

과기정통부, 초거대AI 데이터 인프라 확충한다

- '23년도 「인공지능학습용데이터 구축사업」 2,805억원 투입, 150종 구축 공모
- 초거대AI 등 거대언어모델 지원을 위한 말뭉치 데이터 34종 구축 추진
 - 문화·관광, 교육, 법률 등 다양한 산업분야별 데이터 확대

과학기술정보통신부(장관 이종호, 이하 과기정통부)는 4월 14일(금)부터 5월 19일(금)까지 「인공지능 학습용데이터 구축사업(이하 구축사업)」 참여사업자를 공모한다고 밝혔다.

의료진단, 자율주행, 챗봇, 음성 인식 등 다양한 인공지능 모델 또는 서비스는 데이터 학습을 통해서 구현된다. 대량의 학습데이터는 인공지능의 성능향상과 직결되는 핵심요소이나, 데이터 수집·가공에는 상당한 시간과 비용이 필요하다. 국내 중견·중소기업, 스타트업, 대학 등은 AI 개발·도입을 가로막는 주요 저해 요인으로 학습용데이터 부족 문제를 지속적으로 제기하고 있는 실정이다.

이에 과기정통부는 '17년부터 언어, 이미지, 영상, 음성, 센서 등 다양한 유형의 인공지능 학습데이터를 구축하여 누구나 손쉽게 활용할 수 있도록 지원하고 있다. 현재 AI 허브(aihub.or.kr)를 통해 한국어, 헬스케어, 교통 등 6대 분야 381종의 데이터를 제공 중이며, 이달 말부터는 지난해 구축한 310종을 품질검증을 거쳐 순차적으로 추가 개방할 예정이다.

금년도 구축사업에는 총 2,805억원의 예산을 투입하여 150종의 신규 데이터를 구축한다. 우선 지난해 수요조사를 통해 접수된 약 1천 4백 건의 후보데이터를 토대로 기획된 130종의 데이터는 지정과제 형태로 추진한다. 지정과제는 '22년 1월 수립한 「AI 학습용 데이터 구축·활용 고도화 방안」에 따라 선정된 14대

동 사업은 인공지능 학습데이터를 활용하는 전문기업과 데이터 수집·보유·가공 등의 구축 역량을 갖춘 2개 이상의 기업·기관 등이 컨소시엄을 구성하여 참여할 수 있다. 총 114개의 신규 주관기관을 선정할 예정이며, 선정 기관에는 데이터 종당 구축비용으로 1년간 13.5억원~17억원이 지원된다. 사업 공모와 관련된 상세한 안내는 AI허브(aihub.or.kr), 과학기술정보통신부(www.msit.go.kr)와 한국지능정보사회진흥원(www.nia.or.kr)의 누리집을 통해 확인할 수 있다.

박윤규 제2차관은 “방대한 양의 데이터를 배워 사람처럼 추론하는 초거대 AI 등 새로운 AI 기술이 등장하면서 양질의 대규모 데이터 확보는 어느 때보다 중요해지고 있다.”며, “혁신적인 아이디어를 가진 많은 기업들이 인공지능 도입·개발에 보다 적극적으로 도전할 수 있도록 고품질의 데이터 인프라를 조성함으로써, 우리 국민들이 생활 속 구석구석에서 더욱 다양한 인공지능 혜택을 누릴 수 있는 기반을 다져 나가겠다.”고 밝혔다.

담당 부서	인공지능정책과	책임자	팀 장	최우석 (044-202-6280)
	인공지능확산팀	담당자	사무관	박수정 (044-202-6281)



분야 118종의 데이터와 '22년도 계속과제 12종으로 구성된다. 나머지 20종은 자유과제로 지역특화, 자유주제 등 일반영역과 초거대AI 영역으로 나눠 선정할 예정이다.

- * 범용적 성격의 인공지능 기반기술 및 2대 전략 분야 중심으로 종적 횡적 구축 확대 : 기반기술(한국어, 영상·이미지), 전략분야1(헬스케어, 교통·물류, 재난·안전·환경, 농·축·수산), 전략분야2(제조, 로보틱스, 문화·관광, 스포츠, 교육, 법률, 금융, 지식재산 등)

특히 금년에는 챗GPT와 같은 초거대AI에 기반한 대화형 모델 지원을 위한 데이터 구축도 추진한다. 기존의 지도학습용 데이터(라벨링데이터) 물량은 일부 축소하고, 한국어 21종 등을 포함한 총 34종의 데이터를 '초거대AI 지원 데이터'로 구분하여 자가지도학습에 필요한 비라벨링 데이터 구축을 확대한다. 또한 이미지 데이터 48종은 생성형AI 등의 학습에도 활용할 수 있도록 이미지를 묘사·설명하는 캡션 정보를 추가로 포함한다. '24년도부터는 사업개편을 통해 초거대AI 데이터 구축을 별도 트랙으로 신설하여 본격 지원할 예정이다.

아울러 그간 적용해 온 수행기관의 인건비 편성요건*을 폐지하고 자동라벨링, 합성데이터('22년 1종 → '23년 5종 내외) 등 신기술 도입 적용을 확대한다. 문화·관광, 스포츠, 법률 등 산업특화 분야에서의 인공지능 도입을 유도하기 위하여 관련 데이터 구축 비중을 '22년도 19%→'23년도 25%로 늘린다. 또한 국회·소방 등 공공분야에서 보유한 데이터의 민간이용 확산을 지원하기 위한 공공분야 학습데이터 구축(5종)**도 구축도 강화한다.

- * '국고지원금의 70% 이상 인건비 편성' 시 수행기관 선정단계에 가점 부여
- ② 민간부담금 산출 시 정부지원금에서 클라우드소싱 인건비를 제외

** 법률안 검토 보고서 요약 데이터, 119 지능형 신고접수 음성 인식 데이터 등

< '23년도 인공지능 학습용 데이터 구축사업 공모과제 >

구분	23년	주요내용
지정	130종 (신규과제 94개)	한국어(24종), 영상·이미지(21종), 헬스케어(14종), 교통·물류(7종), 농·축·수산(13종), 재난·안전·환경(18종), 문화·관광(13종), 스포츠(2종), 제조(3종), 로보틱스(3종), 교육(5종), 법률(3종), 금융(2종), 지식재산(2종) ※ '22-23년도 계속과제(12종)는 공모에서 제외, 유사데이터는 1개 과제로 통합공모
자유	20종 (신규과제 20개)	일반(15, 자유주제 2, 지역특화 12(자주제 지역 참여 필수))
계	150종	초거대AI(6, 일반·해외진출 말뭉치 각 3종)

붙임1 '23년도 인공지능학습용데이터 구축사업 신규과제

분야	과제명	데이터명
한국어·문화·관광	1 라이브 스트리밍 영상 통번역 데이터	1 라이브 스트리밍 영상 영어 통번역 데이터 2 라이브 스트리밍 영상 중국어 통번역 데이터 3 라이브 스트리밍 영상 일본어 통번역 데이터
	2 다국어 일상대화 실시간 통번역 데이터	4 다국어 일상대화 실시간 통번역 데이터
	3 국제 학술회의용 전문분야 한영/영한 통번역 데이터	5 국제 학술회의용 전문분야 한영/영한 통번역 데이터
	4 K-콘텐츠 멀티턴 한-영 페르소나 데이터	6 K-콘텐츠 멀티턴 한-영 페르소나 데이터
	5 다국어 기계번역 차명적 오류 탐지 및 교정 데이터	7 다국어 기계번역 차명적 오류 탐지 및 교정 데이터
	6 다국어 번역 품질 평가 데이터	8 다국어 번역 품질 평가 데이터
	7 분야별 한국어 멀티세션 데이터	9 전문분야(기술과학) 한국어 멀티세션 데이터 10 전문분야(사회과학) 한국어 멀티세션 데이터
	8 한국어 SNS 멀티턴 대화 데이터	11 일상대화 한국어 멀티세션 데이터 12 한국어 SNS 멀티턴 대화 데이터
	9 동화 데이터	13 동화 이해도 테스트를 위한 질의응답형 생성 데이터 14 동화 삽화 생성 데이터(문화관광) 15 동화 줄거리 생성 데이터(문화관광)
	10 다양한 소용원에서의 Ground Truth 지식 정보 데이터	16 다양한 소용원에서의 Ground Truth 지식 정보 데이터
	11 한국어 텍스트·비디오·사운드 데이터	17 한국어 텍스트·비디오·사운드 데이터
	12 음성인식에 의한 영상 요약 데이터	18 음성인식에 의한 영상 요약 데이터
	13 한국어 지식그래프 및 다중 이벤트 추출 데이터	19 지식그래프-투-텍스트 데이터 20 한국어 다중 이벤트 추출 데이터
	14 AI 응답 결과에 대한 품질 평가 데이터	21 AI 응답 결과에 대한 품질 평가 데이터
	15 법률안 검토 보고서 및 국회 회의록 데이터	22 법률안 검토보고서 요약 데이터 23 국회 회의록 기반 지식 검색 데이터
	16 아동·청소년 상담 데이터	24 아동·청소년 상담 데이터
영상·이미지	17 고품자 근현대 경험 기반 스토리 구술 데이터	25 고품자 근현대 경험 기반 스토리 구술 데이터
	18 한국인 얼굴 3D 데이터	26 3차원 안면 데이터 27 한국인 표정 3D 데이터
	19 한국인 피부상태 측정 데이터	28 한국인 피부상태 측정 데이터
	20 물체 조작 손 동작 3D 데이터	29 물체 조작 손 동작 3D 데이터
	21 한국인 전신 및 포즈 데이터	30 한국인 전신 및 포즈 데이터
	22 표/차트 이미지 해석 데이터	31 차트 이미지-텍스트 쌍 데이터 32 표 이미지-텍스트 쌍 데이터
	23 다중 언어 OCR 데이터	33 다중 언어 OCR 데이터
	24 한국어 GQA 데이터	34 한국어 GQA 데이터
	25 한국형 텍스트-3D 객체 쌍 데이터	35 한국형 텍스트-3D 객체 쌍 데이터
	26 한국 전통 수묵 채색화 제작 데이터	36 한국 전통 수묵 채색화 제작 데이터
	27 의류 스케치-패턴 도면 쌍 데이터	37 의류 스케치-패턴 도면 쌍 데이터
	28 음성 3D 데이터	38 음성 3D 데이터
	29 가상 실내 공간 3D 합성 데이터	39 가상 실내 공간 3D 합성 데이터

분야	과제명	데이터명
	30 교량 3D 외관점검 영상 데이터	40 교량 3D 외관점검 영상 데이터
	31 낙상사고 위험동작 영상-센서 쌍 데이터	41 낙상사고 위험동작 영상-센서 쌍 데이터
	32 스트레스 상황 바이오 이미지 복합 데이터	42 스트레스 상황 바이오 이미지 복합 데이터
	33 대용량 동영상 콘텐츠 데이터 (고도화)	43 대용량 동영상 콘텐츠 데이터 (고도화)
헬스케어	34 혈관 데이터	44 경동맥 혈관 CT 및 MRI 데이터
		45 말초 혈관 초음파 데이터
		46 경추 CT 데이터
	35 관절 및 관절염 데이터	47 무릎 관절염 켈그렌-로렌스 분류(K-L grade) 데이터
		48 족부측관절 체중부하 CT 데이터
	36 난소 및 배아 데이터	49 난소암 데이터
		50 배아 이미지 합성데이터
	37 내시경 이미지 합성데이터	51 내시경 이미지 합성데이터
	38 구강 이미지 합성데이터	52 구강 이미지 합성데이터
	39 노인 정신건강 영상 데이터	53 노인 정신건강 영상 데이터
	40 뇌출중 경과관찰 데이터	54 뇌출중 경과관찰 데이터
	41 비뇨기계 암 병리이미지 데이터	55 비뇨기계 암 병리이미지 데이터
교통물류	42 약물유도 수면내시경 영상 데이터	56 약물유도 수면내시경 영상 데이터
	43 열화상 카메라 융합 자율주행 데이터	57 열화상 카메라 융합 자율주행 데이터
	44 생활도로 객체인식 자율주행 데이터	58 생활도로 객체인식 자율주행 데이터
	45 전국 도로시설물 영상정보 데이터	59 전국 도로시설물 영상정보 데이터
	46 고해상도 도로노면 이미지 데이터	60 고해상도 도로노면 이미지 데이터
	47 특이 도로 환경 주행 데이터 (고도화) -비포장도로환경주행데이터	61 특이 도로 환경 주행 데이터 (고도화) -비포장도로환경주행데이터
	48 과수원 내 로봇 주행 데이터	62 과수원 내 로봇 주행 데이터(포도, 감귤 류)
		63 과수원 내 로봇 주행 데이터(사과, 배 류)
농축수산	49 식용곤충 사육 자동화 데이터	64 식용곤충 사육 자동화 데이터
	50 원예식물(화분류) 물주기(수분공급 주기) 생육데이터	65 원예식물(화분류) 물주기(수분공급 주기) 생육데이터
	51 국내 재배 아열대열대작물 질병/해충 데이터	66 국내 재배 아열대열대작물 질병/해충 데이터
	52 잔디 생육환경 다분광 이미지 데이터	67 잔디 생육환경 다분광 이미지 데이터
	53 풀벌 질병 진단 이미지 데이터	68 풀벌 질병 진단 이미지 데이터
	54 말뼉 부위 식별 및 이상상태 진단 이미지 데이터	69 말뼉 부위 식별 및 이상상태 진단 이미지 데이터
	55 해조류 및 저서물질 초분광 영상 데이터	70 해조류 및 저서물질 초분광 영상 데이터
	56 지능형 스마트 미꾸리 종자생산 시스템 및 양식 데이터	71 지능형 스마트 미꾸리 종자생산 시스템 및 양식 데이터
	57 해파리 초기유생 데이터	72 해파리 초기유생 데이터
	58 유용식물 유전체 데이터	73 유용 식물 유래 화합물 및 생리활성 데이터
		74 유용식물 유전체 및 전사체 데이터
	59 건물 균열 탐지 이미지 (고도화) -SOC시설물균열패턴이미지데이터	75 건물 균열 탐지 이미지 (고도화) -SOC시설물균열패턴이미지데이터
	60 도로 지상 시설물 안전 관리를 위한 GPR 데이터	76 도로 지상 시설물 안전 관리를 위한 GPR 데이터
재난안전환경	61 제조시설 안전 데이터	77 무인 플랫폼 안전 감시를 위한 데이터
		78 스마트 제조 시설 안전 감시를 위한 데이터
	62 식생탄소 포집 및 탄소흡수원 데이터	79 자연기반 탄소흡수원 시계열 데이터
		80 권역별 식생 탄소 포집량 식별 데이터

분야	과제명	데이터명
	63 공사현장 안전 데이터	81 선박 해양플랜트 스마트 야드 안전 데이터
		82 전기 설비 공사 현장 위험 상태 판단 데이터
	64 수변구역 변화 탐지 데이터	83 수변구역 변화 탐지 데이터 (한강 유역)
		84 수변구역 변화 탐지 데이터 (낙동강 유역)
	65 어린이 보호구역 내 등하교 및 시설물 영상 데이터	85 어린이 보호구역 내 등하교 및 시설물 영상 데이터
	66 놀이기구 및 시설 이용자 위험 상황 인식 데이터	86 놀이기구 및 시설 이용자 위험 상황 인식 데이터
	67 화확물질(유전독성) 유해성 예측 데이터	87 화확물질(유전독성) 유해성 예측 데이터
	68 위급상황 음성/음향 (고도화) -119지능형신고접수음성인식데이터	88 위급상황 음성/음향 (고도화) -119지능형신고접수음성인식데이터
	69 화재 발생 예측 영상 (고도화) -영상기반화재감시및발생위치탐지데이터	89 화재 발생 예측 영상 (고도화) -영상기반화재감시및발생위치탐지데이터
	70 산업 설비 전기 화재 사고 예방 부분방전 데이터	90 산업 설비 전기 화재 사고 예방 부분방전 데이터
	71 가변압기 주조 시료를 이용한 일반 등급 분류 데이터	91 가변압기 주조 시료를 이용한 일반 등급 분류 데이터
	72 하천 유입 농축산계 비점오염원 이미지 데이터	92 하천 유입 농축산계 비점오염원 이미지 데이터
문화관광	73 전통 한복 3D 데이터	93 전통 한복 3D 데이터
	74 관광 특화 맞춤형 데이터	94 관광 특화 맞춤형 데이터
	75 만화웹툰 데이터	95 만화웹툰 생성 데이터
		96 만화웹툰 인물 및 장면 이미지 데이터
	76 탈춤 동작 데이터	97 탈춤 동작 데이터
	77 한국 대중음악 루프 사운드 생성 데이터	98 한국 대중음악 루프 사운드 생성 데이터
	78 관광분야 이미지-텍스트 쌍 데이터	99 관광분야 이미지-텍스트 쌍 데이터
스포츠	79 한국 씨름 동작 데이터	100 한국 씨름 동작 데이터
	80 야구 주요 규칙 판정 영상 데이터	101 야구 주요 규칙 판정 영상 데이터
	81 건설기계 무인 운행 데이터	102 건설기계 무인 운행 데이터
제조	82 배터리 불량량 이미지 데이터	103 배터리 불량량 이미지 데이터
	83 부품 품질 검사 영상 데이터(산박/해양플랜트) (고도화) -LNG탱크품질검사영상데이터	104 부품 품질 검사 영상 데이터(산박/해양플랜트) (고도화) -LNG탱크품질검사영상데이터
	84 물품 적재 및 로봇 행동 데이터	105 로봇 행동 데이터(3D 물건 적재)
로봇/틱스		106 로봇 행동 데이터(소형객체 적재)
	85 로봇 관점 주행 영상 (고도화) - 소셜 내비게이션 로봇 주행	107 로봇 관점 주행 영상 (고도화) - 소셜 내비게이션 로봇 주행
	86 진로문장완성검사 텍스트 데이터	108 진로문장완성검사 텍스트 데이터
교육	87 학습태도 및 성향 관찰 데이터	109 학습태도 및 성향 관찰 데이터
	88 수학 문제 데이터	110 수학 과목 자동 풀이 데이터
		111 수학 과목 문제생성 데이터
	89 공적말하기 실습 및 평가 데이터	112 공적말하기 실습 및 평가 데이터
법률	90 법률 정보 및 다국어 번역 데이터	113 국내 법률 대상 다국어 번역 데이터
		114 법률 지식기반 관계 데이터
	91 법률/규정 텍스트 분석 데이터 (고도화) -상황에따른판례데이터	115 법률/규정 텍스트 분석 데이터 (고도화) -상황에따른판례데이터
금융	92 금융 분야 다국어 병렬 말뭉치 데이터	116 금융 분야 다국어 병렬 말뭉치 데이터
	93 금융 합성데이터	117 금융 합성데이터
지식재산	94 한·중/한·일 특허 및 기술과학 분야 병렬 말뭉치 데이터	118 한·중/한·일 특허 및 기술과학 분야 병렬 말뭉치 데이터