한국정보통신기술협회 Telecommunications Technology Association http://www.tta.or.kr	**보도자료**	*ICT 표준화와 시험인증의 글로벌 리더*
일시	**2025. 6. 16(월) 배포 / 배포시부터 보도해 주시기 바랍니다.**	
담당	TTA AI융합시험연구소 AI신뢰성센터 곽준호 팀장(010-5110-2693), 신예진 책임(010-5110-6093)	

TTA, 생성형 AI 공격 전략에 대한 실증 연구 결과 발표

- 세계 최대 LLM 레드티밍 챌린지 분석을 통해 AI 신뢰성 강화 기반 제시 -

한국정보통신기술협회(회장 손승현, 이하 TTA)는 2025년 6월, TTA 홈페이지를 통해 ‘LLM 유해성 공격 전략에 대한 실증적 분석’ 보고서를 공개했다. 이 보고서는 2023년 미국 라스베이거스에서 개최된 DEF CON 31 Generative AI Red Teaming(GRT) 챌린지의 공개 데이터를 기반으로, 대규모 언어 모델(LLM)을 대상으로 한 실제 공격 사례를 정량적으로 분석한 결과를 담고 있다.

※ 해당 보고서는 TTA 공식 누리집(www.tta.or.kr) 내 공지사항을 통해 내려받으실 수 있습니다.

한양대학교 연구진과 협력하여 수행된 본 연구는 실제 성공한 공격 데이터를 기반으로 LLM에 대한 주요 취약성과 효과적인 공격 전략을 파악하고자 수행되었으며, 프롬프트 유형 및 공격 대상 특성을 라벨링하여 체계적으로 분석했다. 이를 통해 AI 무해성 평가 및 방어 전략 수립에 실질적인 근거를 제공한다.

보고서의 기반이 된 DEF CON 31 GRT 챌린지는 미국 AI Village와 Humane Intelligence, SeedAI 등이 주관한 세계 최대 공개형 LLM 보안 평가 행사다. 행사에는 Anthropic, OpenAI, Meta, Google 등 주요 AI 개발사들이 참여했으며, 전 세계 2,500여 명의 참가자들이 상용 LLM을

대상으로 공격을 수행했다.

연구진은 이 중 공격에 성공한 사례 2,673건을 선별하여, 각각에 대해 ▲공격 타겟(피해 대상), ▲공격 유형(프롬프트 전략)을 별도로 라벨링 하였다. 공격 타겟은 ‘성별 · 인종 · 국적 · 직업 · 정치성향’ 등 총 7개 대분류와 28개 하위 분류로 구성되며, 공격 유형은 ‘질문’, ‘직접 요청’, ‘상황 가정’, ‘편향 주입’, ‘순차/누적 질의’ 등 총 10개 전략 유형으로 분류되었다.

분석 결과, 인종 · 국적 · 성별 등 인구통계학적 속성을 겨냥한 공격이 다수 존재했으며, 이는 LLM이 사회적 고정관념을 재현할 수 있음을 시사한다. 또한 ‘질문’, ‘직접 요청’, ‘순차/누적 질의’ 전략은 높은 빈도로 사용되며, 단순한 방식으로도 가드레일 우회가 가능함을 보여주었다. 특히 사회적 피해(Societal Harm) 카테고리에서는 ‘잘못된 정보 주입(misinformation injection)’을 통해 LLM의 환각(hallucination)을 유도하는 전략이 효과적이었다는 점도 확인되었다.

본 연구에서 가공된 데이터셋은 Hugging Face 플랫폼을 통해 공개된다. 공개되는 데이터셋은 AI 신뢰성 평가, 공격 탐지 알고리즘 개발, 프롬프트 설계 연구 등에 활용될 수 있으며, 학계와 산업계의 다양한 LLM 방어 전략 수립에 기여할 것으로 기대된다.

TTA 손승현 회장은 “이번 보고서는 단순한 공격 탐지 기술을 넘어, AI 시스템이 어떤 사회적 편견과 고정관념을 내재하고 있는지 실증적으로 분석한 연구”라며 “생성형 AI의 가드레일 구축을 위한 실질적인 참고자료로 활용되기를 기대한다”고 밝혔다.