컴퓨터교육학회 논문지 2026년 제29권 제1호
https://doi.org/10.32431/kace.2026.29.1.002

인공지능 환각에 대한 교육적 대응 전략과 연구 방향 - 국내외 연구 분석을 바탕으로 한 시사점*

Educational Response Strategies and Research Directions for AI Hallucinations: Insights from Domestic and International Research Analyses

박윤수† · 박호현†† · 이유미†††
Younsoo Park† · Ho-Hyun Park†† · Yumi Yi†††

요약

2022년 11월 ChatGPT가 등장한 이후로 생성형 인공지능 서비스는 인공지능 교육 분야에서 집중적으로 연구되고 있다. 이는 생성형 인공지 서비스가 인간에 근접한 수준의 상호작용을 제공하는 것이 가능하고, 기존 인공지능 챗봇 서비스보다 그럴듯한 응답을 제공하기 때문이다. 그러나 생성형 인공지능은 학습 데이터와 모델의 편향 등의 요인으로 인해 인공지능 환각을 포함하는 응답을 생성할 수 있다. 이와 관련하여 일부 국내외 연구가 인공지능 환각 현상의 원인을 규명하고, 부정적인 영향을 경감하기 위한 주제를 다루고 있다. 그러나 아직까지는 국내 연구 사례의 수가 해외 연구 사례 대비 매우 적은 편이다. 이에 본 연구에서는 인공지능 환각에 대한 해외 연구 동향을 분석하고, 대표적인 연구 사례를 분석함으로써 인공지능 환각 연구의 현황을 파악하는 한편, 이를 바탕으로 생성형 인공지능의 교육적 활용에 대한 방향성을 제시하고자 한다. 분석 결과를 바탕으로 생성형 인공지능의 교육적 활용에 있어서 인공지능 환각에 대한 명확한 인지와 현실적인 인공지능 환각 경감 방법의 활용이 필요하다는 결론을 도출하였다.

주제어 생성형 인공지능, 연구 동향, 인공지능 환각, 키워드 네트워크 분석, 편향성

ABSTRACT

Since the introduction of ChatGPT in November 2022, generative AI (GAI) services have been intensively studied in the field of AI education. This is because GAI services can provide more human-like interactions and more plausible responses compared to traditional AI services. However, GAI services can generate responses that include AI hallucinations due to factors such as training data bias and model bias. In relation to this subject, some studies investigate the causes of AI hallucinations and explore strategies to mitigate their negative impact. However, the number of domestic studies remains smaller than that of international studies. In this study, we analyze international research trends related to AI hallucinations and review research cases to understand the current status of domestic AI hallucination research. Based on our findings, we propose directions for the educational use of GAI that account for the issue of AI hallucinations. Our analysis highlights the need for both a clear understanding of AI hallucinations and the practical application of methods to mitigate negative impacts for the educational use of GAI.

Keywords Genarative AI, Research Trends, AI Hallucinations, Keyword Network Analysis, Bias

†정회원 MKS 특허경영 전문위원
†정회원 중앙대학교 전자전기공학부 교수
††종신회원 중앙대학교 인공지능인문학연구소 교수 (교신저자)

논문투고 2025년 03월 31일
심사완료 2025년 08월 15일
게재확정 2025년 12월 24일
발행일자 2026년 01월 06일

*본 논문은 농림축산식품부 및 과학기술정보통신부, 농촌진흥청의 재원으로 농림식품기술기획평가원과 재단법인 스마트팜연구개발사업단의 스마트팜다부처패키지혁신기술개발사업(42500903)과 산업통상자원부의 재원으로 한국산업기술진흥원의 2025년 산업 전환형 무기발광 디스플레이 전문인력양성사업 (P0023718)의 지원을 받아 수행된 연구임.

1. 서론

생성형 인공지능 (Generative Artificial Intelligence, GAI) 대규모 데이터셋을 학습하여 음성, 텍스트, 이미지 등 새로운 디지털 콘텐츠를 생성하는 인공지능을 의미한다. 생성형 인공지능은 수십억 개 이상의 매개변수를 학습한 거대 인공지능 모델 (Large Artificial Intelligence Models)을 이용한다. 거대 인공지능 모델은 텍스트 생성에 특화된 거대 언어 모델 (Large Language Models, LLMs)과 텍스트, 음성, 이미지 등 다양한 디지털콘텐츠 생성을 지원하는 거대 멀티모달 모델(Large Multimodal Models, LMMs)로 분류된다[1].

대표적인 거대 인공지능 모델은 Google사의 PaLM-E, OpenAI사의 GPT, Runway AI사의 Stable Diffusion, Midjourney사의 Midjourney, Google사의 Gemini, DeepMind사의 Gopher, Anthropic사의 Claude, OpenAI사의 Codex, Microsoft사의 Vall-E 등이 알려져 있다[2].

생성형 인공지능 서비스는 생성형 인공지능을 서비스 목적에 부합하도록 학습시켜, 사용자에게 웹 또는 인터넷을 통해 서비스하는 것을 의미한다. 대표적인 생성형 인공지능 서비스로는 OpenAI사의 ChatGPT (Chat Generative Pre-trained Transformer)와 Google사의 GEMINI (Generalized Multimodal Intelligence Network)가 있으며[3, 4], 해당 서비스는 인간이 생성하는 텍스트, 음성, 이미지, 동영상, 코드 등의 대규모 데이터를 학습하여 사용자의 입력에 대해 적절한 디지털콘텐츠를 생성하여 응답한다. 세계적으로 다수의 생성형 인공지능 서비스가 공개되어 있는데, 그 중 텍스트, 이미지 및 동영상, 오디오(음성, 음원), 코드, 번역, 디자인 및 모델링 등 다양한 생성 기능에 특화되어 있는 생성형 인공지능 서비스 중 일부를 Table 1에 정리하였다. Table 1에서 볼 수 있듯이 생성형 인공지능 서비스는 챗봇 서비스에 국한되어 있지 않고, 이미지, 영상, 음성 데이터 생성 등 점차 디지털 콘텐츠 전반의 생성 기능을 제공하는 방향으로 발전하고 있다.

생성형 인공지능 서비스가 주목받기 시작한 것은 2022년 11월 30일 ChatGPT가 공개되면서부터이다[3]. ChatGPT는 사용자에게 친숙하면서도 직관적인 대화형 프롬프트를 제공함으로써 이용자의 편의성을 개선하였고, 기존 인공지능 서비스 대비 고차원적 추론 문제 해결 능력이 우수한 것으로 알려져 있다[5]. 이와 같이 ChatGPT의 성능을 비약적으로 발전시킨 생성형 인공지능이 인간이 생성한 산출물과 거의 구분할 수 없을 정도로 높은 품질의 콘텐츠를 생성할 수 있는 성능을 보여주면서 인공지능 기술의 수준이 초기 형태의 범용 인공지능 (Artificial General Intelligence, AGI)에 도달한 것인지에 대한 논쟁까지 생겨났다[6].

이러한 ChatGPT의 등장은 인공지능 교육과 소프트웨어 교육 분야에서도 새로운 도전 과제를 제시한다[7]. 일부 연구자들은 ChatGPT가 상당 수준의 교육 지원 능력을 보유하고 있는 것으로 판단하고 있지만[8], 윤리적 고려 사항과 학습 평가에 있어서 미칠 수 있는 부정적인 영향, 과학적 진실성, 학습자의 사고력에 미치는 영향 등에 대한 위험성에 대해서 논쟁하고 있다[9-12].

이에 대한 논쟁의 원인은 ChatGPT를 교육에 활용해야 하는지 여부에 대한 심도 있는 고찰에서부터 시작된다. Farrokhnia 외 3인은 ChatGPT를 교육 분야에서 활용하지 않고, 금지하는 방안은 실패할 가능성이 높으며, 존재를 부정하는 방안 또한 교육 환경에서 혼란을 초래할 가능성이 크다고 분석하였다[7]. 또한, 생성형 인공지능이 생성한 디지털콘텐츠 검출기를 이용하는 방안은 임시방편에 불과할 것으로 예상하였다. 즉, 교육에서 ChatGPT를 활용하는 것이 옳은지에 대한 완벽한 검증은 되지 않았지만, 광범위한 영역에서 ChatGPT가 사용되고 있고, 그 성능이 비약적으로 발전되고 있으며, 디지털콘텐츠 전반으로 활용 영역이 확대되고 있다(Table 1). 이처럼 생성형 인공지능 관련 기술과 서비스가 빠르게 발전하는 환경에서 신뢰성이 검증되지 않고, 부작용이 우려된다는 이유로 생성형 인공지능 서비스를 교육에서 배제하는 방안은 현실적이지 않다는 것이다. 따라서 인공지능 교육을 포함한 교육 전 분야는 부정적 효과를 제거하여 생성형 인공지능 서비스를 수용할 수 있는 방안을 모색해야 한다.

Table 1. Major GAI Services and Types of Responses

GAI Service	Response Type
ChatGPT (OpenAI)	Text
GEMINI (Google)	Text
Claude (Anthropic)	Text
Notion AI (Notion)	Text
Stable Diffusion (Stable AI)	Image
Midjourney (Midjourney)	Image
Runway ML (Runway)	Video
Runway ML (Runway)	Video & Video style
VALL-E (Microsoft)	Sound (Voice)
Amper Music (Shutterstock)	Sound (Voice)
GitHub Copilot (Microsoft&OpenAI)	Code
Tabnine (Tabnine)	Code
DeepL (DeepL)	Translation
Spline AI (Spline)	Design&Modeling

생성형 인공지능을 교육에 적용하는 데 있어 가장 문제가 되는 것 중 하나는 인공지능 환각 (Artificial Intelligence Hallucinations, AI Hallucinations)이다. 이는 생성형 인공지능 서비스가 잘못된 답변을 생성하거나 존재하지 않는 디지털콘텐츠를 생성하는 현상[13]을 말한다. McIntosh 외 5인의 연구는 생성형 인공지능 서비스에 문헌 검색을 요청했을 때 실존하지 않는 문헌을 실존하는 것처럼 생성하거나, 검색 결과를 마치 실존하는 결과인 것처럼 응답하는 환각 현상을 확인하였다[13]. 인공지능 환각은 GANs

(Generative Adversarial Networks)의 본질적인 특성(Inherent Property of GANs)으로, 모델의 학습 및 조정 과정에서 여러 요인에 의해서 발생할 수 있으며, 대규모의 데이터를 학습한 LLMs에서도 학습 데이터의 불완전성(Imcompleteness in the Training Data), 사용자 질문의 모호함과 불명확한 기대(Vague/Unclear Questions and Expectations by the User) 등의 요인으로 발생할 수 있다[14].

인공지능 환각에 대해서 생성형 인공지능 서비스가 인지하고 있는지 여부를 확인하기 위하여 주요 생성형 인공지능 서비스인 ChatGPT, Gemini, Claude, Notions AI에 각각의 단점에 대해서 기술하도록 프롬프트를 입력하여 확보한 응답을 Table 2에 기술하였다. 각 서비스에 입력한 질의는 “Briefly explain {service_name}'s five disadvantages”로, ChatGPT는 첫 번째 단점을 환각(Hallucinations (False information))으로 응답하였고, Claude는 Potential hallucinations를 네 번째 단점으로 응답하였다. Gemini는 인공지능 환각에 대한 직접적인 단점을 응답하지 않았지만, Bias and fairness, Explainability challenges, Technical complexity를 단점으로 응답한 점은 인공지능 환각으로 인한 효과가 간접적으로 성능에 영향을 미치는 요인으로 판단하고 있는 것으로 해석된다. 예를 들어, Bias and fairness는 생성형 인공지능의 학습 데이터 또는 모델이 편향되어 있어 잘못되거나 틀린 응답을 제공할 가능성을 의미하고, Explainability challenges는 Gemini의 응답이 복잡한 과정을 거쳐 산출되지만 (Technical complexity), 그와 같은 결과가 산출된 이유를 설명할 수 없는 단점을 의미한다. 즉, Gemini가 단점으로 제시한 항목들은 인공지능 환각으로 인한 부정확한 정보의 응답 가능성을 전제로 한 단점이다.

Table 2. Major GAI Services and Types of Responses

GAI Services	Response (5 Disadvantages)
ChatGPT (OpenAI)	1. Hallucination (False information) 2. Lack of real-time data 3. Context misunderstanding 4. Bias and ethical concerns 5. Dependency and creativity limitations
Gemini (Microsoft)	1. Limited availability 2. Technical complexity 3. Bias and fairness 4. Explainability challenges 5. Computational costs
Claude (Anthropic)	1. Knowledge cutoff 2. No web browsing 3. No persistent memory 4. Potential hallucinations 5. Ethical constraints
Notion AI (Notion)	1. Requires human review 2. Limited creativity 3. Context misinterpretation 4. Internet connection dependent 5. Learning curve

Notion AI 또한 인공지능 환각에 대하여 직접 관련이 있는 응답을 하지는 않았지만, 첫 번째 응답으로 Output needs human verification을 응답하였고, 이는 인공지능 환각으로 인해 응답을 완전히 신뢰할 수 없고, 그에 따라 사용자의 주의가 필요하다는 인공지능 환각으로 인한 간접적인 단점을 의미한다.

Table 2에 표현된 것처럼 생성형 인공지능 서비스가 제공하는 응답은 완벽하게 신뢰가능한 응답이 아니고, 특정 응답이 산출된 이유를 기술적으로 설명할 수 없는 단점이 있다. 새로운 지식을 처음으로 접하는 학습자에게 있어서 인공지능 환각은 잘못된 지식과 정보를 바탕으로 개념을 이해하고, 현상을 해석하게 할 가능성이 있으며, 이는 교육 외 분야에서의 인공지능 환각으로 인한 부정적 효과보다 잠재적으로 치명적인 결과를 초래할 가능성이 높다. 따라서 인공지능 교육을 비롯한 교육 전분야에서 ChatGPT를 비롯한 Gemini, Claude, Notion AI와 같은 생성형 인공지능 서비스의 교육적 활용에 대해서 고려한다면, 반드시 인공지능 환각의 영향을 회피 또는 경감할 수 있는 대비책을 마련해야 한다. 그러나 아직까지는 생성형 인공지능 서비스의 교육적 활용에 대한 연구에서 인공지능 환각으로 인한 부정적 영향을 고려한 교육 방법에 대한 논의는 소수만이 확인된다. 구체적으로 해외에서는 체계적이므로 인공지능 환각의 원인을 규명하고, 교육에서 활용을 해야 하는지 여부에 대해서 논의하고 있는 반면[15-34], 국내에서는 인공지능 환각에 대한 일부 사례 연구만이 확인된다[35-42]

이에 본 연구에서는 인공지능 환각에 대한 국내외 연구 동향을 살펴보고, 생성형 인공지능 서비스를 교육에 활용할 때 인공지능 환각을 어떤 측면에서 연구해야 할 것인가에 대해 고찰하고자 한다.

상기 연구를 수행하기 위하여 본 연구의 2절에서는 인공지능 환각을 주제로 한 문헌의 정보를 Web of Science (WoS)에서 수집하여 빈도수와 키워드 네트워크 분석을 수행하였다. 3절에서는 2절에서 분석된 내용을 바탕으로 주요 해외 연구 사례를 정성적으로 리뷰하였고, 키워드를 바탕으로 연구의 주제를 보다 명확하게 분류하였다. 4절에서는 학술지 인용색인(Korea Citation Index, KCI)에서 제목을 바탕으로 검색한 연구 문헌 7건을 선정하여 국내 연구 사례를 분석하였으며, 5절에서는 2절과 3절, 4절의 분석 결과를 바탕으로 인공지능 환각에 대한 교육적 대응 방안과 연구 방향성에 대해서 고찰하였다.

2. 인공지능 환각 해외 연구 문헌정보 분석

인공지능 환각에 대한 동향 분석을 위하여 Web of Science (WoS)에서 ‘Artificial Intelligence Hallucination’, ‘AI Hallucination’을 키워드로, Topic, Title, Author Keywords 필드에서 문헌정보를 검색하여 356건의 문헌 정보를 수집하고[15], 그에 대한 키워드를 추출하여 네트워크 분석을 수행하였다.

인공지능 환각에 대한 연구는 2022년 11월 ChatGPT가 등장하면서 급격히 증가하였다. 이에 데이터의 시간적 범위를 2022년부터 2024년 12월 25일 검색되는 부분으로 지정하였다.

키워드 네트워크 분석은 영문 연구문헌 중에서도 Author Keywords가 존재하는 데이터에 대해서 적용 가능하므로, WoS 데이터 중에서도 Language 필드가 English이고, Author Keywords 컬럼이 존재하지 않는 문헌정보를 제외하여 284건의 데이터를 확보하였다. 2024년 12월 기준 WoS에서 'Generative Artificial Intelligence' 키워드로 검색했을 시 13,427건의 연구문헌이 확인되고, 데이터의 시간적 범위를 2022년부터 2024년 12월 25일 검색되는 부분으로 한정했을 때 9,798건의 연구문헌이 검색된다. 따라서 동기간 동안 생성형 인공지능 대비 인공지능 환각 연구문헌의 비중은 2.90%로 산출된다. 인공지능 환각이 생성형 인공지능에서 나타날 가능성이 높은 현상임을 감안할 시 소수의 연구에서만 인공지능 환각에 대해서 다루고 있어 상대적으로 연구가 저조한 상황이다.

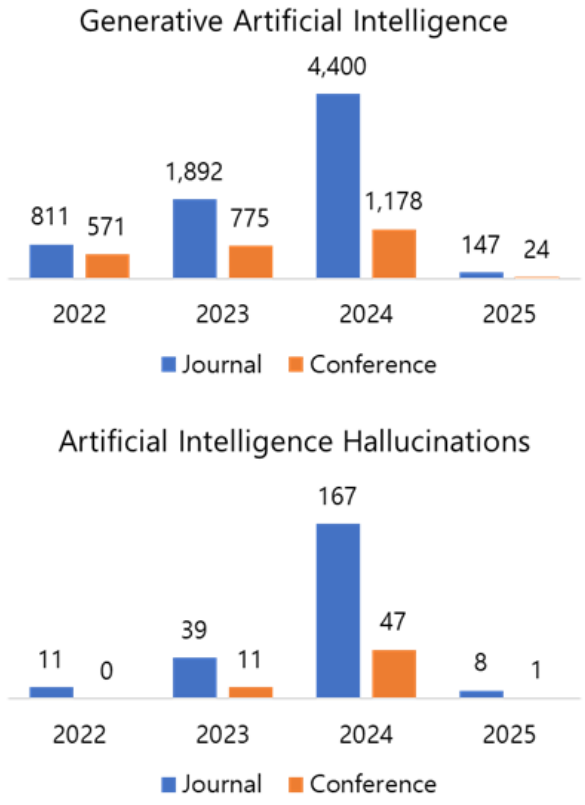

Figure 1. Publication Count by Years (Upper: GAI, Lower: AI hallucination)

해당 문헌을 저널/콘퍼런스, 연도로 구분한 결과를 Fig. 1에 표현하였다. 생성형 인공지능 관련 저널 연구문헌은 2022년 811건을 시작으로 2024년 4,400건으로 연평균 132.92% 증가하였고, 콘퍼런스 문헌은 2022년 571건을 시작으로 2024년 1,178건으로 연평균 43.63% 증가하였다. 2025년에는 각각 147건과 24건이 출판 예정(Early Access)인 것으로 집계되었다. 따라서 생성형 인공지능 관련 연구문헌은 콘퍼런스보다 저널에서 주로 출판되고 있고, 증가하는 수준 또한 저널 출판수가 콘퍼런스 발표수보다 3배가량 빠르게 증가하는 것으로 분석되었으며, 현재의 생성형 인공지능 서비스에 집중된 관심을 고려할 시 이후로도 이와 같은 경향이 지속될 것으로 예상된다.

인공지능 환각 관련 저널 연구문헌은 2022년 11건을 시작으로, 2023년 39건, 2024년 167건이 출판되었으며 분석기간 동안 연평균 289.64% 증가하였다. 콘퍼런스 문헌은 2022년은 0건이고, 2023년 11건, 2024년 47건이 출판되었다. 저널 문헌은 2025년 8건이 출판 예정이고, 콘퍼런스 문헌은 1건이 출판 예정이다.

인공지능 환각 연구 동향을 분석하기 위하여 데이터를 확보한 뒤 동일 개념 또는 유사 개념을 표현하는 방법을 통일시키기 위하여 코사인 유사도가 0.9 이상인 키워드를 동일 그룹으로 그룹화하였다[16]. 또한, 해당 그룹의 대표키워드를 지정하여 그룹 내 리스트와 동일한 키워드는 대표키워드로 치환하도록 프로그램을 작성하였다.

다만, 코사인 유사도는 단어 간 의미의 유사성을 바탕으로 그룹화할 수 있지만, 일부 키워드의 경우에는 전혀 관련이 없는 경우에도 그룹화하여 하나의 그룹으로 표현할 가능성이 있으며, 경우에 따라서는 복수와 단수를 별도의 그룹으로 그룹화하는 경우도 있다. 이에 각 그룹에 속한 키워드의 적정성을 정성적으로 판단하여 그룹화 과정에서 발생하는 오류를 정정하고, 일관성을 유지할 수 있도록 작업하였다.

예를 들어, 'AI chatbots', 'Chatbots', 'chatbots' 키워드는 상기 과정에 의해서 기계적으로 동일 그룹으로 그룹화되었고, 'AI chatbots' 키워드는 나머지 2개 키워드와 유사한 개념이지만, 동일 개념은 아니므로, 그룹에서 제거되어야 한다. 또한, 'Chatbot', 'chatbot'키워드는 'chatbot' 키워드로 그룹화되었는데, 이는 'Chatbots', 'chatbots' 키워드와 동일 개념을 의미하므로, 하나의 그룹으로 그룹화될 필요가 있다. 상기 서술된 방식에 따라서 'Chatbots', 'chatbots', 'Chatbot', 'chatbot' 키워드는 4개 키워드를 포괄할 수 있는 대표키워드인 'Chatbots'으로 그룹화하고, 'AI chatbots' 키워드는 그룹에서 제거하였다.

284건의 데이터에 대하여 중복을 제거하여 1,699건의 키워드를 추출하고, 각 문헌에 순번을 부여하여 {문헌번호, 키워드} 순으로 전처리하였다. 이를 Fig. 2와 같이 키워드 네트워크 그래프로 표현하고, 빈도수 분석을 수행한 결과 중 상위 30개 키워드를 Table 3에 나열하였다. 또한, 연결중심성 (Degree Centrality, DC), 매개중심성 (Betweenness Centrality, BC), 근접중심성(Closeness Centrality, CC), 위세중심성(Eigenvector Centrality, EC)을 분석한 결과를 Table 4에 나열하였다.

연결중심성은 키워드 네트워크상 특정 노드가 다른 노드와 가지는 연결의 수를 의미하며, 매개중심성은 특정 노드

Table 3. Frequency Analysis Results (TOP 30 keywords)

Keyword	Frequency	Keyword	Frequency	Keyword	Frequency
Artificial Intelligence (AI)	143	Ethics	11	eXplainable AI (XAI)	6
Large Language Model (LLM)	118	Education	11	Accuracy	6
ChatGPT	90	Language Model	10	Google Bard	6
Generative Artificial Intelligence (GAI)	57	Knowledge Graph	10	Diagnosis	6
Hallucinations	41	Schizophrenia	9	Bias	6
Machine Learning (ML)	28	Medical Education	9	Patient Education	5
Natural Language Processing (NLP)	25	Clinical Decision Support (CDS)	8	Software Engineering	5
Chatbots	23	Retrieval Augmented Generation (RAG)	8	Systematic Review	5
Deep Learning (DL)	19	Prompt Engineering	7	Academic Integrity	4
GPT	12	Generative Adversarial Networks (GANs)	6	Healthcare	4

를 통해서 다른 노드 간 연결되는 수준을 의미한다. 근접중심성은 특정 노드가 다른 모든 노드와 가까이 위치한 수준을 의미하며, 위세중심성은 근접중심성과 유사하게 다른 노드와의 연결의 수를 의미하지만, 연결된 노드의 중요도까지 고려하는 지표를 의미한다.

연결중심성의 평균은 0.0127, 매개중심성은 0.0015, 근접중심성은 0.3764, 위세중심성은 0.0200으로 분석되었다.

Artificial Intelligence (AI), Chatbots, ChatGPT, Generative Artificial Intelligence (GAI), GPT, Hallucinations, Language Model, Large Language Model (LLM), Machine Learning (ML), Natural Language Processing (NLP) 키워드는 모두 빈도수가 높은 키워드로, 연결중심성, 매개중심성, 근접중심성, 위세중심성이 모두 높은 키워드로 분석되었다. 이는 해당 키워드가 인공지능 환각 연구에서 핵심 주제로 다루어지고 있다는 점을 의미한다.

Deep Learning (DL), Education, Ethics, Generative Adversarial Networks 키워드는 연결중심성, 매개중심성, 근접중심성이 높지만, 위세중심성이 상대적으로 낮은 키워드로 분석되었다. 이는 네트워크 그래프에서 연결의 수가 많고, 다른 키워드를 연결하며, 다른 키워드까지의 거리가 짧은 키워드임을 의미하지만, 연결되어 있는 키워드의 중요도가 상대적으로 낮은 키워드임을 의미한다.

Medical Education 키워드는 연결중심성, 근접중심성, 위세중심성이 높지만, 매개중심성이 상대적으로 낮은 것으로 분석되었다. 또한, Answer, Exam, Examinations, Otolaryngology 키워드는 연결중심성, 위세중심성이 높지만, 매개중심성과 근접중심성이 상대적으로 낮은 것으로 분석되었다. 이는 해당 키워드가 중요도가 높고 다른 키워드까지의 거리가 가까운 키워드이지만, 다른 키워드를 연결하는 키워드는 아니라는 것을 의미한다.

Bias, Fairness, Google Bard, Healthcare, Prompt Engineering 키워드는 상대적으로 근접중심성이 높은 키워드로 분석되었다. 따라서 해당 키워드는 다른 키워드까지의 거리가 가까워 영향력이 높은 키워드임을 의미한다.

Conversational Agents, fMRI, Intelligent Textbooks, MRI, Psychedelics, Schizophrenia 키워드는 상대적으로 매개중심성이 높은 키워드로 분석되었다. 해당 키워드는 상대적인 중요도는 낮지만, 네트워크 그래프로 표현했을 때 다른 키워드까지의 거리가 가까워 영향력이 높은 키워드임을 의미한다.

EAR, Medical Examination, Patient Safety, Response, Surgery 키워드는 상대적으로 위세중심성이 높은 키워드로 분석되었다.

키워드 네트워크 분석 결과에서 중요도가 높은 것으로 분석된 키워드는 인공지능 기술의 개념과 기능에 대한 연구주제를 의미하는 키워드, ChatGPT, Gemini와 같은 LLMs 또는 LMMs 기반의 대화형 챗봇 서비스 (프롬프트 서비스)를 의미하는 키워드, 인공지능 윤리와 관련된 키워드, 편향과 인공지능 환각과 관련된 키워드, 의료 및 교육과 관련된 키워드로 분류된다.

3절에서는 주요 키워드가 연구 주제에서 어떠한 의미로 활용되었는지 살펴보고, 인공지능 환각 해외 연구문헌에 대해서 살펴보고자 한다.

3. 인공지능 환각 주요 해외 연구문헌 리뷰

본 절에서는 앞선 키워드 네트워크 분석을 바탕으로 해외 연구 문헌을 확인하고자 한다. 앞선 키워드 분석 기술적 용어를 제외하고 다빈도 어휘로 등장한 윤리적 측면에서의 공정성과 편향에 대한 연구를 확인하고 환각에 대한 기술적 연구 동향을 살펴보고자 한다. 마지막으로 생성형 인공지능의 교육적 적용에 있어 환각의 문제를 연구한 논문 분석을 통해 생성형 인공지능의 교육적 활용 가능성을 함께 생각해 보고자 한다.

인공지능 환각과 관련된 주요 키워드 중 하나는 편향(bias)이다[17]. Jain 외 1인의 연구에서는 Scheurer 외 2

Table 4. Centrality Analysis Result

Keyword	DC	Keyword	BC	Keyword	CC	Keyword	EC
Artificial Intelligence (AI)	0.5235	Artificial Intelligence (AI)	0.3378	Artificial Intelligence (AI)	0.6499	Artificial Intelligence (AI)	0.3400
Large Language Model (LLM)	0.4540	Large Language Model (LLM)	0.2792	Large Language Model (LLM)	0.6130	Large Language Model (LLM)	0.3082
ChatGPT	0.3543	ChatGPT	0.1287	ChatGPT	0.5699	ChatGPT	0.2802
Generative Artificial Intelligence (GAI)	0.2276	Generative Artificial Intelligence (GAI)	0.1188	Generative Artificial Intelligence (GAI)	0.5287	Natural Language Processing (NLP)	0.1932
Hallucinations	0.1659	Hallucinations	0.0782	Natural Language Processing (NLP)	0.5068	Chatbots	0.1756
Natural Language Processing (NLP)	0.1648	Machine Learning (ML)	0.0415	Hallucinations	0.5043	Generative Artificial Intelligence (GAI)	0.1377
Chatbots	0.1502	Generative Adversarial Networks (GANs)	0.0373	Machine Learning (ML)	0.5016	Language Model	0.1359
Machine Learning (ML)	0.1256	Deep Learning (DL)	0.0292	Chatbots	0.4997	Machine Learning (ML)	0.1344
Language Model	0.0930	Natural Language Processing (NLP)	0.0247	GPT	0.4781	Medical Education	0.1181
GPT	0.0818	Chatbots	0.0178	Deep Learning (DL)	0.4779	Hallucinations	0.1084
Deep Learning (DL)	0.0684	Ethics	0.0174	Medical Education	0.4759	Examinations	0.1061
Accuracy	0.0650	MRI	0.0137	Education	0.4724	Answer	0.0995
Medical Education	0.0650	GPT	0.0129	Ethics	0.4718	Exam	0.0995
Generative Adversarial Networks (GANs)	0.0639	Language Model	0.0123	Generative Adversarial Networks (GANs)	0.4710	Medical Examination	0.0995
Education	0.0617	Conversational Agents	0.0114	Bias	0.4699	Response	0.0995
Ethics	0.0561	fMRI	0.0107	Prompt Engineering	0.4628	GPT	0.0970
Examinations	0.0504	Schizophrenia	0.0100	Healthcare	0.4622	Otolaryngology	0.0957
Otolaryngology	0.0437	Psychedelics	0.0097	Language Model	0.4620	Surgery	0.0951
Answer	0.0426	Education	0.0092	Google Bard	0.4617	EAR	0.0894
Exam	0.0426	Intelligent Textbooks	0.0086	Fairness	0.4594	Patient Safety	0.0884

인의 연구[18], Mehrabi 외 4인의 연구[19], Bender 외 3인의 연구[20]를 근거로 하여 생성형 인공지능이 편향을 가질 수 있고, 그 편향의 예측이 불가능하다고 주장하였다[17]. 또한, Jain 외 1인의 연구에서는 Mehrabi 외 4인의 연구와 Dziri 외 4인의 연구 사례를 근거하여[19, 21], 생성형 인공지능이 편향된 데이터를 학습할수록 인공지능 환각이 나타날 가능성이 높다고 기술하였다[17]. Jain 외 1인의 연구는 인공지능 환각의 원인은 아직까지 명확하게 규명되지는 않았지만, 학습에 사용된 데이터의 편향이 인공지능 환각에 영향을 미칠 가능성이 높음을 의미한다.

Maalek의 연구에서는 인공지능 환각은 GANs에서 나타나는 본질적 특성으로, 네트워크의 훈련과 튜닝 과정에서 자주 나타나며, GANs와 같이 대규모의 데이터를 학습하는 LLMs에서도 GANs의 사례에서와 같은 환각이 나타날 가능성이 높아 이미 해결된 문제 또는 과학적 발견과 논리적 추론이 필요하지 않은 문제에 대해서만 해결책을 제시할 수 있음을 주장하였다[14]. 즉, ChatGPT, Gemini와 같은 생성형 인공지능 서비스를 교육적으로 활용하고자 한다면, 응답에서 인공지능 환각이 나타날 가능성이 높으며, 제한된 해결책만을 제시할 수 있으므로, 이를 종합적으로 고려한 교육적 활용 방안의 마련이 필요하다는 것을 의미한다.

공정성 (fairness)은 인공지능 시스템이 인종, 성적 지향, 나이, 사회경제적 지위과 같은 특정 속성을 기준으로 개인이나 집단에 차별적인 행동을 하지 않아야 한다는 개념을 의미한다[22]. 따라서 Google사와 OpenAI사는 편향된 응답과 잘못된 정보를 방지하고, 공정성을 강화하기 위한 서비스 전략을 마련하였으나, 아직까지는 이를 완벽하게 차단하거나 경감할 수 있을 정도는 아닌 것으로 알려져 있다[23].

Ahmed 외 5인은 ChatGPT와 Bard의 구조와 기능에 대해서 비교하고, 공정성, 윤리적 문제, 한계점에 대해서 논의하였다[24]. 해당 연구에서는 ChatGPT와 Bard 모두 부정확하거나 기만적 콘텐츠의 생성이 가능하고, 그에 따른 콘텐츠 소유권, 저작권 침해, 인간의 일자리 대체로 인한 문제의 발생 가능성이 있다고 기술하였다[24]. 즉, Ahmed 외 5인이 주장하는 바는 LLMs가 제공하는 응답이 인공지능 환각으로 인해 정확성 (accuracy)이 낮고, 그에 따른 윤리적 문제 (ethical issues)의 발생 가능성이 있음을 시사한다.

다음은 생성형 인공지능의 환각 현상에 대한 기술적 차원에서의 연구로 프롬프트 엔지니어링을 통한 성능개선과 생

성형 인공지능 서비스 종류에 따른 환각 현상 차이에 대한 연구를 살펴보았다.

프롬프트 엔지니어링을 이용한 생성형 인공지능의 성능 개선을 위한 연구는 McGowan 외 9인의 연구와 Zinjad 외 2인의 연구에서 확인된다[23, 25].

Zinjad 외 2인은 구직자가 생성형 인공지능 서비스를 이용하여 자신의 이력서를 특정 직무 공고에 적합하도록 수정할 수 있는 파이프라인인 ResumeFlow를 제안하였다[25]. 해당 파이프라인은 사용자가 업로드한 이력서로부터 직무 (work), 교육 (education), 프로젝트 경험 (projects), 스킬 (skills), 기타 (etc.) 정보를 추출하고, JD (job description)로부터 추출한 직무의 상세정보를 바탕으로 지원하고자 하는 직무에 적합한 이력서를 생성하는 기능을 제공한다. 여기서 LLMs는 이력서로부터 구직자의 정보를 추출하거나 JD로부터 직무의 상세내역을 추출하고, 직무에 적합한 이력서를 생성하는 데 활용된다. 파이프라인의 작업 설계에는 LLMs에 풍부한 경험을 가진 이력서 작성 전문가의 페르소나 (persona)를 부여하는 시스템 프롬프트와 주요 직무 세부 사항을 추출하는 작업 프롬프트가 사용되었다. 이는 Zinjad 외 2인이 개발한 파이프라인의 전문성을 개선하는 한편, 환각, 불일치, 편향된 응답 등을 최소화하기 위하여 프롬프트 엔지니어링 방법을 활용했다는 측면에서 의미를 갖는다.

Kuhail 외 4인은 프롬프트 엔지니어링의 구성요소가 지시문 (instruction), 추가 콘텍스트 (additional context), 예제 (example), 제약 조건 (constraints), 역할 (role)이고, LLMs의 블랙박스 (black-box) 특성 등으로 인해 프롬프트 엔지니어링이 주목받고 있다고 주장하였다[26].

2024년 2월 8일부터 Google사는 Bard를 Gemini로 리브랜딩하였다. 따라서 일부 연구에서는 ChatGPT와 Bard를 비교하는 한편, 또 다른 일부 연구에서는 ChatGPT와 Gemini를 비교하는 사례도 있다[23, 24].

McGowan 외 9인의 연구에서는 ChatGPT 3.5와 Bard 2.0을 이용하여 사전에 정의한 참고문헌을 검색하는 실험을 수행하였다[23]. McGowan 외 9인의 연구에서 2023년 3월 ChatGPT를 이용해 참고문헌 검색을 요청했을 때, 총 35개의 참고문헌을 응답했으며, 이 중 단 6%만 실제 논문과 일치하는 것으로 보고하였다. 예를 들어, 33개의 참고문헌 중 12개는 실제 논문의 제목과 유사했지만, 21개는 명확한 출처가 없었다[23]. 즉, ChatGPT가 정확한 응답을 제공할 때도 있지만, 출처를 확인하기 어렵거나 검증되지 않은 정보를 제공할 가능성이 매우 높다는 것을 의미한다. 따라서 McGowan 외 9인의 연구에서 LLMs의 응답 (출력)이 LLMs와 별개의 독립적 검증 수단이 마련돼야 할 필요가 있다고 주장하였다[23].

생성형 인공지능 서비스가 가진 환각 문제는 비단 서비스 성능뿐 아니라 교육에 활용했을 때 더욱 심각할 수 있다. 이에 인공지능 환각 가능성을 최소할 수 있는 기술이나 피교육자의 환각 식별 능력을 높이는 방법에 대한 논의가 이루어졌다. Taneja 외 5인의 연구에서는 ChatGPT의 기능을 이용하여 대화형 가상 교육 지원 (virtual teaching assistant) 기능을 제공하는 Jill Watson 아키텍처를 개발하였다[27]. 해당 아키텍처는 인공지능 환각의 가능성을 최소화할 수 있는 프롬프트 기술 (prompting)을 적용하였으며, 모듈형 설계를 통해 대화형 에이전트 (conversational agents) 형태로 서비스 제공이 가능하여 지능형 교과서 (intelligent textbooks)에도 적용 가능하다.

Pak 외 2인은 18~23세의 공립 대학교 71명과 군사 학교 학생 125명을 대상으로 챗봇의 응답 톤 (tone)에 따라서 잘못된 정보 (환각이 포함된 정보)를 수용하는 수준을 측정하는 실험을 수행하였다[28]. 질문지는 다지선다형이고, 물리학, 역사, 화학, 생물학, 지리학, 심리학 등 다양한 학문 분야에서 문제 은행, 교과서, 웹 자료에서 추출되었으며, 중립 톤 (natural tone)과 정중 톤 (polite tone)의 챗봇을 이용하여 잘못된 정보가 포함된 질문지를 제공하였다. 해당 실험 결과를 바탕으로 Pak 외 2인은 정중 톤 챗봇과 상호작용한 참가자들은 잘못된 정보를 수용할 가능성이 더 낮았고, 생성형 인공지능 서비스의 응답을 믿지 못하는 보수적인 편향 (conservative bias)을 보이는 것으로 보고하였다[28]. 이는 대화형 에이전트가 제공하는 톤이 사용자의 인공지능 환각 식별 능력에 영향을 미칠 가능성을 시사한다. 나아가 Pak 외 2인의 연구결과는 인공지능 교과서 (intelligent textbooks)와 관련된 연구로 볼 수 있다. 이는 해외에서 생성형 인공지능 서비스를 이용한 학습에 대한 연구가 상당 수준 진행되고 있음을 의미한다. 국내에서도 생성형 인공지능 서비스를 이용한 인공지능 디지털 교과서에 대한 정책이 추진되고 있는 점을 고려할 시[34], Pak 외 2인의 연구 사례에서와 같이 인공지능 환각에 대한 보다 구체적이고 실무에 적용 가능한 연구가 수행되어야 한다[28].

앞서 살펴본 연구 사례에서와 같이 다수의 연구문헌에서는 인공지능의 정의와 개념을 설명하기 위하여 Artificial Intelligence (AI), Deep Learning (DL), Machine Learning (ML) 키워드를 사용한다[16-28]. 인공지능 환각 연구에 있어서 인공지능 기술의 개념과 특징, 구조, 모델, 데이터 등 다양한 요소들에 대해서 고려하여 원인과 경감 방법이 연구되고 있음을 시사한다.

이상에서 살펴본 내용 외에 인공지능 환각 연구에 집중한 또 하나의 분야는 의료이다[29-32].

인공지능 환각에 대한 연구는 의료 진단, 의료 교육 등 생성형 인공지능을 적극 활용하고자 하는 의료 진단 및 의료 교육 분야에 일부 집중되어 있다.

Long 외 7인은 2023년 4월 11일 ChatGPT를 대상으로 Royal College of Physicians and Surgeons of Canada의 샘플 시험에서 발췌한 21개의 주관식 문항에 대하여 프롬프트 (현재 질의하는 상황에 대한 설명)의 유무를 구분하여 질의하고, 생성형 인공지능의 응답을 평가하기 위해 일치성 (concordance), 타당성 (validity), 안전성 (safety), 역량 (competency)을 평가하는 CVSC평가모델을 개발하

Table 5. Classification Results of International AI Hallucination Research by Keywords

Research Subjects	Keywords
AI hallucinations problem definition	Accuracy, Ethics, Fairness
Defining cause of AI hallucinations	Bias
AI hallucinations in LLM-based conversational chatbots and other application services	Answer, Chatbots, ChatGPT, Conversational Agents, Generative Adversarial Networks (GANs), Generative Artificial Intelligence (GAI), Google Bard, GPT, Intelligent Textbooks, Language Model, Large Language Models (LLMs), Prompt Engineering, Response, Natural Language Processing (NLP)
AI hallucinations in medical diagnosis and education	EAR, Education, Exam, Examinations, Healthcare, Medical Education, Medical Examination, Otolaryngology, Patient Safety, Surgery

였다[29]. Long 외 7인의 연구 결과에서는 생성형 인공지능이 응답한 결과가 평균 75%의 점수를 기록하였고, 질의를 수행하기 전 프롬프트를 사용했을 때가 사용하지 않았을 때보다 정확도가 높은 것으로 보고되었다[29]. 일부 응답에서는 인공지능 환각이 나타났으며, 이를 근거로 하여 Long 외 7인은 인공지능 환각이 환자의 안전에 위협이 될 수 있으므로, 생성형 인공지능 서비스가 의료 실무에 적용되기 위해서는 안전하고 정확한 응답을 얻을 수 있는 방안의 마련이 필요하다고 주장하였다[29].

Zalzal 외 2인의 연구에서는 생성형 인공지능이 이비인후과 전문의의 임상 시나리오 교육에 있어서 보조 교육 도구로 사용될 수 있는지 여부에 대해서 가설을 세우고 전문가를 대상으로 실험하였다[30]. 실험은 두 명의 이비인후과 전문의를 대상으로 텍스트로 구성된 30개의 주관식과 객관식 질문 두 세트를 주고, 이를 ChatGPT 3.5에 입력하여 응답을 검증하는 형태로 진행되었다. Zalzal 외 2인의 연구 결과에 따르면, 특정 질의에 대하여 ChatGPT가 정답을 응답할 확률은 56.7%이고, 부분적 정답을 응답할 확률은 86.7%, 동일한 질의를 반복했을 때 정확하게 답변할 확률은 73.3%로 보고되었다[30]. Zalzal 외 2인은 이를 근거로 하여 ChatGPT는 이비인후과 분야에서 교육에 활용할 수 있는 수준의 정확도를 확보하기 어렵고, 그에 따라서 부적절한 사용을 방지하기 위한 가이드라인 마련의 필요성을 강조하는 한편, ChatGPT의 의학교육 활용에 있어서 인공지능 환각을 고려한 올바른 판단을 위해서는 아직까지는 전문가의 검토가 필요하다고 기술하였다[30].

Herrmann-Werner 외 7인은 307건의 psychosomatic medicine 객관식 문제를 사용하여 GPT-4의 성능을 평가하였다[31]. Herrmann-Werner 외 7인은 간략한 프롬프트와 상세 프롬프트 2가지 프롬프트를 이용하였고, 생성형 인공지능의 응답은 정량/정성 분석 방법으로 분석하였으며, 오답은 블룸의 텍사노미 계층 프레임을 통해 분류하였다. 분석 결과에 따르면 ChatGPT는 간략한 프롬프트 (short prompt)에서 91%, 상세 프롬프트 (detailed prompt)에서 93%의 정답을 보였고, 오답은 주로 구체적인 사실을 무시하거나 (remember) 비논리적인 근거를 제공하는 데 (understand)에서 발생한 것으로 보고하였다. Herrmann-Werner 외 7인은 오답의 원인 중 하나를 인공지능 환각으로 판단하며, 오답이 나타난 원인은 모델의 편향과 출력을 극대화하기 위한 경향 때문으로 분석하였다.

Chen 외 19인은 생성형 인공지능을 이용하여 Medical College of Wisconsin Ophthalmic Case Studies의 안과 저자 스타일로 진단 결과를 생성하고, 안과 전문의 16인을 대상으로 인간이 생성한 텍스트와 ChatGPT가 생성한 텍스트를 평가하는 실험을 수행하였다[32]. Chen 외 19인의 결과에 따르면, 실험 참가자들은 생성형 인공지능을 이용하여 생성한 진단 결과는 일반적이고, 관련 없는 정보를 포함하며, 인공지능 환각이 자주 발생하고, 독특한 패턴을 갖는 것으로 평가하였다[32].

앞서 Long 외 7인의 연구 결과와 Herrmann-Werner 외 7인의 연구 결과에서 생성형 인공지능이 생성한 결과는 일정 점수를 넘어서서 합격점을 얻은 것으로 보고되었다[29, 31]. 즉, 의료 교육 분야에서 생성형 인공지능이 보이는 지식수준이 인간에 근접한 수준으로 나타나고 있다는 것을 의미한다. 그러나 생성형 인공지능을 활용할 때 인공지능 환각 현상이 나타나고, 부정확한 정보로 인해 심각한 결론에 도달할 수 있는 점을 고려할 시 아직까지는 실무에 적용하기는 어렵다[30, 32]. 이는 인공지능 환각을 포함하여 부정확한 정보를 최소화하기 위한 방안과 안전장치 마련이 필요하다는 것을 의미한다.

앞서 살펴본 것과 같이 인공지능 환각에 대한 연구는 부정확한 정보로 인한 이슈, 공정성과 윤리적 고찰에 대한 이슈를 인공지능 환각으로 인한 주요 이슈로 다루고 있고, 그 원인으로는 편향을 다루고 있다[17-21].

LLMs 기반의 대화형 챗봇 서비스인 ChatGPT, Gemini와 같은 서비스에서 인공지능 환각의 심각성과 대응 방안에 대한 연구가 진행되고 있고, 프롬프트 엔지니어링을 이용한 인공지능 환각을 경감하는 방안에 대한 연구 사례도 확인된다[23, 24]. 또한, 인공지능 디지털 교과서와 관련된 연구도 확인되며[27], 생성형 인공지능을 활용하기 위해서는 환각을 식별할 수 있는 능력이 요구된다는 점을 강조하고 있다[26, 27].

한편, 의료 진단 및 교육에서는 생성형 인공지능이 생성하는 결과가 인간에 근접한 수준으로 나타날 때도 있어 그 활용 가능성에 대해서 지속 연구되고 있고, 일부 연구에서는 프롬프팅 (prompting)을 통해 응답의 정확성을 높일 수 있다는 점을 보고하였다. 그러나 아직까지는 생성형 인공지능을 의료 진단과 의료 교육에 활용하기에는 어렵다는 것이 공통적인 의견이다[29-32].

앞서 리뷰한 연구문헌을 바탕으로 키워드를 연구 주제별

로 분류한 결과를 Table 5에 기술하였다.

연결, 매개, 근접, 위세중심성 분석 결과에서 나타난 키워드는 중복을 제거하였을 때 총 36건의 키워드이다. Table 5에서는 이중 Artificial Intelligence (AI), Deep Learning (DL), fMRI, Machine Learning (ML), MRI, Psychedelics, Schizophrenia 키워드를 제외한 29건의 키워드를 분류에 활용하였다.

WoS에서 인공지능 환각 관련 연구를 검색할 때 'Artificial Intelligence Hallucination' 키워드를 사용한 결과, 일부 인공지능 환각과는 다른 연구가 함께 포함되는 사례가 일부 확인된다. 해당 연구 사례는 주로 정신과 진단과 관련된 연구로, 환각증상, 환각약물, 조현병 등을 의미하는 키워드를 포함하고 있고, 인공지능 모델을 진단에 이용하기는 하지만, 인공지능 환각과는 거리가 있는 연구 주제를 다루고 있다. 대표 연구 사례로 Krishna 외 7인의 연구를 들 수 있다[33]. Krishna 외 7인은 EEG (Electroencephalogram), fMRI, 액티그래프, EHR 등의 데이터를 분석하여 정신분열증을 진단하는 Multi-view learning 모델을 개발하였다[33]. 해당 연구에서는 인공지능 환각이 아닌 정신 진단에서의 환각을 다루고 있고, 인공지능 모델이 이용되고 있어 인공지능 환각 관련 키워드로 검색 시 연구 목적과는 관계없이 포함된 연구 사례에 해당한다. 따라서 Table 5을 분류함에 있어서 fMRI, MRI, Psychedelics, Schizophrenia 키워드는 분류에 활용하지 않았다. 또한, 인공지능의 기본개념과 정의에 대해서 사용하는 Artificial Intelligence (AI), Deep Learning (DL), Machine Learning (ML) 키워드는 제외하였다.

4. 국내 연구문헌 분석

국내에서는 인공지능 기술과 활용, 인공지능 교육과 관련된 연구가 다수 수행되고 있는 것으로 파악된다. 2024년 12월 기준 KCI에서 '인공지능'을 키워드로 검색한 결과 11,591건이 검색되고[35], 인공지능 교육 키워드로 검색한 결과 2,699건의 연구문헌이 검색된다. 반면, '인공지능 환각' 키워드로 검색한 결과 22건의 연구문헌만이 검색된다. 따라서 국내 연구는 소수만 수행되고 있고, Table 5에 정리된 해외 사례와 같이 체계적으로 연구되고 있지는 않은 것으로 파악된다. 따라서 국내 연구문헌 중 제목에 '인공지능 환각', '환각', '할루시네이션' 키워드를 포함하고 있는 7건의 연구문헌을 선정하여 리뷰하였다[36-42].

김형주는 생성형 인공지능에서 나타나는 환각을 정의하고, 부재 증명을 통해 인공지능 환각의 의미에 대해서 비교분석하였다[36]. 김형주의 연구는 인공지능 환각의 정의와 존재에 대해서 철학적으로 접근한 연구 사례로 판단된다.

박형빈은 인공지능 환각의 윤리적 도전과제를 분석하고, 가이드라인을 제시하였다[37]. 박형빈은 인공지능 환각으로 인해 비즈니스 분야에서 부정확한 정보를 바탕으로 의사결정이 이뤄질 수 있고, 그로 인한 손실에 대한 우려, 교육 분야에서 생성형 인공지능의 사용으로 인한 표절, 잘못된 정보, 가상과 현실 세계의 구분이 문제가 될 수 있다고 기술하였다. 또한, 연구 분야에서 생성형 인공지능을 활용할 때 인공지능 환각으로 인해 발생하는 부정확한 정보를 식별하고 검증하기 위한 노력이 필요하며, 이와 유사한 문제가 의료 분야에서 발생 시 생명과 직결되는 치명적인 결과로 이어질 수 있음을 기술하였다. 그리고 게임 산업에서의 정보의 신뢰성과 정확성, 지적 재산권과 저작권, 정보 및 개인 정보보호가 윤리적 도전과제가 될 수 있음을 기술하였다. 상기의 문제들을 종합적으로 고려하여 생성형 인공지능의 사용에 대한 가이드라인이 필요함을 제언하였다[37].

이현승 외 1인은 Ji 외 9인의 연구[43], Maynez 외 3인의 연구[44]를 근거로 하여 인공지능 환각의 완화 방법은 데이터의 품질을 개선하는 방법, 생성된 데이터를 후처리 하는 방법, 언어 모델의 미세조정, 구체적인 프롬프트의 방법이 있다고 기술하였다[38]. 이중 구체적인 프롬프트 방법, 후처리 방법, 미세조정 방법을 생성형 인공지능에 적용하고, 메타버스 미로 게임을 이용하여 게임 콘텐츠 생성에 있어서 생성형 인공지능의 환각 완화 방안을 검증하였다[38]. 이현승 외 1인의 결과에 따르면, 미세조정 모델을 사용하는 것이 인공지능 환각의 완화에 가장 효과적이지만, 이에 소요되는 시간과 비용의 부담이 있고, 후처리 방식은 인공지능 환각의 경감에 효과적이며, 구체적 방식이 가장 손쉽게 적용할 수 있는 방법이라고 기술하였다[38].

이승석 외 1인은 텍스톰을 사용하여 2023년 1월 1일부터 2024년 5월 31일까지 'AI 할루시네이션'을 키워드로 하여 토픽 모델링을 수행하였다[39]. 키워드 상위 20개 키워드로는 데이터, 챗GPT, 위조정보, 거짓, 변조, 모델, 현상, 답변, 역기능, 서비스, 문제, 기술, 활용, 가짜뉴스, 사용, 허위, 사실, 언어모델, 딥페이크, 개인정보, 보안, 오류 등으로, 본 연구에서 분석한 해외 연구의 키워드 빈도수 분석 결과와는 다소 다른 키워드로 구성되어 있다[39]. 이는 토픽 모델링의 경우 인공지능 환각으로 인한 현상, 효과, 기능에 집중하는 반면, 본 연구에서는 학술연구에 초점을 맞추고 있기 때문으로 해석된다. 또한 이승석 외 1인은 토픽별 핵심 키워드를 AI 윤리 및 편향, AI의 미래 전망, 생성형 AI 기술, 의료 분야에서의 AI 활용 및 한계, AI 연구개발, AI 법률 및 규제, 인간-기계 상호작용, AI 보안과 개인정보 보호, 딥페이크, AI 역기능, AI 교육 등 11개 토픽으로 분류하였다.

안진호 외 1인은 감성·경험 전략 서비스에서 사용자 생성 프롬프트가 인공지능 환각을 최소화하는 데 미치는 영향을 평가하기 위하여 5가지 가설을 세우고, 2023년 10월 ~ 11월의 기간 동안 특정 사이트 이용자를 대상으로 설문을 수행하였다[40]. 이를 통해 감성·경험 전략 서비스에서 사용자 생성 프롬프트는 인공지능 환각을 경감하는 데 유의미함을 입증하였다.

박대민 외 1인은 인공지능 환각에 대해서 정의하고, 해외 연구문헌의 사전 출판 사이트인 아카이브(arXiv)으로부터

연구 문헌의 서지 정보를 추출하여 기술통계와 빈도분석, 키워드 분석을 수행하였다[41]. 박대민 외 1인의 연구 결과에 따르면, 아카이브의 연구문헌은 리뷰 문헌, 벤치마크와 성능 평가에 문헌, 환각 탐지 문헌, 환각 완화 문헌으로 주제가 분류된다[41]. 또한, Park 외 1인의 결론에 따르면, 중국계 연구자들의 연구가 활발한 편이고, 연구 문헌의 수가 많고 공동 연구 활성화 수준이 높은 편이며, 연구 주제도 다양하다. 이외에도 사전학습 모델의 미세조정과 강화학습 단계의 환각 완화 적용이 주목 받고 있고, RAG와 결합한 CoT (Chain of Thought)가 고도화되는 추세이며, 텍스트 외에 이미지, 동영상, 음성, 3D 등 다양한 모달리티 및 멀티모달의 환각 문제가 제기될 것으로 예측하였다[41].

조헌국은 물리교육 분야에 있어서 생성형 인공지능을 활용할 수 있고, 그에 따라서 학생 평가, 피드백, 상담 등에 활용하기 어려운 문제점이 있다고 기술하였다[42]. 조헌국은 환각에 대처하기 위한 방법으로 모델의 복잡성을 줄여 환각을 감소시키는 방법, LoRA를 통해 생성된 결과물의 구체적인 해석에 이르도록 안내하는 방법, DPO (Distillation and Pruning Optimization)나 PPO (Proximal Policy Optimization)와 같이 인간의 직관이나 의사결정 과정을 모방해 환각을 경감하는 방법을 검토하였고, RAG는 이러한 한계를 극복하기 위한 효과적인 방법이라고 주장하였다[42].

이처럼 국내 인공지능 환각 연구는 인공지능 환각에 대한 원인과 문제점, 해결 방안 등에 대해서 개략적으로 다루고 있지만, 그 연구의 다양성과 깊이에 있어서는 제한적으로만 수행되고 있다. 특히, 교육 분야에서 생성형 인공지능을 이용하고자 함에 있어서 적극적인 연구가 수행되고 있고, 정부에서도 인공지능 디지털 교과서 도입과 같은 정책을 추진하고 있지만, 이를 교육 현장에 적용하기 위한 연구가 다소 부족한 편이다.

5. 결론 및 제언

2022년 11월 ChatGPT가 등장한 이후로 생성형 인공지능 서비스와 이를 활용하기 위한 연구가 세계적으로 주목받고 있다. 생성형 인공지능 기술이 주목받고 있지만, 생성형 인공지능 서비스에서 나타날 수 있는 인공지능 환각은 소수의 연구만이 확인된다. 대표적으로 2024년 12월 기준 WoS에서 ‘Generative artificial intelligence’ 키워드로 검색한 결과 13,427건의 연구문헌이 검색되고, KCI에서는 2,699건의 연구문헌이 검색된다. 반면 ‘Artificial intelligence hallucination’ 키워드로 검색한 결과는 WoS 356건, KCI 44건의 연구문헌이 확인된다.

사회 전 분야에서 생성형 인공지능 서비스가 활용되고 있고, 특히 생성형 인공지능에 대한 연구가 빠르게 증가하고 있는 상황에서 생성형 인공지능이 가진 치명적인 오류라 할 수 있는 인공지능 환각에 대한 연구가 절실하다.

따라서 인공지능 교육을 비롯한 생성형 인공지능의 교육적 활용에 앞서 인공지능 환각에 대하여 명확하게 파악하고, 그로 인한 부정적 효과를 경감하거나 회피하기 위한 교육 방법과 이를 지원하는 도구를 개발해야 할 필요가 있다.

본 연구의 분석 결과에 따르면, 인공지능 환각은 여러 요인에 의해서 발생하고, 아직까지 원인이 명확하게 규명되지는 않았지만, 데이터의 편향과 모델로 인한 영향이 큰 것으로 파악되며, 데이터의 품질을 개선하는 방법, 생성된 데이터를 후처리 하는 방법, 언어 모델의 미세조정, 구체적인 프롬프트의 이용, RAG 등의 방법으로 경감할 수 있다. 또한, 해외 연구의 경우 인공지능 환각의 원인과 그로 인해 발생하는 문제점, ChatGPT, Gemini와 같은 생성형 인공지능 챗봇 서비스를 이용한 응용 사례 연구, 그리고 의료 진단과 교육 분야에서의 생성형 인공지능의 활용 사례에 대한 연구로 분류할 수 있고, 국내 연구와 비교했을 때 체계적으로 연구되고 있는 것으로 판단된다. 해당 연구 사례에서는 인공지능 환각을 완벽하게 경감하거나 회피할 수 있는 방법을 제시하지는 못했지만, 원인 분석과 부정적 효과를 경감할 수 있는 방안, 실무 적용 가능성 등에 대한 학술적 논쟁을 포함하고 있다. 반면, 국내 연구는 인공지능 환각의 정의, 원인 등에 대해서 개략적으로 논의하고는 있으나, 연구 사례가 부족하고, 근본적인 해결 방안에 대해서는 명확하게 논의되고 있지 않은 것으로 판단된다. 특히, 인공지능 디지털 교과서 정책이 논의되고 있는 현 상황에서 교육 분야에서의 인공지능 환각을 경감하거나 회피하고, 교육에 적용할 것인지에 대해서 논의하고 있는 연구문헌은 찾아보기 어렵다. 따라서 교육 분야에서는 학습자에게 생성형 인공지능에서 발생하는 인공지능 환각의 원인과 해결 방법에 대해서 명확하게 교육하고, 프롬프트 엔지니어링과 같은 비교적 간단하고 빠르게 교육 실무에 적용 가능한 방법을 모색해야 할 필요가 있다.

상기의 분석 결과를 바탕으로 본 연구에서는 생성형 인공지능 서비스 개발과 함께 인공지능 환각 연구가 추구해야 할 방향성에 대해서 제언하고자 한다.

첫째, 인공지능 환각의 문제와 원인에 대한 해외 연구 사례를 참고하여 국내에서도 문제와 원인 규명에 대한 연구가 수행되어야 한다. Jain 외 1인의 연구 사례와 Maalek의 연구 사례에서와 같이 인공지능 환각의 현상에만 집중하지 않고, 원인에 대해서 설명하고, Ahmed 외 5인의 연구 사례에서 드러난 문제의 심각성에 대해서 학습자도 충분히 이해할 수 있는 사례 교육이 필요하다. 이를 통해 학습자가 인공지능 환각이라는 현상을 이해하고, 자신의 업무와 직무, 연구, 학습 등에 활용하고자 할 때 인공지능 환각의 영향을 고려하여 생성형 인공지능 서비스가 제공하는 응답의 진위 여부와 오류를 판단할 수 있는 능력을 배양할 수 있는 교육이 요구된다. 다만, 학습자의 전공, 학습 수준을 고려하여 학습자별 차별화된 교육 방법 또는 교육 모델의 개발이 요구되며, 한편으로는 생성형 인공지능 서비스를 통해서 생성된 데이터의 진위 여부를 판단하기 위한 인공지능 콘텐츠 검출기의 활

용도 고려되어야 할 필요가 있다.

둘째, 상당수 연구가 생성형 인공지능 서비스의 유용함과 활용 방법에 대해서 다루고 있으나, 인공지능 환각으로 인한 부정적 영향에 대해서는 고려하지 않는 것으로 판단된다. 생성형 인공지능 서비스를 활용할 때 인공지능 환각이 나타날 가능성이 높으므로, 이를 고려하지 않고 응용하고자 하는 것은 상당 수준의 잠재적인 위험을 수용하는 결과가 될 가능성이 높다. 따라서 생성형 인공지능을 직무 또는 연구, 교육에 활용하고자 한다면, 강화학습 기반 인간 피드백(RLHF), CoT, AI 신뢰성 검증 시스템, RAG 등 기술적인 환각 완화 기법을 고려하여 활용할 수 있는 방안과 함께 고려해야 한다.

본 연구의 분석 결과 중 RAG 키워드는 빈도수는 높지만, 중심성 분석 결과에서는 상대적으로 중요도가 낮은 키워드로 분석되었다. 그러나 일부 연구 문헌에서는 RAG를 인공지능 환각을 최소화하기 위한 효과적이면서도 현실적인 방법이라 주장한다. 즉, 과거 데이터 분석 결과에서는 나타나지 않았지만, 해당 연구 문헌을 참고할 경우 RAG 구현을 위한 데이터베이스와 데이터베이스를 바탕으로 생성되는 결과를 통제할 수 있는 기술적 방법만 마련된다면 인공지능 환각을 통제할 수 있는 강력한 도구로 활용될 가능성이 높다. 따라서 RAG 기반의 생성형 인공지능의 교육적 활용에 대해서 보다 많은 연구가 수행해야 할 것이다.

셋째, 인공지능 환각의 부정적 영향으로 인해 생성형 인공지능의 응용 또는 생성형 인공지능 서비스를 활용하는 것을 금지하거나 배제하지 않아야 한다. 생성형 인공지능 서비스는 이미 우리의 일상생활에 상당 수준 침투하여 사회 전반에 걸쳐 침투율이 상당 수준 높은 수준이다. 이는 생성형 인공지능이 가진 강력한 성능과 제공하는 편의성 때문이다. 따라서 별도의 가이드라인을 마련하여 공공기관이 DeepSeek 차단하는 예와 같이 의료, 법률, 금융 등 생성형 인공지능의 사용을 상대적으로 강력하게 규제해야 하는 고위험군과 규제를 완화해야 하는 저위험군의 경우를 평가하여 이를 바탕으로 규제의 강도와 방법을 고려해야 할 것이다.

참고문헌

[1] Wu, J., Gan, W., Chen, Z., Wan, S., & Philip, S. (2023). Multimodal Large Language Models: A Survey. *Proceedings of the IEEE International Conference on Big Data (BigData)*, Sorrento, Italy, 2247-2256. https://doi.org/10.1109/BigData59044.2023.10386743

[2] Granda, B., Inzhivotkina, Y., Apolo, M., & Fajardo, J. (2024). Educational Innovation: Exploring the Potential of Generative Artificial Intelligence in Cognitive Schema Building. Edutec. *Revista Electrónica de Tecnología Educativa*, (89), 44-63. https://doi.org/10.21556/edutec.2024.89.3251

[3] OpenAI. (2025). ChatGPT (Feb. version) [Large Language Model]. https://openai.com/chatgpt/

[4] Microsoft. (2025). Gemini (Feb. version) [Large Language Model], https://gemini.google.com/

[5] Liu, J., Hui, K., Al Zoubi, F., Zhou, Z., Samartzis, D., Yu, C., Chang, J., & Wong, A. (2024). The Great Detectives: Humans Versus AI Detectors in Catching Large Language Model-generated Medical Writing. *International Journal for Educational Integrity, 20*(8), 1-14. https://doi.org/10.1007/s40979-024-00155-6

[6] Banh, L., & Strobel, G. (2023). *Generative Artificial Intelligence. Electronic Markets, 33*(63), 1-17. https://doi.org/10.1007/s12525-023-00680-1

[7] Farrokhnia, M., Banihashem, S., Noroozi, O., & Wals, A. (2024). A SWOT Analysis of ChatGPT: Implications for Educational Practice and Research. *Innovations in Education and Teaching International, 61*(3), 460-474. https://doi.org/10.1080/14703297.2023.2195846

[8] Qadir, J. (2023, May). Engineering Education in the Era of ChatGPT: Promise and Pitfalls of Generative AI for Education. *Proceedings of the 2023 IEEE Global Engineering Education Conference (EDUCON)*, Kuwait, Kuwait, 1-9. https://doi.org/10.1109/EDUCON54358.2023.10125121

[9] Mhlanga, D. (2023). Open AI in Education, the Responsible and Ethical Use of ChatGPT Towards Lifelong Learning. *FinTech and Artificial Intelligence for Sustainable Development: The Role of Smart Technologies in Achieving Development Goals*, 387-409. https://doi.org/10.1007/978-3-031-37776-1_17

[10] Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit Spewer or the End of Traditional Assessments in Higher Education?. *Journal of applied learning and teaching, 6*(1), 342-363. https://doi.org/10.37074/jalt.2023.6.1.9

[11] Cotton, D., Cotton, P., & Shipway, J. (2024). Chatting and Cheating: Ensuring Academic Integrity in the Era of ChatGPT. *Innovations in education and teaching international, 61*(2), 228-239. https://doi.org/10.1080/14703297.2023.2190148

[12] Susnjak, T., & McIntosh, T. (2024). ChatGPT: The End of Online Exam Integrity?. *Education Sciences, 14*(6), 656. https://doi.org/10.3390/educsci14060656

[13] McIntosh, T., Liu, T., Susnjak, T., Watters, P., Ng, A., & Halgamuge, M. (2023). A Culturally Sensitive Test to Evaluate Nuanced GPT Hallucination. *IEEE Transactions on Artificial Intelligence, 5*(6), 2739-2751. https://doi.org/10.1109/TAI.2023.3332837

[14] Maalek, R. (2024). Integrating Generative Artificial Intelligence and Problem-Based Learning into the Digitization in Construction Curriculum. *Buildings, 14*(11), 3642. https://doi.org/10.3390/buildings14113642

[15] Clarivate (2024). Web of Science (Dec. version). http://www.clarivate.com/

[16] Ashok, A., Natarajan, G., Elmasri, R., & Smith-Stvan, L. (2020). SimsterQ: A Similarity based Clustering Approach to Opinion Question Answering. *Proceedings of the 3rd Workshop on e-Commerce and NLP*, 69–76. https://doi.org/10.18653/v1/2020.ecnlp-1.11

[17] Jain, R., & Jain, A. (2024). Generative AI in Writing Research Papers: A New Type of Algorithmic Bias and Uncertainty in

Scholarly Work. *Proceedings of the 2024 Intelligent Systems Conference (IntelliSys)*, 656-669. https://doi.org/10.1007/978-3-031-66329-1_42

[18] Scheurer, J., Balesni, M., & Hobbhahn, M. (2024). Large Language Models can Strategically Deceive their Users when Put Under Pressure. *Proceedings of the ICLR 2024 Workshop on Large Language Model (LLM) Agents*, Vienna, Austria, 1-28. https://doi.org/10.48550/arXiv.2311.07590

[19] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys (CSUR), 54*(6), 1-35. https://doi.org/10.1145/3457607

[20] Bender, E., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 610-623. https://doi.org/10.1145/3442188.3445922

[21] Dziri, N., Milton, S., Yu, M. Osmar, Z. Reddy, S. (2022). On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models?. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States, 5271–5285. https://doi.org/10.18653/v1/2022.naacl-main.387

[22] Xivuri, K., & Twinomurinzi, H. (2021). A Systematic Review of Fairness in Artificial Intelligence Algorithms. *Proceedings of the 20th IFIP WG 6.11 Conference on e-Business, e-Services and e-Society*, 271-284. https://doi.org/10.1007/978-3-030-85447-8_24

[23] McGowan, A., Gui, Y., Dobbs, M., Shuster, S., Cotter, M., Selloni, A., Goodman, M., Srivastava A., Cecchi G., & Corcoran, C. (2023). ChatGPT and Bard Exhibit Spontaneous Citation Fabrication During Psychiatry Literature Search. *Psychiatry Research, 326*, 115334. https://doi.org/10.1016/j.psychres.2023.115334

[24] Ahmed, I., Kajol, M., Hasan, U., Datta, P., Roy, A., & Reza, M. (2024). ChatGPT Versus Bard: A Comparative Study. *Engineering Reports, 6*(e12890), 1-18. https://doi.org/10.1002/eng2.12890

[25] Zinjad, S., Bhattacharjee, A., Bhilegaonkar, A., & Liu, H. (2024). ResumeFlow: An LLM-facilitated Pipeline for Personalized Resume Generation and Refinement. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Washington DC, USA, 2781-2785. https://doi.org/10.1145/3626772.3657680

[26] Kuhail, M., Berengueres, J., Taher, F., Khan, S., & Siddiqui, A. (2024). Designing a Haptic Boot for Space With Prompt Engineering: Process, Insights, and Implications. *IEEE Access, 12*, 134235-134255. https://doi.org/10.1109/ACCESS.2024.3449396

[27] Taneja, K., Maiti, P., Kakar, S., Guruprasad, P., Rao, S., & Goel, A. (2024, July). Jill Watson: A Virtual Teaching Assistant Powered by ChatGPT. *Proceedings of the International Conference on Artificial Intelligence in Education*, 324-337. https://doi.org/10.1007/978-3-031-64302-6_23

[28] Pak, R., Rovira, E., & McLaughlin, A. (2024). Polite AI Mitigates User Susceptibility to AI Hallucinations. *Ergonomics, 68*(10), 1735–1745. https://doi.org/10.1080/00140139.2024.2434604

[29] Long, C., Lowe, K., Zhang, J., Santos, A., Alanazi, A., O'Brien, D., Wright, E., & Cote, D. (2024). A Novel Evaluation Model for Assessing ChatGPT on Otolaryngology–Head and Neck Surgery Certification Examinations: Performance Study. *JMIR Medical Education, 10(e49970)*, 1-8. https://doi.org/10.2196/49970

[30] Zalzal, H., Cheng, J., & Shah, R. (2023). Evaluating the Current Ability of ChatGPT to Assist in Professional Otolaryngology Education. *OTO open, 7*(4), 1-8. https://doi.org/10.1002/oto2.94

[31] Herrmann-Werner, A., Festl-Wietek, T., Holderried, F., Herschbach, L., Griewatz, J., Masters, K., Zipfel, S., & Mahling, M. (2024). Assessing ChatGPT's Mastery of Bloom's Taxonomy Using Psychosomatic Medicine Exam Questions: Mixed-Methods Study. *Journal of Medical Internet Research, 26(e52113)*, 1-13. https://doi.org/10.2196/52113

[32] Chen, J., Reddy, A., Al-Sharif, E., Shoji, M., Kalaw, F., Eslani, M., Lang, P., Arya, M., Koretz, Z., Bolo, K., Arnett, J., Roginiel, A., Do, J., Robbins, S., Camp, A., Scott, N., Rudell, J., Weinreb, R., Baxter, S., & Granet, D. (2025). Analysis of ChatGPT Responses to Ophthalmic Cases: Can ChatGPT Think like an Ophthalmologist?. *Ophthalmology Science, 5*(1), 100600. https://doi.org/10.1016/j.xops.2024.100600

[33] Krishna, B., Ritwika, S., Javadi, L., Prashanthi, B., Babu, B., Nemova, D., Joshi, A., & Al-Farouni, M. (2024). Early Identification of Schizophrenia Using Multi-view Learning Model. *Cogent Engineering, 11*(1), 1-6. https://doi.org/10.1080/23311916.2024.2384649

[34] Lim W. (2024). Critical Reflections on AI-based Personalized Education: Expectations, Concerns, and Policy Recommendations for AI Digital Textbooks. *Education Review (ER)*, (56), 178-224. https://doi.org/10.23119/er.2024..56.178

[35] NRF (2024). Korea Citation Index (Dec. version). https://www.kci.go.kr/

[36] Kim, H. (2023). Hallucinatory 'AI Hallucination'. *PHILOSOPHY·THOUGHT·CULTURE, 43*, 131-154. https://doi.org/10.33639/ptc.2023..43.006

[37] Park, H. (2023). Hallucination Issues and Ethical Challenges of Generative AI: Focusing on Topics Applicable to Elementary AI Ethics Education. *Korean Journal of Elementary Education, 34*(4), 21-36. https://doi.org/10.20972/Kjee.34.4.202312.21

[38] Lee, H., & Kim, J. (2023). Case Study on Mitigating Hallucinations in Generative AI for Game Content Generation. *Journal of Korea Game Society, 23*(5), 121-129. https://doi.org/10.7583/JKGS.2023.23.5.121

[39] Lee, S., & Lee, S. (2024). Big Data Analysis on AI Hallucination : Focusing on LDA Topic Modeling and Sentiment Analysis. *Korean Journal of Industrial Security, 14*(2), 151-168. https://doi.org/10.33388/kais.2024.14.2.151

[40] Ahn, J., & Jung, W. (2024). Exploring Factors to Minimize Hallucination Phenomena in Generative AI - Focusing

on Consumer Emotion and Experience Analysis -. *Journal Of Service Research and Studies, 14*(1), 77-90. https://doi.org/10.18807/jsrs.2024.14.1.077
[41] Park, D., & Lee, H. (2024). Literature Review of AI Hallucination Research Since the Advent of ChatGPT: Focusing on Papers from arXiv. *Informatization Policy, 31*(2), 3-38. https://doi.org/10.22693/NIAIP.2024.31.2.003
[42] Jho, H. (2024). Leveraging Generative AI in Physics Education: Addressing Hallucination Issues in Large Language Models. *New Physics: Sae Mulli, 74*(8), 812-823. https://doi.org/10.3938/NPSM.74.812
[43] Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On Faithfulness and Factuality in Abstractive Summarization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1906-1919. https://doi.org/10.18653/v1/2020.acl-main.173
[44] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys, 55*(12), 1-38. https://doi.org/10.1145/3571730

박윤수

· 2014년 중앙대학교 전기전자공학부(공학사)
· 2016년 중앙대학교 전기전자공학과(공학석사)
· 2018년 중앙대학교 전기전자공학과(공학박사)
· 2016년~2020년 중앙대학교 다빈치교양대학 강사
· 2021년~2021년 중앙대학교 인문콘텐츠연구소 연구원
· 2022년~2023년 NICE평가정보 전문연구원
· 2023년~2024년 MKS 특허경영 전문위원

관심분야 : O-RAN, AI Literacy, Data Valuation, Secure Authentication Protocols, Passwordless Authentication
26874624@hanmail.net

박호현

· 1987년 서울대학교 계산통계학과(이학사)
· 1995년 KAIST 정보통신공학과(공학석사)
· 2001년 KAIST 전자전산학과(공학박사)
· 1987년~2003년 삼성전자 수석연구원
· 2003년~현재 중앙대학교 전자전기공학부 교수

관심분야 : 빅데이터, 딥러닝, 머신 비전, 정보 보안, 실시간 시스템, 임베디드 시스템 등
hohyun@cau.ac.kr

이유미

· 1998년 중앙대학교 국어국문학과(학사)
· 2001년 중앙대학교 국어국문학과(석사)
· 2006년 중앙대학교 국어국문학과(박사)
· 2019년~현재 중앙대학교 인공지능인문학연구소 교수

관심분야 : 인공지능인문학, 화용론
joystu@cau.ac.kr