연구데이터 관점에서 본 거대언어모델 품질 평가 기준 제언*

A Proposal of Evaluation of Large Language Models Built Based on Research Data

한나은 (Na-eun Han)**
서수정 (Sujeong Seo)***
엄정호 (Jung-ho Um)****

초 록

본 연구는 지금까지 제안된 거대언어모델 가운데 LLaMA 및 LLaMA 기반 모델과 같이 연구데이터를 주요 사전학습데이터로 활용한 모델의 데이터 품질에 중점을 두어 현재의 평가 기준을 분석하고 연구데이터의 관점에서 품질 평가 기준을 제안하였다. 이를 위해 데이터 품질 평가 요인 중 유효성, 기능성, 신뢰성을 중심으로 품질 평가를 논의하였으며, 거대언어모델의 특성 및 한계점을 이해하기 위해 LLaMA, Alpaca, Vicuna, ChatGPT 모델을 비교하였다. 현재 광범위하게 활용되는 거대언어모델의 평가 기준을 분석하기 위해 Holistic Evaluation for Language Models를 중심으로 평가 기준을 살펴본 후 한계점을 논의하였다. 이를 바탕으로 본 연구는 연구데이터를 주요 사전학습데이터로 활용한 거대언어모델을 대상으로 한 품질 평가 기준을 제시하고 추후 개발 방향을 논의하였으며, 이는 거대언어모델의 발전 방향을 위한 지식 기반을 제공하는데 의의를 갖는다.

ABSTRACT

Large Language Models (LLMs) are becoming the major trend in the natural language processing field. These models were built based on research data, but information such as types, limitations, and risks of using research data are unknown. This research would present how to analyze and evaluate the LLMs that were built with research data: LLaMA or LLaMA base models such as Alpaca of Stanford, Vicuna of the large model systems organization, and ChatGPT from OpenAI from the perspective of research data. This quality evaluation focuses on the validity, functionality, and reliability of Data Quality Management (DQM). Furthermore, we adopted the Holistic Evaluation of Language Models (HELM) to understand its evaluation criteria and then discussed its limitations. This study presents quality evaluation criteria for LLMs using research data and future development directions.

키워드: 거대언어모델, 품질평가, 연구데이터, 데이터품질관리, 품질평가기준
Large Language Model (LLM), Quality Evaluation for LLM, Research Data Quality Management (DQM), evaluation criteria for LLM

* 본 논문은 한국과학기술정보연구원 연구사업(과제번호: K-23-L01-C03-S01 및 K-23-L03-C02-S01)의 지원에 의해 이루어진 것임.
** 한국과학기술정보연구원(KISTI) 박사 후 연구원(betterhan@kisti.re.kr) (제1저자)
*** 한국과학기술정보연구원(KISTI) 기술원(seos@kisti.re.kr) (공동저자)
**** 한국과학기술정보연구원(KISTI) 책임연구원(jhum@kisti.re.kr) (공동저자)
■ 논문접수일자 : 2023년 8월 16일 ■ 최초심사일자 : 2023년 9월 4일 ■ 게재확정일자 : 2023년 9월 18일
■ 정보관리학회지, 40(3), 77-98, 2023. http://dx.doi.org/10.3743/KOSIM.2023.40.3.077

1. 서 론

1.1 연구의 배경 및 목적

최근 ChatGPT로 대표되는 거대언어모델(Large Language Model, 이하 LLM)에 대한 관심이 집중되면서 여러 LLM을 통한 산출물들이 기하급수적으로 생산되고 있다. 다양한 LLM들은 이용자들이 원하는 결과물을 도출 받는데 있어서 강력한 편리성을 기반으로 각광받고 있다. 그러나, LLM을 통해 결과물을 도출하는 행위에 관하여 학습자의 학습 윤리에 대한 문제가 지속적으로 논의되고 있을 뿐만 아니라, 또 다른 측면에서는 해당 모델들이 제공하는 결괏값에 대한 정확성, 신뢰성, 공정성, 투명성 등에 대한 문제점이 지속적으로 제기되고 있다(안성원 외, 2023; Dwivedi et al., 2023). LLM은 기본적으로 사전학습(pre-trained)된 언어 모델로써 추가학습 혹은 정밀학습(fine-tuning)이 가능하다는 특성을 갖는데, 추가학습을 통해 LLM의 활용 분야 및 범위는 다양하게 확장될 수 있다. 하지만 이는 동시에 사전학습에 활용하는 데이터의 품질 및 신뢰성 등에 따라서 LLM이 제공하는 결괏값의 품질 및 신뢰성 등이 영향을 받을 수 있다는 것을 의미한다. 사전학습에 활용된 데이터가 정확하지 않으면 자연스레 해당 LLM이 제공하는 결괏값 역시 정확성을 확보할 수 없고, 사전학습에 활용된 데이터를 포함하여 LLM이 인용한 참고문헌이 정확하게 제시되지 않은 경우 결괏값으로 제공받은 데이터의 신뢰도 및 투명성 등을 판단할 수 없다. 그러므로 LLM이 제공하는 데이터의 품질 평가를 수행하기 위해서는 결괏값에 대한 품질 평가뿐만 아니라 활용된 사전학습데이터에 대한 출처 및 품질 평가 역시 진행되어야 할 필요성이 존재한다.

특히나 다양한 LLM 가운데에서도 LLaMA와 같이 연구자들을 주요 이용 대상으로 선정하여 학술 연구를 목적으로 개발된 모델들이 존재하는데, 해당 모델들은 연구데이터를 주요 사전학습데이터로 활용한다는 특징을 갖는다. 이와 같은 모델들은 연구자들이 다양한 연구를 수행하는 데 활용될 수 있을 것으로 기대할 수 있는데, 연구 결과물의 품질을 확보하기 위해서는 연구 과정에서 활용하는 연구데이터에 대한 정확성, 신뢰성 등에 대한 품질 검증이 필수적이다. 그러나 현재 활용되는 다양한 LLM을 대상으로 진행하는 성능 평가 및 품질 평가는 일관된 기준이 존재하지 않을 뿐 아니라 학술 연구를 목적으로 한 LLM과 같이 특정 목적 및 분야를 중심으로 개발된 모델의 품질 평가를 진행하는 데 필수적인 중요 요인들이 누락되는 경우도 빈번하다. 이에 본 연구는 여러 LLM 가운데 LLaMA 및 LLaMA 기반 모델과 같이 연구데이터를 주요 사전학습데이터로 활용한 모델이 생산하여 제공하는 데이터 품질에 중점을 두어 현재의 평가 기준을 분석하고 연구데이터 관점에서 품질 평가 기준을 제안하여 추후 개발 방향을 논의하고자 한다.

1.2 연구의 내용 및 방법

본 연구는 먼저 LLM의 특성 및 한계점을 이해하기 위해 메타(Meta)에서 개발한 LLaMA(Touvron et al., 2023), 스탠퍼드 대학의 연구팀이 개발한 오픈 소스 언어 모델인 Alpaca(Taori et al., 2023), 버클리 대학, 카네기 멜런 대학,

스탠퍼드 대학, 샌디에이고 대학의 연구원들이 개발한 Vicuna(Chiang et al., 2023), 그리고 OpenAI에서 딥러닝 기술을 사용하여 개발한 언어 모델인 ChatGPT(OpenAI, 2023) 총 4개의 모델을 비교하였으며, 현재 LLM 성능 평가 및 품질 평가의 표준화가 이루어져 있지 않은 상황에서 가장 폭넓은 언어 모델 평가 지표를 가지고 있는 Holistic Evaluation for Language Models(HELM)를 중심으로 LLM 성능 평가 및 제공하는 데이터의 품질 평가 방식을 분석하고 한계점을 논의하였다. 이를 바탕으로, 추후 개발 방향을 논의하기 위해 연구데이터를 주요 사전학습데이터로 활용한 LLM 제공 데이터의 품질 평가에 적용할 수 있는 요인을 연구데이터를 대상으로 하는 품질 평가 기준 가운데 중요성과 필요성을 기준으로 추출하였다. 즉, 연구데이터를 평가하는데 있어, 다양한 선행연구의 품질 평가 수행 시 중요 요인으로 논의한 평가 기준(중요성)과 LLM 특성 기반 여러 LLM이 생산하는 데이터 품질 평가 시 필요한 평가 기준(필요성)을 기반으로 추출하였음을 의미한다. 본 연구에서는 연구데이터의 관점에서 유효성, 기능성, 신뢰성을 학술 연구를 목적으로 개발된 LLM의 주요 품질 평가 기준으로 제시하였다.

2. 이론적 배경

2.1 거대언어모델(Large Language Model, LLM)

자연어(natural language)란 우리가 일상 생활에서 사용하는 언어를 말하고, 자연어 처리(natural language processing)란 이러한 자연어의 의미를 분석하여 처리할 수 있도록 구현하는 것을 말한다. 언어 모델(Language Model, 이하 LM)은 연속 단어(혹은 문장)에 확률을 할당하는 확률 분포를 의미하는데(Jurafsky & Martin, 2021), 기존의 모델보다 성능이 월등하게 좋아서 딥러닝 기반 자연어 처리 모델이 각광받고 있다. 특히나 그 중에서도 GPT(Generative Pre-trained Transformer) 모델이 2022년 후반부터 관심을 받고 있다(이기창, 2021).

2006년 제프리 힌튼 교수의 발표 이래 학계에서는 딥러닝 방법론이 활발하게 연구되면서 하드웨어 및 데이터 규모의 성장과 함께 연구개발이 가속화되었다. 학술 연구에서 활용되던 인공 지능 기술이 산업화된 계기는 2016년 구글 딥마인드의 알파고와 이세돌 8단의 바둑 대국으로 보는 시각이 존재하는데, 이를 계기로 매우 큰 언어 데이터를 트랜스포머(transformer) 방식으로 학습한 엘모(ELMO), GPT(GPT-1, GPT-2, GPT-3), BERT 모델들이 개발되었다. 이러한 매우 큰 언어 데이터를 학습한 모델들은 OpenAI사에서 GPT-3가 나오게 되면서 거대언어모델(LLM)로 불리게 되었다.

이와 같은 트랜스포머 모델은 다음과 같은 두 가지의 특징을 보인다(이경님, 조은경, 2023; Brown et al., 2020). 첫째로, 사전학습 모델링(pre-trained modeling)은 범용 모델(general model)을 목표로 자기 지도 학습(self-supervised learning)과 같은 레이블이 없는 혹은 데이터에 대한 평가가 없는 샘플 데이터를 가지고 학습을 진행하는 기계학습방식(Yarowsky, 1995)을 사용한 것을 의미하며, 둘째로, 해당 모델은

정답 지표를 나타내주는 레이블이 있는 적은 양의 테스트 데이터를 사용하여 정밀 학습(fine-tuning)을 범용 모델에 적용시켜 높은 성능의 분류 결과를 나타내는 특징을 갖는다. 이에 대한 예시로는 다음 세 가지를 들 수 있는데 첫째, 길고 상세한 버전의 텍스트를 짧고 이해할 수 있는 텍스트로 압축하는 텍스트 요약(summarization), 둘째, 국가명, 회사명, 사람 이름 등 이름을 가진 개체(Named Entity)를 인식하는 개체명 인식(name entity recognition), 셋째, 컴퓨터가 원문을 문장 단위로 읽어 들인 다음 해당 문장에 대응되는 최적의 번역문을 통째로 생성해 내는 과정인 기계 번역(machine translation) 등이 있다(Wilcox et al., 2020).

LLM은 단순히 데이터 크기만을 의미하는 것은 아니며, 모델 자체의 크기가 비약적으로 커진 것 역시 포함한다. LLM의 축소화 기법에 대해 논의한 Kaplan et al.(2020)에 따르면 GPT-3부터 과학 분야에 초점을 맞춰 제안된 Galactice (Taylor et al., 2020)까지 파라미터(parameters)가 1,000억개가 넘는 모델들이 소개되기 시작하였다. 700억 개의 파라미터를 가진 Chinchilla (Rae et al., 2021)는 데이터를 중요도에 따라서 각기 다른 비중으로 전달시키기 위해 사용되는 기법인 모델 가중치(我妻 幸長, 2018)가 아닌 필요에 따라 텍스트를 잘게 나누어서 의미가 있는 최소 단위인 토큰(token)(조태호, 2022) 수를 기준으로 스케일링 법칙을 재정의한 바 있는데, 여기서 스케일링 법칙이란 모델 크기 혹은 데이터 크기에 따라 성능이 향상하는 현상을 의미한다(Kaplan et al., 2020). 이는 기존의 빅데이터에 딥러닝을 적용한 것뿐만 아니라 모델의 파라미터 수를 증가시켜 모델 자체의 복잡성이나 규모를 키운 것이라고 이해할 수 있다. 파라미터 수는 모델의 고지능 혹은 고성능을 판단하는 평가 지표로 사용되는데, 이는 파라미터가 인간 뇌의 뉴런 및 뉴런 간의 연결에 해당되는 매개 변수로 여겨지기 때문이다. 이러한 파라미터 개수가 많을수록 더 많은 정보를 저장하고 처리할 수 있으므로, 파라미터의 개수가 많은 모델일수록 고지능 혹은 고성능 모델로 여겨진다(안성원 외, 2023).

GPT 및 Chinchilla 모델은 활용하는 데이터 및 해당 모델의 파라미터에 대한 정보가 제한적이기 때문에 본 연구에서는 LLaMA(Touvron et al., 2023), Alpaca(Taori et al., 2023), Vicuna (Chiang et al., 2023)를 중심으로 연구데이터의 관점에서 위의 LLM의 특성, 성능 및 품질 평가와의 연관성, 한계점 등 여러 논의 사항을 살펴봄과 동시에, 연구데이터가 LLM에서 주요 사전학습데이터로 사용됨에 있어서 무분별하게 편집되어 활용되는 상황 및 품질 평가가 제대로 이루어지지 못하는 점을 ChatGPT(OpenAI, 2023) 모델의 내용과 함께 논의하고자 한다.

2.2 연구데이터 및 품질 평가 기준

연구데이터란 일반적으로 “연구개발과제 수행 과정에서 실시하는 각종 실험, 관찰조사 및 분석 등을 통하여 산출된 사실 자료로서 연구결과의 검증에 필수적인 데이터”를 의미하며(국가연구개발정보처리기준, 과학기술정보통신부고시 제2020-102호), 연구데이터의 형태는 정형 및 비정형 데이터를 포함한다(국가과학기술연구회, 2019). 데이터는 “데이터를 활용하는 사용자의 다양한 활용 목적이나 만족도를 지속적

으로 충족시킬 수 있는 수준"을 만족하였을 때 품질이 보장되는데(English, 2009), 본 연구의 대상인 연구데이터의 품질관리는 "정형 및 비정형 데이터를 포함하는 형태의 연구데이터의 품질을 확보하고, 유지, 관리 및 개선함으로써 사용자에게 유용한 가치를 제공할 수 있도록 하는 일련의 활동"이라고 정의된다(한나은, 2023).

데이터의 품질을 관리하고 측정하기 위해 다양한 관리지표 및 평가지표 등이 여러 선행연구를 통해 제안된 바 있다. 한국데이터베이스진흥원의 DQC(Database Quality Certification)는 데이터를 정형 데이터와 비정형 데이터로 구분하여 정형 데이터는 완전성, 유일성, 유효성, 일관성, 정확성을 만족하여야 하고, 비정형 데이터 기능성, 신뢰성, 사용성, 효율성, 이식성을 만족하여야 한다고 제안하였다(한국데이터베이스진흥원, 2006). 데이터의 특성에 상관없이 안전행정부 품질관리 매뉴얼의 품질기준에서는 준비성, 완전성, 일관성, 정확성, 유효성, 보안성, 적시성, 유용성의 기준이 충족되어야 한다고 제시되었으며(안전행정부, 2014), CODATA(Committee on Data for Science and Technology)에서 작성된 연구 논문은 접근성, 정확성, 사용 방법의 적절성, 메타데이터의 적절성, 완전성, 일관성, 포괄성, 데이터 포맷의 개발성, 데이터 라이센스의 개방성, 재사용 가능성의 기준이 충족되어야 함을 주장하였다(Kindling & Strecker, 2022). 앞서 언급한 세 가지의 품질기준에서 공통적으로 중요하다고 언급된 지표는 완전성, 일관성, 정확성이며, 유효성 및 필요성은 이 중 두 가지의 품질기준에서 충족되어야 한다고 언급된 바 있다. 연구데이터의 품질기준은 연구데이터를 대상으로 한 품질 수준을 측정하기 위한 내용에 초점을 맞춘 것으로써 유효성, 완전성, 일관성, 접근성, 정확성, 고유성, 적시성, 유용성, 보안성이 만족하여야 함이 논의된 바 있다(한나은, 2023).

3. 연구데이터 기반 거대언어모델 비교 및 품질평가 기준 제안

3.1 거대언어모델 비교

본 장에서는 LLM의 특성 및 한계점을 이해하기 위해 4개의 LLM을 선정하여 해당 모델 개발 단체, 데이터의 공개 여부, 코드 공개 여부, 상용화의 여부 및 평가 방식의 공개 여부 등을 분석하였으며, 분석을 위해 선정한 LLM 모델은 LLaMA, Alpaca, Vicuna, ChatGPT로, 분석 내용은 다음 〈표 1〉과 같다.

〈표 1〉 거대언어모델 비교 표

모델명	LLaMA	Alpaca	Vicuna	ChatGPT(GPT-4)
모델 개발 단체	Meta	Stanford	LMSYS org.	OpenAI
데이터 공개 여부	○	○	○	X
코드 공개 여부	○	○	○	X
상용화 여부	Open-source	Open-source	Open-source	Closed-source
평가방식 공개 여부	○	○	○	○

메타의 LLaMA 모델은 사전 학습 과정에서 활용된 자료 중 다수가 연구데이터로 이해할 수 있는데, 전체 데이터 가운데 67%가 CommonCrawl에서 수집되었으며, 그다음으로 15%는 C4, 4.5%는 오픈 소스 소프트웨어 인터넷 호스팅을 지원하는 서비스인 Github에서 수집되었다. 그뿐만 아니라, 4.5%는 위키피디아, 4.5%는 서적, 2.5%는 ArXiv, 그리고 2%는 StackExchange로부터 수집되었음이 밝혀진 바 있는데, 이를 기반으로 연구데이터가 LLaMA의 전체 사전 학습데이터 가운데 최대 83%를 차지하는 것으로 추정할 수 있다(Touvron et al., 2023). 연구데이터로 활용된 출처 가운데 대중적으로 널리 알려진 ArXiv를 살펴보면, 서적, 위키피디아, 출판 전(preprint) 논문 아카이빙 사이트인 코넬대학교의 ArXiv는 2023년 9월 기준 현재 물리, 수학, 컴퓨터과학, 계량생물, 계량재무, 통계, 전기공학, 시스템과학, 경제학 분야에 해당하는 2,322,119건의 오픈 액세스(open access) 논문을 보유하고 있고, 이는 대표적인 연구데이터라고 볼 수 있다. 이를 기준으로 ArXiv, github, CommonCrawl에 담긴 연구 자료들은 연구데이터로 인정될 수 있다. 그뿐만 아니라, Github에는 연구 과정에서 만들어진 많은 코딩 자료 및 연구 과정에 대한 설명이 담긴 자료들이 저장되어 있다. 또한 CommonCrwal은 웹 기반 데이터이지만 저작권법에 영향을 받은 많은 데이터를 포함하기 때문에 미국에서는 저작권이 있는 작업이 포함되어 있다는 표기가 필수적이며, 유럽과 같은 다른 여러 국가에서는 저작권법을 피하고자 문장 섞기나 일반적인 크롤링 데이터 섞기 등을 하는 경우가 빈번하다(Schäfer, 2016). 따라서 CommonCrawl의 모든 데이터가 연구데이터라고 단정할 수는 없지만 저작권을 가진 연구데이터를 포함하고 있음을 추정할 수 있다. 미국 행정 관리 예산국(Office of Management and Budget, 이하 OMB)의 자료에 따르면, 과학 커뮤니티에서 연구 발견을 검증하는데 필요하다고 인정되는 '기록된 사실 자료(recorded factual material)'를 연구데이터로 정의하고 있다. 연구데이터에 대한 정의는 다양하게 존재하는데, 앞선 2.2장에서 제시한 내용에 국한되지 않고, 퀸스랜드 대학에서는 사실(facts), 관찰(observations), 이미지(images), 컴퓨터 프로그램 결과(computer program results), 음성(recordings), 측정 지표(measurements), 경험(experiences)까지도 연구데이터로 정의하고 있으며, 멜본 대학에서는 사실(facts), 관찰(observations or experiences), 출력물 혹은 컴퓨터 파일로 된 기초연구자료, 질문 및 응답, 비디오 혹은 오디오 테이프, 실험 결과 같은 실험 노트를 모두 연구데이터로 정의한다(김선태, 이정훈, 정한민, 2017). LLaMA 모델은 2023년 2월에 오픈 소스 모델로 공개되면서 연구 개발 목적으로만 활용할 수 있도록 상업적 이용을 제한하고 연구자들로 이용 대상을 한정하였다. LLaMA 모델은 문장 생성 및 요약, 번역, 질의응답 등 다양한 자연어 처리 작업에 사용될 수 있으며, 문장 단위의 처리에 집중할 수 있도록 개발되었다. 특히 해당 모델은 대화형 모델과는 다르게 지시와 명령에 대한 이해 및 특정 작업 수행에 강점을 가진다는 특징을 갖는다(Touvron et al., 2023).

이후, LLaMA 모델을 활용하여 스탠퍼드 대학에서는 GPT-3.5가 만들어낸 5만 2천 개의 지시(instruction-following) 샘플 데이터인 다빈

치(davinci) 파일을 이용하여 자기 교수법(self-instruction) 기반의 정밀 학습(fine-tuning)을 진행하여 Alpaca 모델을 만들어 배포하였다(Peng et al., 2023). Alpaca 모델 역시 학술 연구에 한정하여 사용할 수 있도록 배포되고 있는데, LLaMA가 상업적 기업인 Meta에서 개발되었다면, Alpaca 모델은 LLaMA 모델과 비슷한 성격을 띄고 있지만 연구를 직접 수행하는 대학의 연구자들이 개발했다는 측면에서 차이를 갖는다.

Large Model Systems Organization(LMSYS)사의 Vicuna 모델의 경우는 LLaMA 모델에 ShareGPT 혹은 Share user-ChatGPT(OpenAI, 2023)로 불리는 7만 개의 대화형 샘플 데이터를 활용하여 정밀 학습을 진행시킨 모델을 생성하였다. Vicuna 모델은 LLaMA 모델과 Alpaca 모델을 활용하여 개발된 모델이라고 할 수 있으나, 앞선 두 모델과는 다르게 대화형 모델 혹은 어플리케이션으로만 기능하게끔 개발되었다는 것이 차이점이라고 할 수 있다.

결과적으로 GPT와 LLaMA를 제외한 두 모델은 LLaMA를 범용 모델 기반으로 활용하여 정밀 학습을 진행한 모델이라고 할 수 있다. 본 연구에서는 연구데이터의 관점에서 학술 연구를 목적으로 기업에서 개발된 LLaMA와 해당 모델을 발전시켜 대학의 연구자들이 개발한 Alpaca 모델, 그리고 LLaMA와 Alpaca를 기반으로 대화형 모델로 개발된 Vicuna 모델을 연구데이터를 주요 사전학습데이터로 활용한 모델로 선정하여 비교하였으며, 해당 LLM이 갖는 특징을 파악하기 위해 가장 대중적으로 널리 알려진 LLM 중 하나인 ChatGPT를 함께 비교함으로써 이해를 도울 수 있도록 하였다.

ChatGPT의 경우는 2022년 11월, 일반에 공개되었을 때 GPT-3.5 모델을 기반으로 작동하였는데, 이는 문장을 완성할 때 단어를 순차적으로 '출력'하는 것이 아닌 출력된 단어들을 '재입력'한 후 다음 단어를 출력하는 방식인 자동회귀 방식을 추가하여 570기가바이트 분량의 학습 데이터를 학습시킨 모델이라고 할 수 있다. 이는 GPT-3 모델(Brown et al., 2020)을 기준으로 학습 데이터를 설명한 것으로써 현재까지 GPT-3.5와 GPT-4 모델의 학습 데이터 정보는 공개된 것이 없으며, 이전 모델보다 좀 더 최신의 자료를 사용했다라고만 언급된 바 있다(OpenAI, 2023). Brown et al.(2020)에 따르면 정제된 ChatGPT에서 활용하는 자료는 CommonCrawl이라는 웹 데이터가 전체 데이터셋의 60%를 차지하고 있는 것으로 나타난다.

각각의 모델들은 해당 모델의 성능을 검증하기 위해 각자 성능 평가를 시행하여 결과를 공개하고 있는데, 일관된 기준으로 평가한 것은 아니며 각기 다른 평가 방식을 통해 공통되지 않은 영역을 평가하였다는 한계를 갖는다. 먼저, LLaMA는 여러 작업을 진행할 수 있는지에 대한 평가를 진행하였는데, 평가 작업에 관한 항목은 상식 추론(common sense reasoning), 문맥적 의미가 제시되지 않은 질의 응답(closed-book question answering), 독해(reading comprehension), 수학적 추론(mathematical reasoning), 코드 생성(code generation), 거대 다중작업 언어 이해(massive multi-task language understanding)으로 구성되어 있다. 다음으로 Alpaca는 5명의 사람이 직접 실시한 평가를 바탕으로 하고 있으며,

이메일 작성, 소셜미디어 및 도구 활용 등 이용자 중심의 다양한 항목을 중심으로 평가를 진행하였다고 밝혔다(Taori et al., 2023). Vicuna는 해당 모델을 통해 생성된 답변 결과를 ChatGPT에 제공하여 GPT가 평가할 수 있도록 하는 방식을 취하였으며(Chiang et al., 2023), GPT-4 모델은 언어 모델 품질 평가 방법 중 하나인 truthfulQA (Lin, Hilton, & Evans, 2021)의 내용을 참고하여 총 9가지 영역(learning, technology, writing, history, math, science, recommendation, code, business)에 대해 일반적으로 사람에게는 정상적으로 보이지만 기계 학습(machine learning) 모델에는 오 분류(wrong classification)를 유발하도록 설계된 특수하게 조작된 입력값의 질문을 진행함으로써 평가하였다(OpenAI, 2023).

하지만 이러한 평가들이 곧 해당 LLM을 활용하여 도출 및 출력한 값의 품질을 평가한 것은 아니라는 한계를 갖는다. 위의 내용들은 여러 기준을 활용하여 평가를 시행하였다는 사실 혹은 다른 LLM과 비교할 때 어떠한 점이 더 나은지에 대해서 강조하는 것에 초점을 두고 있을 뿐, 해당 LLM을 활용하여 출력한 답변의 신뢰성 또는 정확성에 대해서는 평가하지 않는다. 또한 LLaMA(Taori et al., 2023)와 GPT 모델(Brown et al., 2020)의 경우 학습 데이터를 필요에 따라서 임의대로 잘라내어 사용하였다는 내용을 담고 있으므로 데이터 품질을 보장하기 어렵다는 한계를 갖는다. 또한, Alpaca와 Vicuna 모델의 경우에는 정밀 학습의 데이터를 제외하고, 개발의 기본으로 활용한 기본 모델인 LLaMA가 학습 데이터의 품질을 보장할 수 없는 상황이기 때문에, 역시 같은 한계를 가진 것으로 추정한다.

3.2 Holistic Evaluation for Language Models(HELM)

미국 스탠퍼드 인간중심 인공지능 연구소(Stanford Institute for Human-Centered Artificial Intelligence, HAI)와 기초 모델 연구센터(Center for Research on Foundation Models, CRFM)는 영어 모델을 중심으로 거대 언어 모델들을 전체론적으로 평가하는 방안(Holistic Evaluation for Language Models, 이하 HELM)을 소개하였으며(Percy et al., 2022), 2023년 여름을 기준으로 61개의 모델과 각 모델 생성 목적에 따라서 바뀌는 표적 시나리오 42가지를 검증하였다. HELM은 세 가지의 요소를 가지고 있는데, 이는 다음과 같다.

(1) 광범위한 적용 범위와 불완전성 인식(Broad coverage and recognition of incompleteness): 언어 모델의 광범위한 능력과 위험을 고려할 때, 우리는 광범위한 시나리오에 걸쳐 LLM을 평가할 필요가 있다. 평가를 넓히는 것은 소규모 데이터 세트 모음에서 대규모 데이터 세트 모음에 이르기까지 자연어 처리 커뮤니티에서 지속적인 추세였다. 그러나 모든 시나리오 및 LM과 관련된 전체의 요구 사항을 고려할 수는 없으므로 전체 평가는 하향식 분류법을 제공하고 누락된 모든 주요 시나리오와 계량적으로 평가하는 방식인 메트릭을 명시해야 한다.

(2) 다중 메트릭 측정(Multi-metric measurement): 사회적으로 유익한 시스템은 정확성뿐만 아니라 많은 값을 반영한다. 전인적 평가는 고려된 각 시나리오에 대한 모든 면모를 평가하

는 이러한 여러 가지의 평가 지표들의 모음을 보여주는 복수의 요구 사항서를 보여줄 수 있어야 한다.

(3) 표준화(Standardization): 평가 대상은 언어 모델이지 시나리오별 시스템이 아니기 때문에 서로 다른 LLM을 의미 있게 비교하기 위해서는 표준화된 내용을 적용할 수 있는 전략이 필요하다. 또한 각 LLM은 가능한 동일한 시나리오에서 평가되어야 한다.

HELM은 모델의 기능을 평가하는 지표를 제시하고 표준화하는 것을 목적으로 하고 있지만 연구데이터가 제대로 활용되었는지에 대한 평가는 진행하고 있지 않다. HELM의 평가 요소에는 수학 문제 해결 능력, 언어(이해), 최소 대립쌍, 지식, 추론(수학을 제외한), 기억력 및 저작권, 허위 정보 등이 포함된다. 이러한 평가 요소의 구성 이유는 대부분의 LLM이 생성형 모델이라는 특성에 기반하기 때문이다. LLM의 바탕이 되는 트랜스포머(transformer) 모델은 실제 인간과 유사한 문장을 생성하기 위한 모델로 서로 떨어져 있는 데이터나 단어들의 의미가 문맥에 따라 미묘하게 달라지는 점을 반영하여 구현하고, 주어진 문장 다음에 올 단어를 예측하는 것에 그치지 않고 단어 사이에 올 수 있는 단어도 예측하기 위한 양방향(bidirectional) 예측 모델로 설계되었다. 이러한 특성 중 문장을 생성하는 능력에 초점을 맞춘(decoder) 모델이 LLM이다. 특히 GPT 모델은 사람이 이해하는 순서대로 단어 배열을 계산하는 모델이며, 기계 번역 능력이 디코더 모델과 결합하면 압축된 정보를 받아 요구되는 작업을 수행하는 능력이 한층 발전되어 한글에서 영어 혹은 반대로 영어에서 한글로 번역하는 능력뿐만 아니라 특정 데이터를 가지고 결과를 예측하는 것과 같은 시퀀스(sequence) 전환에 해당되는 작업 수행이 가능하게 된다(안성원 외, 2023). 예를 들면 과거에 축적한 데이터를 기반으로 날씨를 예측하거나 빈칸을 추론하는 등의 데이터를 LLM이 스스로 생성하여 사람을 보조하는 수준으로 머물던 기존의 인공지능 모델을 뛰어넘었다고 할 수 있다.

그러나 이처럼 LLM이 인간처럼 언어를 이해한다고 해서 문제가 없는 것은 아니다. 사용된 학습 데이터의 품질에 따라서 편향성과 유해성이라는 부정적인 면모를 보여주기도 하며, 이를 완화하기 위해서 HELM에서는 여러 내용을 점검할 수 있는 시나리오를 확립할 것을 권고하고 있다. HELM은 LLM의 기능을 평가하기 위해 다양한 지표를 제시하고 있는데, 그 중 주요 내용을 간략하게 정리하여 제시하면 다음 〈표 2〉와 같다.

〈표 2〉 HELM 주요 평가 지표

주요 평가 지표	세부 평가 지표
수학 문제 해결 능력	
언어(이해)	언어 모델링(language modeling)
	최소 대립쌍(minimal pairs)
지식	질문-답변(question-answering)
	문서 완성(text completion)
추론 (수학을 제외한)	기본 요소(reasoning primitives)
	현실적 추론(realistic reasoning)
기억력 및 저작권	
허위 정보	이야기 반복(narrative reiteration)
	이야기 삽입(narrative wedging)

먼저 언어(이해)는 다음 언어를 얼마나 잘 예측하는가에 대한 평가를 의미한다. 영어를 기반

으로 하는 다양한 LLM을 대상으로 평가가 진행되었는데, 언어모델링(language modeling)과 최소 대립쌍(minimal pairs)을 중심으로 평가 시나리오가 작성되었다. 두 가지의 접근법 모두 언어학에서 광범위하게 연구된 부분인데, 언어 모델링은 특히 심리 언어학 분야에서 다양하게 연구된 바 있다(Hale, 2001; Levy, 2008). 다음 언어 혹은 단어를 예측하는 성능은 모델 또는 인간이 특정한 영역에서 사용하는 언어 분포를 얼마나 잘 학습했는가를 보여주는 것인데, 최소 쌍 비교는 특히나 특정 언어 현상에 관하여 세분화하여 이해를 확장하도록 하는 방법이다(Chomsky, 1957). 이러한 언어모델링과 최소 쌍 비교는 다른 해답의 관점에서 언어를 이해하고 문장 간의 단어 구성이나 해당 언어의 특성을 공유하는 방법이라고 할 수 있다.

둘째로, HELM은 모델의 지식을 측정하기 위해 질문-답변(question answering)과 문서 완성(text completion)을 바탕으로 평가를 진행한다. 질문-답변은 학교에서 진행하는 시험과 같은 가벼운 언어적 이해와 추론이 필요한 지식을 평가하기 위해 개발된 질문을 재사용하기도 하며, 문서 완성을 통해서는 특정한 사실 기반의 지식을 분리하여 평가할 수 있도록 한다. 특히나 질문-답변은 인터넷 검색, 챗봇(chatbot) 등 다양한 어플리케이션의 기반이 되는 자연어 처리의 기본 작업이라고 할 수 있다. LLM의 지식을 평가하는 데 주로 사용되는 데이터는 위키팩트(WikiFact)로, 위키팩트는 LLM의 사실적인 지식을 평가 및 확인할 수 있도록 하는 위키피디아(Wikipedia) 기반의 사실 완성 시나리오를 의미한다(Petroni et al., 2019; Yasunaga et al., 2022). 이때, 입력은 부분 문장으로 하고, 출력은 연속적인 문장이 되도록 한다. HELM은 인문학, 사회과학, 그리고 기타 일반적인 사실을 다루는 12개의 도메인을 식별하였다. 여기에서 말하는 추론은 학습 데이터를 기반으로 언어적 관점에서의 추론 능력을 의미한다. 이러한 부분을 평가한 방식 중 하나로 GPT-4 모델이 미국 로스쿨 입학시험(LSAT) 등 여러 전문 시험을 치렀을 때 상위 10%에 포함될 정도로 긴 글을 완성하는 능력을 갖춘 것을 확인하였다. 또한, 미국의 의사면허시험(USMLE) 문제들에서 90% 이상의 정답률을 보이며 합격한 사례도 존재한다(Lee, Goldberg, & Kohane, 2023).

셋째로, 추론과 관련해서 HELM은 추론의 기본 요소(reasoning primitives)와 현실적 추론(realistic reasoning)으로 나누어 평가를 진행하였다. 추론의 기본 요소에는 비 증폭 추론, 증폭 추론, 재귀적 계층 구조, 상태 추적 등이 포함되는데, 이와 같은 작업을 실시함으로써 언어와 지식에서 추론을 분리하여 평가한다. 그 다음으로, 이와 같은 기본 요소를 기반으로 수학적 추론, 법적 추론, 논리적 추론, 데이터 추론 등과 같은 현실적인 맥락에서 모델을 테스트함으로써 기본 요소들을 통합적으로 평가할 수 있다. 이렇게 두 가지의 다른 접근 방식을 종합하면 모델이 추론을 수행하는데 필요한 주요 역량을 어느 정도 보유하고 있는지, 그리고 추론에 의존해야 하는 실제 사용 사례에서 모델이 어떻게 활용될 수 있는지에 대한 평가가 가능하다.

넷째는 기억력과 저작권에 관련된 부분이다. 저작권을 갖는 자료를 활용하여 모델을 훈련하고 결과물을 생성하는 것에 대해서 공정사용(fair use)과 같은 원칙에 따라 법적 허용이 가능한 상황들은 이전 연구에서 논의된 바 있다

(Lemley & Casey, 2020). HELM은 다양한 모델의 축어적 콘텐츠 생성 능력을 조사하기 위한 실험을 진행하였는데, 높은 수준에서 모델과 일반적으로 저작권이 있는 자료의 서론이나 첫머리 일부분인 프롬프트(prompt)가 주어진 상황에서 모델이 완성된 내용을 얼마나 재현할 수 있는지를 측정하는 것이다. 구체적으로 세 가지의 출처로부터 프롬프트를 수집하였는데, 이는 북스코퍼스(Book's corpus)에서 무작위로 추출한 저작권이 있는 1,000권의 책, 북스코퍼스에서 베스트셀러 목록에 포함된 저작권이 있는 20권의 책, 그리고 리눅스 커널 소스(Linux kernel source)에서 무작위로 추출한 GPL(General Public License)가 있는 2,000개의 함수이다. 1,000권의 책의 경우는 무작위로 추출한 문단 시작 부분에서 다양한 수의 토큰을 프롬프트로 사용하고 있음이 확인되었고, 20권의 책에서는 첫 문단만 고려한 것으로 나타났다. 마지막으로 2,000개의 함수를 프롬프트로 활용한 경우에는 각 함수의 맨 위부터 시작되는 줄의 수를 다양하게 하여 프롬프트를 구성하였다. 책과 소스 코드와 같은 한정적인 콘텐츠만 추출하는 것에 초점을 맞추었기 때문에 다른 출처로부터 내용을 추출할 때의 정성적 동작을 충분히 반영하지 못한다는 한계를 지닌다. Meta가 개발한 LLaMA 모델의 경우 위와 같은 특성을 보이고 있는데 단순한 프롬프트를 넣어줬을 때, 학습 데이터에서 추출한 관련 정보를 바탕으로 문장을 만들어내는 것이 가능하다. 추출한 문장을 살펴보면 여러 논문이나 책의 인용구들을 볼 수 있는데 데이터의 원출처는 알 수 없지만 저작권이 있는 자료들에서 프롬프트의 지시에 맞게 연관 인용구와 참고 문헌을 뽑아내는 특성을 보여줌으로써 해당 부분이 학습 데이터에 포함된 연구데이터의 양과 품질에 직접적으로 연관성을 갖는다고 이해할 수 있다.

다음으로 HELM은 허위 정보에 대한 내용을 평가하였는데, 허위 정보는 남을 속이거나 호도하여 대상의 행동에 영향을 미치려는 의도로 유포되는 거짓 정보를 의미한다. LLM의 생성 능력이 점점 향상되고 있는 상황에서 허위 정보에 악의적으로 활용되는 것은 LLM의 잠재적인 위험으로 여겨진다. 특히 이미 여러 연구를 통해 LLM이 허위 정보를 생성하기 위해 효율적이면서 동시에 효과적인 수단이 될 수 있다는 내용이 확인된 바 있다(Buchanan et al., 2021; Radford et al., 2019). 모델 생성에 있어서 반복적인 편집 및 입력의 수정, 그리고 시스템을 정밀하게 학습시킬 수 있는 편집자와 팀을 구성하면 언어 모델은 인간이나 모델이 스스로 관리할 수 있는 것 이상의 성능을 가진 결과물을 생성할 수도 있다(Buchanan et al., 2021). 허위 정보를 평가하는 것은 모델이 생성해 내는 콘텐츠의 신뢰성과 연결될 수 있는데, HELM은 이야기 반복(narrative reiteration)과 이야기 삽입(narrative wedging)을 기준으로 허위 정보와 관련된 내용을 평가하였다. 이 두 가지의 기준은 허위 정보와 관련되어 연구자들이 우려하는 환경과 가장 밀접하게 연관된 내용이라고 할 수 있다(Pennycook et al., 2021). 이야기 반복은 LLM이 특정 이야기를 발전시키는 능력을 평가하는 것인데, 이는 주어진 논제를 뒷받침하는 제목을 생성하도록 모델의 성능을 조정하여 진행하였다. 이와 같은 흐름에서 이야기 반복은 의역 능력 및 LM 생성의 제

어 가능성과 관련된 핵심 기능을 재사용한다. 반면, 이야기 삽입은 LLM이 인종이나 종교와 같은 집단의 정체성에 따라 사람들을 구분하는 메시지를 생성해 내는 능력을 평가한다. 이와 같은 이야기의 삽입은 개인이 사회 집단에 소속됨으로써 느끼는 감정에 압력을 가함으로써 공동체를 분열하게 만들고 파편화하거나 양극화할 수 있다. 이러한 허위 정보는 LLM의 최대 약점으로 꼽히는 할루시네이션(hallucination)과도 연결점을 갖는다. 그럴듯한 정보를 제공하면서 실제로는 존재하지 않는 내용을 제시하는 것을 할루시네이션이라고 하는데 구글이 처음 Bard 모델을 시연할 때 역사적 사실을 틀리게 발표한 것과 연관시킬 수 있다. 이러한 특성은 이야기 삽입으로 인해서 허위 정보를 생성해 내는 LLM의 한계점과 연결이 된다.

HELM에서는 전체적인 모델 평가를 실시하기 위한 여러 방면의 시나리오와 평가 방식을 제시하였지만, 이 역시 표준화되지 않은 요인들이 존재하는 한계점을 가지고 있다. 연구데이터를 기반으로 개발된 모델들인 LLaMA, Alpaca, Vicuna의 경우 HELM에서 평가 요인으로 활용하는 기억력, 저작권, 질문-답변, 언어 이해의 관점에서 좋은 성능을 보여준 바 있다. 해당 평가 내용에 따르면, Alpaca의 경우 이메일이나 소셜 미디어 글 작성 등 90개 항목에서 GPT보다 성능이 앞선 것으로 나타났다. LLaMA의 경우에는 다중 처리능력에서 월등한 특성을 보이고 있지만 상식 문제를 평가하는 부분에서 Chinchilla 모델보다 낮은 성능이 보였는데, 이는 책과 논문 데이터가 Chinchilla 모델보다 적은 것이 이유로 분석되었다(Touvron et al., 2023). 위의 내용들을 종합해볼 때, LLM의 모델 특성과 성능을 평가할 때 활용된 학습 데이터의 절대적인 양과 학습 데이터에서 연구데이터가 차지하는 비중의 영향을 많이 받는다는 것을 알 수 있다.

3.3 연구데이터 기반 거대언어모델 품질 평가 기준

이론적 배경에서 논의한 바와 같이 데이터의 품질을 관리하고 측정하기 위해 다양한 관리지표 및 평가 지표 등이 여러 선행연구를 통해 제안된 바 있다. 일반적으로 완전성, 일관성, 정확성, 유효성, 사용성, 기능성, 신뢰성, 효율성, 유일성, 보안성, 적시성, 유용성, 재사용 가능성 등이 여기에 포함된다. 본 연구에서는 해당 품질 평가 지표 가운데 '유효성', '기능성', '신뢰성'을 연구데이터 기반 LLM의 품질 평가 기준으로 도출하였는데, 해당 평가 기준의 내용은 다음 〈표 3〉과 같다.

〈표 3〉 연구데이터 기반 LLM 품질 평가 기준

평가 기준	내용
유효성(Validity)	제공되는 데이터가 사실의 왜곡 없이 정확한 값을 일관적으로 보여주고 있는지를 나타내는 개념(정확성 및 일관성을 포함) (ISO 8000)
기능성(Functionality)	특정 조건에서 사용할 때 명시된 요구와 내재된 요구를 만족할만한 기능을 제공하는지를 나타내는 개념(적절성 및 기능순응성을 포함) (ISO/IEC 25000)
신뢰성(Reliability)	규정된 신뢰 수준을 유지하며, 이용자로 하여금 오류를 방지할 수 있도록 일정 수준 이상의 신뢰도를 유지하는지를 나타내는 개념(한국데이터베이스진흥원 DQC)

유효성은 정확성(accuracy) 및 일관성(consistency)를 포함하는 개념인데, 정확성은 실제에 존재하는 개념, 사물, 사건 등의 객체가 오류 없이 나타나고 있는가를 나타내는 것으로 정확성을 평가하는 기준은 구문 데이터의 정확성, 의미 데이터의 정확성, 데이터셋의 부정확성의 위험, 그리고 데이터 범위의 정확성을 포함한다(ISO/IEC 25024). 일관성은 데이터가 실세계에서 갖는 사실과 동일한 값을 나타내고 있는지를 의미하는 것으로, 일관성의 평가 기준은 참조 무결성, 데이터 불일치의 위험, 의미론적인 일관성을 포함한다(ISO/IEC 25024).

다음으로 기능성은 적절성(appropriateness)와 기능순응성(functional adapability)를 포함하는 내용인데, 적절성은 제공되는 내용이 목적에 적합하고 적절한 내용인가를 의미하며, 기능순응성은 해당 모델이 가지고 있는 구축 표준과 명령 규칙에 맞게 기능하는가를 의미한다.

끝으로 신뢰성은 규정된 신뢰 수준을 유지하고, 이용자에게 오류를 방지할 수 있도록 일정 수준 이상의 신뢰도를 유지하는가를 평가하는 기준이다.

본 연구에서 연구데이터의 품질을 평가하는 여러 기준 가운데 '유효성', '기능성', '신뢰성'을 추출하여 지표로 삼은 이유는 다음과 같다.

첫째, LLM이 제공하는 데이터의 특성에 기인한다. 일반적으로 연구데이터는 정형 및 비정형 데이터를 모두 포함하지만, LLM에서 제공하는 데이터는 설명적인(narrative) 성격을 띠기 때문에 수치, 동영상, 텍스트, 이미지 등을 모두 포함하는 일반적인 데이터를 평가하는 기준을 적용하는 것에는 한계가 존재하며 평가가 불가능한 때도 있다.

다음으로 해당 평가 기준은 중요도를 기준으로 선정되었다. LLM에서 제공하는 데이터가 일반적으로 오픈 액세스 기반의 다양한 온라인 소스 데이터를 바탕으로 조합되어 제공되는 산출물임을 고려할 때, 해당 산출물이 얼마나 정확한 사실을 나타내고 있는지, 이용자의 요구에 얼마나 기능적으로 적합하고 적절한 답을 제공하는지, 또한 해당 내용이 얼마나 신뢰할 수 있는지에 대한 부분은 추후 제공받은 데이터를 활용하면서 중요한 부분이 될 수 있다. 특히나 유효성, 기능성, 신뢰성과 같은 품질관리 지표는 이미 다양한 선행연구에서 데이터 품질 평가의 기준으로 언급된 바 있으며, 특히나 김형섭(2020)은 데이터 품질관리의 중요성을 논의하면서 정확성, 일관성의 중요성을 가장 높은 순위로 주장한 바 있는데, 정확성과 일관성은 모두 유효성을 평가하는 기준에 포함되는 내용이다. 그뿐만 아니라 앞선 이론적 배경에서 언급한 바와 같이 여러 품질기준에서 공통으로 완전성, 일관성, 정확성, 유효성에 대한 중요성을 강조한 바 있다.

셋째로, 현재 LLM이 생산하고 산출하는 데이터를 평가하는 데 활용되고 있는 기준에서 미흡한 부분을 보완할 수 있는 평가 요인을 선정하였다. 앞서 논의한 바와 같이 HELM과 같은 평가 방법에서는 해당 내용들이 정확하게 평가되고 있지 않으며, 이미 선행 논의에서 LLM이 제공하는 데이터의 유효성 및 신뢰성 등에 대한 비판이 제기된 바 있다(Dwivedi et al., 2023).

그러나, 해당 평가 요인만으로는 LLM이 생산하고 산출하는 데이터의 정확한 품질 평가는 어려우며, 추출한 평가 요인들은 일반적인 데이터를 평가하는 기준으로부터 추출한 것이기 때

문에 계속해서 정밀 학습을 통해 정제되는 LLM이 제공하는 데이터의 특성에 완전하게 부합하는 기준이라고 논의하기는 어렵다. 그러므로 추가로 평가해야 할 요인들이 남아있다는 한계점을 갖는다.

3.4 연구데이터 기반 거대언어모델 품질 평가 논의

LLM은 지시에 대한 응답(instruction/answering or question answering), 정보 재창출(information retrieval), 요약(summarization), 감성 분석(sentiment analysis), 언어(번역과 독해), 추론(reasoning, 수학이나 상식) 등과 같은 도구성 및 기능을 갖는다.

연구데이터 기반 LLM으로 알려진 LLaMA는 상식 추론(common sense reasoning), 문맥적 의미가 제시되지 않은 질의 응답(closed-book question answering), 코드 생성(code generation), 거대 다중작업 언어 이해(massive multi-task language understanding)를 바탕으로 자체 평가를 실시하였다(Touvron et al., 2023).

데이터 활용의 측면에서 품질 평가는 저작권이 침해되지 않는 자료를 사용하였는지에 대한 검증이 필요하다. 많은 연구데이터들이 웹 텍스트(web text)라는 이름으로 무분별하게 추출되어 사용되는 경우가 많이 발생한다(박성호, 2020; 박현경, 2020; 이수현, 전상홍, 2023). 제공되는 데이터가 정확하면서 일부분이 아닌 전체의 정보를 가지고 있으며(validity), 신뢰할 수 있는 정보를 가지고 있는 자료(reliability)라는 것을 입증해야 할 필요가 있다. 메타의 LLaMA 모델의 경우에는 일관성을 확보하기 위해 과학 분야 아카이브 연구데이터인 ArXiv 데이터를 사용하면서 참고 문헌과 초록을 제거하였다(Touvron et al., 2023; Lewkowycz et al., 2022). 또한, GPT-4 모델의 경우는 다양한 외국어를 이해하기 위해 큰 토큰 값을 사용하였고 CommonCrawl 데이터 역시 임의로 정제해서 사용하였다고 밝혔다. 이처럼 대량의 정보를 이용하면서 발생할 수 있는 문제는 개인 정보의 문제, 저작물에 대한 저작권 침해, 불량(toxic)한 내용을 포함하고 있는 정보들이 모델에 끼치는 영향 등이 있다(Gehman et al., 2020). 이는 데이터의 유효성과 연결되는데 첫째로 데이터의 양이 증가할수록 개인 정보 또는 저작권을 침해할 위험이 있는 정보를 모두 제거할 수 없는 경우가 발생할 수 있고, 둘째로, 자료에 대한 정확성을 떨어뜨리는 결과를 낳을 수 있다. 또한, GPT 모델과 같이 데이터의 공정 과정을 밝히지 않거나 web-text, books1, books2 같은 모호한 용어를 사용하여 데이터에 대한 정보 공개를 차단하여 명확하게 공개하지 않을 때는 해당 데이터에 대한 신뢰성 확보를 어렵게 한다.

LLM의 특성상 훈련 데이터로 입력된 정보는 출력값으로 나오는 생성 언어에 지대한 영향을 미친다. Touvron et al.(2023)은 LLaMA 모델이 책과 과학 논문에서 데이터를 추출하면서 추론과 폭넓은 언어 이해를 위해 데이터 공정 작업을 진행한다고 밝혔다. 따라서 기능적 평가(functionality)는 연구데이터가 가장 영향을 끼칠 것으로 예상되는 상식 추론, 수학적 추론과 같은 추론 부분 및 질의응답 부분에 대한 평가를 중심으로 이루어져야 한다.

본 연구에서는 이미 앞선 장에서 여러 논의를 통해 연구데이터 품질 평가 기준 가운데 연

구데이터 기반 LLM 평가에 중요한 요인으로 활용될 수 있는 요인으로 유효성, 신뢰성, 기능성을 추출한 바 있다. 유효성은 정확성과 일관성을 포함하는 개념이며, 신뢰성은 특정 신뢰 수준을 유지하고 이용자의 오류 방지를 가능케 하는 것을 의미한다. 기능성은 적절성 및 기능 순응성을 포함함과 동시에 평가의 대상이 설명적 언어 도출을 기반으로 하는 언어적 모델임과 동시에 연구데이터를 기반으로 하고 있어서 추론 및 언어 이해에 대한 기능을 포함하도록 하였다. 해당 기준을 중심으로 본 연구에서 분석의 대상으로 삼은 4개의 LLM이 생산 및 산출하는 데이터의 품질 평가에 대한 논의를 진행하였는데, 이를 표로 정리하면 다음 〈표 4〉와 같다.

이 장에서는 본 연구에서 분석한 4개의 LLM이 활용하는 사전학습데이터의 특성을 기반으로 해당 LLM이 생산 및 도출하는 결과 데이터의 품질을 어떻게 평가할 수 있는지에 대해 논의한다.

먼저, 유효성은 정확성 및 일관성을 포함하는 개념인데, LLaMA 및 LLaMA기반 모델인 Alpaca와 Vicuna는 CommonCrawl이라는 웹 데이터가 수집 데이터에서 차지하는 비율이 가장 높을 뿐 아니라 연구데이터를 포함하여 수집 데이터를 임의로 편집하여 사용하기 때문에 해당 데이터가 왜곡 없이 정확한 값을 일정하게 도출하고 있는가에 대한 평가를 만족하기 어렵다. ChatGPT 모델의 경우는 다른 모델들과 다르게 활용하는 데이터에 대한 정보 자체를 공개하고 있지 않기 때문에 결괏값에 대한 정확성 및 일관성을 나타내는 유효성을 평가할 수 없다. 활용하는 사전학습데이터의 형태를 기반으로, 4개의 LLM을 통해 산출된 데이터의 유효성을 평가하기 어렵다는 결론이 도출되는데, 이는 해당 모델들의 성능을 기반으로 도출된 내용이라기보다는 사전학습데이터 자체에 대한 정보가 거의 없거나, 해당 사전학습데이터 자체를 편집하여 사용함으로써 왜곡이 발생할 위험으로부터 기반한다고 할 수 있다.

다음으로, 본 연구의 논의 대상인 4개의 LLM에서 산출되는 데이터의 신뢰성은 해당 LLM이 참고 자료를 얼마나 정확하게 명시하고 있는가를 기준으로 논의하고자 한다. 신뢰성과 관련해서는 LLaMA 및 LLaMA 기반의 모델들을 활용하는 데이터를 명확하게 제공하고 있으며, 참고 자료에 대한 비율까지 정확하게 제시하고 있다. LLaMA를 개발한 메타에서 활용 데이터에 대한 근거 자료 및 비율을 제공한

〈표 4〉 연구데이터 기반 LLM 산출 데이터 품질 평가표

모델명	평가 요인			
	유효성	신뢰성	기능성	
			추론	언어 이해
LLaMA	X	○	○	○
Alpaca	X	○	○	○
Vicuna	X	○	X	○
ChatGPT(GPT-4)	X	X	○	○

바 있고, Alpaca에서도 LLaMA 데이터를 기본적으로 활용함과 동시에 해당 모델에서 추가로 사용한 데이터에 대한 정보를 지속해서 제공하고 있다. 학습 데이터 및 활용 데이터의 명확한 제공이라는 측면에서 LLaMA 기반 모델들은 일정 수준 이상의 신뢰성을 갖는다고 평가할 수 있다. 반면에, ChatGPT 모델은 활용 데이터에 대한 정보를 전혀 공개하고 있지 않거나, web-text, books1, books2 같은 모호한 용어를 활용하여 제공하고 있으므로 신뢰성 검증이 불가능하다.

끝으로 기능성의 관점에서 평가한 추론과 언어 이해는 대부분 충족하였다는 결론에 이르는데, 이는 본연구에서 실질적으로 해당 LLM의 성능 평가를 시행한 것은 아니며, 4개의 모델을 개발하면서 개발 주체가 직접 해당 모델의 성능에 대해 공식적으로 언급한 부분과 HELM과 같이 기존에 실시된 평가 내용을 참조하여 도출하였다. 다만, 다른 LLM과 달리 Vicuna의 경우에는 상세하고 잘 구성된 응답을 생성하는 챗봇으로 사용할 목적으로 개발되었기 때문에 언어 이해 능력과 질문 응답에 대한 능력은 있지만 추론 기능은 없다고 평가되며, 결과적으로 본 연구의 평가 기준에 충족하지 못한다고 판단하였다. LLM에서 추론이란, 기존의 데이터를 바탕으로 새로운 데이터를 생성하거나 생성한 데이터를 기반으로 재학습하는 능력을 의미하는데, Vicuna 모델은 응답을 생성하는 것에만 집중하기 때문에 다른 LLM에서 보여주는 추론 능력을 기준으로 평가할 때, 언어적 추론 기능 외의 다른 추론 기능이 없다고 할 수 있다.

4. 결론 및 제언

본 연구는 4개의 LLM 모델을 비교・분석하고 연구데이터 품질 평가 기준 가운데 LLM 특성에 알맞은 요인을 추출하여 유효성, 신뢰성, 기능성을 기준으로 해당 LLM이 생성 및 산출하는 데이터의 품질에 대한 논의를 진행하고 평가 기준을 제시하였다. 많은 모델이 온라인상의 데이터를 기반으로 결과물을 도출하고 있으며, GPT 모델의 경우에는 데이터에 대한 정보를 공개하고 있지 않다. 그러므로 대부분은 유효성을 검증하기 어려우며, 연구데이터를 임의로 편집하여 사용하는 경우 역시 유효성을 측정할 수 없다. GPT 모델과는 다르게 LLaMA 기반의 LLM은 활용하는 학습 데이터를 명확하게 제공하고 있으므로 참고 자료를 정확하게 제공하고 있다는 내용을 기준으로 봤을 때, LLaMA 기반의 LLM이 생산해 내는 데이터는 일정 수준 이상의 신뢰성을 갖는다고 볼 수 있다. 그리고 대부분은 기능적인 관점에서 추론과 언어 이해는 일정 수준 이상의 성능을 기반으로 결괏값을 제공하고 있는 것으로 확인되었다.

LLM은 목적에 따라 서로 다른 기능을 갖는다. 예를 들어, Vicuna의 경우 상세하고 잘 구성된 응답을 생성하는 챗봇을 목적으로 제작되었기 때문에 언어 이해와 질문 응답에 중점을 두어 개발되었다. 그러므로 해당 모델은 추론할 수 없는 특징을 갖는다. 여러 모델에 대해 기능적인 평가를 확장하기 위해서는 연구데이터를 기반으로 한 모델이 제공하는 출력값을 대상으로 한 영향력에 관한 연구가 추가로 필요하다. 그뿐만 아니라 본 연구는 LLM이 제공하는 데이터의 특성에 기인하여 일반적으로 데

이터를 평가하는 요인 가운데 3개 요인을 추출하여 LLM이 생산하는 데이터의 품질에 관한 논의를 시행하였는데, 해당 평가 요인만으로는 LLM이 생산 및 산출해 내는 데이터의 품질을 정확하게 평가하기는 어렵다는 한계를 갖는다. 추출한 평가 요인들은 정형 및 비정형 데이터를 모두 포함하는 데이터를 평가하는 기준으로부터 추출한 것이기 때문에 LLM이 제공하는 데이터의 특성에 완전하게 부합하는 기준이라고 논의하기는 어려우며 추후 추가로 평가해야 할 요인들에 대한 논의가 필요하다.

이후 후속 연구를 통해서는 CommonCrawl과 같은 데이터뿐만 아니라 보다 정제화된 연구데이터를 학습시켰을 경우 LLM이 제공하는 결괏값의 차이와 같은 품질 평가 비교 연구, 또는 국가연구데이터 DataON에서 제공하는 분석 서비스인 CANVAS와 같은 정확성과 신뢰성, 기능성을 충분히 보장받을 수 있는 분석 방법을 활용한 결괏값을 LLM에서 도출한 결괏값 등과 비교함으로써 LLM의 성능 평가 및 품질 평가를 진행할 수 있을 것으로 기대하는 바이다.

참 고 문 헌

국가과학기술연구회 (2019). 연구데이터 관리 가이드라인 (2019-07).

국가연구개발정보처리기준, 과학기술정보통신부고시 제2020-102호.

김선태, 이정훈, 정한민 (2017). 연구데이터 이해와 관리. 대전: 한국과학기술정보연구원.

김형섭 (2020). 데이터 품질관리 평가 모델에 관한 연구. 한국융합학회논문지, 11(7), 217-222. https://doi.org/10.15207/JKCS.2020.11.7.217

박성호 (2020). 텍스트 및 데이터 마이닝을 목적으로 하는 타인의 저작물의 수집・이용과 저작재산권의 제한: 인공지능의 빅데이터 활용을 중심으로. 인권과 정의, 494, 39-69. http://doi.org/10.22999/hraj..494.202012.003

박현경 (2020). 인공지능 학습과정에서 저작물의 이용에 관한 소고. 스포츠엔터테인먼트와 법, 23(1), 129-152. http://doi.org/10.19051/kasel.2020.23.1.129

안성원, 유재흥, 조원영, 노재원, 손효현 (2023). 초거대언어모델의 부상과 주요 이슈: ChatGPT의 기술적 특징과 사회적, 산업적 시사점. 경기: 소프트웨어정책연구소.

안전행정부 (2014). 공공데이터 관리지침, 제2014-13호.

이경님, 조은경 (2023). 초거대 언어 모델을 기반으로한 AI 대화 인터페이스-AI 대화 모델의 현황과 언어적 연구의 모색. 국어학, 105, 345-374. https://doi.org/10.15811/jkl.2023..105.010

이기창 (2021). (Do it!) BERT와 GPT로 배우는 자연어 처리: 트랜스포머 핵심 원리와 허깅페이스 패키지 활용법. 서울: 이지스퍼블리싱.

이수현, 전상홍 (2023). ChatGPT 기술 산업 현황 보고서. 한국저작권위원회.

조태호 (2022). 모두의 딥러닝 - 누구나 쉽게 이해하는 딥러닝. 서울: 도서출판길벗.

한국데이터베이스진흥원 (2006). 데이터 품질관리 지침(Ver 2.1).

한나은 (2023). 연구데이터 품질관리를 위한 프로세스 모델 제안. 정보관리학회지, 40(1), 51-71. https://doi.org/10.3743/KOSIM.2023.40.1.051

我妻 幸長 (2018). はじめてのディープラーニング -Pythonで學ぶニューラルネットワークとバックプロパゲーション- (Machine Learning). 최재원 옮김(2019). 실체가 손에 잡히는 딥러닝, 기초부터 실천 프로그래밍. 서울: 책만.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

Buchanan, B., Lohn, A., Musser, M., & Sedova, K. (2021). Truth, lies, and automation. Center for Security and Emerging Technology, 1(1), 2.

Chiang, W. L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., & Xing, E. P. (2023). Vicuna: An Open-source Chatbot Impressing Gpt-4 with 90%* Chatgpt Quality. Available: https://lmsys.org/blog/2023-03-30-vicuna/

Chomsky, N. (1957). Logical structure in language. Journal of the American Society for Information Science, 8(4), 284.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., Carter, L., & Wright, R. (2023). "So what if ChatGPT wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy. International Journal of Information Management, 71, 102642.

English, L. P. (2009). Information Quality Applied: Best Practices for Improving Business Information, Processes and Systems. New Jersey: Wiley.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). Realtoxicityprompts: Evaluating Neural Toxic Degeneration in Language Models. https://doi.org/10.48550/arXiv.2009.11462

Hale, J. (2001). A Probabilistic Earley Parser as a Psycholinguistic Model. In Second Meeting of the North American Chapter of the Association for Computational Linguistics.

International Organization for Standardization (2015). ISO/IEC 25024: 2015: Systems and Software Engineering-Systems and Software Quality Requirements and Evaluation (SQuaRE)-Measurement of Data Quality. ISO/IEC.

Jurafsky, D. & James H. M. (2021). Speech and Language Processing (3rd ed.). California: Standford University.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling Laws for Neural Language Models. https://doi.org/10.48550/arXiv.2001.08361

Kindling, M. & Strecker, D. (2022). Data Quality Assurance at Research Data Repositories. Data Science Journal, 21(1). http://doi.org/10.5334/dsj-2022-018

Lee, P., Goldberg, C., & Kohane, I. (2023). The AI Revolution in Medicine: GPT-4 and beyond. London: Pearson.

Lemley, M. A. & Casey, B. (2020). Fair learning. Texas Law Review, 99(4), 743-785.

Levy, R. (2008). Expectation-based syntactic comprehension. Cognition, 106(3), 1126-1177. https://doi.org/10.1016/j.cognition.2007.05.006

Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., Wu, Y., Neyshabur, B., Gur-Ari, G., & Misra, V. (2022). Solving quantitative reasoning problems with language models. Advances in Neural Information Processing Systems, 35, 3843-3857.

Lin, S., Hilton, J., & Evans, O. (2021). Truthfulqa: Measuring How Models Mimic Human Falsehoods. https://doi.org/10.48550/arXiv.2109.07958

OpenAI (2023). GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774

Peng, B., Li, C., He, P., Galley, M., & Gao, J. (2023). Instruction Tuning with Gpt-4. https://doi.org/10.48550/arXiv.2304.03277

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. Nature, 592(7855), 590-595.

Percy, L., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Re, C., Acosta-Navas, D., Hudson, D. A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., Wang, J., Santhanam, K., Orr, L., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N., Guha, N., Chatterji, N., Khattab, O., Henderson, P., Huang, Q.,

Chi, R., Xie, S. M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., & Koreeda, Y. (2022), Holistic Evaluation of Language Models. https://doi.org/10.48550/arXiv.2211.09110

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Language Models as Knowledge Bases?. https://doi.org/10.48550/arXiv.1909.01066

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., Rutherford, E., Hennigan, T., Menick, J., Cassirer, A., Powell, R., Driessche, G., Hendricks, L. A., Rauh, M., Huang, P., Glaese, A., Welbl, J., Dathathri, S., Huang, S., Uesato, J., Mellor, J., Higgins, I., Creswell, A., McAleese, N., Wu, A., Elsen, E., Jayakumar, S., Buchatskaya, E., Budden, D., Sutherland, E., Simonyan, K., Paganini, M., Sifre, L., Martens, L., Li, X. L., Kuncoro, A., Nematzadeh, A., Gribovskaya, E., Donato, D., Lazaridou, A., Mensch, A., Lespiau, J., Tsimpoukelli, M., Grigorev, N., Fritz, D., Sottiaux, T., Pajarskas, M., Pohlen, T., Gong, Z., Toyama, D., d'Autume, C. M., Li, Y., Terzi, T., Mikulik, V., Babuschkin, I., Clark, A., Casas, D. L., Guy, A., Jones, C., Bradbury, J., Johnson, M., Hechtman, B., Weidinger, L., Gabriel, I., Isaac, W., Lockhart, E., Osindero, S., Rimell, L., Dyer, C., Vinyals, O., Ayoub, K., Stanway, J., Bennett, L., Hassabis, D., Kavukcuoglu, K., & Irving, G. (2021). Scaling Language Models: Methods, Analysis & Insights from Training Gopher. https://doi.org/10.48550/ARXIV.2112.11446

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023). Stanford Alpaca: An Instruction-following Llama Model. Available: https://github.com/tatsu-lab/stanford_alpaca

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., & Stojnic, R. (2022). Galactica: A Large Language Model for Science. https://doi.org/10.48550/arXiv.2211.09085

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., 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. https://doi.org/10.48550/arXiv.2302.13971

Wilcox, E., Qian, P., Futrell, R., Kohita, R., Levy, R., & Ballesteros, M. (2020). Structural Supervision Improves Few-shot Learning and Syntactic Generalization in Neural Language Models. https://doi.org/10.48550/arXiv.2010.05725

Yarowsky, D. (1995, June). Unsupervised word sense disambiguation rivaling supervised methods. In 33rd Annual Meeting of the Association for Computational Linguistics, 189-196.

Yasunaga, M., Bosselut, A., Ren, H., Zhang, X., Manning, C. D., Liang, P. S., & Leskovec, J. (2022). Deep Bidirectional language-knowledge graph pretraining. Advances in Neural Information Processing Systems, 35, 37309-37323.

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

An, Seong-Won, Yu, Jae-Hong, Jo, Won-Young, No, Jae-Won, & Son, Ho-Hyun (2023). Rise of Hyper-scale LLM(Large Language Model) and issues. Gyeonggi: Software Policy Research Institute.

Azma Yukinaga (2018). Deep Learning that is Tangible, Practical Programming from the Basics. Tokyo:SBクリエイティブ.

Han, Na-Eun (2023). Proposal of process model for research data quality management. Korean Society for Information Society, 40(1), 51-71.
https://doi.org/10.3743/KOSIM.2023.40.1.051

Jo, Tae-Ho (2022). Deep Learning for Everyone - Deep Learning that Anyone can Easily Understand. Seoul: Gilbut.

Kim, Hyung-Sub (2020). A study on the data quality management evaluation model. Journal of the Korea Convergence Society, 11(7), 217-222.
https://doi.org/10.15207/JKCS.2020.11.7.217

Kim, Seon-Tae, Lee, Jeong-Hoon, & Jeong, Han-Min (2017). Understanding and Managing Research Data. Daejeon: Korea Institute of Science and Technology Information.

Korea Data Agency (2006). Data Quality Management Guidelines (Ver 2.1).

Lee, Gi-Chang (2021). (Do it!) Learning Natural Language Processing with BERT and GPT: Transformer Core Principles and How to Use the Hugging Face Package. Seoul: Easyspublishing.

Lee, Kyong-NIm & Ho, Eun-Kyoung (2023). AI dialogue interface based on large language models: the state of the art AI dialogue models and seeking linguistic research topics. The Society of Korean Linguistics, 105, 345-374. https://doi.org/10.15811/jkl.2023..105.010

Lee, Su-Hyeon & Jeon, Sang-Hong (2023). ChatGPT State of the Technology Industry Report. Korea Copyright Commission.

Ministry of Security and Public Administration (2014). Government Data Management Guidelines.

No. 2014-13.

National Research and Development Information Processing Standards, Ministry of Science and ICT Notice No. 2020-102.

National Research Council of Science and Technology (2019). Research Data Management Guidelines (2019-07).

Park, Hyung-Kyung (2020). A study on the use of copyrightable works in machine learning. The Korean Association of Sports and Entertainment Law, 23(1), 129-152. http://doi.org/10.19051/kasel.2020.23.1.129

Park, Seong-Ho (2020). A study on whether collecting and using other people's copyrighted works for the purpose of text and data mining falls under the copyright limitations: focusing on the use of big data in artificial intelligence. Human Rights and Justice, 494, 39-69. http://doi.org/10.22999/hraj..494.202012.003