제타(zeta) 유해 대화 생성 건 경위 및 개선 계획

문제 상황, 문제 원인, 긴급 조치, 근본 개선 계획

2026. 10.

주식회사 스캐터랩

스캐터랩은 황정아 의원실의 점검을 통해 확인된 제타(zeta)의 유해 대화 생성 문제를 엄중히 받아들이고 있습니다. 성인인증이 되지 않은 이용 환경에서 미성년으로 설정된 캐릭터를 대상으로 유해한 대화가 생성되고, AI가 차단을 피하는 표현까지 제시한 것에 대해 깊은 책임을 느낍니다.

이번 사안은 유해 대화를 걸러내는 어뷰징 탐지 모델의 탐지 기준이 다소 느슨하게 설정되어 있었고, 어뷰징 탐지 모델이 AI의 응답을 개별 단위로 판정하여 대화 전체의 맥락을 충분히 반영하지 못한 데에서 비롯되었습니다. 스캐터랩은 문제 확인 직후 어뷰징 탐지 기준 강화와 반복적인 어뷰징 시도에 대한 제재 도입 등 두 가지 긴급 조치에 착수하였으며, 10월 10일 일부 이용자부터 적용을 시작해 10월 18일까지 적용 대상 이용자 전체로 확대할 예정입니다. 나아가 이야기를 생성하는 AI가 처음부터 이용자 연령에 맞게 전개와 묘사를 조절하도록 하는, 보다 구조적이고 근본적인 개선을 2026년 4분기 중 추진할 계획입니다.

※ 플레이 모델: 이용자와 대화하며 이야기를 생성하는 제타의 자체 AI 모델
※ 어뷰징 탐지 모델: 이용자 입력과 AI 응답을 별도로 검사해 부적절한 대화를 차단하는 모델

1. 문제 상황

황정아 의원실은 성인인증을 하지 않은 계정으로, 청소년 이용자와 동일한 안전 기준이 적용되는 '세이프 모드'에서 제타를 점검하였으며, 그 과정에서 다음 문제가 확인되었습니다.

- **유해 대화 생성:** 미성년으로 설정된 캐릭터가 등장하는 2 개 플롯(콘텐츠)에서 성폭력·협박 묘사가 생성되었습니다. 어뷰징 탐지 모델이 문제 발화 200 여 건을 차단하였으나, 이용자가 표현을 바꿔 가며 반복적으로 시도하는 과정에서 차단되었어야 할 발화가 일부 생성 및 노출되었습니다.
- **차단 회피 표현 제시:** 이용자의 요청에 따라 AI 가 필터링을 피할 수 있는 표현 방식을 제시하였습니다. 해당 내용은 이용자의 요청에 맞추어 생성된 허구의 서술(hallucination)이기는 하였으나, AI 가 우회 요청에 대하여 응답을 제시한 것 자체가 개선이 필요한 문제라고 인식하고 있습니다.

2. 문제 원인

제타의 대화 단계는 이야기를 생성하는 플레이 모델과, 그 결과를 별도로 검사해 부적절한 대화를 차단하는 어뷰징 탐지 모델로 구성됩니다. 이번 사안은 정상적인 창작 표현이 과도하게 차단되지 않도록 어뷰징 탐지 모델의 기준이 다소 느슨하게 설정되어 있었고, 어뷰징 탐지 모델이 AI의 응답을 개별 단위로 판정하여 대화 전체의 맥락을 충분히 반영하지 못한 데에서 비롯되었습니다.

3. 긴급 조치

스캐터랩은 문제 확인 직후 아래 두 가지 긴급 조치에 착수하였습니다. 두 조치는 서비스 전체에 한 번에 적용하지 않고, 서비스 운영에 부작용이 없는지 확인하면서 적용 범위를 단계적으로 넓히는 방식으로 순차 적용하고 있습니다. 2026년 10월 10일부터 적용 대상 이용자의 약 5~10%에 적용하는 것을 시작으로, 적용 비율을 순차적으로 확대하여 10월 18일까지 적용 대상 이용자 전체에 적용을 완료할 예정입니다.

조치	주요 내용
어뷰징 탐지 기준 강화	• 세이프 모드 이용자 전체를 대상으로 탐지 기준을 강화하여, 직접적인 묘사뿐 아니라 간접적·은유적 표현도 더 잘 차단되도록 함 • 황정아 의원실의 점검 대화 기록을 강화된 어뷰징 탐지 기준으로 재검사한 결과, 차단되는 문제 발화가 200 여 건에서 400 여 건으로 약 2 배 증가

조치	주요 내용
반복적인 어뷰징 시도에 대한 제재 도입	• 같은 플롯에서 부적절한 대화로 차단이 반복되면 '경고 → 대화 정지 → 대화 초기화' 순으로 단계별 제재 • 모든 제재 안내에 판정 오류 신고 기능을 제공하고, 신고 사례는 어뷰징 탐지 모델 개선에 반영

4. 근본 개선 계획

긴급 조치에 더하여, 유해한 서사가 생성되는 구조 자체를 바꾸는 근본적인 개선을 추진합니다. 핵심은 안전 통제의 1차 주체를 별도의 어뷰징 탐지 모델에서 이야기를 만드는 플레이 모델로 옮기고, 추가적인 청소년 보호 장치를 도입하는 것입니다.

과제	주요 내용	시점
청소년 이용시간 및 결제 한도 도입	청소년 이용자의 기본 이용시간을 하루 3시간, 기본 결제 한도를 월 10만원으로 설정	2026년 4분기
플레이 모델 고도화	어뷰징 탐지 모델뿐만 아니라 이야기를 만드는 단계에서부터 이용자 연령에 맞게 전개와 묘사 수위를 조절할 수 있도록 플레이 모델 고도화. 대화가 부적절한 방향으로 전개될 시 장면을 전환하고, 이용자의 차단 우회 요청에는 응하지 않도록 모델 학습 진행	2026년 4분기
신규 어뷰징 탐지 모델 배포	대화 전체의 맥락을 보다 더 포괄적으로 반영하여 어뷰징 여부를 판정하고, 이번에 확인된 우회 표현 유형을 학습에 반영. 플레이 모델 고도화 이후에도 어뷰징 탐지 모델은 2차 안전장치로 유지	2026년 4분기
플롯(콘텐츠) 연령등급 세분화	플롯을 2~3단계 연령등급으로 세분화하여 등급에 맞지 않는 이용자에게는 플레이를 제한	2026년 4분기

스캐터랩은 이번 사안을 계기로 청소년 이용자 보호 체계를 전반적으로 점검하고, 개선 결과를 지속적으로 보완해 나가겠습니다. 앞으로도 관계 기관의 의견에 귀 기울이며, 청소년이 안전하게 이용할 수 있는 AI 서비스 환경을 만드는 데 최선을 다하겠습니다.