거대언어모델의 환각현상 완화 및 평가 기술 동향

A Survey of Trends in Hallucination Mitigation and Evaluation for Large Language Models

이동수 (D.S. Lee, d-soolee@etri.re.kr) SoC인력양성실 책임연구원
이하영 (H.Y. Lee, underzero11@etri.re.kr) 지능정보융합연구실 연구원
김원영 (W.-Y. Kim, wykim@etri.re.kr) 지능정보융합연구실 책임연구원
김낙우 (N.W. Kim, nwkim@etri.re.kr) 콘텐츠지능화연구실 책임연구원
이문영 (M.Y. Lee, munyounglee@etri.re.kr) 지역ICT융합연구실 선임연구원

ABSTRACT

Since the release of OpenAI's ChatGPT, global attention has turned to the development of artificial intelligence, particularly artificial general intelligence. Leading technological companies worldwide have invested heavily in this race, anticipating a transformative impact on the markets and industries. Although large language models (LLMs) have advanced rapidly over the past three years, they have yet to deliver the anticipated profitability gains. This gap is largely attributable to the inherent limitations of LLMs, most notably hallucinations, which remain one of the greatest barriers to their widespread adoption. In this paper, we examine the conceptualization of hallucinations in the context of LLMs, review their negative impacts, and survey current research trends in hallucination mitigation and evaluation approaches.

KEYWORDS AI, hallucination, LLM, 거대언어모델, 인공지능, 환각현상

I. 서론

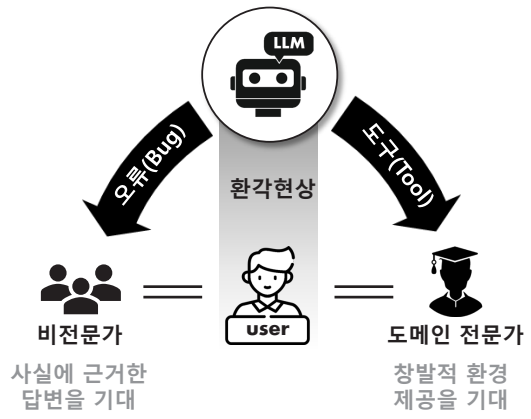
2022년 11월, OpenAI의 ChatGPT가 출시되면서 인공지능, 특히 거대언어모델(LLM: Large Language Model)은 인류의 관심과 기대를 한 몸에 받으면서 모든 산업을 아울러 영향력을 미치는 킬러 기술로 인식되고 있다. OpenAI, Google, Anthropic, Meta와 같은 글로벌 AI 선도기업들을 중심으로 LLM이 경쟁적으로 개발되고 있고, 그 결과로 매달 평균 20개의 신규 LLM이 공개되고 있으며 허깅페이스 플랫폼에서는 매달 평균 10,667개의 파생모델이 등록되고 있다[1].

하지만 현재 시장에서는 AI를 통한 직접적인 수익을 창출하기보다는 GPU(Graphics Processing Unit)

* DOI: https://doi.org/10.22648/ETRI.2025.J.400607
* 이 연구는 2025년도 한국전자통신연구원 기본사업의 지원을 받아 수행되었음[25ZT1100, 수도권 지역산업 기반 ICT융합기술 고도화 지원사업].

등 핵심 인프라에 대한 투자가 기업의 자본 가치를 끌어올리는 데 이바지하고 있는 양상이다. Deloitte의 보고서에 의하면, 2024년 기준으로 LLM을 포함한 생성형 AI에 투자한 기업 중 80%는 ROI(Return on Invest)가 30% 이하의 낮은 수준에 머물러 있다고 한다[2]. 또한, 기업의 AI 도입이 빠르게 확산됨에 따라 이로 인한 다양한 위험 요소도 함께 증가하고 있는 것으로 보고되고 있다. 2024년 McKinsey AI 현황 보고서에 따르면 마케팅 및 영업, 제품 및 서비스 개발 분야에서 생성형 AI의 사용이 두 배 이상 증가함에 따라 생성형 AI로 인한 부정확성, 지식재산권 침해, 사이버보안에 대한 우려가 증가하고 있다[3].

AI 전문가들은 기업이 AI를 도입하는 데 있어서 가장 큰 걸림돌로 환각현상(Hallucination)을 지적하고 있다[4,5]. 환각현상은 LLM 모델이 고도화되더라도 본질적으로 나타나는 특성이며[6], 이는 응용 분야에 따라 창의성을 발휘하는 도구가 될 수도 있지만 때로는 잘못된 정보를 사실처럼 제공하는 치명적 오류로 작용할 수도 있다(그림 1). 환각현상을 '도구'로 보는 관점에서는 우연한 발견, 창의성 등의 잠재력을 활용하되, 도메인 전문가의 통제하에 그 부작용을 최소화함으로써 긍정적 기능으로 LLM이 활용될 수 있다. 반면, 환각현상을 '오류'로 간주하는 관점에서는 LLM의 출력을 사실 또는 확정된 결과로 받아들일 가능성이 높은 비전문가들에게 심각한 오해와 부정적 영향을 초래할 수 있다. 특히, 금융, 의료, 법률과 같은 고정확성·고신뢰성이 요구되는 도메인에서는 환각현상이 치명적인 결과로 이어질 수 있다.



아이콘 출처 게티이미지뱅크, 무단 전재 및 재배포 금지

그림 1 사용자에 따른 환각현상의 영향

이러한 환각현상과 관련하여, 최근 Anthropic의 CEO인 다리오 아모데이(Dario Amodei)는 "AI가 왜 특정 단어를 선택하거나 실수를 저지르는지 그 이유를 설명할 수 없는 기술적 한계를 갖고 있다."라며, 현재의 AI 기술은 내부 작동 방식에 대한 충분한 이해 없이 작동되고 있다고 밝혔다. 이에 따라 Anthropic은 AI 내부를 해부학적으로 분석할 수 있는 'AI용 MRI'를 개발 중이지만, 범용 인공지능(AGI: Artificial General Intelligence)의 등장보다 먼저 이러한 해석 기술을 완성할 수 있을지 우려하고 있다[7].

본고에서는 LLM의 환각현상에 대한 산업적 영향 및 기술적 연구 동향을 살펴보고자 한다. Ⅱ장에서 환각현상에 대한 개념과 산업적 영향을 살펴보고, Ⅲ장에서 환각현상의 완화 방법으로 지식 증강 기술과 응답 검증 기술에 대해 논의한다. Ⅳ장에서는 환각현상에 대한 평가 기술에 대한 동향을 소개하고, 마지막으로 Ⅴ장에서 결론을 제시한다.

II. LLM의 환각현상과 산업적 영향

1. LLM의 환각현상

심리학에서 환각은 "신체 외부 세계의 적절한 자극이 없는 상태에서 깨어있는 개인이 경험하는 지각"으로 정의하고 있으며, 실제처럼 느껴지는 비현실적 자각을 의미한다[8]. 자연어처리(NLP: Natural

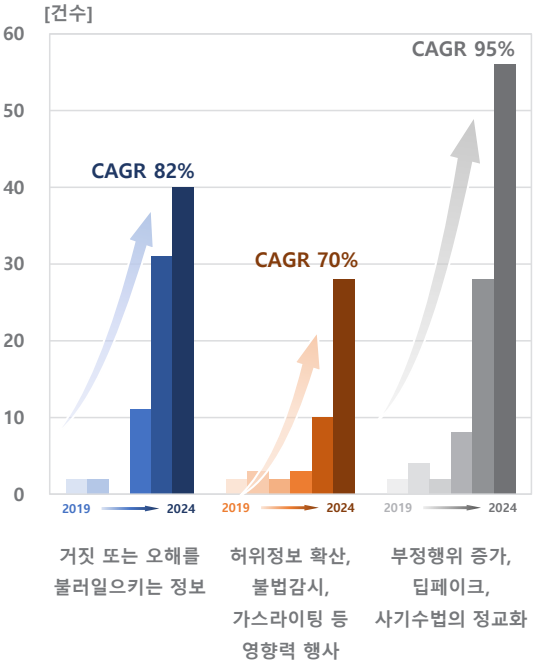
Language Processing) 분야에서 환각은 유창하고 자연스럽지만 불성실하거나 터무니없는 텍스트를 생성하는 현상을 지칭하며, 원본 입력 데이터에 충실하지 않거나 앞뒤가 맞지 않는 출력을 의미한다[8]. NLP 분야에서 환각의 반대말은 충실성(Faithfulness)이며, 제공된 출처에 대해 일관되고 진실되게 유지되는 것을 충실성으로 정의하고 그렇지 않은 것을 환각으로 정의한다[8]. 입력소스와의 관계를 고려하는 NLP 분야와 다르게 LLM 분야에서는 입력소스와 학습 데이터와의 관계 모두를 고려하여 다양한 분야에서 사용자 지시와의 정렬(Alignment) 및 사실 수준에서 환각이 정의되며 이에 따라 사실성(Factuality) 환각과 충실성(Faithfulness) 환각으로 구분할 수 있다[9]. 사실성 환각은 사실과의 불일치 또는 근거 없는 결과를 생성하는 경향을 의미하고 사실적 모순(Factual Contradiction)과 사실 조작(Factual Fabrication)이 포함된다. 충실성 환각은 사용자의 지시나 정보에 일치하지 않는 결과를 생성하거나 결과 자체에 논리적 모순이 있는 경향을 의미한다[9].

LLM에서 환각현상이 발생하는 원인은 학습 단계에서 허위 및 편향된 정보, 최신 지식(Up-to-Date Knowledge)이나 희소 지식(Long-Tailed Knowledge) 같은 지식경계(Knowledge Boundary)의 누락, 사람의 불안전하거나 부정확한 응답으로 만들어진 데이터셋 등 잘못된 정보를 학습하거나, 추론 과정에서 확률에 의한 단어 선택으로 부적절하거나 비논리적인 문장을 만들 수 있기 때문이다[9].

2. 환각현상의 산업적 영향

환각현상으로 인한 피해는 최근 들어 급증하고 있다. 대표적인 사례로는 OpenAI의 음성인식 모델인 Whisper가 40개 의료기관에서 700만 건 이상의 의료 상담 기록을 텍스트로 변환하는 서비스를 제공하는 과정에서 존재하지 않는 약물 이름이나 증상을 추가하는 등 약 40%의 환각 사례가 있었음이 확인되었으며, 이는 환자의 발언 내용을 왜곡하거나 오해를 일으킬 수 있는 수준인 것으로 나타났다[10]. 그리고 에어캐나다의 생성형 AI챗봇이 고객에게 잘못된 환불 규정을 안내한 사례가 소송으로 이어져, 법원에서 AI 환각현상에 대한 기업의 배상 책임을 인정하는 사례도 있었다[11]. 또 다른 사례로는 홍콩의 뉴스 웹사이트인 BNN Breaking이 아일랜드 출신 DJ에 대해 생성형 AI로 작성된 허위 기사를 게재한 혐의로 명예훼손 소송을 당했으며, 현재 해당 웹사이트는 폐쇄된 상태다[12].

AI의 유해한 행동에 대한 보고를 수집하는 AI 사건 데이터베이스(AI Incident Database)를 기반으로 MIT AI Risk Repository에서 분석한 결과에 따르면, 2022년부터 AI에 의한 거짓 또는 오해를 불러일으



출처 Reproduced from MIT AI Risk Repository Website. https://airisk.mit.edu/

그림 2 증가율이 높은 AI Risk 유형

키는 정보의 제공과 사기 및 딥페이크 관련 AI의 오용에 따른 사건이 급격히 증가하고 있다. AI가 거짓 또는 오해를 불러일으키는 정보를 생성하여 신체적, 정신적, 물질적 피해를 유발하는 경우는 2024년에 발생한 AI 사건 중 18%를 차지하고 있으며, 2019년부터 연평균 82%씩 급증하고 있다(그림 2) [13].

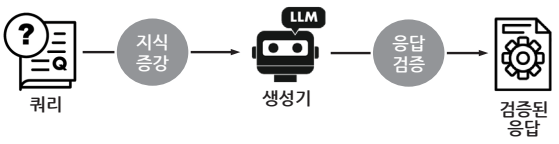
환각현상은 특히, 높은 정확성과 신뢰성이 요구되는 금융, 의료, 법률 등의 특정 도메인에서 생성형 AI 모델을 적용할 때, 허위 정보 생성, 도메인 지식 부족 등의 문제가 치명적으로 작용하여 그 활용이 제한될 수 있다. 금융 분야에서는 생성형 AI 모델이 복잡한 계산 및 세부 금융 지식이 요구되는 상황에서 부정확성이 두드러지며, 잘못된 수치 제시, 부정확한 예측, 과도하게 일반화된 투자 조언 등의 오류가 발생할 수 있다. 또한, 금융 규제나 업계 특성에 대한 이해 부족으로 인해 잘못된 결론에 도달하는 사례도 보고되고 있다[14]. 의료 분야에서 허위 정보는 환자 건강에 심각한 위험을 초래할 수 있다. Med-halt 테스트 결과, AI 모델에 의해 생성된 결과가 잘못된 진단과 치료법, 약물 정보 등을 생성하는 경향이 높았으며, 특히 병리학적 정보나 약물 상호작용 관련 질문에서 오류가 자주 발생하였다[15]. ChatLaw는 법률 분야에 특화된 LLM으로 법적 문서 해석, 법률 상담 등을 지원하지만, 여전히 법적 해석에 오류를 범하거나 법적 선례를 잘못 인용하며 일관되지 않은 조언을 제공하는 문제가 있었다[16]. 스탠퍼드대학교 연구진은 법률 분야 AI 도구를 공급하는 일부 기업에서 자사 제품이 환각을 제거하거나 회피할 수 있다고 홍보하고 있었지만, 17~33%의 환각이 있음을 실험적으로 증명하면서 환각에 대한 과장 광고를 경고하고, AI 도구에 대한 투명한 벤치마킹과 공개 평가가 필요하다고 주장하였다[17].

III. 환각현상 완화 기술

1. 환각현상 완화를 위한 접근 방식

거대언어모델(LLM) 관련 연구 중 환각현상 관련 연구에 대한 관심이 높아지고 있다. LLM의 환각현상 주제로 arXiv에 등록된 1,327편의 논문을 분석해 보면 2022년 5편, 2023년 282편, 2024년에는 1,040편이 게시되어, 2023년을 기점으로 관련 연구가 급격히 증가하고 있음을 보여주고 있다[18]. 이러한 환각현상 관련 연구주제는 1) 환각 자체를 유형화하고 원인을 분석, 2) 환각 탐지나 완화 방법, 3) 환각 여부를 평가할 벤치마크 개발에 관한 연구로 분류된다[19]. 이와 같은 연구 흐름 속에서 환각현상을 완화하는 방법으로 최근에 Chain of Thought(CoT)와 같은 추론 방법을 사용하여 더 정확한 답변을 생성하는 연구가 중점적으로 진행되었지만, OpenAI에서 2025년 4월에 공개한 추론 모델인 o3 및 o4-mini에서 이전 모델보다 더 많은 환각을 유발하는 것으로 보고되어, 추론 모델이 오히려 환각을 심화시킬 수 있다는 가능성도 드러났다[20].

LLM의 환각현상 완화 방법은 모델 내부 접근 방식과 모델 외부 접근 방식으로 구분해 볼 수 있다. 모델 내부 접근 방식은 쿼리 입력과 응답 출력을 제외한 LLM 자체의 구조 또는 디코딩 과정에서 환각현상을 줄이려는 접근 방식으로, 학습 단계에서 LLM의 아키텍처와 훈련 방식에 내재된 한계를 최적화하거나, 추론 단계에서 디코딩 방법을 사실성 및 충실성 관점에서 개선하는 연구들이 이에 해당한다[21]. 반면, 싱가포르국립대학교 연구진은 LLM이 범용적으로 문제를 해결하려 할수록 환각을 피하기 어려운 근본적인 한계를 갖고 있음을 수학적으로 설명하고[22], 모델 내부에서 환각을 완전히 제거하려는 시도보다 프로그램이 가능한 가드레일[23]을 활용하거나 RAG(Retrieval-Augmented



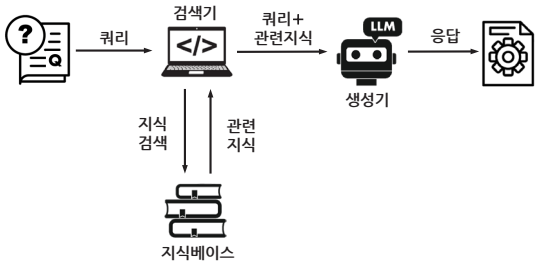
아이콘 출처 게티이미지뱅크, 무단 전재 및 재배포 금지

그림 3 환각현상 완화를 위한 모델 외부 접근 방식

Generation)와 같이 외부 지식기반을 통해 LLM에 추가 정보를 제공하고 응답을 검증하는 방식이 더욱 현실적인 대안이라고 제안하였다. 이처럼, 모델 외부 접근 방식은 외부에 구축한 지식기반을 활용하여 입력 쿼리의 정보를 증강하거나 LLM이 생성한 응답의 사실 여부를 사후적으로 검증하는 방식으로 구현된다(그림 3).

2. 지식 증강 기술

LLM 모델 외부 접근 방식 중에서는 별도의 지식기반을 통해 쿼리와 관련된 지식을 증강하여 LLM 모델의 입력으로 사용하는 RAG 방식이 대표적이다. RAG는 LLM에서 쿼리에 응답하기 전에 별도의 지식기반에서 관련 정보를 검색하고 이를 이용하여 응답을 생성하는 방법으로, 출처를 인용하여 답변의 정확성을 높이고 최신 지식 및 도메인별 지식을 쉽게 활용할 수 있어 환각을 완화하는데 효과적인 방법이다[24]. RAG는 그림 4와 같이 벡터 인코딩 모델을 활용하여 BM25, DPR, ColBERT 같은 기법을 활용하여 문서를 검색하는 검색 단계와 검색된 문서를 조건으로 사용하여 응답을 생성하는 생성 단계로 구분된다[24]. 2020년에 처음 소개[25]된 이후, RAG는 데이터를 인덱싱하고 검색해서 응답을 생성하는 기본 방식(Naive RAG)에서, 검색 및 생성의 품질을 개선하기 위해 검색 전에 데이터 인덱싱을 포함한 임베딩을 최적화하고 검색 후에 검색된 문서를 LLM의 컨텍스트 윈도우에 최적화하는



아이콘 출처 게티이미지뱅크, 무단 전재 및 재배포 금지

그림 4 RAG 개요도

방식(Advanced RAG)으로 발전하였고, 이후 다양성과 유연성을 제공하는 모듈형 방식(Modular RAG)으로 발전하였다[24]. 또 다른 한편으로, RAG는 검색을 통해 얻은 외부 지식을 LLM 프롬프트에 직접적으로 추가하는 방식에서 지식그래프를 통해 추론 통로를 강화하거나 CoT 프롬프팅처럼 단계별 지식검색을 제공하여 추론 연결에서 사실적 오류를 줄이는 방식(Iterative RAG)으로 발전하고 있다[24].

그러나 RAG에서도 환각현상이 발생할 수 있다. 검색 단계에서는 지식베이스의 오류가 있을 수 있으며 청킹이나 임베딩에 의해 결정되는 검색기의 성능에 따라 검색실패로 환각이 발생할 수 있다. 더불어 환각현상을 완화하기 위해서는 입력되는 사용자 쿼리에 대해서도 검색이 불필요한 쿼리인지를 판단해야 하고, 모호하거나 복잡한 쿼리를 재구성하는 것이 필요하다[21]. RAG의 생성 단계에서도 쿼리와 관련 없는 자료가 검색된 정보로 입력되거나 검색된 정보가 많은 경우에 주요 정보를 활용하지 않아서 발생하는 생성병목 문제로 환각현상이 발생할 수 있다[21].

RAG를 통해 더 많은 정보를 제공할수록 성능이 향상되는 경향을 보이지만, 일정 토큰 크기 이후에는 오히려 성능이 감소하는 현상이 나타난다. Quinn Leng 등의 연구진은 RAG 검색을 통해 긴 문장(2천~2백만 토큰)의 정보를 입력받은 LLM 모델들

이 올바른 답변을 하는지 평가한 결과에서 128k 토큰 이후에 성능이 저하되는 것으로 보고하였다[26]. 이를 보완하기 위해 Ali Behrouz 등 Google의 연구진은 장기기억 메모리를 갖는 아키텍처인 Titans를 제안하였으며, 이를 통해 2백만 토큰을 포함하는 긴 컨텍스트 윈도우를 갖거나 장기적인 정보 종합 능력을 평가하고 시계열 데이터를 분석하는 경우에도 기존 트랜스포머 모델보다 높은 성능을 유지할 수 있다고 발표하였다[27].

3. 응답 검증 기술

LLM 개발을 주도하는 글로벌 AI 선도기업은 주로 인간 피드백의 강화(Enhancing Human Feedback), 기계적 설명 가능성(Mechanical Interpretability), 견고성(Robustness) 등 AI 서비스의 성능을 향상시키기 위한 기술 연구에 집중해 왔다[28]. 그러나 대부분의 일반 기업은 선도기업들이 제공하는 LLM을 활용하는 처지에 있기 때문에, LLM을 안전하게 배포하기 위해서는 응답 생성 결과의 사실성 및 입력 정보에 대한 충실성을 검증하거나, 답변할 수 없는 질문에 대해서는 명시적으로 응답을 거부할 수 있는 기능 등 부정확하거나 신뢰할 수 없는 응답을 제어할 메커니즘이 필요하다.

카네기멜런대학교와 Google의 연구진은 언어모델을 이용하여 언어모델의 결과물에 대한 출처를 탐색(Research)하고 출처를 기반으로 결과물을 변경(Revise)하는 방법을 제안하였고, 이를 통해 언어모델의 처음 결과물의 문장 구조를 유지하면서 출처 정보를 강화할 수 있음을 확인하였다[29]. 싱가포르 난양공과대학교(NTU)와 알리바바그룹의 DAMO 아카데미 연구진은 CoT 프롬프팅의 사실성을 향상시키기 위해 Self-Consistency 기반으로 불확실한 예측을 식별한 뒤 핵심 정보를 검증(Verify)할 질문을 생성하고, 외부 지식 검색을 통해 신뢰할 수 있는 출처를 확보하여 기존의 논리를 수정함으로써 보다 사실적인 예측을 생성하는 방법을 제안하였다[30].

CoT 방식은 문제 해결을 위해 모델이 중간 단계를 명시적으로 표현함으로써 사고 과정을 나타내는 방법으로, 질문에 대한 답을 바로 제공하는 대신 그 답을 도출하기 위해 수행된 단계적인 사고 과정을 통해 더 논리적이고 구조적인 답을 생성할 수 있게 한다. Meta AI 연구진은 기존 CoT 방식을 확장하여 LLM이 생성한 답변을 체계적으로 검증하는 절차를 포함한 Chain-of Verification 방식을 제안해 CoT보다 약 30% 환각을 감소시켰다[31]. 마이크로소프트 연구진은 LLM이 생성한 문장을 외부 지식과 자동화된 피드백을 결합해 사실성을 향상시킨 Plug-and-Play 프레임워크인 LLM Augmenter를 제안하였고[5], 또 다른 연구에서는 별도의 파인튜닝이나 도메인 특화된 프롬프트 없이도 근거 없는 환각을 탐지 및 제어하는 Chain-of-NLI 기법을 도입해 CoT 대비 약 35% 환각 감소 효과를 확인하였다[32].

LLM의 환각현상은 종종 훈련 데이터의 범위를 벗어난 질문에 대해서도 답변을 제공하려는 '유용성(Usefulness)'을 고려하는 과정에서 발생한다. 이에 따라, LLM이 "I don't know"라고 응답하며 답변을 거부하는 방식이 환각을 제어하는 방법으로 연구되고 있다. Xinxi Chen 등 연구진은 작은언어모델(SLM: Small Language Model)을 미세조정하여 정보가 충분하지 않을 경우에 "I don't know"라고 답할 수 있도록 학습시키는 방법을 제안하고 RAG와 결합하여 잘못된 정보가 포함될 가능성을 줄이고 신뢰할 수 있는 응답을 생성할 수 있도록 하였다[33]. 포츠담대학교(HPI)와 텔아비브대학교 연구진은 모델의 불확실성을 표현하도록 확률을 조정하는 새로운 모델 캘리브레이션 방식으로, LLM의 어휘에 "I don't know([IDK])" 토큰을 추가하고, 모델이 정답을

맞히지 못할 경우 [IDK] 토큰을 선택할 수 있도록 유도하는 방법을 제안하였다[34]. 홍콩대학교 연구진은 사전 학습된 모델의 지식과 교육 데이터 간의 불일치를 파악하여, 모델이 모르는 질문에 대해서는 답하지 않도록 조정하는 R-Tuning(Refusal-Aware Instruction Tuning)을 제시하였다[35].

IV. 환각현상 평가 기술

LLM의 환각현상을 평가하기 위한 벤치마크는 1) 환각현상의 정량적 지표를 평가하는 평가 벤치마크(Evaluation Benchmark)와 2) 환각현상을 탐지하는 방법의 성능을 평가하는 탐지 벤치마크(Detection Benchmark)가 있다[21]. 평가 벤치마크는 주어진 문맥에 따른 사실적 부정확성이나 불일치성을 평가하기 위해 다지선다형의 질문과 추출된 응답을 평가하는 Multi-Choice QA, 또는 문장 형태로 요약된 답변을 생성하고 대부분 사람에 의해 평가하는 Generative QA 방식으로 수행되며, 여기에는 TruthfulQA, HalluQA, REALTIMEQA 등의 벤치마크가 포함된다. 탐지 벤치마크는 SelfCheckGPT-Wikibio, HaluEval, BAMBOO, FELM 등이 포함된다[21].

환각현상과 관련한 다양한 벤치마크가 존재하지만, 스탠퍼드대학교의 인간중심AI연구소(Human-Centered Artificial Intelligence)는 AI Index Report에서 AI 안전성 및 책임성 평가에 대한 표준 벤치마크가 부족하다고 강조하였다. 개발자들은 주로 수학, 코딩, 언어능력을 포함한 LLM의 범용 성능에 대해서는 일정 수준의 합의에 도달한 벤치마크를 활용하여 테스트하고 있지만, 안전성 및 책임성 관련한 AI 평가에 대해서는 벤치마크에 대한 일정 수준의 합의가 아직 형성되지 않았다고 보고하였다[36].

이러한 상황 가운데, 환각현상에 대한 새로운 벤치마크로 LLM 모델의 환각현상을 비교해 볼 수 있는 시도가 진행되고 있다. 미국의 스타트업인 Vectara는 환각현상에 대한 신뢰할 수 있는 모델을 식별하기 위해 자체 개발한 환각 평가 모델 HHEM(Hughes Hallucination Evaluation Model)을 기반으로 환각 리더보드를 운영 중이며, 최근까지 공개된 모델 중 Google의 Gemini 2.0 Flash 모델이 0.7%의 환각율로 가장 좋은 성능으로 일관되게 사실 정보를 제공하는 것으로 보고되었다[37]. Google DeepMind는 LLM이 제공된 소스 자료에 대한 응답을 얼마나 정확하게 근거로 삼고 환각을 피하는지에 대해 측정하기 위해 FACTS Grounding 벤치마크를 제공하고 있다[38]. Kaggle에서 FACTS Grounding 리더보드를 운영하고 있으며, 2025년 7월 기준, Google Gemini 2.5 모델들이 상위권(87.8% 이하)을 차지하고 있다[39]. 리더보드에는 포함되지 않았지만, Douwe Kiela 스탠퍼드대학교 교수에 의해 2023년에 설립된 ContextualAI는 검색된 소스 데이터에 강력하게 기반하여 정확한 응답을 제공하는 GLM(Grounded Language Model)을 소개하고 FACTS Grounding 벤치마크에서 최고점인 88%의 성능을 얻었다고 발표하였다[40].

V. 결론

LLM에서 환각현상을 완전히 제거할 수 있다고 주장하는 AI 기업들이 출현하고 있지만[41-43], 이러한 기업들은 구체적인 적용 사례들이 아직 보고되지 않아 신뢰받지 못하는 상황이다. 더불어 환각현상에 대한 AI 평가 벤치마크는 범용 성능을 평가하는 벤치마크와 달리 아직 보편적 기준이 될 만한 표준 벤치마크가 정립되어 있지 않다. 즉, 환각현상을 평가하는 절대 기준이 없어 환각현상을 제거했다는 객관적 검증이 어려운 상황이다.

글로벌 AI 선도기업들은 환각현상에 관한 연구를

지속하고 있지만, 이들이 제공하는 LLM을 운용하는 일반 기업의 입장에서는 LLM의 환각현상을 객관적으로 평가하면서 LLM으로부터 독립적인 외부 보조수단을 통해 환각현상을 제어하는 기술을 확보하는 것이 필요하다. 이를 위해 기업이 자체 구축한 지식베이스를 통해 사용자 쿼리에 대한 지식 증강으로 응답을 생성하고, 생성된 응답을 다시 지식베이스를 기반으로 검증하여 환각현상을 제어할 수 있어야 한다.

인간 사회의 질서가 개인의 본성에 의지하기보다는 교육 및 법체계를 기반으로 한 규범적 구조 속에서 사회적 평가와 검증을 통해 유지되듯이, AI가 고도화될수록 이를 평가하고 검증할 수 있는 제어 기술 역시 함께 마련되어야 한다. 특히, 국내 AI 기업은 글로벌 AI 기업이 설정한 기준에 의존하기보다 국내 환경에 적합하게 수립된 기준을 기반으로 한국 사회의 특수성에 부합하는 안전하고 책임 있는 LLM을 배포하고 활용하는 것이 필요하다.

용어해설

환각현상 LLM 분야에서 환각현상은 생성된 출력이 실제 사실과 일치하지 않는 현상(Factuality Hallucination), 또는 사용자 입력이나 문맥에 충실하지 않거나 결과 자체가 논리적으로 모순되는 현상(Faithfulness Hallucination)을 의미함

참고문헌

[1] A.D. Thompson, "Integrated AI: The sky is steadfast (2024 AI retrospective)," LifeArchitect.ai, 2024. 12. https://lifearchitect.ai/the-sky-is-steadfast/

[2] J. Rowan et al., "Deloitte's State of Generative AI in the Enterprise Q4 report," Deloitte, 2025. 1. https://www.deloitte.com/us/en/what-we-do/capabilities/applied-artificial-intelligence/content/state-of-generative-ai-in-enterprise.html

[3] A. Singla et al., "The State of AI in early 2024: Gen AI adoption spikes and starts to generate value," McKinsey&Company, 2024. 5. 30. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-2024

[4] C. Huyen, "Open Challenges in LLM research," Chip Huyen's Blog, 2023. 8. 16. https://huyenchip.com/2023/08/16/llm-research-open-challenges.html

[5] X. Chen et al., "Honest AI: Fine-tuning "Small" Language Models to Say "I don't Know" and Reducing Hallucination in RAG," arXiv preprint, 2024. doi: 10.48550/arXiv.2410.09699

[6] K. Hammond, "The hallucination Problem: A Feature, Not a Bug," Center for Advancing Safety of Machine Intelligence (CSAMI), 2024. 8. 26. https://casmi.northwestern.edu/news/articles/2024/the-hallucination-problem-a-feature-not-a-bug.html

[7] D. Amodei, "The Urgency of Interpretability," 2025. 4. https://www.darioamodei.com/post/the-urgency-of-interpretability

[8] Z. Ji et al., "Survey of Hallucination in Natural Language Generation," arXiv preprint, 2024. doi: 10.48550/arXiv.2202.03629

[9] L. Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," arXiv preprint, 2024. doi: 10.48550/arXiv.2311.05232

[10] A. Koenecke et al., "Careless Whisper: Speech-to-Text Hallucination Harms," in Proc. ACM Conf. Fairness, Accountability, and Transparency, (Rio de Janeiro, Brazil), Jun. 2024, pp. 1672-1681.

[11] 권택경, "챗봇 엉뚱한 답변에 기업이 배상, AI 활용에 신중해야," 동아닷컴, 2024. 2. 19. https://www.donga.com/news/It/article/all/20240219/123592776/1

[12] https://incidentdatabase.ai/

[13] https://airisk.mit.edu/

[14] H. Kang and X-Y. Liu, "Deficiency of large language models in finance: An empirical examination of hallucination," arXiv preprint, 2023. doi: 10.48550/arXiv.2311.15548

[15] A. Pal et al., "Med-halt: Medical domain hallucination test for large language models," arXiv preprint, 2023. doi: 10.48550/arXiv.2307.15343

[16] J .Cui et al., "Chatlaw: Open-source legal large language model with integrated external knowledge bases," arXiv preprint, 2024. doi: 10.48550/arXiv.2306.16092

[17] V. Magesh et al., "Hallucination Free? Assessing the Reliability of Leading AI Legal Research Tools," arXiv preprint, 2024. doi:

10.48550/arXiv.2405.20362

[18] arXiv 홈페이지(https://arxiv.org/)에서 Abstract 필드에 LLM 및 hallucination을 포함하는 컴퓨터과학(Conputer Science)분야 논문 검색 후 연도별 정리

[19] 박대민, 이한종, "챗GPT 등장 이후 인공지능 환각 연구의 문헌 검토: 아카이브(arXiV)의 논문을 중심으로," 정보화정책 제31권 제2호, 2024. 6, pp. 3-38.

[20] M. Zeff, "OpenAI's new reasoning AI models hallucinate more," TechCrunch, 2025. 4. 18.

[21] L. Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," arXiv preprint, 2024. doi: 10.48550/arXiv.2311.05232

[22] Z. Xu et al., "Hallucination is inevitable: An innate limitation of large language models," arXiv preprint, 2024. doi: 10.48550/arXiv.2401.11817

[23] T. Rebedea et al., "NeMo Guardrails: A Toolkit for controllable and Safe LLM applications with Programmable Rails," arXiv preprint, 2023. doi: 10.48550/arXiv.2310.10501

[24] Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv preprint, 2023. doi: 10.48550/arXiv.2312.10997

[25] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," arXiv preprint, 2020. doi: 10.48550/arXiv.2005.11401

[26] Q. Leng et al., "The Long Context RAG Capabilities of OpenAI o1 and Google Gemini," Databricks blog, 2024. 10. 8. https://www.databricks.com/blog/long-context-rag-capabilities-openai-o1-and-google-gemini

[27] A. Behrouz et al., "Titans: Learning to Memorize at Test Time," arXiv preprint, 2024. doi: 10.48550/arXiv.2501.00663

[28] 장진철, "주요 기업의 AI 안전 대응 동향 및 시사점," KISTEP 이슈분석 273호, 2024. 10, pp. 1-15.

[29] L. Gao et al., "RARR: Researching and Revising What Language Model Say, Using Language Models," arXiv preprint, 2023. doi: 10.48550/arXiv.2210.08726

[30] R. Zhao et al., "Verify-and-Edit: A Knowledge-Enhanced Chain-of-Thought Framework," arXiv preprint, 2023. doi: 10.48550/arXiv.2305.03268

[31] S. Dhuliawala et al., "Chain-of-verification reduces hallucination in large language models," arXiv preprint, 2023. doi: 10.48550/arXiv.2309.11495

[32] B. Peng et al., "Check your facts and try again: Improving large language models with external knowledge and automated feedback," arXiv preprint, 2023. doi: 10.48550/arXiv.2302.12813

[33] D. Lei et al., "Chain of natural language inference for reducing large language model ungrounded hallucinations," arXiv preprint, 2023. doi: 10.48550/arXiv.2310.03951

[34] R. Cohen et al., "I Don't Know: Explicit Modeling of Uncertainty with an [IDK] Token," arXiv preprint, 2024. doi: 10.48550/arXiv.2412.06676

[35] H. Zhang et al., "R-tuning: Teaching large language models to refuse unknown questions," arXiv preprint, 2024. doi: 10.48550/arXiv.2311.09677

[36] N. Maslej et al., "The AI Index 2025 Annual Report," Stanford University, Human-Centered Artificial Intelligence, 2025. 4. https://hai.stanford.edu/ai-index/2025-ai-index-report

[37] Github, "Hallucination Leaderboard," 2025. 4. 29. https://github.com/vectara/hallucination-leaderboard

[38] Google DeepMind FACTS team, "FACTS Grounding: A new benchmark for evaluating the factuality of large language models," Google DeepMind, 2024. 12. 17.

[39] Kaggle, "FACTS Grounding Leaderboard," 2025. 6. 20. https://www.kaggle.com/facts-leaderboard

[40] I. Sinha, "Introducing the most grounded language model in the world," ContextualAI, 2025. 3. 4. https://contextual.ai/blog/introducing-grounded-language-model/

[41] M.C. Wood and A.A. Forbes, "100% Hallucination Elimination Using Acurai," arXiv preprint, 2024.

[42] M. Nuñez, "Exclusive: Alembic Debuts hallucination-free AI for enterprise data analysis and decision support," VentureBeat, 2024. 5. 6. https://venturebeat.com/ai/exclusive-alembic-debuts-hallucination-free-ai-for-enterprise-data-analysis-and-decision-support/

[43] G. Rao, "Hallucination free AI: A path to modern explainability in Enterprise," Howso, 2024. 10. 1. https://www.howso.com/hallucination-free-ai/