안전하고 책임감 있는 AI, Ally (Safe and Responsible AI)

Ally 는 여러분의 게임 플레이를 함께하는 AI 동반자입니다. 저희는 Ally 를 만드는 단계에서부터 안전을 핵심 기준으로 고려했으며, 한 가지 장치에 의존하지 않고 학습부터 점검, 개선에 이르기까지 여러 단계에 걸쳐 안전을 지키고자 노력하고 있습니다.

Ally 가 지키도록 학습한 기준

저희는 아래 영역에서 Ally 가 상황에 맞게, 안전하게 대응하도록 기준을 세우고 학습시켰습니다.

- **민감한 사회적 주제** — 정치·종교·역사·문화·영토 등 사회적 갈등을 유발할 수 있는 주제에 대해서는 특정 입장이나 이념을 대변하지 않고, 중립적이고 신중하게 대응합니다.
- **혐오와 차별** — 인종·국적·성별·종교·장애·성적 지향 등 정체성에 기반한 혐오·차별 표현에 동조하지 않으며, 모든 이용자가 존중받을 수 있는 환경을 지향합니다.
- **자해 및 자살** — 자기 위해 행위를 조장하거나 돕지 않으며, 이용자의 안전이 우려되는 상황에서는 전문 지원 기관 안내를 우선합니다.
- **현실 세계의 위해 행위** — 현실의 폭력·범죄·위협·괴롭힘·불법 행위를 조장하는 요청에는 응하지 않습니다.
- **성적 콘텐츠** — 성적 대상화·성희롱·노골적인 성적 표현에 응하지 않으며, 특히 미성년자와 관련된 콘텐츠를 가장 엄격하게 다룹니다.
- **개인정보 보호** — 개인의 연락처·주소·위치·계정 등 민감한 정보를 수집·추적·공개하거나 공유하지 않습니다.

동시에, 저희는 안전을 지키는 방식도 중요하게 고려했습니다. 전투·생존·전략 과정에서 자연스럽게 등장하는 표현은 현실의 위해 행위와 구분하여, 게임 플레이를 지원하는 방향으로 우선 해석합니다.

Ally 를 안전하게 만드는 과정

학습 — 위 기준에 따라 학습 데이터를 정제하고, 게임 속 맥락과 현실의 위해를 구분하여 더 안전하게 응답하도록 Ally 를 학습시킵니다.

점검 — 내부 평가와 함께, 내·외부 여러 참여자가 직접 다양한 상황에서 Ally 를 시험하며 취약한 지점을 찾아냅니다.

개선 — 점검 결과와 실제 플레이에서 얻은 피드백을 바탕으로 기준과 모델을 꾸준히 보완합니다. 또한 부적절한 콘텐츠가 노출되지 않도록 돕는 실시간 보호 장치도 함께 운영하며 고도화해나갑니다.

한계와 지속적인 노력

AI 는 정해진 답을 그대로 반복하는 프로그램이 아니라 학습을 통해 스스로 응답을 만들어내는 모델입니다. 그렇기에 위와 같은 노력에도 불구하고 저희가 의도한 기준과 다른 응답이 나타날 수 있으며, Ally 의 응답은 회사의 입장이나 의견을 대변하지 않습니다. 저희의 안전 기준은 이용자 피드백과 운영 경험, 법적 검토 및 지역별 요구사항을 반영하여 지속적으로 개선되며, 저희는 안전성과 유용성, 그리고 즐거운 플레이 경험 사이의 균형을 유지하기 위해 노력하고 있습니다.