언어만의 문제가 아니다:
ROK-FORTRESS가 보여주는 다국어 AI 안전의 과제

2026년 9월 17일

프론티어 AI 시스템이 전 세계적으로 점차 널리 배포됨에 따라, 여러 언어 환경에서 이들 시스템의 안전성을 평가하는 것이 핵심적인 과제로 부상하고 있다. 대부분의 다국어 안전성 벤치마크는 원문 프롬프트가 설명하는 시나리오를 그대로 유지한 채 프롬프트만 다른 언어로 번역한다. 그러나 실제 환경에서는 언어의 변화가 종종 맥락의 변화, 즉 서로 다른 제도, 주체, 법률, 지정학적 현실의 변화와 함께 나타난다. 선행 연구에서는 유해한 요청(harmful requests)을 영어가 아닌 다른 언어로 번역했을 때 일부 모델이 해당 요청에 응할 가능성이 더 높아질 수 있음을 보여주었다. ROK-FORTRESS는 프롬프트의 언어와 해당 프롬프트가 참조하는 국가별 인물, 기관 및 사건을 서로 분리하여 변경할 경우 모델의 응답이 어떻게 달라지는지를 탐구한다.

Scale AI와 **한국 인공지능안전연구소(Korea AI Safety Institute)**가 공동 개발한 새로운 벤치마크인 ROK-FORTRESS는 바로 이러한 질문에 답하기 위해 설계되었다. 그 결과는 다국어 AI 안전성에 관한 기존의 통념이 시사하는 것보다 더욱 복잡하며, 동시에 더 많은 시사점을 제공한다.

ROK-FORTRESS는 Scale AI와 한국 인공지능안전연구소 간의 더 폭넓은 협력의 일환으로 개발된 최신 벤치마크이다. 양 기관의 협력에는 공동 연구, 대규모 언어모델(LLM) 평가, 그리고 AI 시스템의 취약점과 잠재적인 적대적 악용방식을 식별하기 위한 레드티밍 활동에 관한 지속적인 협력이 포함된다.

번역에만 의존하는 평가의 한계

대부분의 다국어 안전성 벤치마크는 동일한 기본 요청을 서로 다른 언어로 번역하여 모델을 평가한다. 이러한 설계는 언어 자체의 효과를 분리하여 측정하는 데 도움이 되지만, 동일한 의도가 서로 다른 지정학적 맥락에 놓였을 때 모델의 응답이 어떻게 달라지는지는 보여주지 못한다. 예를 들어 대규모 인명 피해를 초래하는 공격을 다루는 적대적 요청은

근본적인 의도가 동일하더라도, 미국의 1995년 오클라호마 연방정부 청사 폭탄 테러를 언급할 때와 한국의 1987년 대한항공 858편 폭파 사건을 언급할 때 모델이 매우 다르게 받아들일 수 있다. 따라서 번역에만 의존하는 평가는 언어와 지정학적 맥락 설정이 어떻게 상호작용하는지를 밝혀내기 어렵다.

ROK-FORTRESS는 언어와 지정학적 맥락 설정 각각의 효과 뿐 아니라 두 요소 간의 상호작용까지 측정할 수 있도록 설계되었다. 이 벤치마크는 단순한 번역이 아니라, 언어 형식과 맥락 설정 모두를 체계적으로 조정하는 통제된 트랜스크리에이션 매트릭스(controlled transcreation matrix)를 활용하여, 각 적대적 프롬프트를 최대 4가지 변형으로 평가한다.

이 과정에 (1) 언어는 영어와 또는 한국어, (2) 지정학적 맥락설정은 미국 또는 한국의 주체, 기관 및 실행 관련 세부 사항으로 각각 독립적으로 변경한다. 또한 각 적대적 프롬프트에는 과잉 거부(over-refusal)를 측정하기 위하여 무해한 대응 프롬프트(benign counterpart)를 짝으로 구성하였다. 전체 데이터셋은 국가안보 및 공공안전의 4가지 영역인 화학·생물학·방사능·핵·폭발물(CBRNE) 위협, 정치적 폭력과 테러리즘, 범죄 및 금융 활동, 정보 유출에 걸쳐 총 1,235개의 과제를 포함한다.

이 벤치마크는 보정된 LLM기반 판정 패널(LLM-as-judge)을 사용해 모델 응답을 채점한다. 이 패널은 전문가가 작성한 참조 레이블(reference labels)과 대조해 검증을 거쳤으며, 채점에는 전문 레드티머가 개발한 프롬프트별 이진 평가 루브릭을 적용한다. 이러한 방법론은 단순히 모델이 요청을 거부했는지 여부뿐 아니라, 어떤 유형의 유해한 정보를 제공했는지 또는 제공하지 않았는지까지 평가할 수 있도록 한다.

언어와 맥락을 분리하여 확인한 결과

가장 핵심적인 결과는 다국어 안전성 연구에서 널리 받아들여져 온 설명과 상반된다. 평가한 14개 모델 중 거의 모든 모델에서 한국어로 작성되고 한국의 맥락에 기반한 프롬프트가 일관되게 더 낮은 유해성과 연관되어 있었다. 영어 프롬프트는 전반적으로 가장 높은 등급 가중 위험 점수(tier-weighted risk score)를 보였고, 언어와 맥락을 모두 한국에 맞게 완전히 트랜스크리에이션한 프롬프트는 가장 낮은 점수를 보였으며, 나머지 중간 조건의 변형들은 그 사이에 위치했다. 이러한 패턴은 한국어 특화 지역 모델에서도 나타났다.

그 차이는 상당하다. 가장 유해성이 높은 모델과 가장 낮은 모델 간 위험 점수는 거의 9배 차이가 났다. 이는 현재 프론티어 모델 간 안전성 대응 양상의 편차가 얼마나 큰지

보여준다. 하지만 이러한 편차에도 불구하고, 한국어로 작성되고 한국의 맥락에 기반한 프롬프트가 전반적으로 더 낮은 유해성을 유발하는 패턴은 유지되었다. 프론티어 AI 모델 평가의 관점에서는 이러한 패턴을 이른바 '한국어·한국 맥락에서의 상대적 이점(Korean advantage)'로 설명할 수 있다.

몇 가지 추가적인 결과는 이러한 양상을 더욱 구체적으로 보여준다.

"한국어가 더 안전하다"는 효과가 모두 실질적인 안전성 향상을 의미하는 것은 아니다. 주요 실험에서는 역할극, 가상의 배경 이야기, 감정적 호소 등의 형태로 유해한 요청을 감추는 정교한 적대적 프롬프트를 사용했다. 이러한 기법을 제거하고 동일한 유해 정보를 평범하고 직접적인 방식으로 요청했을 때, 한국어와 한국어 맥락 이점(Korean advantage)은 대부분 사라졌고 모델 유형에 따라 뚜렷하게 다른 양상이 나타났다.

OpenAI, Anthropic, Google 등의 독점형 모델(proprietary models)은 이러한 적대적 기법을 제거한 뒤에도 한국어에서 다소 더 안전한 경향을 유지했다. 반면 5개의 오픈소스 프론티어 모델에서는 반대 현상이 나타나 한국어 요청에 응할 가능성이 오히려 더 높아졌다. 이는 요청을 저자원 언어(lower-resource language)로 번역하는 것 자체가 모델의 안전장치를 우회할 수 있다는 기존 연구 결과와도 일치한다[1]. 이러한 결과는 초기에 관찰된 유해 응답 억제 효과의 일부가 한국어 환경에서의 내재적인 안전성 정렬(safety alignment)이 더 강하기 때문이라기보다, 트랜스크리에이션 과정에서 적대적 프롬프트의 포장 구조(adversarial wrappers)가 효과를 잃었기 때문일 가능성을 시사한다.

전반적으로 보면 언어는 맥락 변화보다 더 큰 유해 응답 억제 효과를 유발했다. 언어를 영어에서 한국어로 전환할 때 나타나는 언어적 효과는, 미국의 주체와 기관을 한국의 주체와 기관으로 변경할 때 나타나는 맥락적 효과보다 약 2.5배 크게 작용한다(각각 약 10%p 4%p). 이러한 결과에 부합하는 한 가지 해석은 한국어가 모델로 하여금 위험을 보다 보수적으로 판단하게 하는 신호로 작용할 수 있다는 것이다.

또한 언어와 맥락은 서로 상호작용하며, 그 상호작용은 모든 모델에 공통적으로 나타나는 것이 아니라 모델별로 다르게 나타났다. 14개 모델 중 4개에서는 한국의 문화적 맥락을 추가했을 때 언어에 따른 유해 응답 억제 효과가 유의미하게 감소했다. 이는 중복 효과(redundancy effect)와 일치하는 결과로 볼 수 있다. 즉, 모델의 내부 위험 표상에서 두 개의 "한국" 관련 신호가 서로 중첩되어 있기 때문에 하나의 신호가 추가되면 다른

[1] Yong et al. (2023) Low-Resource Languages Jailbreak GPT-4; Deng et al. (2024) Multilingual Jailbreak Challenges in Large Language Models

신호의 한계 효과(marginal impact)가 감소한다는 것이다. 주목할 점은 반대 방향, 즉 한국의 맥락 설정이 억제 효과를 증폭시키는 방향에서 통계적으로 유의한 결과를 보인 모델은 없었다는 것이다. 이러한 모델별 차이는 번역만을 활용한 평가가 단순히 불완전한 데 그치는 것이 아니라 실제 환경에서의 안전성 대응을 잘못 해석하게 만들 수도 있음을 의미한다.

마지막으로, 한국어에서 더 낮게 나타난 유해성 점수는 실질적인 응답의 유해성이 감소한 효과와 모델이 전반적으로 보다 보수적으로 반응하는 경향(broader conservatism)을 모두 반영한다.

첫째, 안전성 향상이 관찰되는 경우 그 효과는 실제로 존재한다. 모델이 거부하지 않고 실제로 응답한 사례만 분석하더라도, 14개 모델 중 12개는 한국어 질문에서 여전히 더 낮은 유해성의 응답을 제공했다.

둘째, 이러한 신중한 대응에는 비용이 따른다. 모델들은 무해한 한국어 요청도 훨씬 더 자주 거부하며, 일부 모델에서는 거부 빈도가 약 두 배에 달했다. 결과적으로 관찰된 안전성 향상이 단순히 거부율 증가만으로 설명되는 것은 아니지만, 그렇다고 해서 모델이 실제 위협과 무해한 질문을 더 정확하게 구별하게 되었다는 것을 반드시 의미하는 것도 아니다.

이것이 중요한 이유

이러한 결과는 프론티어 AI 평가가 직면한 보다 광범위한 과제를 보여준다. 즉, 실제 배포 환경을 더욱 충실하게 반영하는 조건에서 모델의 대응 양상을 측정할 필요가 있다는 것이다. 이러한 시사점은 한국에만 국한되지 않는다. 언어와 지정학적 맥락이 함께 변화하는 모든 환경에 적용되며, 실제 국제적 AI 배포 환경의 상당수가 이에 해당한다.

동맹국 정부의 활용 환경에서는 영어 기반 안전성 평가에서 우수한 성능을 보인 모델이라 하더라도, 현지 언어로 현지 맥락에 기반한 위협에 대해 질의할 경우에는 다른 응답 양상을 보일 수 있다. 이는 단순히 이론적으로 가능한 예외적 상황이 아니다. 미국 외 국가에서 국가안보 및 공공안전 분야에 배포되는 AI 시스템에서 점차 흔해지고 있는 조건이다.

벤치마크 설계의 관점에서 ROK-FORTRESS의 결과는 번역만을 활용한 평가가 실제 배포 환경에서의 안전성 격차를 잘못 추정할 수 있음을 시사한다. 배포 조건을 더 잘 반영하려면, 벤치마크는 기저 의도를 보존하면서 언어와 지정학적 맥락 설정을 모두 조정한 트랜스크리에이션 프롬프트를 평가할 필요가 있다.

안전성 정렬의 관점에서는 영어 적대적 프롬프트를 단순히 번역한 데이터뿐 아니라 문화적 맥락을 반영한 변형을 사후학습(post-training) 데이터와 레드티밍 과정에 점차 적극적으로 포함할 필요가 있다. 이를 통해 모델이 실제로 배포될 다양한 맥락에서도 안전에 관한 대응이 일관되게 일반화될 수 있도록 해야 한다.

보다 현실적인 AI 안전성 평가를 향하여

ROK-FORTRESS는 Scale AI의 프론티어 모델 대상 국가안보 및 공공안전 벤치마크인 FORTRESS를 기반으로 하며, Scale AI와 한국 인공지능안전연구소의 파트너십을 통해 공식화된 연구 협력을 반영한다. 데이터셋의 공개용 일부는 이미 Hugging Face와 같은 공개 데이터셋 저장소에서 이용할 수 있으며, 문화적 맥락을 반영한 AI 안전성 평가에 관한 후속 연구를 지원하고 있다.

더 큰 목표는 측정하기 가장 쉬운 환경이 아니라, 있는 그대로의 현실 세계를 반영하는 안전성 평가 체계를 확립하는 것이다. 전 세계의 서로 다른 언어와 다양한 문화적 환경에 배포되는 프론티어 AI의 경우, 이는 단순히 영어 환경에서 안전장치가 유효한지를 확인하는 데 그치지 않고, 모델이 실제로 사용될 다양한 맥락에서도 안전성이 일반화되는지를 검증해야 한다는 것을 의미한다.

ROK-FORTRESS는 Scale AI와 한국 인공지능안전연구소가 공동으로 개발하였다. 논문은 *arXiv* 에서 사전 공개 논문(Priprint)으로 제공된다.