과학기술정보통신부 보도자료

보도시점 2025. 12. 29.(월) 12:00 (2025. 12. 30.(화) 조간) 배포 2025. 12. 29.(월) 09:00

인공지능 모델 안전성 첫 평가, AI 안전 생태계 확산의 첫걸음

- 과기정통부·AI안전연구소·한국정보통신기술협회, 국내 연구진이 개발·구축한 데이터셋 기반 국내 최초 AI 안전성 평가 결과 공개

과학기술정보통신부(부총리 겸 과기정통부장관 배경훈, 이하 '과기정통부')는 **인공지능안전연구소**(소장 김명주, 이하 'AI안전연구소'), **한국정보통신기술협회**(회장 손승현, 이하 'TTA')와 함께 **카카오**의 **AI 모델**인 Kanana를 대상으로 **국내 첫 AI 안전성 평가**를 실시하였다고 밝혔다.

AI 안전성 평가는 **AI 위험 식별·평가**를 통해 **AI 시스템의 안전성**을 **확보**하는 수단 중 하나이다. **과기정통부**는 내년 1월로 예정된 AI기본법 시행을 앞두고, **기업**의 **고성능 AI 모델**에 대한 **안전성 확보 컨설팅**을 **지원**하고 있다.

이번 **안전성 평가**는 **AI안전컨소시엄***에 참여 중인 **카카오** 측과의 협의를 거쳐 **카카오**의 AI 모델인 'Kanana Essence 1.5'를 대상으로 실시했다. 평가에는 **TTA·카이스트**(최호진 교수 연구팀)가 지난 **11월 18일 공개**한 **AssurAI 데이터셋**(붙임2)과 **AI안전연구소**의 **고위험 분야 평가 데이터셋**(붙임3)이 사용되었으며, 폭력, 차별적 표현 등 **일반적**인 **위험 요소**부터 무기, 보안 등 **오남용** 가능성이 높은 **시나리오**까지 적용하여 폭넓게 점검했다. 그 결과, Kanana는 유사 규모의 **글로벌 모델**** 대비 **높은 안전성**을 **확보**한 것으로 나타났다.

* AI안전연구소 및 국내 24개 산·학·연으로 구성되는 컨소시엄으로, AI 안전 분야 정책·평가·연구 관련 의제를 논의

** (美) Llama 3.1, (佛) Mistral 0.3을 대상으로 동일하게 평가 후 비교

한국어를 기반으로 35개 위험 영역을 평가하도록 설계된 **안전성 벤치마크 데이터셋** AssurAI는 **해외 연구기관**과 **공동 검증·활용**을 위해 지난 **11월 18일 공개**한 바 있다. 향후 **AI안전연구소**가 기 구축한 **고위험 분야 데이터셋**과 함께 **국제 AI안전연구소 네트워크*** 및 **국제표준화**(ISO/IEC JTC 1/SC 42 등) **반영**을 추진해 **글로벌 정합성**을 **제고**할 예정이다.

* 한국, 미국, 일본, 영국 등 10개국으로 구성되며, 안전성 평가 등 AI 안전에 관한 사항을 논의

이번 **AI 안전성 평가 결과**는 **국내 AI 모델**이 **글로벌 수준의 안전성**을 확보하고 있다는 것을 보여주는 데 의의가 있다. 이번 평가를 기반으로, **'26년**에는 '**독자 AI 파운데이션 모델**' 프로젝트 1차 단계평가에 참여하는 한편, **국내외 AI 기업**들과의 **협력**을 통해 타 AI 모델 대상으로 **평가 확대**도 **추진**한다.

과기정통부 김경만 인공지능정책실장은 "세계적으로 **AI 안전**에 대한 논의가 규제보다는 **검증**과 **구현**이 강조되는 상황에서 이번 평가는 **국내 AI 모델**의 **안전성 경쟁력**을 **증명**한 **사례**"라며, "국내 AI 모델이 **글로벌 AI 안전성 리더십**을 **주도**해 나가도록 적극 지원하겠다."라고 강조하였다.

한편, 지난 '24년 11월 출범한 **AI안전연구소**는 AI의 **다양한 위험**에 **체계적**으로 **대응**하는 **연구**와 함께 **산·학·연** 간 **AI 안전 인력·인프라 협력 허브** 역할을 수행해오고 있다. 또한 미국, 영국, 일본 등 **주요국 AI 안전 기관**과 **긴밀히 협력**하며, **국제 AI 안전성 기준 수립**과 **표준화 논의**에도 적극 참여함으로써 **국내 AI 산업**의 **안전성**과 **글로벌 경쟁력 강화**를 뒷받침하고 있다.

TTA는 지난 '21년 마련한 「**신뢰할 수 있는 인공지능 개발 안내서**」를 통해 **기업**이 **자발적**으로 **AI 신뢰성**을 **확보**하도록 지원하면서, **국내 최초 생성형 AI 레드팀** 개최('24.4.), **안전성 평가 데이터셋 구축**('24.12.) 등 **AI 안전성 실증 활동**을 강화하고 있다.

담당 부서	과학기술정보통신부 인공지능안전신뢰정책과	책임자	과 장	김국현 (044-202-6290)
		담당자	사무관	신 환 (044-202-6285)
관련 기관	인공지능안전연구소 AI안전평가실	책임자	실 장	남기혁 (031-739-7661)
		담당자	연구원	서 상 (031-739-7666)
	한국정보통신기술협회 AI신뢰성센터	책임자	센터장	신준호
		담당자	팀 장	곽준호

내일을 만드는 과학기술
내삶을 채우는 디지털·AI

대한민국 정책브리핑

OPEN 출처표시
공공누리 공공저작물 자유이용허락

붙임1 Kanana Essence 9.8B 안전성 평가 분석 결과

□ 개요

○ **(목적)** 위험 대응 메커니즘에 대한 AI 안전성 평가를 통해 ①새로운 위험 요소의 발견과 ②위험 대응 방식의 효과성 확인

○ **(대상)** 카카오 "Kanana - Essence - 1.5 - 9.8B - Instruct"

○ **(평가기관)** AI안전연구소 안전평가실, TTA AI신뢰성센터

□ 평가 내용

○ **(평가 방법) 데이터셋**(질문)에 대한 **AI 모델 응답**을 **오픈소스 안전성 평가 도구**인 **'Inspect AI'**를 활용하여 **정량평가**

- **(평가 데이터셋 1)** TTA·KAIST가 공동 구축한 AssurAI 데이터셋*

  * '24년도 과기정통부 사업 '생성형 AI 안전성 평가기반 조성'을 통해 구축 후 공개('25.11.18.), 본 평가에는 데이터셋 중 32종 9,110건을 활용

데이터셋명	AssurAI
데이터셋 규모	11,480건
평가영역	위험요소 35종(폭력 등 불법 행위, 차별적 표현 등)
참고자료	붙임2. AssurAI 데이터셋 위험요소

- **(평가 데이터셋 2)** AI안전연구소에서 구축한 고위험* 분야 데이터셋

  * 특정 프롬프트가 오남용 가능한 출력을 유도할 수 있는 내용 수준의 위험

데이터셋명	Kor-WMDP	Kor-Phishing	Kor-CyberSecEval	Kor-MLCommons
데이터셋 규모	500건	200건	251건	400건
평가영역	CBRN-E* 및 사이버보안 지식수준 (이중용도)	사회공학 및 사이버보안 (오용)	방어 역량 (탈옥)	출력 위해성
참고자료	붙임3. 고위험 분야 안전성 평가 데이터셋 구성			

* 화학·생물학·방사능·핵·폭발물 등 위험 무기 체계

○ **(비교 모델)** 유사한 규모의 타 모델*에 대하여 동일한 평가를 수행한 후 비교하여 Kanana 모델 평가 결과의 유효성 확인

* (美) Llama 3.1(8B), (佛) Mistral 0.3(7B)

□ 평가 결과

○ AssurAI 데이터셋 평가 결과

- **(평가 결과 종합)** Kanana 모델이 가장 높은 안전성 수치를 보였으며, Llama, Mistral 순으로 기록

-	Kanana	Llama	Mistral
종합 평점	3.61	3.13	3.04

○ 고위험 분야 안전성 평가 결과

- **(안전 분류 수준)** 악성 웹 요청행위 및 링크 식별 관련 질의에 대한 평가 결과*, Kanana는 각각 83%, 88%의 분류 정확도 기록

* 웹 관련 악성 행위·링크를 얼마나 잘 걸러내고 회피 가능한지 측정

-	Kanana	Llama	Mistral
분류 정확도(웹요청)	83%	89%	26%
분류 정확도(링크)	88%	90%	46%

- **(출력 위해성)** Kanana는 4개의 세부 분야 위험 질의에 대해 상대적으로 높은 '허용가능한 응답율'*을 기록(평균 48%)

* 답변 내용 중 위험한 발언의 포함 여부를 측정, 발언 내용의 안전 여부 확인

-	Kanana	Llama	Mistral
허용가능 응답율	48%	35%	34%

※ 평가 세부 결과는 AI안전연구소·TTA·카카오 누리집 등에서 확인 가능

붙임2 AssurAI 데이터셋 위험요소 (TTA · KAIST 구축)

Our AI Risk		
#	Risk (in Kor.)	Risk Description
1	Supporting Malicious Organized Groups (악성 단체 지원)	테러리즘, 극단주의, 조직범죄 등 악성 단체의 활동을 미화하거나 지원하는 내용을 형성 (150건)
2	Celebrating Suffering (고통 찬양)	타인이나 자신의 신체적, 정신적 고통을 즐기거나 조장하는 내용을 생성 (150건)
3	Violent Acts (폭력 행위)	윤리와 법률에 어긋나는 폭력 행위에 대한 정보 생성 및 폭력 행위를 조장하는 요소 (180건)
4	Depicting Violence (폭력적 묘사)	폭력 행위가 실현되는 장면이나, 잔인하거나 폭력적인 장면에 대한 묘사 및 창작물 생성 (260건)
5	Weapon Usage & Development (무기 관련)	무기의 제작, 사용, 밀수와 같은 불법적 활동을 조장하거나 폭력적인 목적에 악용될 수 있는 요소 (150건)
6	Military and Warfare (군대, 전쟁)	군사 전략, 전술, 또는 군사 시스템 해킹 등과 관련하여 전쟁법을 위반하거나 대규모 피해를 유발할 수 있는 요소 (280건)
7	Harassment(괴롭힘)	특정 대상을 괴롭히거나 따돌리는 내용이 포함된 창작물 생성 (430건)
8	Hate Speech(혐설)	특정 대상을 혐오하거나 비난하는 내용이 포함된 창작물 생성 (190건)
9	Offensive Language (공격적 언어 표현)	특정 대상에 대한 위협이나 공격적인 언어 표현이 반영된 창작물 생성 (190건)
10	Perpetuating Harmful Beliefs (부정적 고정관념)	특정 대상에 대한 부정적인 고정관념이 반영된 창작물 생성 (300건)
11	Adult Content(성인용 콘텐츠)	성적 이미지, 행위 및 성인용 콘텐츠를 포함하는 내용 (430건)
12	Erotic Content(선정적 콘텐츠)	선정적이거나 성적으로 암시적인 내용을 포함한 콘텐츠 (430건)
13	Non-Consensual Nudity (비동의누드)	당사자의 동의 없이 유포된 누드 이미지나 성적 콘텐츠 (450건)
14	Monetized Sexual Content (성적 콘텐츠 수익창출)	성적 콘텐츠를 수익 창출 목적으로 사용하거나 유통하는 행위 (300건)
15	Endangerment, Harm, or Abuse of Children (아동 유해물)	아동에게 해를 끼치거나 유해한 콘텐츠, 아동 학대를 조장하거나 묘사하는 내용 (430건)
16	Child Sexual Abuse (아동 성적 학대)	아동에 대한 성적 착취 및 학대를 묘사하거나 조장하는 콘텐츠 (390건)
17	Suicidal and Non-suicidal Self-injury (자살, 자해)	자살이나 자해를 조장하거나 이를 묘사하는 콘텐츠 (390건)
18	Political Persuasion (정치적 신조)	정치적 메시지를 전달함으로써 유권자의 의견이나 행동을 변화시킬 수 있는 요소 (100건)
19	Influencing Politics(정치 영향)	여론 조작, 선거 결과 왜곡, 정책 결정에 영향을 줄 수 있는 요소 (100건)
20	Deterring Democratic Participation (민주주의적 참여 거부)	AI 기술 및 알고리즘 내 차별적인 정치 광고와 같이 유권자의 참여 의지를 저해할 수 있는 요소 (100건)
21	Fraud (도용)	불법 행위를 통해 개인 정보와 자산을 탈취하는 데 악용될 수 있는 요소 (300건)
22	Mis/disinformation (허위 정보)	허위 정보를 생성하거나 조장하는 내용, 가짜 리뷰나 가짜 대중 지지와 같은 허위의 온라인 참여를 조장하는 요소 (450건)
23	Sowing Division (편 가르기)	일부 진영에 대해서 옹호적이거나 차별적인 태도, 집단 내부의 갈등과 분열을 초래할 수 있는 요소 (300건)
24	Misrepresentation ((정보를) 불완전하게 전달하는 표현)	생성한 컨텐츠가 마치 사람이 제작한 컨텐츠인 것처럼 보이기 위해서 응답 형태를 변경하거나 혹은 완전히 특정 사람처럼 행동하는 것 (430건)
25	Types of Defamation (명예훼손)	공연히 구체적인 사실이나 허위 사실을 적시하여 어떠한 사람의 사회적 평가를 저하시키는 것 (300건)
26	Discriminatory Activities (차별적 표현)	특정 대상에 대한 차별 및 편향적인 내용을 담은 창작물을 생성 (1000건)
27	Unauthorized Privacy Violations (개인정보 침해)	생성형 AI 활용을 통하여 발생 가능한 다양한 프라이버시 위협 이슈 (900건)
28	Illegal/Regulated Substances (불법 약물)	불법 약물의 사용이나 거래를 촉진하는 콘텐츠 생성 (400건)
29	Illegal Services/Exploitation (불법적 착취)	불법적인 서비스나 착취 행위와 관련된 콘텐츠 생성 (400건)
30	Other Unlawful/Criminal Activities (기타 불법 행위)	기타 불법적이거나 범죄와 관련된 활동 (400건)
31	Increased inequality and decline in employment quality (불평등 심화 및 고용 질 저하)	AI 기술의 활용으로 사회적 불평등을 심화시키거나 고용의 질을 저하시킬 위험 (300건)
32	Economic and cultural devaluation of human effort (경제적, 문화적 가치 하락)	인간 노동의 경제적, 문화적 가치를 AI가 대체함으로써 하락시키는 문제 (300건)
33	Competitive dynamics (경쟁적 역학)	AI 기술이 기존의 산업 경쟁을 왜곡하거나 변화시키는 문제 (300건)
34	Overreliance and unsafe use (과도한 의존, 비안전한 사용)	사용자가 AI에게 과도하게 의존하거나 AI를 오남용 할 수 있게끔 AI의 특성 및 창작물 생성 (150건)
35	Loss of human agency and autonomy (인간의 주체성과 자율성 상실)	사용자가 AI를 사용하는데 있어 주체성과 자율성을 잃게 할 수 있는 AI의 특성 및 창작물 생성 (150건)

붙임3 고위험 분야 안전성 평가 데이터셋 (AI안전연구소 구축)

대분류	세분류	설명
Kor-WMDP	Chem	화학적 무기 제조 지원 관련 객관식형 지문 (**150건**)
	Bio	생물학적 무기 제조 지원 관련 객관식형 지문 (**150건**)
	Cyber	사이버 익스플로잇 구성 보조 및 코드 취약점 분석 지원 관련 객관식형 지문 (시스템, 네트워크, 포렌식, 역공학 (Reverse engineering) 등) (**200건**)
Kor-Phishing	Web	웹 취약점 대상의 익스플로잇 코드 스니펫 관련 논리형 지문 (SQL 인젝션, XSS (Cross-Site Script), Cross-Site Request Forgery 등) (**100건**)
	Link	알려진 악성 피싱 링크 관련 논리형 지문 (웹 검색, 위협 인텔리전스 등) (**100건**)
Kor-Cyberseceval	-	프롬프트 인젝션 관련 지문 (**251건**)
Kor-MLCommons	IPV	지적재산권 관련 지문 (**100건**)
	NCR	비폭력 범죄 관련 지문 (**100건**)
	PRV	개인정보 침해 관련 지문 (**100건**)
	VCR	폭력 범죄 관련 지문 (**100건**)