SHIN&KIM
법무법인(유) 세종

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

2026.09.03

과학기술정보통신부(이하 '과기정통부')와 한국지능정보사회진흥원(이하 'NIA')은 지난 8월 27일 「독자 AI 파운데이션 모델」 프로젝트에 참여한 5개 정예팀이 1차 단계평가 과정에서 구축한 총 29종(약 3,544만 건, 약 1.56조 토큰 규모)의 AI 학습용 데이터를 품질검증 등을 거쳐 'AI 허브'(aihub.or.kr)에 개방한다고 밝혔습니다.

아래에서는 이번 데이터 개방의 배경과 주요 내용, 그리고 그 시사점 등에 대해 살펴보도록 하겠습니다.

1. 데이터 개방의 배경 및 개요

5개 정예팀은 각자의 개발 전략에 맞춰 확보할 데이터의 성격 및 범위를 기획하고 이에 따라 구축된 AI 학습용 데이터를 개방하였습니다. 1차 평가단계 데이터 구축에는 정부의 데이터 구축·가공 예산('25년, 150억 원)이 투입되었으며, 사전학습에 필요한 대규모 데이터셋부터 영상·음성 등 멀티모달 데이터, 안전성 확보를 위한 레드티밍 데이터 등으로 구성되어 있습니다.

개방 규모는 토큰으로 환산할 경우 약 1.56조 토큰(추정)에 달하며, 이는 이론적으로 약 70~80B 수준의 대형 AI 모델 학습에 활용될 수 있는 규모입니다. 이번 개방은 정부 예산으로 구축한 데이터를 민간에 환류하여 AI 생태계 전반의 질적 성장과 자생력 강화를 도모한다는 취지로, 'AI 3대 강국 도약을 위한 AI 고속도로 구축', '세계에서 AI를 가장 잘 쓰는 나라 구현' 등 관련 국정과제의 이행 성격을 함께 갖습니다.

2. 개방 데이터의 주요 내용

각 정예팀별로 구축한 데이터의 주요 내용과 활용 분야는 다음과 같습니다.

정예팀	주요 구축 데이터	주요 활용 분야
네이버클라우드팀	공개 영상(234만 건)·방송 영상(52만 건)과 이를 기반으로 한 텍	영상 이해능력 학습, 이미지 생성 모

	스트(1,550만 건)·음성 질의응답(1,200만 건). 텍스트는 장면을 문장 단위로 기술한 캡션 등으로 구성	델 학습
업스테이지팀	1조 토큰 규모의 대규모 사전학습 데이터, 사후학습 데이터(50만 건)	대규모 AI 모델 신규 학습(from scratch), 추론·판단·실행 등 AI 에이전트 개발
에스케이텔레콤팀	수학·과학·법률 등 전문 도메인의 고난도 다단계 추론데이터, 음성·이미지 등 사후학습용 데이터, 국내 법령과 한국 사회의 보편적 가치관을 반영한 한국형 레드티밍 데이터(약 1만 건)	실무 해결역량 강화, 안전성·신뢰성 향상
엔씨에이아이팀	제조 분야 기술문서, 민원 상담 음성 등 현장 데이터에 기반한 긴 문맥 이해·단계별 추론·멀티턴 대화·질의응답·멀티모달리티 등 LLM 핵심 역량 데이터 7종	사후학습 성능 고도화, 텍스트-이미지-음성 통합학습
엘지AI연구원팀	한국 가정환경에 최적화된 피지컬 AI 멀티모달 데이터셋(국내 50개 실제 가정환경에서 50종 이상 가사활동 촬영, 17만 개 이상 영상 클립을 객체·분할·자세·맥락 정보 등으로 다층 라벨링)	비전·VLM 학습, 연구·상용서비스 개발

3. 품질검증 절차 및 데이터 이용 방법

NIA와 한국정보통신기술협회(이하 'TTA')는 1차 단계평가 완료 후 정예팀으로부터 구축·가공 예산을 통해 확보한 데이터 전체를 제공받아, 데이터 품질검증, 개인정보·유해성 검증, 라이선스 제약 데이터 제외 등의 조치를 진행하였습니다. 사업 공고상 참여조건에 따라 확보 데이터 중 50% 이상을 개방하여야 하는 가운데, 네이버클라우드·업스테이지·에스케이텔레콤·엔씨에이아이는 품질검증 등을 거친 데이터 전체를 개방하였고, 엘지AI연구원은 의무개방량(50% 이상)을 준수하는 범위에서 공개 대상 데이터를 통계적으로 선별하여 개방하였습니다. 이번에 공개된 29종의 데이터는 국내 기업·연구자·학생 등 대한민국 국민이라면 누구나 AI 허브(aihub.or.kr)의 '독자AI모델 데이터' 항목에서 팀별·데이터 종류별로 검색하여 무료로 내리받아 활용할 수 있으며, 일부 데이터는 보안이 보장된 온라인 개발 환경인 '안심존'을 통한 별도 신청 절차를 거쳐 이용할 수 있습니다.

4. 향후 계획

과기정통부는 앞으로도 「독자 AI 파운데이션 모델」 고도화를 위한 지원을 지속하는 한편, 2차 단계평가 과정에서 데이터 구축·가공 예산을 통해 구축되는 데이터에 대해서도 품질검증 등을 거쳐 개방을 진행할 예정입니다. 과기정통부 김경만 인공지능정책실장은 "「독자 AI 파운데이션 모델」 프로젝트는 단순히 AI 모델 개발을 넘어, 개발 과정에서 축적된 경험과 노하우를 통해 AI 생태계 전반의 질적 성장에 기여한다는 점에서 큰 의미가 있다"라면서, "오늘 개방된 데이터는 정예팀의 AI 학습 전략이 담긴 귀중한 자산인 만큼 AI 생태계의 성장과 자생력 강화를 이끄는 밑거름이 될 것"이라고 밝혔습니다.

5. 시사점

(양질의 AI 학습데이터를 무상으로 확보할 수 있는 기회) 대규모·고품질 학습데이터의 확보는 그간 AI 개발의 핵심 애로 사유로 지적되어 왔습니다. 이번 개방으로 국내 기업·연구자·학생은 사전학습용 대규모 데이터셋부터 멀티모달·추론·레드티밍 데이터에 이르기까지 다양한 데이터를 무료로 활용할 수 있게 되었습니다. 특히 자체적으로 데이터를 구축할 여력이 충분치 않은 스타트업·중소기업 등은 자사의 개발 방향에 부합하는 데이터셋이 있는지 확인하고 이를 자사의 AI 모델 개발에 적극 활용함으로써 개발에 소요되는 시간과 비용을 크게 줄일 수 있을 것으로 기대됩니다.

(데이터 활용 시 이용조건·라이선스 등 준수사항의 사전 확인 필요) NIA·TTA가 개인정보·유해성 검증 및 라이선스 제약 데이터 제외 등의 절차를 거쳤으나, 실제 활용 단계에서는 AI 허브의 이용약관과 각 데이터셋별 이용조건을 확인할 필요가 있습니다. 또한 일부 데이터는 '안심존'을 통해서만 이용이 가능한 만큼, 데이터의 상업적 이용 가능 여부, 재배포·2차적 활용의 허용 범위, 이용 방식 등을 사전에 점검하여 향후 분쟁이나 규제 리스크를 예방할 필요가 있습니다.

(민·관 협력을 통한 정부 주도 데이터 공유 흐름의 확산) 이번 개방은 정부 예산으로 구축한 데이터를 민간에 환류하여 생태계 전반의 성장을 도모하는 것으로, 정부가 표준·데이터를 마련하고 산·학·연이 이를 공유·활용하도록 지원하는 최근의 정책 흐름과 궤를 같이합니다. 2차 단계평가 과정에서 구축되는 데이터의 추가 개방도 예정되어 있는 만큼, 관련 기업으로서는 개방 데이터의 확대 동향을 지속적으로 주시하면서 활용 방안을 모색할 필요가 있습니다.

(AI 안전성·신뢰성 관련 데이터가 포함된 점에 유의) 이번 개방 데이터에는 한국형 레드티밍 데이터, AI Safety Alignment 데이터 등 AI의 안전성·신뢰성 확보를 위한 데이터가 포함되어 있습니다. AI의 안전성과 신뢰성이 정책적으로 강조되는 흐름 속에서, 이러한 데이터는 기업이 자사 AI 서비스의 안전성 확보 체계를 정비하는 데에도 참고가 될 수 있으므로 그 활용 가능성을 함께 검토해 볼 필요가 있습니다.

About Shin & Kim's ICT Group

법무법인(유) 세종 ICT그룹은 ICT 분야에서 독보적인 전문성과 인적 네트워크를 보유하고 있으며, 고객들로부터 최근 수년간 가장 높은 평가를 받고 있습니다. 방송과 통신, 개인정보, 인터넷 IT 분야에서 축적된 역량을 바탕으로 방송·통신·ICT 규제 동향 파악 및 대관, 법제개선·입법컨설팅, 규제영향 분석과 기업의 전략 수립 등에 대한 종합적인 법률자문을 제공하고 있으며, AI Compliance, 데이터 활용 및 계약, 침해사고 대응 등과 관련하여서도 다양한 업무경험과 전문성을 보유하고 있습니다. 보다 전문적인 내용이나 궁금하신 사항이 있으면 언제든지 연락 주시기 바랍니다.

관련구성원

강신욱
대표변호사

02-316-4059
sokang@shinkim.com

노진홍
변호사

02-316-1639
jhnoh@shinkim.com

김현이
변호사

02-316-4070
hiekim@shinkim.com

강지현
변호사

02-316-1518
jhykang@shinkim.com

정소정
변호사

02-316-1877
sjjeong@shinkim.com