데이터 품질과 생성형 AI 기술 동향

김혜진 한국전자통신연구원 책임연구원, 과학기술연합대학원대학교 조교수

1. 머리말

ChatGPT가 소개된 이후, AI로 인해 세상이 가파르게 달라지고 있다. 2024년 5월 22일엔 "오픈AI(OpenAI)가 배우 스칼릿 조핸슨(Scarlett Johansson)의 음성을 무단 학습·도용했다"는 기사가 나왔다[1]. 사람과 음성으로 대화할 수 있는 GPT-4o, 특히 스카이(Sky)의 목소리가 조핸슨의 그것과 너무나 닮아 논란이 불거진 것이다. 해당 기사엔 성우의 목소리를 무단으로 사용한 음성 AI 스타트업의 이야기도 함께 실려 있다.

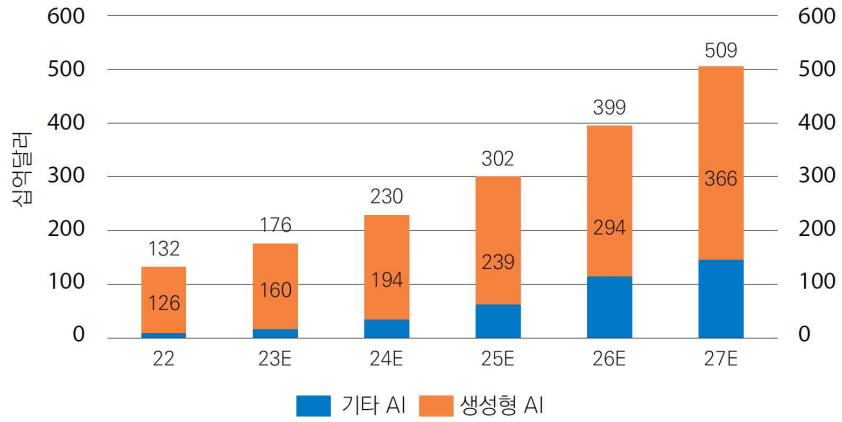
생성형 AI의 핵심은 학습에 필요한 데이터다. 최근의 논란은 AI 모델 학습 시 허락을 받지 않고 사용했기에 벌어진 일이다. 오픈AI는 ChatGPT 출시 이후 다양한 저작권 침해 논란에 휩싸였다. 뉴욕타임스는 "오픈AI가 자사 콘텐츠를 무단 사용했다"며 지난해 저작권 침해 소송을 걸었다[2]. 또 IT 매체 더버지에 따르면, GPT-4 훈련을 위해 100만 시간이 넘는 유튜브(YouTube) 영상이 무단으로 사용됐다고 한다[3].

반면 일론 머스크가 설립한 AI 기업 x AI는 8조 2,000억 원 규모의 대규모 투자 유치에 성공했다. 이는 일론 머스크가 보유한 SNS 서비스 X(구 트위터)를 바탕으로, 고화질 영화 53만 편 분량의 데이터를 확보할 수 있기 때문이다[4]. 오픈AI와 xAI의 상반된 사례는, 좋은 품질의 학습 데이터를 확보하는 것이 얼마나 어렵고 중요한지를 보여주고 있다.

컨설팅 업체 IDC(International Data Corporation)에 따르면, 글로벌 AI 시장에서 가장 빠른 성장세를 보이는 것은 생성형 AI 솔루션이다. 예상되는 연평균 성장률은 73.3%로서, 전체 AI 시장 성장률인 30.4%를 2배를 훌쩍 넘는다[5].

이는 학습 데이터가 AI의 핵심이기 때문이다. 최근 한 분석기관의 추정에 따르면, 고품질 텍스트 데이터는 2026년, 이미지 데이터는 2030년대 후반이 되면 부족해질 것으로 전망된다[6]. 그 해결책으로서, 생성형 AI가 만들어 낸 데이터를 사용해 AI 성능을 높이려는 노력이 이어지고 있다[7].

1) 실리콘밸리=오로라 특파원, 김민기 기자, "목소리도 훔치나…뜨거운 감자 된 'AI음성,' 조선일보, B01면 경제종합 2024.05.22
2) 장유미 기자, "법 어겨도 모르쇠…불법 판 치는 빅테크,AI 학습 데이터 무단 사용," 지디넷코리아, 2024.04.08. https://zdnet.co.kr/view/?no=20240408091423
3) AI리포터, "오픈AI, GPT-4 학습에 유튜브 영상 100만시간 활용했다," 디지털투데이, 뉴스위드AI, 2024.04.08. https://www.digitaltoday.co.kr/news/articleView.html?idxno=512935
4) 윤상언 기자, "X 데이터 무기 쥔 머스크, xAI에 8조원 투자유치 성공," 중앙일보 2024.05.29.
5) 박세라, "생성형 AI, AI 에이전트 시대, "대신 증권, Issue Report 2023.12.19
6) 김병필, "멀티모달리티 인공지능," 중앙일보, 김병필의 인공지능 개척시대 칼럼, 2024.04.20
7) Teng Hu et al., "AnomalyDiffusion: Few-Shot Anomaly Image Generation with Diffusion Model," CVPR 2024



[그림 1] 글로벌 AI 솔루션 지출 금액 추이 및 전망

산업계는 향후 생성형 AI를 핵심 동력으로 재편될 예정이다[8]. 전 세계 데이터의 70%는 아마존과 마이크로소프트, 구글 등의 플랫폼에 저장되어 있고, 유럽은 가이아-X 프로젝트를 통해 데이터의 주권을 지키기 위해 노력하고 있다[9]. 또한, 제조공정 분야에선 생성형 AI를 결함 검출 등에 활용함으로써 그 효율성을 높이고 있다[10].

2. AI 경쟁력의 중심, 데이터

2.1 AI는 왜 데이터가 중요한가?

AI를 간단히 수식화하면 다음과 같다.

$$y = Wx + b$$

이 식에서 y는 AI의 출력값을, x는 AI의 데이터를, (W, b)는 AI 모델을 의미한다. 모델은 사람이 디자인하기 따름이나, 트랜스포머(Transformer)의 등장 이후 모델의 새로운 디자인에 대한 의존도는 줄어들고 있다. 그렇다면 이 수식에서 남은 것은 x, 즉 데이터다. 이제, AI 성능을 결정하는 것은 바로 x를 얼마나 많이, 잘 수집하는가에 달렸다. 이것이 현재 전 세계 빅테크 기업들이 양질의 학습 데이터 확보에 사활을 거는 이유다.

뉴욕타임스는 4월 초 오픈AI, 구글(Google), 메타(Meta)가 데이터 확보를 위해 저작권 회피에 대해 논의했다는 기사를 발표했다[11]. 이 기사에 따르면, 오픈 AI는 인터넷상의 고품질 텍스트 데이터를 모두 사용하는 한계점에 부딪혔다. 이에 추가 텍스트 데이터 확보를 위해 위스퍼(Wisper)라는 툴을 만들어, 유튜브(YouTube)에서 대화 텍스트를 만들었다고 한다. 구글의 경우, 자사 AI 모델 학습을 위해 유튜브 동영상을 복사했으며, 더 나아가 서비스 약관을 바꿔 구글 문서, 구글 지도의 레스토랑 리뷰, 기타 온라인 자료를 활용할 수 있도록 허용했다. 한편 메타는 사이먼&슈스터(Simon&Schuster) 출판사 인수를 논의했는데, 이는 소송에 직면하더라도 인터넷을 통해 저작권이 있는 데이터를 수집하기 위함이다.

8) 한정호 기자, "[2023데이터컨퍼런스⑥] 생성형 AI와 빅데이터 도입 전략," 아이티데일리, 2023.11.30
9) 김동영, "제조업 데이터 활용을 위한 생태계," 매일경제 2024.06.12
10) Mingyu Lee et al., "Text-Guided Variational Image Generation for Industrial Anomaly Detection and Segmentation," CVPR 2024
11) Cade Metz et al., "How Tech Giants Cut Corners to Harvest Data for A.I.," The New York Times, April, 6, 2024

국내서도 데이터 주도권 논의가 이뤄지고 있다. 최근 라인야후 사태는 외국 데이터 확보에 대해 우리가 어떤 스탠스를 취해야 하는지, 진지한 고민을 우리에게 던진다. 이는 알테쉬(알리·테무·쉬인)와 국내 소비자 개인정보 데이터에 대한 문제이기도 하다[12]. 즉, 대한민국을 포함해 전 세계가 데이터 확보 전쟁에 이미 뛰어들었고, 데이터 보호와 확보가 매우 중요한 시점임이 분명하다.

2.2 AI 학습을 위해 공개된 데이터 현황

그 중요성이 높아짐에 따라, 현재 많은 데이터들이 수집·공개되고 있다. 주간기술동향은 영상분야에서의 공개된 데이터들[13], 국내외 공개 데이터 수집 플랫폼인 AI-HUB, KAMP, Open-X, Hugging Face에 대해 소개하고 있다[14]. 이 중 AI-HUB와 KAMP는 국내 데이터 플랫폼으로, 국가 지원을 받아 수집된 데이터 셋이다. AI-HUB는 한국어, 영상이미지, 헬스케어, 교통물류, 재난안전환경, 농축수산, 문화관광, 스포츠, 교육, 로보틱스, 제조, 지식재산, 법률, 금융 등 광범위한 데이터들을 포함한다. KAMP는 주로 제조에 특화된 데이터를 다룬다.

한편, Hugging Face와 Open-X는 해외 공개 데이터 셋이다. Open-X에선 구글을 중심으로 세계 연구자들이 자발적으로 데이터 셋을 수집하고 있다. 이들은 로봇에 필요한 데이터를 전 세계가 함께 구축해보자는 취지를 공유한다. Hugging Face는 이미 공개된 다양한 데이터 셋과 함께 AI 모델들도 제공한다. 이를 바탕으로 사용자는 손쉽게 관련 데이터에 접근하고 활용할 수 있다.

우리나라는 2014년 공공데이터의 제공 및 이용 활성화에 관한 법률을 시행하면서 공공데이터 개방을 본격화했다. 이에 따라 공공데이터포털(www.data.go.kr)을 운영하고, 공공기관에서 만들어내는 공적인 정보·자료를 국민에게 공개하고 있다.

이와 같이 전 세계적으로 국가 차원의 공개 데이터 수집이 활발하다. 미국의 경우, 2009년 버락 오바마(Barack Obama) 대통령이 '개방정부 운영각서'를 발표한 이후 공공 데이터 개방을 적극 추진하고 있다. 현재 미 정부는 연방 공공데이터 포털(www.data.gov)을 통해 각 정부기관들이 제공하는 다양한 주제의 데이터 셋을 공개하고 있다. 주된 주제로는 기후, 에코, 교육, 재무, 건강 등이 꼽힌다.

[그림 2] 미 연방 공공데이터 포털

[그림 3] 영국 공공데이터 포털

12) 최경진, "AI시대 데이터 주권과 글로벌 데이터 안전지대 주도," 전자신문 ET 시론, 2024.05. 24. 25면 오피니언한정호 기자, "[2023데이터컨퍼런스⑥] 생성형 AI와 빅데이터 도입 전략," 아이티데일리, 2023.11.30

13) 김혜진, "영상 분야에서의 인공지능 발달 단계에 따른 데이터와 모델의 변화," 주간기술동향2071호, 2022년 12월 1일

14) 김혜진, "영상 분야에서의 인공지능 발달 단계에 따른 데이터와 모델의 변화," 주간기술동향2071호, 2022년 12월 1일

영국은 2010년 정보 자유법(Freedom of Information Act) 개정을 통해 데이터 공개를 강화했다. 이에 따라 정부기관은 데이터 공개 요청에 대해 적극적으로 응답해야 하며, 현재 공공 데이터 포털(www.data.gov.uk)을 운영하고 있다. 역시 범죄, 교육, 운송, 환경 등 분야별 데이터들이 공개돼 있다.

EU(유럽연합, European Union)는 통 계청(Eurostat)에서 각 회원국 통계데이터 및 유럽 차원의 광범위한 주제 데이터를 공개하고 있다. 더불어 EU는 2019년 GDPR(일반개인정보보호법, General Data Protection Regulation)을 시행하며 데이터 공개 및 활용에 대한 규제를 강화하고, 유럽연합 공공데이터 포털(https://data.europa.eu/)을 운영하고 있으며, 회원국들의 데이터 공개를 촉진하고 있다.

OECD(경제협력개발기구, O rganisation for Economic Co-operation and Development)는 OECD Data(https://data.oecd.org/)에서 경제, 교육, 고용, 환경 등 다방면 데이터를 공개하고 있다.

세계은행에선 World Bank Open Data(https://data.worldbank.org/)를 통해 전 세계 국가들의 개발지표, 경제지표 등 통계데이터를 공개하고 있다. 여기엔 여러 출처, 지표, 국가, 지역에 대한 데이터가 포함된다.

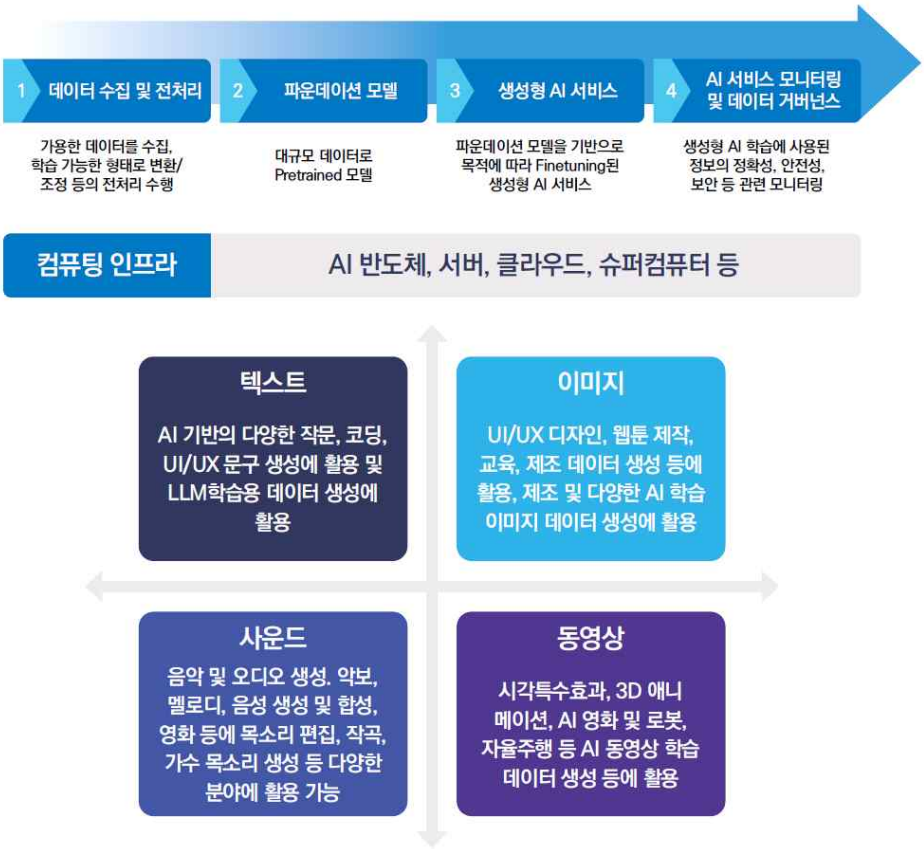
한편, 통계사이트 아워월드인데이터(Our World in Data, https://ourworldindata.org/)는 지속가능발전 목표(SDG)를 추적하는 데이터를 제공하고 있다. 이는 UN(국제연합, United Nations) 및 기타 국제기구의 공식 통계를 바탕으로 한다.

[그림 4] 유엔개발계획(United Nations Development Programme)의 지속가능발전목표 (Sustainable Development Goals) 17가지

3. 생성형 AI의 SOTA 모델들과 생성 데이터 형태

2.2절에서 논의한 바와 같이 공개된 데이터가 많음에도 불구하고, 2.1절에서처럼 좋은 데이터는 고갈되고 있다. 이를 극복하기 위해 현재 생성형 AI로부터 데이터를 생성하고자 하는 노력이 활발하다. 이 절에서는 SOTA(State-of-the-Art)로 알려진 생성형 AI 모델들, 그리고 해당 모델로부터 생성될 수 있는 학습 데이터의 형태를 소개한다.

최근 발표된 한 보고서에선[15] [그림 5]와 같이 생성형 AI의 밸류체인, 활용되는 데이터 형태를 보여주고 있다. 생성형 AI SOTA는 텍스트, 이미지, 사운드, 동영상 등 다양한 데이터를 생성하고 있음을 알 수 있다. 생성형 AI 모델은 대용량 데이터를 수집해 파운데이션 모델을 만든다. 이 모델로부터 생성된 모달들은 서비스뿐만 아니라 학습용 데이터로도 활용 가능하다.



[그림 5] 생성형 AI의 밸류체인 및 데이터 활용

생성형 AI 중 가장 빠르게 발전하는 분야는 LLM(거대언어모델, Large Language Model)이라 불리는 텍스트 생성 AI다. 현재 오픈AI의 ChatGPT, 구글의 제미나이(Gemini), 메타의 라마(LLama), 엔트로픽(Anthropic)의 클로드(Claude), 네이버의 클로바X(CLOVA X) 등은 인간과 유사한 방식으로 텍스트를 생성할 수 있는 수준에 도달한 것으로 보인다. 이들은 높은 수준의 텍스트를 생성하기에, 최근 자연어 연구자들은 학습 데이터를 만들기 위해 LLM을 활용하고 있는 추세다. LLM으로 높은 성능에 도달한 텍스트는 특정 목적에 활용될 수 있도록 미세조정(Fine-turning)된다.

15) 류승희, 이효정, 최창환, 이종민, "창작 영역에 뛰어든 생성형 AI 투자 현황과 활용 전망," 삼정 KPMG Issue Monitor 제163호 2024.05

생성형 AI로서 LLM 다음으로 널리 알려진 것은 이미지 생성 AI이다. 텍스트 정보, 스케치, 영역 분할 정보와 같은 메타 정보를 바탕으로 이미지를 생성한다. 달리-3(DALL-E 3), 미드저니(Midjourney), 스테이블 디퓨전(Stable Diffusion), 어도비 파이어플라이(Adobe Firefly), 구글 이마젠(Google Imagen) 등이 있다. 최신 모델들은 사진과 분간하기 힘들 정도의 이미지를 생성하며, 특정 인물이나, 스케치, 특정 자세 등 조건에 따라 생성이 가능한 수준이다. 현재 이를 활용해, 제조 분야에서 결함검출을 위한 학습 데이터로 활용하려는 연구가 활발히 진행 중이다.

사운드(Sound) 생성 AI도 활발히 연구되고 있다. 이 분야는 음악 생성과 오디오 생성으로 나눌 수 있다. 음악 생성 AI는 악보, 리듬, 멜로디 등 음악적 요소를 학습해 새로운 음악을 생성할 수 있다. 이 기술은 작곡가들에게 영감을 주거나 새로운 음악적 스타일을 탐색하는 데 활용되고 있다. SOTA 기술로는 스태빌리티AI(StabilityAI)의 스테이블 오디오(Stable Audio) 서비스, Suno.ai의 수노(Suno), 구글의 리리아(Lyria), 메타의 뮤직젠(MugicGen) 등이 있다.

오디오 생성 AI는 음성인식, 음성합성 등의 분야에서 활용되며, 자연스러운 인간 목소리 데이터 생성이 가능하다. 오디오 생성에는 오픈AI의 다국어 목소리 생성 기술, 메타의 오디오젠(AudioGen), 일레븐랩스(ElevenLabs), 타입캐스트(TypeCast) 등이 있다.

비디오 생성 AI는 동영상을 생성하는 분야다. 이는 이미지 생성, 사운드 생성, 텍스트 생성 등 생성형 모델의 종합 형태라 볼 수 있다. 현재 텍스트 입력으로 이미지, 사운드를 포함한 동영상이 생성 가능한 수준이다. 비디오 생성 AI는 자율주행과 로봇 등을 위한 학습 동영상 생성에 활용되고 있다[16].

<표 1> 텍스트 생성형 모델

텍스트 생성형 모델	회사/기관	비고
GPT-4o	오픈AI	
제미나이, 젬마	구글	젬마 Open Source
라마 3	메타	Open Source
클로드 3	엔트로픽	오픈AI 직원이 퇴사 후 만든 회사
그록-1(Grok-1)	xAI	일론 머스크 설립

출처: 류내원, "생성형 AI 현황 및 전망" 주간기술동향 2127호 2024.03.27 활용 및 재구성

<표 2> 이미지 생성형 모델

이미지 생성형 모델	회사/기관	비고
달리-3	오픈AI	-
미드저니	미드저니	-
스테이블 캐스케이드 (Stable Cascade)	Stability.ai	Open Source
이마젠 2	구글	-
에뮤(Emu)	메타	-

출처: 류내원, "생성형 AI 현황 및 전망" 주간기술동향 2127호 2024.03.27 활용 및 재구성

<표 3> 비디오 생성형 모델

비디오 생성형 모델	회사/기관	비고
젠2(Gen2)	런웨이(Runway)	높은 선명도 영상
이마젠 비디오(Imagen Video)/루미에르(Lumiere)	구글	5초 분량 고화질 영상/미공개
소라(Sora)	오픈AI	1분 분량 영상
SVD(스테이블 비디오 디퓨전, Stable Video Diffusion)	스태빌리티 AI	2초 분량 고화질 영상
메이크 어 비디오(Make-A-Video)/ 에뮤 비디오(Emu Video)	메타	4초 분량 고화질 영상

출처: 류승희 외, "창작 영역에 뛰어든 생성형 AI 투자현황과 화룡 전망," 삼정 KPMG Issue Monitor 제163호 2024.05 활용 및 재구성

4. 맺음말

AI 시대에서 데이터는 국부(國富)다[17]. 얼마 전 마이크로소프트(Microsoft)가 진행한 '데이터 품질과 AI 성능의 연관성 실험'은 '데이터의 질'이 얼마나 중요한지를 보여준다. 교과서로 학습한 모델이 좀 더 많은 데이터를 사용한 거대 모델의 성능을 뛰어넘었기 때문이다[18]. 파라미터 차이가 각각 1.3억 개와 13억 개로 약 10배 차이를 보였음에도 불구하고, 교과서라는 고품질 데이터를 학습한 모델이 상식, 언어 이해 및 지식, 다단계 추론 등 다양한 영역에서 더 뛰어난 결과를 보였다.

양질의 데이터가 곧 국가경쟁력이 되는 시대다. 양질의 데이터를 생성할 수 있는 생성형 AI 모델 개발이 중요하다. 이뿐만 아니라 양질의 데이터를 확보하고, 이를 잘 보호·관리하기 위해 관련 표준을 수립하려는 노력이 절실하다.

※ 출처: TTA 저널 제213호

※ 본 연구는 2024년도 산업통상자원부 및 한국산업기술진흥원(KIAT) 연구비 지원에 의한 연구임('P0023760') [EUV반도체 생산성 향상을 위한 펠리클/마스크 결함검출용 데이터 구축 및 AI솔루션 개발]

16) Yufe Wang et al., "RoboGen: Towards Unleashing Infinite Data for Automated Robot Learning via Generative Simulation," ICML 2024
17) 김대기, "생성형AI시대…결국엔 똘똘한 데이터가 경쟁력 가른다," 매일경제 2023.12.04
18) Yuanzhi Li et al., "Textbooks Are All You Need II: phi-1.5 technical report", 2023