

보도시점 2026.8.27.(목) 14:00
(2026.8.28.(금) 조간)

배포 2026.8.27.(목) 09:00

「독자 AI 파운데이션 모델」 프로젝트로 확보한 AI 학습용 데이터 29종 본격 개방

- 5개 정예팀이 1차 단계평가 과정에서 구축한 약 3,544만건,
약 1.56조 토큰(추정) 규모의 데이터, 품질검증을 거쳐 'AI 허브'에 개방

【관련 국정과제】 20. AI 3대 강국 도약을 위한 AI 고속도로 구축
21. 세계에서 AI를 가장 잘 쓰는 나라 구현

과학기술정보통신부(부총리 겸 과학기술정보통신부 장관 배경훈, 이하 '과기정통부')와 한국지능정보사회진흥원(원장 김형철, 이하 'NIA')은 「독자 AI 파운데이션 모델」 프로젝트 5개 정예팀*이 1차 단계평가 과정에서 확보한 총 29종(약 3,544만건, 약 1.56조 토큰 규모)의 데이터를 AI허브 누리집(aihub.or.kr)에 공개한다고 밝혔다.

* 컨소시엄 별 주관기관(가나다順) : 네이버클라우드, 업스테이지, 에스케이텔레콤, 엔씨에이아이, 엘지경영개발원 AI연구원(이하 엘지AI연구원)

■ 정예팀별 AI개발 노하우가 담긴 대규모 AI학습 데이터 개방

이번 개방 데이터는 정예팀이 각자의 개발 전략에 맞춰 기획하여 데이터 구축·가공 예산('25년, 150억원)을 통해 확보한 데이터로, 사전학습에 필요한 대규모 데이터셋부터 영상·음성 등 멀티모달, 안전성 확보를 위한 레드티밍 데이터 등으로 구성되어 있다. 토큰으로 환산하면 약 1.56조 토큰(추정)에 달하며 이론적으로 약 70~80B 수준의 대형AI모델 학습에 사용될 수 있는 규모이다.

먼저, 네이버클라우드는 공개 영상 234만건과 방송 영상 52만건, 해당 영상 클립 기반의 텍스트 1,550만 건과 음성 질의응답 1,200만 건을 함께 구축하였다. 텍스트는 장면을 문장 단위로 기술한 캡션 등으로 구성되어 AI모델의 영상 이해능력 학습과 이미지 생성 모델 학습에 활용될 수 있다.

업스테이지는 1조 토큰 규모의 대규모 사전학습 데이터와 50만건의 사후 학습 데이터를 구축하였다. 사전학습 데이터셋은 대규모 AI모델을 새로학습 (from scratch)하는 데 쓰일 수 있으며, 사후학습 데이터는 단순 질의 응답을 넘어 추론, 판단, 실행 등 AI에이전트 개발을 위해 활용될 수 있다.

에스케이텔레콤은 수학·과학·법률 등 전문 도메인 영역의 고난도 다단계 추론 데이터, 음성·이미지 등 사후학습용 데이터와 국내 법령과 한국 사회의 보편적 가치관을 반영한 한국형 레드티밍 데이터 약 1만건을 구축하였다. AI모델의 실무 해결역량 강화와 안전성·신뢰성 향상 등에 활용될 수 있다.

엔씨에이아이는 제조 분야 기술문서, 민원 상담 음성 등 현장의 실제 데이터를 기반으로 긴 문맥 이해, 단계별 추론, 멀티턴 대화, 질의 응답, 멀티 모달리티 등 LLM 핵심 역량 데이터 7종을 구축하였다. 사후학습 성능 고도화와 산업 현장에 필요한 텍스트-이미지-음성 통합학습에 활용될 수 있다.

엘지AI연구원은 한국 가정환경에 최적화된 피지컬AI 멀티모달 데이터셋을 구축하였다. 국내 50개 실제 가정환경에서 50종 이상의 가사 활동을 촬영하여 17만개 이상의 영상 클립 등을 구축하고, 객체·분할·자세·맥락 정보 등을 다층 라벨링하였으며, 비전·VLM 학습, 연구·상용서비스 개발 등에 활용될 수 있다.

■ 품질검증·선별 등을 거쳐 고품질·고신뢰 데이터 개방

이번 데이터 개방을 위해 NIA·한국정보통신기술협회(회장 손승현, 이하, 'TTA')는 1차 단계평가 완료 후 구축·가공 예산을 통해 확보한 데이터 전체를 각 정예팀으로부터 제공받았으며, 이후 데이터 품질검증, 개인정보·유해성 검증, 라이선스 제약 데이터 제외 등의 조치를 진행하였다.

「독자 AI 파운데이션 모델」 사업 공고상 참여조건에 따라 데이터 구축·가공 예산을 통해 확보한 데이터 중 50% 이상 개방이 필요한 상황에서, 네이버 클라우드, 업스테이지, 에스케이텔레콤, 엔씨에이아이는 품질검증 등을 거친 데이터 전체를 개방하기로 하였다. 엘지AI연구원은 의무개방량(50% 이상)을 준수하여 공개 대상 데이터를 통계적으로 선별·개방*하기로 하였다.

* 세부 데이터셋 별로 일정간격에 따라 통계적 추출(TTA 검증)

이번에 공개된 29종의 AI학습용 데이터는 국내 기업, 연구자, 학생 등 대한민국 국민이라면 누구나 AI허브(aihub.or.kr)를 통해 무료로 다운로드 받아 활용할 수 있다. AI허브의 ‘독자AI모델 데이터’ 항목에서 팀 별, 데이터 종류별로 원하는 데이터를 검색하여 확인할 수 있으며, 일부 데이터는 홈페이지 내 별도 신청(안심존*)을 거쳐 이용할 수 있다.

* 안전한 데이터 활용을 위해 보안이 보장된 온라인 환경을 제공하는 AI모델 개발 공간

과기정통부는 앞으로도 「독자 AI 파운데이션 모델」 고도화를 위한 지원을 지속해 나가는 한편, 향후 2차 단계평가 과정에서 데이터 구축·가공 예산을 통해 구축된 데이터에 대해서도 품질검증 등을 거쳐 개방을 진행할 예정이다.

과기정통부 김경만 인공지능정책실장은 “「독자 AI 파운데이션 모델」 프로젝트는 단순히 AI모델 개발을 넘어, 개발과정에서 축적된 경험과 노하우를 통해 AI생태계 전반의 질적 성장에 기여한다는 점에서 큰 의미”라면서, “오늘 개방된 데이터는 정예팀의 AI학습 전략이 담긴 귀중한 자산인 만큼 AI생태계의 성장과 자생력 강화를 이끄는 밑거름이 될 것”이라고 밝혔다.

담당 부서	인공지능정책실 인공지능데이터정책과	책임자	과 장	이소라 (044-202-6580)
		담당자	서기관	나상민 (044-202-6565)



내일을 만드는 과학기술
내 삶을 채우는 디지털·AI

대한민국
지정브리핑



붙임1

정예팀별 데이터 구축 · 활용 현황 및 개방(안)

정예팀명	데이터셋명	형식	개방량	
네이버 클라우드	공개 영상 데이터	영상	2,340,000건	7.05TB
	공개 영상 데이터	텍스트	9,000,000건	7.5GB
	공개 영상 데이터	음성	6,420,000건	717.5GB
	방송 영상 데이터	영상	520,000건	985GB
	방송 영상 데이터	텍스트	6,500,000건	6GB
	방송 영상 데이터	음성	5,580,000건	709GB
업스테이지	사전학습용 데이터	텍스트	1조 토큰	1.61TB
	지식/지능 데이터	텍스트	300,000건	265MB
	사용자 선호 반영 데이터	텍스트	100,000건	224MB
	에이전트 행위 능력 데이터	텍스트	100,000건	310MB
	벤치마크 데이터	텍스트	2,084건	2MB
SKT	AI 파운데이션 모델(LLM/LAM) 사후 학습용 데이터	텍스트	63,000건	328MB
	LMM(speech, text) 학습용 데이터	음성	7,845건	127GB
	LMM(speech, text) 학습용 데이터	텍스트	7,845건	30MB
	LMM(image, text) 학습용 데이터	이미지	5,000건	7GB
	LMM(image, text) 학습용 데이터	텍스트	5,000건	7MB
	AI 파운데이션 모델(LLM/LAM) 레드티밍 데이터	텍스트	10,000건	40MB
NC AI	고품질 멀티턴 데이터 집합	텍스트	61,641건	226MB
	Synthetic & Instruction 데이터 집합	텍스트	858,960건	1.7GB
	생각과정 씨앗 학습/검증 데이터	텍스트	55,976건	1.1GB
	다국어 사전학습 데이터	텍스트	2,504,351건	8.63GB
	다국어 기계독해 데이터	텍스트	200,000건	370MB
	AI Safety Alignment 데이터	이미지	10,000건	11MB
	대화생성형AI 학습용 데이터	이미지	179,054건	5.33GB
	행동생성형AI 학습용 데이터	이미지	145,452건	5.3GB
	의료 AI 학습용 데이터	이미지	307,842건	28.6GB
	민원 분야 대화 음성 및 영상	음성	6,868건	35.4GB
LG AI	휴머노이드 로봇을 위한 Scene Understanding 데이터셋	이미지	151,952건	38GB
	휴머노이드 로봇을 위한 Scene Understanding 데이터셋	텍스트	1,302건	0.5GB
			35,444,173건	11.3TB

AI 데이터 찾기
AI 서비스 소개
리더보드
커뮤니티
AI 개발자원
고객지원
로그인
회원가입

[홈](#) |
 [AI 데이터 찾기](#) >
 독자 AI 모델 데이터

전체 ▼

검색어를 입력해주세요.

검색 조건 열기 ▼

🔍 검색

전체

한국어

영상이미지 멀티모달

LG AI 연구원

업스테이지

SKT

네이버 클라우드

NC AI

※ 데이터 다운로드 PC에서만 가능합니다.

전체 ▼

최신순 ▼

20 ▼

조회

📄 다운로드

한국어

Synthetic & Instruction
데이터 집합

305
 0
 3

📄 다운로드

한국어

고품질 멀티턴 데이터 집
합

161
 0
 4

📄 다운로드

영상이미지 멀티모달

방송 영상 데이터

133
 1
 3

📄 다운로드

영상이미지 멀티모달

공개 영상 데이터

106
 0
 2

+ 신청하기

한국어
(인식론(문리언))

AI Safety Alignment 데
이터

138
 0
 3

📄 다운로드

한국어

다국어 기계독해 데이터

136
 0
 3

📄 다운로드

한국어

다국어 사전학습 데이터

132
 0
 2

📄 다운로드

한국어

생각과정 씨앗 학습 데이
터

176
 0
 5

📄 다운로드

영상이미지 멀티모달

홍 휴머노이드 로봇용
Scene Understanding...

140
 0
 4

📄 다운로드

한국어

사전 학습 데이터

140
 0
 2

📄 다운로드

한국어

벤치마크 데이터

133
 0
 4

📄 다운로드

한국어

에이전트 행위 능력 데이
터

144
 0
 6