructions with Human Feedback." *Advances in neural information processing systems, 35*, 27730-27744.

Pan, L., Saxon, M., Xu, W., Nathani, D., Wang, X. & Wang, W. Y. (2023). *Automatically Correcting Large Language Models: Surveying the Landscape of Diverse Self-Correction Strategies.* arXiv preprint. https://arxiv.org/abs/2308.03188

Park, D. (2024). *Media Artificial Intelligence.* Seoul: Yulgokbook Publishing Company.

{박대민 (2024). 〈미디어 인공지능: 영상 분야의 딥러닝 활용을 중심으로〉. 서울: 율곡출판사.}

Park, D. (2023a). "Journalism Artificial Intelligence Based on Trustworthy Artificial Intelligence : Toward a Commensurability between Media Trust and Trustworthiness of Artificial Intelligence System." *Media & Society, 31*(4), 5-47.

{박대민 (2023a). 신뢰할 수 있는 인공지능 기반의 저널리즘 인공지능: 언론 신뢰와 인공지능 신뢰성 간 통약가능성을 바탕으로. 〈언론과 사회〉, 31권 4호, 5-47.}

Park, D. (2023b). "Topology of Media Bias : Fat-Tailed Distribution as Universal Distribution of

Quotation by Analyzing News Source Networks with 16.5 Million Articles." *Korean Journal of Journalism & Communication Studies, 67*(6), 189-222.

{박대민 (2023b). 편향의 위상학 : 1650만 건 기사의 뉴스 정보원 연결망 분석을 통해 파악한 인용 방식의 보편적 분포로서 두터운 꼬리 분포. 〈한국언론학보〉, 67권 6호, 189-222.}

Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., Huang, Q., Liden, L., Yu, Z., Chen, W. & Gao, J. (2023). *Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback.* arXiv preprint. https://arxiv.org/abs/2302.12813

Press, O., Zhang, M., Min, S., Schmidt, L., Smith, N. A. & Lewis, M. (2022). *Measuring and Narrowing the Compositionality Gap in Language Models.* arXiv preprint. https://arxiv.org/abs/2210.03350

Qiu, Z., Liu, W., Xiao, T. Z., Liu, Z., Bhatt, U., Luo, Y., Weller, A. & Schölkopf, B. (2022). *Iterative Teaching by Data Hallucination.* arXiv preprint. https://arxiv.org/abs/2210.17467

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. (2019). "Language Models are Unsupervised Multitask Learners." https://openai.com/research/better-language-models (Retrieved on April 27, 2024).

Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K. & Shoham, Y. (2023). *In-Context Retrieval-Augmented Language Models.* arXiv preprint. https://arxiv.org/abs/2302.00083

Rawte, V., Chakraborty, S., Pathak, A., Sarkar, A., Tonmoy, S. M. T. I., Chadha, A., Sheth, A. P. & Das, A. (2023a). *The Troubling Emergence of Hallucination in Large Language Models - An Extensive Definition, Quantification, and Prescriptive Remediations.* arXiv preprint. https://arxiv.org/abs/2310.04988

Rawte, V., Sheth, A. & Das, A. (2023b). *A Survey of Hallucination in Large Foundation Models.* arXiv preprint. https://arxiv.org/abs/2309.05922

Rehman, T., Mandal, R., Agarwal, A. & Sanyal, D. K. (2023). *Hallucination Reduction in Long Input Text Summarization.* arXiv preprint. https://arxiv.org/abs/2309.16781

Rejeleene, R., Xu, X. & Talburt, J. (2024). *Towards Trustable Language Models: Investigating Information Quality of Large Language Models.* arXiv preprint. https://arxiv.org/abs/2401.13086

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. (2022). *High-Resolution Image Synthesis with Latent Diffusion Models.* Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 10684-10695.

Saha, S., Yu, X. V., Bansal, M., Pasunuru, R. & Celikyilmaz, A. (2022). *MURMUR: Modular Multi-Step Reasoning for Semi-Structured Data-to-Text Generation.* arXiv preprint. https://arxiv.org/abs/2212.08607

Sarkar, D., Bali, R. & Ghosh, T. (2018). *Hands-On Transfer Learning with Python: Implement Advanced Deep Learning and Neural Network Models using TensorFlow and Keras.* Packt Publishing.

Saunders, W., Yeh, C., Wu, J., Bills, S., Long, O., Ward, J. & Leike, J. (2022). *Self-Critiquing Models for Assisting Human Evaluators.* arXiv preprint. https://arxiv.org/abs/2206.05802

Schulman, J. (2023). "Reinforcement Learning from Human Fheedback: Progress and Challenges." https://www.youtube.com/watch?v=hhiLw5Q_UFg. (Retrieved on April 27, 2024).

Shi, W., Han, X., Lewis, M., Tsvetkov, Y., Zettlemoyer, L. & Yih, S. W. T. (2023). *Trusting Your Evidence: Hallucinate Less with Context-Aware Decoding.* arXiv preprint. https://arxiv.org/abs/2305.14739

Shi, Z., Wang, Z., Fan, H., Yin, Z., Sheng, L., Qiao, Y. & Shao, J. (2023). *ChEF: A Comprehensive Evaluation Framework for Standardized Assessment of Multimodal Large Language*

Models. arXiv preprint. https://arxiv.org/abs/2311.02692

Si, C., Gan, Z., Yang, Z., Wang, S., Wang, J., Boyd-Graber, J. L. & Wang, L. (2022). *Prompting GPT-3 To Be Reliable*. arXiv preprint. https://arxiv.org/abs/2210.09150

Song, M. & Lee, S. (2024). "What Concerns Does ChatGPT Raise for Us?: An Analysis Centered on CTM (Correlated Topic Modeling) of YouTube Video News Comments." *Informatization Policy, 31*(1), 3-31.

{송민호 · 이수범 (2024). ChatGPT는 우리에게 어떤 우려를 초래하는가?: 유튜브 영상 뉴스 댓글의 CTM(Correlated Topic Modeling) 분석을 중심으로. 〈정보화정책〉, 31권 1호, 3-31.}

Sun, Z., Shen, Y., Zhou, Q., Zhang, H., Chen, Z., Cox, D.D., Yang, Y. & Gan, C. (2023a). *Principle-Driven Self-Alignment of Language Models from Scratch with Minimal Human Supervision.* arXiv preprint. https://arxiv.org/abs/2305.03047

Sun, Q., Yin, Z., Li, X., Wu, Z., Qiu, X. & Kong, L. (2023b). *Corex: Pushing the Boundaries of Complex Reasoning through Multi-Model Collaboration.* arXiv preprint. https://arxiv.org/abs/2310.00280

Tam, D., Mascarenhas, A., Zhang, S., Kwan, S., Bansal, M. & Raffel, C. (2022). *Evaluating the Factual Consistency of Large Language Models Through News Summarization*. arXiv preprint. https://arxiv.org/abs/2211.08412

Tian, K., Mitchell, E., Yao, H., Manning, C. D. & Finn, C. (2023). *Fine-tuning Language Models for Factuality*. arXiv preprint. https://arxiv.org/abs/2311.08401

Tonmoy, S. M. T. I., Zaman, S. M. M., Jain, V., Rani, A., Rawte, V., Chadha, A. & Das, A. (2024). *A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models. arXiv* preprint. https://arxiv.org/abs/2401.01313

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E. & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models.* arXiv preprint. https://arxiv.org/abs/2302.13971

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). "Attention is All You Need." *Advances in neural information processing systems, 30*.

Verma, S., Goel, T., Tanveer, M., Ding, W. & Sharma, R. (2024). *Machine learning techniques for the Schizophrenia diagnosis: A comprehensive review and future research directions.* arXiv preprint. https://arxiv.org/abs/2301.07496

Vu, T., Iyyer, M., Wang, X., Constant, N., Wei, J., Wei, J., Tar, C., Sung, Y., Zhou, D., Le, Q. & Luong, T. (2023). *FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation.* arXiv preprint. https://arxiv.org/abs/2310.03214

Wan, D., Zhang, S. & Bansal, M. (2023). *HistAlign: Improving Context Dependency in Language Generation by Aligning with History.* arXiv preprint. https://arxiv.org/abs/2305.04782

Wang, Y., Zhong, W., Li, L., Mi, F., Zeng, X., Huang, W., Shang, L., Jiang, X. & Liu, Q. (2023a). *Aligning Large Language Models with Human: A Survey.* arXiv preprint. https://arxiv.org/abs/2307.12966

Wang, Z., Mao, S., Wu, W., Ge, T., Wei, F. & Ji, H. (2023b). *Unleashing the Emergent Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration*. arXiv preprint. https://arxiv.org/abs/2307.05300

Wang, J., Wang, Y., Xu, G., Zhang, J., Gu, Y., Jia, H., Yan, M., Zhang, J. & Sang, J. (2023c). *An LLM-free Multi-dimensional Benchmark for MLLMs Hallucination Evaluation.* arXiv preprint. https://arxiv.org/abs/2311.07397

Wang, B., Wu, F., Han, X., Peng, J., Zhong, H., Zhang, P., Dong, X., Li, W., Li, W., Wang, J. & He, C. (2023d). *VIGC: Visual Instruction Generation and Correction.* arXiv preprint. https://arxiv.org/abs/2308.12714

Wang, J., Zhou, Y., Xu, G., Shi, P., Zhao, C., Xu, H., Ye,

Q., Yan, M., Zhang, J., Zhu, J. & Sang, J. (2023e). *Evaluation and Analysis of Hallucination in Large Vision-Language Models.* arXiv preprint. https://arxiv.org/abs/2308.15126

Wang, F. (2024). *LightHouse: A Survey of AGI Hallucination.* arXiv preprint. https://arxiv.org/abs/2401.06792

Wang, J., Chang, Y., Li, Z., An, N., Ma, Q., Hei, L., Luo, H., Lu, Y. & Ren, F. (2024a). *TechGPT-2.0: A large language model project to solve the task of knowledge graph construction.* arXiv preprint. https://arxiv.org/abs/2401.04507

Wang, X., Zhou, Y., Liu, X., Lu, H., Xu, Y., He, F., Yoon, J., Lu, T., Bertasius, G., Bansal, M., Yao, H. & Huang, F. (2024b). *Mementos: A Comprehensive Benchmark for Multimodal Large Language Model Reasoning over Image Sequences.* arXiv preprint. https://arxiv.org/abs/2401.10529

Wei, J. W., Huang, D., Lu, Y., Zhou, D. & Le, Q. (2023). *Simple Synthetic Data Reduces Sycophancy in Large Language Models.* arXiv preprint. https://arxiv.org/abs/2308.03958

Wilie, B., Xu, Y., Chung, W., Cahyawijaya, S., Lovenia, H. & Fung, P. (2023). *PICK: Polished & Informed Candidate Scoring for Knowledge-Grounded Dialogue Systems.* arXiv preprint. https://arxiv.org/abs/2309.10413

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J. & Hooi, B. (2023). *Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs.* arXiv preprint. https://arxiv.org/abs/2306.13063

Xu, P., Shao, W., Zhang, K., Gao, P., Liu, S., Lei, M., Meng, F., Huang, S., Qiao, Y. & Luo, P. (2023). *LVLM-eHub: A Comprehensive Evaluation Benchmark for Large Vision-Language Models.* arXiv preprint. https://arxiv.org/abs/2306.09265

Xue, T., Wang, Z., Wang, Z., Han, C., Yu, P. & Ji, H. (2023). *RCOT: Detecting and Rectifying Factual Inconsistency in Reasoning by Reversing Chain-of-Thought.* arXiv preprint. https://arxiv.org/abs/2305.11499

Yang, Z., Dai, Z., Salakhutdinov, R. & Cohen, W.W. (2018). *Breaking the Softmax Bottleneck: A High-Rank RNN Language Model.* Paper presented at 6th International Conference on Learning Representations, ICLR 2018, April 30 - May 3, 2018.

Yang, S., Sun, R. & Wan, X. (2023a). *A New Benchmark and Reverse Validation Method for Passage-level Hallucination Detection.* arXiv preprint. https://arxiv.org/abs/2310.06498

Yang, L., Zhang, S., Yu, Z., Bao, G., Wang, Y., Wang, J., Xu, R., Ye, W., Xie, X., Chen, W. & Zhang, Y. (2023c). *Supervised Knowledge Makes Large Language Models Better In-context Learners.* arXiv preprint. https://arxiv.org/abs/2312.15918

Yao, J., Ning, K., Liu, Z., Ning, M. & Yuan, L. (2023a). *LLM Lies: Hallucinations are not Bugs, but Features as Adversarial Examples.* arXiv preprint. https://arxiv.org/abs/2310.01469

Yao, Y., Xu, X. & Liu, Y. (2023b). *Large Language Model Unlearning.* arXiv preprint. https://arxiv.org/abs/2310.10683

Ye, H., Liu, T., Zhang, A., Hua, W. & Jia, W. (2023a). *Cognitive Mirage: A Review of Hallucinations in Large Language Models.* arXiv preprint. https://arxiv.org/abs/2309.06794

Yu, X., Cheng, H., Liu, X., Roth, D. & Gao, J. (2023). *Automatic Hallucination Assessment for Aligned Large Language Models via Transferable Adversarial Attacks.* arXiv preprint. https://arxiv.org/abs/2310.12516

Yun, H. S., Marshall, I. J., Trikalinos, T. A. & Wallace, B. C. (2023). *Appraising the Potential Uses and Harms of LLMs for Medical Systematic Reviews.* arXiv preprint. https://arxiv.org/abs/2305.11828

Zha, Y., Yang, Y., Li, R. & Hu, Z. (2023). *AlignScore: Evaluating Factual Consistency with A Unified Alignment Function.* Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 11328–11348.

Zhai, Y., Tong, S., Li, X., Cai, M., Qu, Q., Lee, Y. J. & Ma, Y. (2023). *Investigating the Catastrophic Forgetting in Multimodal Large Language Models.* arXiv preprint. https://arxiv.org/abs/2309.10313

Zhang, H., Duckworth, D., Ippolito, D. & Neelakantan, A. (2020). *Trading off diversity and quality in natural language generation.* arXiv preprint. https://arxiv.org/abs/2004.10450

Zhang, J., Li, Z., Das, K., Malin, B. & Sricharan, K. (2023c). SAC3: *Reliable Hallucination Detection in Black-Box Language Models via Semantic-aware Cross-check Consistency.* arXiv preprint. https://arxiv.org/abs/2311.01740

Zhang, M., Press, O., Merrill, W., Liu, A. & Smith, N. A. (2023b). *How Language Model Hallucinations Can Snowball.* arXiv preprint. https://arxiv.org/abs/2305.13534

Zhang, S., Pan, L., Zhao, J. & Wang, W.Y. (2023d). *The Knowledge Alignment Problem: Bridging Human and External Knowledge for Large Language Models.* arXiv preprint. https://arxiv.org/abs/2305.13669

Zhang, T., Qiu, L., Guo, Q., Deng, C., Zhang, Y., Zhang, Z., Zhou, C., Wang, X. & Fu, L. (2023e). *Enhancing Uncertainty-Based Hallucination Detection with Stronger Focus.* arXiv preprint. https://arxiv.org/abs/2311.13230

Zhang, Y., Cui, L., Bi, W. & Shi, S. (2023f). *Alleviating Hallucinations of Large Language Models through Induced Hallucinations.* arXiv preprint. https://arxiv.org/abs/2312.15710

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A.T., Bi, W., Shi, F. & Shi, S. (2023a). *Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models.* arXiv preprint. https://arxiv.org/abs/2309.01219

Zhao, R., Li, X., Joty, S.R., Qin, C. & Bing, L. (2023a). *Verify-and-Edit: A Knowledge-Enhanced Chain-of-Thought Framework.* Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 5823–5840.

Zhao, Z., Wang, B., Ouyang, L., Dong, X., Wang, J. & He, C. (2023b). *Beyond Hallucinations: Enhancing LVLMs through Hallucination-Aware Direct Preference Optimization.* arXiv preprint. https://arxiv.org/abs/2311.16839

Zhong, Z., Wu, Z., Manning, C. D., Potts, C. & Chen, D. (2023). *Mquake: Assessing Knowledge Editing in Language Models via Multi-Hop Questions.* arXiv preprint. https://arxiv.org/abs/2305.14795

Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., Zhang, S., Ghosh, G., Lewis, M., Zettlemoyer, L. & Levy, O. (2023a). *LIMA: Less Is More for Alignment.* arXiv preprint. https://arxiv.org/abs/2305.11206

Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M. & Yao, H. (2023b). *Analyzing and Mitigating Object Hallucination in Large Vision-Language Models. arXiv* preprint. https://arxiv.org/abs/2310.00754

Zhu, J., Qi, J., Ding, M., Chen, X., Luo, P., Wang, X., Liu, W., Wang, L. & Wang, J. (2023). *Understanding Self-Supervised Pretraining with Part-Aware Representation Learning.* arXiv preprint. https://arxiv.org/abs/2301.11915

Zong, M. & Krishnamachari, B. (2022). *A Survey on GPT-3.* arXiv preprint. https://arxiv.org/abs/2212.00857