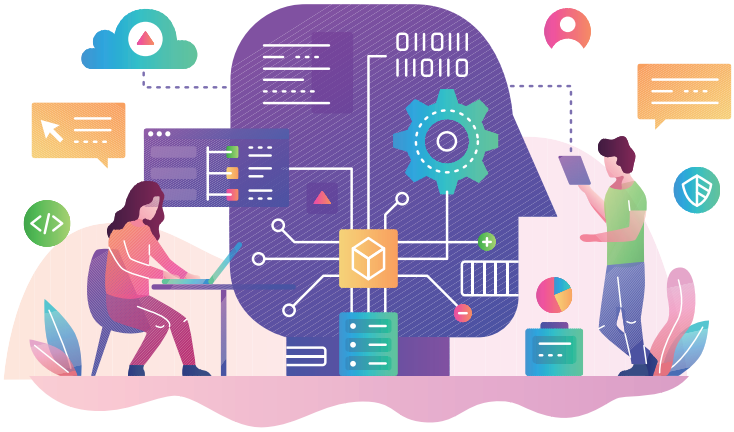
ASTI MARKET INSIGHT 2023-205

ASTI MARKET INSIGHT

데이터 라벨링



데이터분석본부 호남지원 선임연구원 **이은지** Tel: 062-951-7704 e-mail: eunji_lee@kisti.re.kr

KEY FINDING

1. 인공 지능(AI) 모델의 품질 향상과 신뢰성 제고를 위해서 양질의 데이터 확보가 중요하다는 인식이 자리잡게 되면서 AI 학습 데이터세트 구축, 품질 관리, 데이터 라벨링 자동화 도구 개발 등 데이터 라벨링 시장이 형성되었다.
2. 데이터 라벨링의 세계 시장 규모는 2022년 8억7,530만 달러이며, 연평균 33.2 %로 성장해 2027년 36억6,550만 달러가 될 것으로 전망된다.
3. 정부는 데이터 실효성 제고와 데이터 품질 향상을 위해 '제1차 데이터 산업 진흥 기본 계획(2023~2027)'을 발표해 사회의 모든 데이터의 혁신적 생산·개방·공유를 추진하고, 민간 주도의 데이터 유통 거래 생태계 마련을 위한 중장기 계획을 공개하였다.
4. 데이터 라벨링 시장은 AI 서비스 시장과 밀접한 연관성을 가지고 있어 AI 서비스 시장의 확대는 데이터 라벨링 시장을 지속적으로 견인할 것으로 전망된다.
5. AI가 인간과 더욱 자연스럽게 소통하기 위한 멀티모달 연구가 활발하게 전개되면서 텍스트뿐만 아니라 이미지, 음성, 행동, 표정 등 여러 모달리티 데이터에 대한 수요가 증가해 향후 데이터 라벨링 시장의 성장을 촉진할 것으로 예상된다.

1) 시장의 개요

데이터 라벨링(Data Labeling)은 원천 데이터(Raw Data)에 인공 지능(AI) 학습에 활용할 수 있도록 기능이나 목적에 부합하는 정보를 부착하는 활동을 의미한다. 데이터 라벨링을 통해 구축된 데이터는 AI 모델 학습에 활용되며, AI 모델의 정확도를 개선하고 서비스 개발에 활용된다. 데이터의 수집, 가공, 정제 등 데이터 구축 작업은 AI 모델 개발 과정 중 약 80 %의 업무를 차지하지만, 모델 개발의 주요 업무로 인식되기 보다 보조적인 작업으로 여겨지는 경우가 많았다. 그러나 양질의 데이터가 AI 모델의 품질 향상과 신뢰성 제고를 위한 중요한 부분임을 공감하는 분위기가 형성되면서, AI 학습 데이터세트 구축, 품질 관리, 데이터 라벨링 자동화 도구 개발 등 데이터 라벨링 산업이 주목받게 되었으며, 점차 AI 기술 성숙도의 향상으로 다양한 기술 구현이 가능해지면서 다양한 산업과 서비스 영역에서 AI 도입에 대한 긍정적인 인식이 제고되며, 데이터 라벨링 시장도 함께 성장하고 있다.

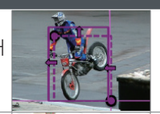
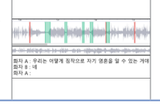
데이터 라벨링 산업의 밸류체인은 데이터 구축 및 라벨링을 위한 설

계 및 컨설팅, 데이터 라벨링 솔루션, 데이터 판매 및 제공 영역으로 나눌 수 있다. 데이터 라벨링을 위한 설계 및 컨설팅 영역의 경우, 데이터 설계, 품질, 성능개선, 분석 및 활용에 대한 컨설팅 서비스를 제공한다. 데이터 라벨링 솔루션 영역은 수집된 원천 데이터를 정제하고, 목적에 맞게 데이터 라벨링 작업을 수행할 수 있는 솔루션을 제공하는 분야이며, 대부분 플랫폼 형태로 개발되어 클라우드 환경에서 배포 및 서비스되고 있다. 데이터 판매 및 제공 영역은 수요자 맞춤형 데이터 재가공을 통한 데이터 제공, 데이터 판매 등 데이터 중개 서비스를 제공하는 영역으로 최근에는 플랫폼을 통해 데이터 수요자와 공급자가 직접 데이터를 거래하는 방식으로 변화하고 있다.

그림 1 데이터 라벨링 시장의 밸류체인

데이터 라벨링은 활용 분야에 따라 텍스트, 이미지, 비디오, 오디오 등과 같이 데이터의 유형이 달라진다. 이미지와 영상 데이터의 라벨링은 의료 영상 진단, 자율 주행 자동차의 도로 상황 인식, 콘텐츠 유통 플랫폼 내 영상 분석 등 시각 지능이 필요한 분야에 활용되고, 텍스트와 음성 데이터의 라벨링은 문서 분류, 문서 요약, 음성 인식, 대화형 AI, 자동 응답 시스템, 회의록 작성 등 AI가 필요한 분야에 활용된다. 최근 OpenAI가 ChatGPT를 공개한 이후 글로벌 빅테크 기업의 대규모 언어 모델 구축 경쟁이 촉발되었으며, 자국어 대규모 언어 모델 개발을 위한 각국의 적극적인 노력과 함께 대용량 텍스트 데이터의 수요가 증가하고 있다. 또한 인간처럼 사고하고, 추론하는 능력을 기반으로 창작물을 생성할 수 있는 생성형 AI 시대로 진화하면서 언어, 이미지, 오디오 정보를 복합적으로 반영할 수 있는 멀티모달 데이터에 대한 수요가 더욱 증가할 것으로 판단된다.

표 1 데이터 유형별 데이터 라벨링 기법

데이터 유형	라벨링 기법(예시)			
이미지	• 바운딩 박스 : 작업 대상에 박스를 드래그해 그린 다음 대표 키워드를 태깅하는 작업		• 폴리 라인 : 작업 대상의 외곽선에 여러 점을 찍어 선으로 선택하는 기법	
	• 폴리곤 : 작업 대상의 외곽선을 따라 점을 찍어 라벨링 하는 기법		• 폴리 라인 : 작업 대상의 외곽선에 여러 점을 찍어 선으로 선택하는 기법	
영상	• 스켈레톤 추출 : 작업 대상의 행동 패턴 분석을 위해 대상의 특정 부위에 점을 찍어 연결		• 시맨틱 세그멘테이션 : 작업 대상의 장면, 상황적 특징을 의미있는 단위로 분할하는 작업	
텍스트	• 태깅/라벨링 : 카테고리, 감정, 긍·부정 분류를 위해 텍스트가 갖는 정보를 태깅하는 작업		• 문서 요약 : 원본 데이터에 대한 추출적/생성적 요약문을 구축	
음성	• 화자 구분 : 제시된 음성을 듣고, 동일한 사람의 목소리인지 판단하고, 화자를 태깅하는 작업		• 전사(transcript) : 주어진 음성을 듣고 받아쓰는 작업	

2) 정책 및 규제 현황

미국에서는 민간이 주도하고 정부의 투자 협력을 통해 데이터 라벨링을 비롯해 데이터 가공 기술이 적용된 AI 학습 데이터세트를 공개해 AI 모델 및 서비스의 개발에 이용할 수 있도록 하였다. 또한 미국 정부는 데이터법(Data Act)을 제정해 데이터 구축을 위한 표준을 만들어 규격화하고, 데이터 활용에 대한 혜택과 제도를 선제적으로 정립하였으며, 2021년에는 '연방 데이터 전략 실행 계획'을 발표해 농업·기후·소비·교육·에너지·의료 등 14 개 분야로 공공 데이터 전략을 이행하는 등 국가 데이터 저장 네트워크 구축과 데이터 개방 및 활성화를 위한 제도와 기술의 토대를 마련하였다.

중국에서는 중국과학원(CAS)을 중심으로 데이터 경제 활성화 정책을 추진하고 있는 가운데 스마트 시티 및 디지털 차이나 등 사회경제적 디지털화가 추진되면서 정부가 주도해 데이터 거래 활성화를 위해 노력하고 있으며, 알리바바, 텐센트 등 중국 기업은 결제, 금융, 보험, 의료 등 데이터를 기반으로 하는 생활 전반 서비스 플랫폼 생태계를 구축하고 있다.

우리나라에서는 데이터 산업의 초기 시장을 빠르게 형성하기 위하여 공공 데이터의 개방, 재정 투입을 통한 AI 학습용 데이터 구축 사업, AI 바우처 지원 사업 등 정부가 주도해 데이터 산업 기반의 조성에 힘써 왔다. 하지만 정부 정책의 관련 산업 촉진에도 불구하고 데이터 공급자 중심의 데이터 구축으로 데이터 판매자와 수요자의 니즈의 불균형에 대한 한계가 제기되었다. 이에 정부는 데이터 실효성 제고와 데이터 품질 향상을 위해 2023년 1월 '제1차 데이터 산업 진흥 기본 계획' 발표를 통해 사회의 모든 데이터의 혁신적 생산·개방·공유를 추진하고, 민간 협력을 바탕으로 데이터 유통 거래 생태계 조성을 위한 중장기 계획을 공개하였다. 이는 AI·데이터 교육을 통한 데이터 산업의 기초 체력 강화 방안 모색과 AI 일상화 및 산업 고도화를 위한 다각적인 측면에서의 데이터 산업 정책으로 2027년까지 데이터 시장을 50조 원까지 육성하는 것을 목표로 하고 있다. 또한 ChatGPT로 상징되는 초거대 생성형 AI 산업에서 국가 경쟁력 확보를 위해 기존 AI 학습용 데이터 정책을 초거대 AI 학습용 핵심 데이터 구축으로 전환해 지원할 계획임을 밝힘으로써 초거대 AI 기술과 함께 AI 데이터 산업은 지속 성장할 것으로 판단된다.

3) 시장 동향

시장 규모 및 전망

AI 기술의 효용성과 확장 가능성이 여러 적용 사례를 통하여 입증되면서, AI 기술을 활용한 혁신적인 제품 및 서비스 개발을 위한 데이터 라벨링 시장은 함께 성장하고 있다. 데이터 라벨링의 세계 시장 규모는 2022년 8억7,530만 달러이며, 연평균 33.2 %로 성장해 2027년 36억6,550만 달러가 될 것으로 전망된다.

지역별로 살펴보면, 북미 지역은 33.8 %로 세계 시장에서 가장 큰 점유율을 차지한다. 이는 미국의 독보적인 AI 기술력과 IBM, Oracle, Google, TELUS, AWS, Adobe 등 글로벌 빅테크 기업을 보유하고 있기 때문이다. 북미 지역은 2022년 2억9,620만 달러에서 연평균 31.8 %로 성장해 2027년 11억7,650만 달러에 이를 것으로 전망된다. 유럽 지역은 2022년 2억3,380만 달러에서 연평균 32.6 %로 성장해 2027년 9억5,670만 달러에 이를 것으로 전망된다. 아시아 지역은 2022년 2억1,050만 달러에서 연평균 34.9%로 성장하여 2027년 9억4,200만 달러에 달할 것으로 예측되며, 지역별 시장 규모 중 가장 높은 성장이 예측된다. 아시아 지역의 데이터 라벨링 시장의 높은 성장 배경에는 의료 및 생명 과학 분야 전반에 AI 혁신 기술을 적극적으로 도입하는 분위기에 영향을 받은 것으로 판단된다.

표 2 데이터 라벨링 지역별 시장 규모 및 전망(2022~2027) (단위:백만달러)

구 분	2022	2023	2024	2025	2026	2027	CAGR(%)(2022~2027)
미주	296.2	413.8	562.3	739.3	942.5	1,176.5	31.8
유럽	233.8	328.2	448.3	592.7	759.9	956.7	32.6
아시아	210.5	300.9	418.4	563.1	735.0	942.0	34.9
중동/아프리카	78.0	110.9	153.4	205.4	266.6	342.7	34.4
중남미	56.8	80.8	111.7	149.6	194.2	247.6	34.2
계	875.3	1,234.6	1,694.1	2,250.1	2,898.2	3,665.5	33.2

출처 : Markets&Markets, Data Annotation & Labeling Market, 2022

데이터 라벨링을 활용 분야별로 살펴보면, 금융, IT, 의료, 유통/고객 서비스, 모빌리티, 공공, 기타 분야로 구분할 수 있으며, 그중에 금융, IT, 의료 분야는 실생활과 밀접하게 관련이 있으면서도 활용도가 높아 시장점유율도 높다. 국내의 경우 금융 분야에서 마이데이터 산업이 시작되면서 은행, 카드, 증권 등 기존 금융 분야뿐만 아니라 데이터 기반 핀테크 업체, 카카오, 네이버 등의 빅테크 기업까지 금융 데이터를 기반으로 서비스 시장 선점에 총력을 기울이고 있으며, 소비자는 금융 상품을 추천받거나 재무 상태를 컨설팅받을 수 있고, 신용 정보와 재무 상황을 한눈에 확인할 수 있게 되었다. 현재 의료 분야에서는 개인 의료 정보의 거래가 금지되어 있어 마이데이터로서 유통되고 있지는 않지만, 의료 기술 연구나 신약 의료 기기 개발을 위한 의료 데이터 분석 및 진단 소프트웨어 솔루션 등 의료 AI 플랫폼 개발에 활용되고 있다. AI 기술에 대한 가시적인 성과가 나타나면서 다양한 산업이 AI로 귀결되고 있어 데이터 라벨링의 시장 성장은 계속될 것으로 전망된다.

표 3 데이터 라벨링의 분야별 세계 시장 규모 및 전망(2022~2027) (단위:백만달러)

구 분	2022	2023	2024	2025	2026	2027	CAGR(%)(2022~2027)
금융	239.8	333.0	449.5	586.9	742.6	921.2	30.9
IT	156.3	219.1	298.7	394.2	504.6	634.1	32.3
의료	106.1	152.3	212.6	287.4	376.6	484.7	35.5
통신	107.4	151.9	208.8	277.9	358.6	454.5	33.4
유통/고객 서비스	89.9	128.6	178.9	241.0	314.8	403.7	35.0
모빌리티	85.9	122.6	170.4	229.1	298.8	382.6	34.8
공공분야	55.9	79.6	110.3	147.9	192.3	245.6	34.5
기타	34.0	47.6	65.0	85.8	109.9	139.1	32.5
합계	**875.3**	**1,234.7**	**1,694.2**	**2,250.2**	**2,898.2**	**3,665.5**	**33.2**

출처 : Markets&Markets, Data Annotation & Labeling Market, 2022

표 4 데이터 라벨링의 분야별 활용 사례

구 분	활용 사례
금융	사기 방지, 금융 거래, 리스크 관리, 자동 상담 시스템 및 챗봇, 고객 이탈 방지, 신용 평가 및 대출 자동화
IT	업무 생산성 향상, 소프트웨어 개발 및 제공(오토코딩), 감성 분석을 통한 신제품 기획, 셀프서비스 채널의 효율성 향상
의료	의료 영상 분석, 고위험 질환 진단, 환자 전자 건강 기록 관리, 원격 환자 진료, 진단 예측, 신약 개발 가속화, 로봇 수술
통신	네트워크 이상 감지, 서비스 모니터링, 신규 서비스에 대한 감성 분석, 고객 이탈 방지, 자동 상담 시스템 및 챗봇
유통/고객 서비스	재고 분석, 구매 이력 분석, 웹사이트 트래픽 분석, 추천 시스템, AR/VR 기반 제품 미리보기 애플리케이션, 검색 엔진 개선
모빌리티	자율 주행 자동차, 첨단 운전자 보조 시스템(ADAS), 운전자 피로 감지 및 차내 모니터링, 차선 이탈 감지, 차량 간 통신, 스마트 주차, 차량 번호판 인식
공공분야	정밀 농업, 식품 안전, 민간 및 군사 정보, 국가 안보, 콜드체인 관리(백신), 전자 거버넌스 개선
기타	에너지 관리, 미디어/엔터테인먼트, 제조, 교육 등

경쟁 현황

미국 등에서는 빅테크 기업을 중심으로 AI 학습 데이터 품질 관리를 통한 신뢰성 있는 AI 모델을 개발하기 위해 데이터 라벨링 솔루션 및 플랫폼을 제공하는 기업을 인수·합병하는 전략으로 AI 산업을 확장하고 있다. 미국의 메타(Meta, 구 페이스북)의 경우 합성 데이터 생성 기업인 AI리버리(AI Reverie)를 인수해 데이터 라벨링, 데이터 생성을 통한 AI 모델을 개발하는데 활용 하고 있다. 호주의 에펜(Appen)은 크라우드 소싱으로 학습용 데이터를 생산 및 제공하고, AI 데이터 기업 인수에 3,500억 원을 투자하였고, 미국의 스케일AI(Scale AI)는 로봇·자율 주행차·드론 이미지에 주석을 추가하는 소프트웨어를 개발해 1,800만 달러 규모의 펀딩을 유치하였다.

국내의 경우 스타트업을 중심으로 크라우드 소싱 방식의 데이터 라벨링 플랫폼 기업이 다수 등장하고 있으며, 데이터 라벨링 자동화 및 데이터 신뢰성 검증에 대한 기술력을 갖춘 기업의 경우 국내외 빅테크 기업의 투자를 유치하는 등 성장을 가속화하고 있다.

표 5 데이터 라벨링 전문 기업

구 분	주요 내용
에펜(Appen) (호주)	• 유니콘 기업 • 자율 주행, 로보틱스, 챗봇 등 다양한 유형의 학습 데이터를 지원하며, 전세계 140 개 이상 국가의 작업 인력 보유 • 원스톱 AI 데이터 어노테이션 플랫폼
스케일AI(Scale AI) (미국)	• 로봇·자율차·드론용 이미지에 주석을 추가하는 소프트웨어 개발 • 데이터 라벨링 과정을 API 형태로 제공 • 웨이모(Waymo), 우버(Uber) 자율 주행 기술 기업의 파트너 기업
크라우드웍스 (한국)	• 크라우드 소싱 방식, 데이터 품질 향상을 위한 전수 검수 시스템 도입 • 2022 가트너 하이프 사이클 리포트 '데이터 라벨링 및 가공 부문' 벤더로 등재
슈퍼브에이아이 (한국)	• 실리콘밸리 와이콤비에이터 투자 유치 • 기업형 소프트웨어형 서비스(SaaS) 데이터 플랫폼 '슈퍼브에이아이 스위트' 공개
셀렉트스타 (한국)	• AI 데이터 구축 올인원 솔루션 'DATUMO(feature space tool)' 공개 • 삼성, 네이버, LG, SK 등 국내 빅테크 기업의 파트너 기업

4) 분석자 인사이트

AI 기술은 컴퓨팅 자원의 발전으로 대용량 데이터 학습이 가능해지고, 효율적인 학습 모델의 개발로 정확도가 견고해지면서 자율 주행, 로보틱스, 음성 인식 인터페이스를 기반으로 사용자 편의성을 향상시키기 위한 서비스가 일반화되는 AI 대중화 시대에 이르렀다. 데이터 라벨링 시장은 AI 서비스 시장과 밀접한 연관성을 가지고 있어 AI 서비스 시장의 확대는 데이터 라벨링 시장을 견인할 것으로 전망되며, 특히 AI가 인간과 더욱 자연스럽게 소통하기 위한 멀티모달 연구가 활발하게 전개되면서 텍스트뿐만 아니라 이미지, 음성, 행동, 표정 등 여러 모달리티 데이터에 대한 수요가 증가하고 데이터 라벨링 시장의 성장을 촉진할 것으로 예상된다.

데이터 라벨링 기업은 대규모 데이터 가공에서부터 개인 정보 비식별화, 라벨링, 검수 등 전과정을 수행할 수 있는 플랫폼을 개발해 데이터 라벨링 작업 환경의 개선과 데이터 품질 관리 기술 개발에 힘쓰고 있으며, 최근에는 크라우드 소싱 형태의 플랫폼 확장을 통해 데이터 공급 기업과 수요 기업 또는 작업자를 연결해 주는 매개체 역할을 수행하고 있다. 앞으로의 데이터 라벨링 시장은 초거대 AI 모델 학습을 위한 대규모의 데이터의 수요에 대응하고, 데이터 구축에 소요되는 시간적, 경제적 비용을 절감하기 위한 기술 개발에 주력할 것으로 판단되며, 데이터 생성 기술을 통해 생성한 가상 데이터로 실제 데이터를 대체하거나 보완하는 기술 개발이 주목받을 것이다. 데이터 라벨링 시장 진입을 위한 주요 영향 요인은 〈표 6〉과 같이 정리할 수 있다.

표 6 시장 영향 요인 분석

촉진 요인

- 초거대 모델의 등장으로 대량의 인공지능 학습 데이터 수요 증가
- 맥락 이해, 상황 인지 등 AI 고도화를 위한 멀티모달 데이터 수요 증가

저해 요인

- 의료 영상, 자율 주행 등 과학적 전문 지식이 필요한 분야의 경우 도메인 전문가의 역량 필요
- 수작업이 필요한 분야가 아직 많음

기회 요인

- 마이데이터 서비스 확대를 통한 초개인화 서비스의 발달 (금융, 의료 분야 등)
- AI 기술의 성숙으로 자율 주행, 로봇 등 AI 도입에 대한 긍정적 인식 향상
- 기술력을 갖춘 SW기업의 데이터 라벨링 시장 진출이 용이

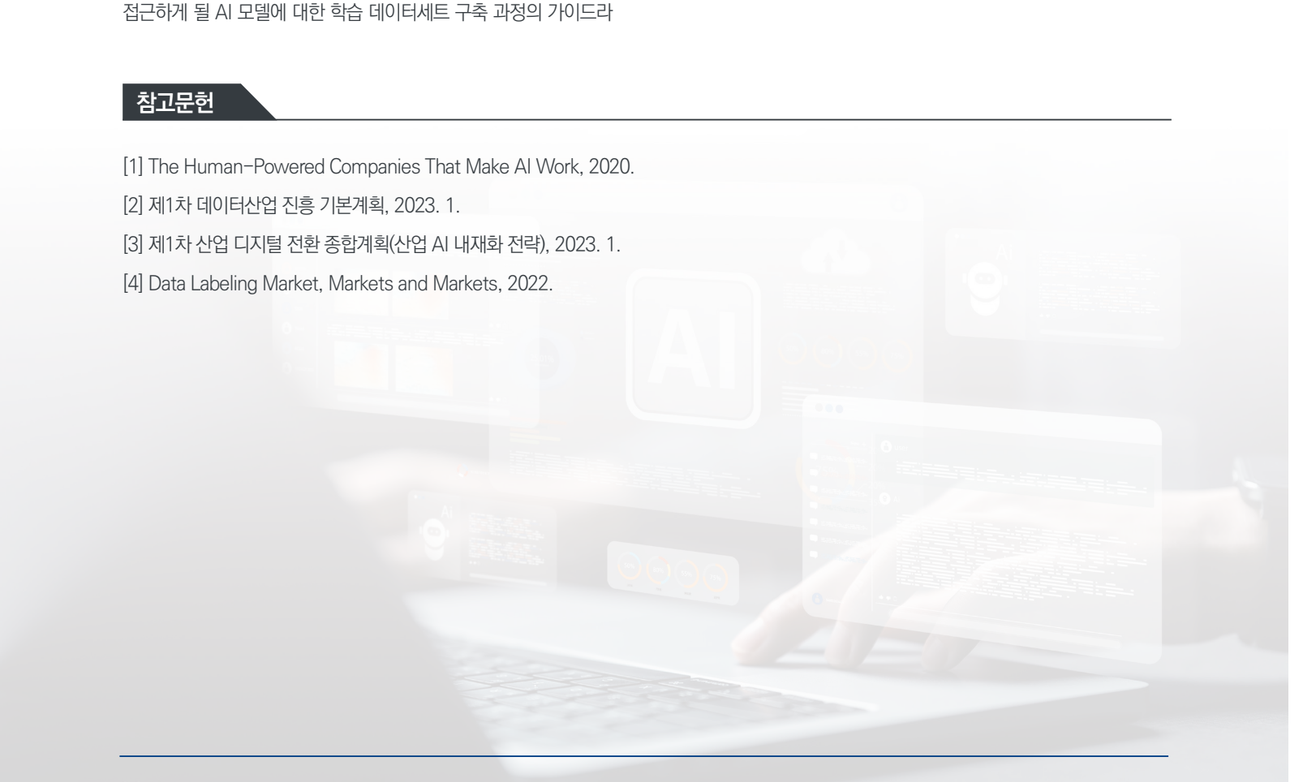
위협 요인

- AI를 활용한 스마트 라벨링 기술 개발 (자동화)을 통한 데이터 라벨링 구축 시간/비용 절감 및 작업 환경 개선
- 라벨링 결과에 대한 품질 관리 기술 개발
- AI 모델 발전에 따른 학습 데이터세트 변화 대응

데이터 라벨링은 AI 기술과 비즈니스 영역을 모두 이해하고 있어야 제대로 된 설계가 가능하고, 데이터에 대한 신뢰성을 높일 수 있으며, 작업에 대한 효율성을 제고할 수 있다. 따라서 앞으로 새롭게 접근하게 될 AI 모델에 대한 학습 데이터세트 구축 과정의 가이드라인이나 표준 규격 마련을 통한 작업 효율과 품질 향상이 필요할 것으로 사료되며, 개인 정보에 대한 법적 권한이나 관리 방안 등 운영 리스크 관리에 대한 노력이 필요할 것이다. ASTI

참고문헌

[1] The Human-Powered Companies That Make AI Work, 2020.
[2] 제1차 데이터산업 진흥 기본계획, 2023. 1.
[3] 제1차 산업 디지털 전환 종합계획(산업 AI 내재화 전략), 2023. 1.
[4] Data Labeling Market, Markets and Markets, 2022.



www.astinet.kr
에서 원문을 다운로드
받으실 수 있습니다.

ASTI MARKET INSIGHT

한국과학기술정보연구원
Korea Institute of Science and Technology Information
www.kisti.re.kr

본원 (우)34141 대전광역시 유성구 대학로 245 한국과학기술정보연구원
T. 042) 869-1004,1234 **F.** 042) 869-1091

분원 (우)02456 서울특별시 동대문구 회기로 66 한국과학기술정보연구원
T. 02) 3299-6114 **F.** 02) 3299-6244

비매품/무료
93560
9 788929 415266
ISBN 978-89-294-1526-6