NIA 한국지능정보사회진흥원	보 도 자 료	
보 도 일 시	온라인 배포시점부터 보도하여 주시기 바랍니다.	
배 포 일 시	담당팀	AI데이터확산팀
2025. 8. 8.(금) 09:00	담 당 자	신선영 팀장 (053-230-4281) 박수영 수석(053-230-4214)

한국지능정보사회진흥원-인공지능안전연구소, AI 모델 안전성 확보 시동

- LLM 안전성 및 신뢰성 평가 데이터 구축 사업 착수보고회 개최 -

□ 한국지능정보사회진흥원(원장 황종성, 이하 진흥원)과 과학기술정보통신부(장관 배경훈, 이하 과기정통부)는 AI 모델의 안전성 및 신뢰성 평가 기반 구축을 위한 「LLM 안전성 및 신뢰성 평가 데이터셋 구축」 사업 착수보고회를 8월 7일(목) 진흥원 서울사무소에서 개최하였다.

□ 이 사업은 진흥원과 과기정통부가 추진 중인 '독자 AI 파운데이션 모델 프로젝트'를 통해 구축되는 **AI 모델 등에 대한 안전성과 신뢰성 성능을 평가하기 위한 기반을 마련**하는 것이다. 특히 AI 모델의 위험 요소에 대한 체계적이고 정량적인 검증 체계를 구축하는데 중점을 두고 있다.

□ 진흥원은 그동안 AI 허브를 통한 대규모 AI 학습데이터 구축, AI 바우처 사업을 통한 중소기업 AI 도입 지원, 신뢰할 수 있는 인공지능 실현을 위한 다양한 정책 연구 등을 통해 국내 AI 생태계 발전을 주도해왔다.

○ 이번 사업을 통해 진흥원은 AI 모델의 성능 평가를 넘어 안전성과 신뢰성까지 아우르는 종합적인 AI 평가 체계를

구축하며, 글로벌 AI 거버넌스 논의에서 한국의 위상을 높이는 역할을 담당하게 된다.

○ 특히 이번 과제에서는 **사이버 보안, AI 자율성, 사실 기반 정확성, 사회적 가치 편향** 등 AI 안전성과 신뢰성 핵심 항목을 아우르는 총 2만 건 이상의 한국어 특화 벤치마크 데이터셋을 구축할 계획이며, 글로벌 기준에 부합하는 AI 모델 평가 체계로 발전시킬 예정이다.

○ 이날 착수보고회에는 주무부처인 과기정통부를 비롯하여 진흥원과 데이터 구축을 수행하는 인공지능안전연구소(소장 김명주) 관계자가 참석하여 데이터 구축 방향, 평가 체계 설계, 글로벌 연계 전략 등을 논의하였다.

○ 인공지능안전연구소 김명주 소장은 “이번 사업은 단순 성능 비교를 넘어, AI가 우리 사회에서 신뢰받으며 활용될 수 있는 기준을 세우는 첫 시도”라며, “인공지능안전연구소가 공정하고 객관적인 평가체계를 만들어가며, 국내외 AI 신뢰성 생태계를 선도해 나가겠다”고 말했다.

□ **진흥원 황종성 원장**은 “AI 안전성 평가는 선택이 아닌 필수가 된 시대“라며, “진흥원이 그동안 AI 허브를 통해 축적한 1,300여 종의 AI 학습데이터와 평가 체계 전문성을 총동원해, 세계 최고 수준의 AI 안전성 검증 체계를 구축하겠다. 이를 통해 대한민국이 글로벌 AI 신뢰성 표준을 선도하는 국가로 자리매김할 것”이라고 강조했다. <끝>

붙임. 사진자료

붙 임 사진자료

지난 7일 LLM 안전성 및 신뢰성 평가 데이터 구축 사업 착수보고회가 과기정통부 이소라 과장(왼쪽에서 다섯 번째)과 인공지능안전연구소 김명주 소장(왼쪽에서 네 번째) 등이 참석한 가운데 개최되었다.