과학기술정보통신부 보도자료

보도시점 2026. 5. 7.(목) 12:00 (2026. 5. 8.(목) 조간) 배포 2026. 5. 7.(목) 09:00

기존 AI허브 데이터, 업사이클링 통해 추론용 학습 데이터로 새롭게 혁신

- LLM(대규모 언어 모델) 및 피지컬AI 분야 각각 15종씩 기존 데이터셋 재가공
- 재가공된 데이터는 'AI Hub(aihub.or.kr)'를 통해 개방

【관련 국정과제】 20. AI 3대 강국 도약을 위한 AI 고속도로 구축

과학기술정보통신부(부총리 겸 과기정통부 장관 배경훈, 이하 과기정통부)와 **한국지능정보사회진흥원**(원장 김형철)은 기존 AI 허브 데이터를 **최신 생성형 AI 기술 환경에 맞게 재가공**하는 **'AI 학습용데이터 업사이클링*'** 사업 공고가 **4월 30일**부터 시작된다고 밝혔다.

* 'Upgrade'와 'Recycling'을 합친 단어로, 본 사업에서는 기존에 구축하여 AI허브에서 제공 중이던 AI 학습용데이터를 최신 기술 환경에 맞게 다시 가공하는 것을 의미

최근 빠르게 변화하는 기술 환경에 대응하여 **기존 판별형 AI 중심 라벨링 데이터를 재가공**하여 **추론 과정과 행동 정보를 포함하는 생성형 AI용 데이터로 전환**함으로써 보다 복잡한 문제 해결이 가능하도록 AI 기술 경쟁력 확보를 지원할 계획이다.

이번 사업은 **LLM(Large Language Model, 대규모 언어 모델)**과 **피지컬 AI** 분야를 중심으로 **총 30종의 데이터셋을 재가공(30억원 규모)**하며, 학습 데이터를 신규 구축하는 것에 비해 **예산 투입 대비 정책 효과**를 높일 수 있을 것으로 기대된다.

■ AI허브 데이터 전수 분석 기반 업사이클링 대상 선정

이번 사업은 2022년까지 구축*된 AI허브 데이터 691종을 생성형 AI용 데이터로의 확장 가능성, 데이터 활용도 등을 기준으로 전수 분석하고, **외부 전문가 검토**를 다시 거쳐 최종적으로 30종을 선정하였다.

* 2023년부터는 AI 학습용데이터가 생성형 AI용으로 구축되어 업사이클링 대상에서 제외

■ 판단 과정 학습을 위한 ‘LLM 데이터’

LLM 데이터 분야에서는 기존 텍스트 데이터를 기반으로 **질문-근거 검토-오류 검증-답변 확정에 이르는 추론 과정을 포함**하도록 데이터를 재구성한다. 이를 통해 단일 정답 제시에 그치지 않고, **다양한 판단 경로와 자기 검증 과정을 학습할 수 있는 데이터로 확장**할 계획이다. 특히 동일한 문제에 대해 **복수의 추론 경로를 구성**하고 **근거 기반 판단 및 오류 수정 과정을 포함**함으로써, 복잡한 문제 해결이 가능한 추론형 AI 학습 기반을 마련할 예정이다.

■ 시각·언어·행동 통합 ‘피지컬 AI 데이터’

피지컬 AI 분야에서는 **기존 이미지·영상 데이터를 기반으로 시각 정보(V), 언어명령(L), 행동 및 제어(A)를 통합한 구조로 데이터**를 고도화한다. 이를 통해 객체 인식 수준을 넘어, **시간 흐름에 따른 상황 변화와 객체 간 상호작용을 이해**하고 **목표 기반 행동을 생성할 수 있는 데이터로 확장**할 계획이다. 특히 연속적인 장면 정보와 객체 움직임 데이터를 활용하여 행동 경로와 작업 목표를 정의할 수 있는 형태로 재구성한다.

업사이클링된 데이터는 향후 ‘AI Hub(aihub.or.kr)’를 통해 공개되어 기업, 연구기관, 스타트업 등이 자유롭게 활용할 수 있도록 제공될 예정이다.

과기정통부는 본 사업을 통해 데이터의 품질과 적합성을 높이는 동시에 다양한 AI 환경에서 활용 가능한 구조로 개선하여, 최신 AI 시대에 대응하는 데이터 인프라를 지속적으로 확충해 나갈 계획이다.

과기정통부 최동원 인공지능인프라정책관은 “이번 업사이클링 사업을 통해 **적은 비용**으로도 **최신 생성형 AI 기술 환경**에 맞는 **AI 학습용데이터를 확보**할 수 있을 것”이라며, “**이미 축적된 데이터 자산**이 낭비되지 않도록 **활용 가치**를 끌어올려 나가겠다”고 밝혔다.

담당 부서	인공지능정책실 인공지능데이터정책과	책임자	과 장	이소라 (044-202-6580)
		담당자	사무관	문나윤 (044-202-6583)

붙임 2026년 AI 학습용데이터 업사이클링 대상

분야	연번	AI허브 후보 데이터셋	정부지원금
피지컬 AI	1	실내 자율주차용 데이터 (2022)	15억 (1종당 1억)
	2	로봇 핸드용 객체 특성 식별 데이터(2022)	
	3	장애인 길안내 자율주행 휠체어 융합센서 데이터(2021)	
	4	드론 자율항법을 위한 영상 및 센서 데이터(SLAM DATA)(2021)	
	5	배송용 로봇 시각 환경 인식 주행 영상(2021)	
	6	Ego-Vision 관점의 2D, 3D 손 움직임 데이터(2021)	
	7	특이 도로 환경 주행 데이터(2021)	
	8	1인칭 시점 보행영상(2020)	
	9	일상생활 작업 및 명령 수행 데이터(물체)(2022)	
	10	인도보행영상(2019)	
	11	손·팔 협조에 의한 파지-조작 동작 데이터(2022)	
	12	자율주행드론 비행 영상(2020)	
	13	K-pop 안무 영상(2020)	
	14	교통법규 위반 상황 데이터(2022)	
	15	사람 인체/자세 3D(2020)	
LLM	1	추상 요약 사실성 검증 데이터(2022)	15억 (1종당 1억)
	2	지식검색 대화(2022)	
	3	외부 지식 기반 멀티모달 질의응답 데이터(2022)	
	4	기계독해(2018)	
	5	수학분야 학습자 역량 측정 데이터(2020)	
	6	도서자료 기계독해(2020)	
	7	한국어 지식기반 관계 데이터(2022)	
	8	자연어 분석 후처리용 과교정 검증 데이터(2022)	
	9	논문자료 요약(2020)	
	10	비디오 네러티브 질의응답 데이터(2021)	
	11	일반상식 문장 교정 데이터(2022)	
	12	도서자료 요약(2020)	
	13	생활 및 거주환경 기반 VQA(2020)	
	14	민원(콜센터) 질의-응답 데이터(2020)	
	15	특허 지식베이스(2017)	