

# 이미지 및 텍스트 생성형 AI의 현황과 한계

임해찬

한국정보통신기술협회

ohaechano@tta.or.kr

## Current Status and Limitations of Image and Text Generative AI

Im Hae Chan

Telecommunications Technology Association

### 요약

생성형 AI는 인간을 모방하여 다양한 창의적인 작업을 수행할 수 있는 능력이 있지만, 할루시네이션과 같은 문제를 내포하고 있다. 본 논문은 'ChatGPT'의 할루시네이션과 'Krea.ai'의 이미지 생성 오류를 도출하여 생성형 인공지능의 현황과 한계에 대해 다루었다.

### I. 서론

2022년 11월 30일, Open AI가 발표한 'ChatGPT'는 단 2개월 만에 월간 사용자 1억을 돌파하여 생성형 인공지능(Generative Artificial Intelligence)이 화두에 오르게 되었다. 생성형 AI는 인간을 모방해 대화, 이미지, 음성, 코드, 동영상 등 다양한 결과물을 생성해 주는 인공지능이다 [1]. 생성형 AI를 잘 활용한다면 전문 지식이나 배경이 없더라도 다양한 결과물을 생성해 낼 수 있다는 장점이 있다.

반면 이러한 장점으로 인해 그 누구라도 생성형 AI를 악용할 수도 있다. 생성형 AI로 인해 거짓 정보 생성, 개인정보 유출, 딥페이크(Deepfake)로 인한 사회적 혼란 등 다양한 문제들이 발생하고 있다. 이러한 생성형 AI의 문제점을 개선하기 위해서는 다양한 방법의 해결책이 요구되며 본 논문에서는 이러한 생성형 AI의 문제점을 도출하기 위해 이미지, 텍스트 생성형 AI를 활용하여 생성형 AI의 한계와 이를 해결하는 방법에 대한 방향성을 살펴본다.

### II. 본론

생성형 AI 모델은 편향, 투명성, 고비용, 거짓 정보 생성 등 다양한 문제가 존재한다. 이러한 문제들은 인공지능이 훈련 데이터의 한계나 처리 과정의 복잡성에서 비롯되며, 결과적으로 사용자에게 잘못된 정보를 제공하거나 예측 불가능한 행동을 유발할 수 있다. 대표적인 사례로 'ChatGPT'의 할루시네이션, 'Krea.ai'의 생성 이미지의 오류 등을 예로 들 수 있다 [2]. 이는 AI 기술이 사회적으로 널리 활용됨에 따라, 이러한 문제들이 더 심각한 결과를 초래할 수 있음을 시사한다.

'ChatGPT'는 대화형 인공지능으로 문서의 요약, 문장 생성, 문제 해결 등 다양한 작업을 수행할 수 있으며 DALL-E2와 연계하여 이미지 생성 작업도 수행할 수 있다. 하지만 'ChatGPT'는 LLM(Large Language Model) 모델로써 할루시네이션이 자주 발생하는 단점이 있다. 할루시네이션이란 대화형 인공지능에서 존재하지 않거나 사실이 아닌 답을 마치 진실인 듯이 답변하는 것을 뜻한다[3]. 이러한 증상은 훈련 데이터의 한계, 문맥 이해의 한계, 질문의 모호성, 예측 기반 생성, 사용자의 입력에 따른

영향으로 구분할 수 있다[4]. 할루시네이션은 사용자의 질문에 대해 유연하게 답변하는 'ChatGPT'의 장점을 보여주는 사례이기도 하지만, 역사적 왜곡 또는 거짓 정보를 사실처럼 이야기하여 사회적 혼란을 초래할 수 있다.

본 논문에서는 텍스트 생성형 AI의 한계를 확인하기 위해 <표 1>과 같이 'ChatGPT'를 이용해 할루시네이션을 도출하였다. 할루시네이션을 도출하기 위해 잘못된 정보를 진실인 듯이 확정하고 그 뒤에 정답에 필요한 질문을 하였다. 이러한 현상은 LLM 모델이 문맥 이해 능력에 기반하여 주어진 질문에 대해 가장 정답일 확률이 높은 답변을 하려고 하기 때문이다. 해당 질문을 분석해 보면 출생 연도를 알려달라는 질문에 대한 정답은 정확히 답변한 것을 볼 수 있다.

<표 1> ChatGPT의 할루시네이션

| 질문                       | 답변  |
|--------------------------|---|
| 해시계를 발명한 정약용의 출생 연도를 알려줘 | 정약용은 1762년에 태어났습니다. 그는 조선 후기의 실학자이자 개혁가로, 다양한 학문과 발명에서 뛰어난 업적을 남겼습니다. 해시계도 그의 발명 중 하나입니다. |

데이터베이스에 데이터가 없을 때도 잘못된 결과를 낼 수 있기에 이를 확인하고자 <표 2>와 같이 질문하였다. 해당 결과는 올바르게 답함을 확인하였고, 이를 통해 데이터베이스에 있는 정보인테도 불구하고 거대 언어 모델의 한계로 인한 텍스트 생성형 LLM 모델의 문제점 중 하나인 할루시네이션 현상임을 확인하였다.

<표 2> 데이터베이스의 데이터 존재 여부 확인에 대한 질의응답

| 질문              | 답변   |
|-----------------|--|
| 해시계를 만든 사람이 누구야 | 해시계(해시계, 석국)를 발명한 사람은 고려 시대의 과학자 장영실입니다. 장영실은 조선 세종 대왕 시대에 활동했으며, 조선 초기의 가장 유명한 과학자 중 한 명입니다....(생략)                                       |
| 정약용이 해시계를 만들었어? | 아니요, 정약용이 해시계를 만든 것은 아닙니다. 해시계는 고려 시대의 과학자 장영실이 발명했습니다. 장영실은 조선 세종대왕 시대에 활동한 과학자로, 해시계뿐만 아니라 자력루(물시계)와 같은 다양한 천문학적 및 측시 기기를 발명했습니다....(생략) |

이러한 텍스트 생성형 AI의 문제점을 해결하기 위해 다양한 방면의 노력이 이루어지고 있다. 대표적인 예로 'AI 레드팀 챌린지'가 있다. '생성형 AI 레드팀 챌린지'는 AI 모델의 위험·취약점을 찾는 클라우드 소싱(Crowd Sourcing)이다. 23년 8월에 미국에서 구글, 오픈 AI, 엔비디아 등 빅테크 8개 사가 참여하여 공개 챌린지를 개최하였으며 24년 4월 국내에서도 생성형 AI 레드팀 챌린지가 개최되었다[5, 6].





<표 3> 위험·취약점 제출 수 및 비중(%)

| 위험·취약점  | 건수    | 비율      |
|---------|-------|---------|
| 탈옥      | 698   | 8.64%   |
| 편견 및 차별 | 1,156 | 14.31%  |
| 인권 침해   | 642   | 7.94%   |
| 사이버 공격  | 681   | 8.43%   |
| 불법 콘텐츠  | 1,005 | 12.44%  |
| 잘못된 정보  | 3,370 | 41.71%  |
| 일관성     | 528   | 6.53%   |
| 합계      | 8,080 | 100.00% |

국내에서 행해진 '레드팀 챌린지'의 위험·취약점의 제출 결과는 <표 3>과 같이 나타났으며 잘못된 정보(할루시네이션)가 총 8,080건 중 3,370건(41.71%)으로 가장 많이 제출되었다. 이러한 방법을 통해 기업은 AI 모델의 문제점을 파악하고 안전성과 견고성을 평가할 수 있으며 오용될 가능성을 줄일 수 있다.

'Krea.ai'는 텍스트를 입력하면 이미지를 생성해 주는 사이트이다. 본 논문에서는 이미지 생성형 AI의 한계를 확인하기 위해 해당 사이트에서 'hand'를 입력하여 이미지를 반복 생성해 보았다. 그 결과 <표 4>와 같이 정상인 이미지와 손가락이 누락, 변형, 추가된 다양한 오류를 가진 이미지가 도출되었다. 이미지 생성형 AI는 주로 이미지 데이터를 기반으로 학습을 진행하며, 비록 픽셀의 패턴은 인식할 수 있지만, 손의 움직임이나 복잡한 구조까지 완전히 파악하는 데는 한계가 있다. 손은 다양한 각도, 손가락의 길이와 두께, 그리고 주름 등 복잡한 요소들로 구성되어 있어, AI가 이를 정확하게 재현하기 어려운 손가락이 누락되거나 추가되고 변형되는 등의 오류가 자주 발생한다.

<표 4> 'Krea.ai'의 손 생성 이미지 결과

| 구분 | 설명   | 이미지   |
|----|--|---|
| 정상 | 다섯 손가락과 손바닥의 주름 등의 특징이 모두 정상인 경우           |  |
| 누락 | 손가락이 1개가 누락 되어 4개 이하로 나타나는 경우              |  |
| 변형 | 손가락 앞 뒷면이 반대로 되거나 붙는 등 손가락의 형태에 변형이 일어난 경우 |  |
| 추가 | 손가락이 1개가 추가되어 6개 이상으로 나타나는 경우              |  |

### III. 결론

본 논문은 텍스트 및 이미지 생성형 인공지능 모델의 한계와 그로 인한 문제점을 도출하였다. LLM 모델이 문맥 이해 능력에 기반하여 주어진 질문에 대해 최적의 답변 도출 시 발생하는 '할루시네이션' 현상을 살펴본 것으로, 이는 잘못된 정보의 전파로 이어질 수 있다. 'ChatGPT'와 같은 대화형 인공지능에서는 질문의 모호성이나 문맥 이해의 한계로 인해 잘못된 답변을 할 가능성이 있는 것으로 나타났다. 이러한 텍스트 생성형 AI의 문제점을 해결하는 방안으로 AI 레드팀 챌린지와 같은 클라우드 소싱 방법이 활용되고 있으며, AI 모델의 취약점을 평가하고 개선하려는 노력이 이뤄지고 있음을 확인하였다.

이와 함께, 이미지 생성형 AI의 한계를 확인하기 위해 'Krea.ai'에서 특정 텍스트 입력(hand)을 기반으로 반복적인 이미지 생성 실험을 수행하였다. 분석 결과, AI가 생성한 손 이미지는 여러 구조적 오류를 드러냈는데, 정상적인 이미지와 달리 손가락이 누락되거나 추가되는 현상, 그리고 손가락의 변형 등의 다양한 문제가 발견되었다. 이러한 오류는 더욱 정밀하고 다양한 학습 데이터셋의 구축, 학습 알고리즘의 고도화 등을 통해 보완할 수 있지만 완벽하지는 않기에 이에 대한 향후 추가 연구가 필요하다.

### 참 고 문 헌

- [1] 양은영. (2023.06). 생성형 AI의 개발 및 이용에 관한 규제 필요성 - 대규모 언어모델에 기반한 대화형 인공지능 서비스(LLMs AI)를 중심으로 -. 성균관법학, 35(2), 293-325.
- [2] 허준행. (2023.05). 인공지능 시대, 한국 현대사 연구의 과제와 전망 - 생성형 인공지능 챗GPT를 중심으로. 한국사학연구, (74), 45-80.
- [3] 환각 (인공지능). (2024.3). 위키백과. 12:33, 2024년 5월 7일에 확인. [https://ko.wikipedia.org/w/index.php?title=%ED%99%98%EA%B0%81\\_\(%EC%9D%B8%EA%B3%B5%EC%A7%80%EB%8A%A5\)&oldid=36617638](https://ko.wikipedia.org/w/index.php?title=%ED%99%98%EA%B0%81_(%EC%9D%B8%EA%B3%B5%EC%A7%80%EB%8A%A5)&oldid=36617638) 에서 찾아볼 수 있음.
- [4] 안진호, 정육환. (2024.3). 생성형AI의 환각현상 최소화를 위한 요인 탐색 연구 - 소비자의 감성·경험 분석을 중심으로 -. 서비스 연구, 14(1), 77-90.
- [5] Generative AI Red Teaming Challenge. 2024년 05월 07일에 확인. URL: <https://www.humane-intelligence.org/grt>
- [6] 2024 생성형 AI 레드팀 챌린지. 2024년 5월 7일에 확인. URL: <https://aiignite.org/HOME>