

# 생성형 인공지능 서비스
# 이용자 보호 가이드라인


방송통신위원회
Korea Communications Commission

KISDI 정보통신정책연구원
KOREA INFORMATION SOCIETY DEVELOPMENT INSTITUTE

GUIDELINES FOR USER PROTECTION

# 생성형 인공지능 서비스
# 이용자 보호 가이드라인

# Contents


생성형 인공지능 서비스 이용자 보호 가이드라인 / 4

생성형 인공지능 서비스 이용자 보호 가이드라인 실행 방식 / 10

Ⅰ. 배경 및 필요성 / 12

Ⅱ. 목적 및 지향점 / 13

Ⅲ. 용어의 정의 / 13

Ⅳ. 실행을 위한 기본 원칙 / 14

Ⅴ. 생성형 인공지능 서비스 이용자 보호를 위한 실행 방식 / 15

부록 / 실행 사례 / 28

1. 생성형 인공지능 서비스의 이용자 인격권 보호 / 30

2. 생성형 인공지능 서비스의 결정 과정을 알리려는 노력 / 35

3. 생성형 인공지능 서비스의 다양성 존중 노력 / 44

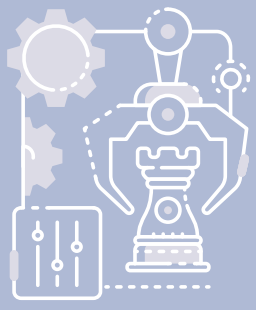
4. 생성형 인공지능 서비스의 입력데이터 수집·활용 과정에서의 관리 / 58

5. 생성 콘텐츠 활용에서 발생할 수 있는 문제 해결을 위한 책임과 참여 / 66

6. 생성 콘텐츠의 건전한 유통·배포를 위한 노력 / 74

GUIDELINES FOR USER PROTECTION











# 생성형 인공지능 서비스
# 이용자 보호 가이드라인

# 생성형 인공지능 서비스 이용자 보호 가이드라인

## Ⅰ. 목적

이 가이드라인은 생성형 인공지능 서비스 이용 과정에서 발생할 수 있는 잠재적 위험들을 사전에 방지하여 안전하고 신뢰할 수 있는 생성형 인공지능 서비스 이용 환경을 마련함으로써 이용자의 권익을 보장하고, 생성형 인공지능의 올바른 사용으로 얻을 수 있는 혜택을 모든 사회구성원이 고루 누릴 수 있도록 하는 것을 목적으로 한다. 이때 생성형 인공지능 생태계 구성원의 책임감을 바탕으로 자율적인 참여를 통한 효과적인 이용자 보호 체계를 구축하고자 한다.

## Ⅱ. 용어의 정의

1. **"생성형 인공지능(generative AI)"**이란 대규모 데이터를 학습하여 만들어진 모델을 사용하여 이용자의 요구에 따라 텍스트, 이미지, 비디오, 오디오, 코드 등 새로운 콘텐츠를 생성하는 인공지능 기술을 말한다.

1-1. **"생성형 인공지능 서비스"**란 생성형 인공지능을 활용하여 공공 또는 민간에게 제공되는 서비스를 말한다.

2. **"개발사"**란 생성형 인공지능의 파운데이션 모델을 개발하고 시장에 출시하여 서비스 제공자가 모델을 사용할 수 있게 하거나 자체 서비스를 제공하는 사업자를 말한다.

3. **"서비스 제공자"**란 생성형 인공지능 모델을 활용해 디지털 도구 및 서비스를 제공하는 사업자를 말한다.

4. **"이용자"**란 생성형 인공지능 서비스를 이용하거나 이용하려는 자를 말하며 해당 서비스로 생성된 콘텐츠를 유통하거나 소비하는 자를 모두 포함한다.

5. **"생성형 인공지능 산출물"**은 이용자가 입력한 프롬프트에 기반하여 생성형 인공지능이 산출한 최종 결과물을 말한다.

## Ⅲ. 적용 범위

이 가이드라인은 생성형 인공지능 개발사 및 서비스 제공자 등에 적용된다.

## Ⅳ. 기본 원칙

이 가이드라인에서 생성형 인공지능 서비스의 전반에 걸쳐 추구해야 할 기본적인 방향성을 다음과 같이 제시한다.

1. 생성형 인공지능 서비스는 인간의 존엄성을 보호하고 개인의 자유와 권리를 보장하며, 인간이 적절하게 통제하고 감독할 수 있는 방식으로 제공되어야 한다.

2. 생성형 인공지능 서비스는 인공지능시스템 사용에 따른 작동 원리 및 결과, 이용자에게 미치는 영향 등에 대하여 이용자에게 알기 쉽게 설명되도록 제공되어야 한다.

3. 생성형 인공지능 서비스는 안전하게 작동되어야 하며, 예상치 못한 피해를 최소화하고 악의적으로 이용되거나 변경되지 않도록 제공되어야 한다.

4. 생성형 인공지능 서비스는 이용자에게 차별 또는 불공정한 결과가 발생하지 않도록 제공되어야 한다.

## Ⅴ. 생성형 인공지능 서비스 이용자 보호를 위한 실행 방식

위의 기본 원칙이 제시하는 목표를 실현하기 위한 방식으로 다음과 같은 것들을 제시한다.

### 1. 생성형 인공지능 서비스의 이용자 인격권 보호

가. 개발사는 생성형 인공지능 기반 시스템이 이용자의 인격권을 침해할 수 있는 요소를 발견하고 통제할 수 있는 알고리즘을 구축하도록 노력한다.

나. 서비스 제공자는 서비스를 구축하는 과정에서 사전에 이용자가 타인의 인격권을 침해하는 결과물을 생성하지 않도록 관리할 책임을 인지하고, 내부 모니터링 체계나 이용자 신고 프로세스 등 관련 방안을 마련하도록 노력한다.

다. 서비스 제공자는 생성형 인공지능의 산출물이 이용자의 권리를 침해하지 않도록 서비스 제공 범위와 방법을 고려한다.

## 2. 생성형 인공지능 서비스의 결정 과정을 알리려는 노력

가. 서비스 제공자는 생성형 인공지능 산출물이 생성형 인공지능으로 생성되었음을 자율적으로 선택한 방식으로 알린다.

나. 서비스 제공자는 해당 생성형 인공지능 서비스와 관련한 영업활동을 하는 사업자의 정당한 이익을 침해하지 않는 범위에서 생성형 인공지능의 결정 과정을 이해할 수 있는 기본적인 정보를 이용자가 쉽게 접근하고 이해할 수 있는 방식으로 제공한다.

## 3. 생성형 인공지능 서비스의 다양성 존중 노력

가. 개발사와 서비스 제공자는 생성형 인공지능을 개발하고 운영하는 과정에서 알고리즘 설계, 데이터의 수집과 입력 및 알고리즘 실행 등 기술적으로 실현 가능한 범위 안에서 편향과 차별을 줄이기 위해 노력한다.

나. 개발사와 서비스 제공자는 자사 서비스에서 생성형 인공지능 산출물의 편향성을 줄이고 다양성을 보장하는 자체 원칙이나 기준을 마련한다.

다. 서비스 제공자는 생성형 인공지능의 이용자들이 차별적 사용을 하지 않도록 이에 대한 방지와 필터링 기능을 제공한다.

라. 서비스 제공자는 이용자가 생성형 인공지능의 산출물이 편향적이라고 판단하였을 때 바로 신고할 수 있는 직관적인 방법을 제공하고 신고 내용을 처리하는 절차를 마련한다.

## 4. 생성형 인공지능 서비스의 입력데이터 수집·활용 과정에서의 관리

가. 서비스 제공자는 서비스 이용 과정에서 발생하는 이용자의 입력 및 생성 데이터를 학습데이터로 활용할 경우, 이를 사전에 고지하고 이용자가 동의 혹은 거부할 수 있는 절차를 마련한다.

나. 서비스 제공자는 서비스 이용 과정에서 발생하는 이용자의 입력 및 생성 데이터를 학습데이터로 활용할 경우, 해당 활용 여부에 대한 이용자의 선택권을 보장하도록 노력한다.

다. 이용자의 입력 및 생성 데이터를 학습데이터로 활용하는 것과 관련하여, 서비스 제공자는 기업 내 감시 조직을 갖추거나 책임자를 선정하여 이용자의 입력 및 생성 데이터가 안전하고 정당하게 활용될 수 있도록 관리한다.

### 5. 생성 콘텐츠 활용에서 발생할 수 있는 문제 해결을 위한 책임과 참여

가. 서비스 제공자는 생성형 인공지능 산출물과 관련하여 이용자의 이해를 돕기 위해 서비스 제공자 및 이용자 각자의 책임 범위를 정의하고 이용 단계에서 이용자가 자기 책임을 명확히 인지하도록 안내한다.

나. 서비스 제공자는 이용자가 인지하지 못한 피해의 발생을 최소화하기 위해 모니터링 시스템을 갖추거나 다른 방식의 위험관리 체계를 마련한다.

### 6. 생성 콘텐츠의 건전한 유통·배포를 위한 노력

가. 서비스 제공자는 이용자가 생성형 인공지능 서비스를 이용하여 부적절한 콘텐츠를 생성하거나 공유하지 않도록 안내한다.

나. 서비스 제공자는 이용자의 프롬프트 입력값과 생성된 산출물이 도덕적·윤리적 기준을 준수하는지 검토하고 관리한다.

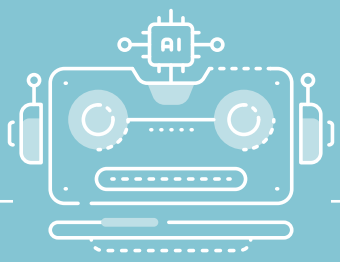
다. 서비스 제공자는 이용자가 의도적 혹은 비의도적으로 유해한 콘텐츠를 청소년에게 유포하지 않도록 노력한다.

## 부 칙 〈2025.2.28. 제정〉

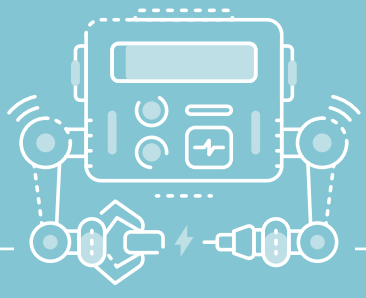
1. 이 가이드라인은 제정 후 1개월이 경과한 날부터 시행한다.

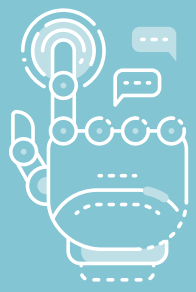
2. 방송통신위원회는 시행일을 기준으로 매 2년이 되는 시점(매 2년째의 시행일 전까지를 말한다)마다 이 가이드라인에 대하여 타당성을 검토하여 개선 등의 조치를 취하여야 한다.

GUIDELINES FOR USER PROTECTION







# 생성형 인공지능 서비스
# 이용자 보호 가이드라인
# 실행 방식

# 생성형 인공지능 서비스 이용자 보호 가이드라인 실행 방식

## Ⅰ. 배경 및 필요성

생성형 인공지능을 이용하여 다양한 콘텐츠를 생성할 수 있게 되면서 비즈니스 영역에서는 새로운 서비스를 개발하고 사용자 경험을 향상하고 있습니다. 이용자들은 생성형 인공지능을 검색부터 창작까지 다양한 기능으로 활용하고 있고, 업무 효율성을 높이는데도 적극 수용하고 있습니다. 다른 한편에서는 새로운 기술의 빠른 확산 과정에서 늘 나타났던 기술의 불완전성으로 인한 이슈들, 그리고 신규 서비스에 대한 이용자의 경험과 이해 부족에 따른 이슈들이 사회 문제로 주목받았습니다.

이에 대한 대응으로 개발사는 인공지능 모델 개발에서 윤리 원칙 준수를 실천하고, 서비스 제공자는 생성형 인공지능 서비스에 대한 이용 정책을 마련하거나 기존 이용 정책을 보완하며, 활용성이 높은 대학이나 공공기관 등에서는 해당 영역별 이용자 활용 가이드라인을 개발·배포하는 등 건전한 활용 환경을 조성하는 데 노력하고 있습니다.

생성형 인공지능 이용자 보호를 위한 본 가이드라인은 그러한 노력의 하나입니다. 먼저 현 시점에서 생성형 인공지능의 기술적 한계와 제도적 미비, 산출물이 개인과 사회에 미치는 영향 등을 고려하였습니다. 그리고 기존과는 달리 서비스 활용 과정에서 이용자와의 상호작용이 미치는 영향도 염두에 두었습니다.

본 가이드라인을 통해서 생성형 인공지능 생태계에서 기술·서비스 특성에 따라 이용자들이 겪을 수 있는 문제들을 상기하는 한편, 사업자들의 모범적인 처리 방안을 널리 공유하여 향후 후발주자의 서비스 론칭을 지원하는 동시에 이용자 권익 침해 요인의 최소화와 안전한 생성형 인공지능 서비스 안착에 기여하고자 합니다.

## Ⅱ. 목적 및 지향점

- 이 가이드라인은 생성형 인공지능 서비스 이용 과정에서 발생할 수 있는 잠재적 위험들을 사전에 방지하고 이용자 권익을 보호하기 위해 개발사 및 서비스 제공자 등이 염두에 두어야 할 실천 방안을 제시합니다.

- 안전하고 신뢰할 수 있는 생성형 인공지능 서비스 이용 환경을 마련함으로써 이용자의 권익을 보장하고, 생성형 인공지능의 올바른 사용으로 얻을 수 있는 혜택을 모든 사회구성원이 고루 누릴 수 있도록 하는 것을 추구합니다.

## Ⅲ. 용어의 정의

### 생성형 인공지능

"생성형 인공지능(generative AI)"이란 대규모 데이터를 학습하여 만들어진 모델을 사용하여 이용자의 요구에 따라 텍스트, 이미지, 비디오, 오디오, 코드 등 새로운 콘텐츠를 생성하는 인공지능 기술을 말합니다.

### 생성형 인공지능 서비스

"생성형 인공지능 서비스"란 생성형 인공지능을 활용하여 공공 또는 민간에게 제공되는 서비스를 말합니다.

### 개발사

"개발사"란 생성형 인공지능의 파운데이션 모델을 개발하고 시장에 출시하여 서비스 제공자가 모델을 사용할 수 있게 하거나 자체 서비스를 제공하는 사업자를 말합니다.

### 서비스 제공자

"서비스 제공자"란 생성형 인공지능 시스템을 활용해 디지털 도구 및 서비스를 제공하는 사업자를 말합니다.

### 이용자

"이용자"란 생성형 인공지능 서비스를 이용하거나 이용하려는 자를 말하며 해당 서비스로 생성된 콘텐츠를 유통하거나 소비하는 자를 모두 포함합니다.

### 생성형 인공지능 산출물

"생성형 인공지능 산출물"은 이용자가 입력한 프롬프트에 기반하여 생성형 인공지능이 산출한 최종 결과물을 말합니다.

## Ⅳ. 실행을 위한 기본 원칙

본 가이드라인은 신기술의 도입이 초래할 수 있는 기술적·사회적 위험으로부터 안전한 지능정보서비스 환경을 조성하기 위한 방송통신위원회의 정책적 노력으로 마련한 「이용자 중심의 지능정보사회를 위한 기본 원칙」('19)을 기조로 하며, 그밖에 국내 기준으로 범부처 인공지능 윤리기준('20)을, 글로벌 기준으로는 OECD의 「인공지능에 대한 권고(Recommendation of the Council on Artificial Intelligence, '24)」와 유네스코(UNESCO)의 「인공지능 윤리에 대한 권고(Recommendation on the Ethics of Artificial Intelligence), '21」를 참조합니다.

이에 생성형 인공지능 서비스의 전반에 걸쳐 추구해야 할 기본 방향성을 다음과 같이 제시합니다.

1. 생성형 인공지능 서비스는 인간의 존엄성을 보호하고 개인의 자유와 권리를 보장하며, 인간이 적절하게 통제하고 감독할 수 있는 방식으로 제공되어야 합니다.

2. 생성형 인공지능 서비스는 인공지능시스템 사용에 따른 작동 원리 및 결과, 이용자에게 미치는 영향 등에 대하여 이용자에게 알기 쉽게 설명되도록 제공되어야 합니다.

3. 생성형 인공지능 서비스는 안전하게 작동되어야 하며, 예상치 못한 피해를 최소화하고 악의적으로 이용되거나 변경되지 않도록 제공되어야 합니다.

4. 생성형 인공지능 서비스는 이용자에게 차별 또는 불공정한 결과가 발생하지 않도록 제공되어야 합니다.

# V. 생성형 인공지능 서비스 이용자 보호를 위한 실행 방식

다음은 개발사 및 서비스 제공자가 생성형 인공지능 서비스를 제공하는 데 필요한 실행 지침입니다. 생성형 인공지능 서비스 제공 과정에서 발생할 수 있는 이슈를 짚어보고, 부정적 영향을 최소화하기 위한 지침과 현재 모범적으로 서비스를 운영하는 사업자들의 실천사례를 공유합니다.

## 1. 생성형 인공지능 서비스의 이용자 인격권 보호

생성형 인공지능을 기반으로 한 시스템과 광범위하고도 상세한 데이터 분석으로 이용자의 생활에 밀착한 섬세한 서비스를 제공하는 것이 가능해졌습니다. 그러나 그 결과물의 밀착성으로 인하여 거꾸로 이용자의 프라이버시를 비롯한 인격권을 침해할 수 있다는 위험성 역시 제기됩니다.

대화형 서비스 등을 통해 이용자가 원치 않는 혐오 발언이 표출되는 사례가 일부 서비스에 나타난 바 있습니다. 또한 피싱을 통해 이용자의 민감 정보를 수집하거나 재산상의 피해를 주는 결과를 초래할 수도 있습니다. 생성형 인공지능 기반의 검색 서비스는 어떤 내용이든 상세하게 설명해 줄 수 있다는 장점이 있지만, 정교한 데이터의 조합을 통해 오히려 거짓 정보에 해당하거나 편향적이거나 타인의 명예를 훼손하는 결과물을 만들어 낼 수도 있을 것입니다. 시스템의 섬세함을 이용하여 특정인을 연상케 하는 사진이나 영상을 만들어 낸다면, 경우에 따라 해당 특정인이 다른 활동을 통해 얻을 수 있었던 경제적 이익을 침해하는 문제로도 이어질 가능성도 존재합니다.

생성형 인공지능의 알고리즘을 통해 만들어진 표현물이 명예훼손 등의 결과를 가져온다 해도, 행위자가 특정되지 않을 가능성이 있습니다. 이 경우 기존 제도를 통한 해결이 어렵다는 문제에 직면합니다. 개발사 혹은 서비스 제공자가 대개는 생성형 인공지능 기반 시스템을 악의적으로 구성하지 않았을 것이므로, 그들에게 책임을 묻기도 어렵습니다. 생성형 인공지능이 만들어 낸 정보는 다시 학습 데이터로 사용되어 명예훼손 등의 반복적인 재생산이 가능하기 때문에, 사전 단계에서 이를 관리할 수 있는 방법을 강구하는 일이 중요합니다.

### ✿ 실행을 위한 방안

- 개발사는 생성형 인공지능 기반 시스템이 이용자의 인격권을 침해할 수 있는 요소를 발견하고 통제할 수 있는 알고리즘을 구축하도록 노력합니다.

- 서비스 제공자는 서비스를 구축하는 과정에서 사전에 이용자가 타인의 인격권을 침해하는 결과물을 생성하지 않도록 관리할 책임을 인지하고, 내부 모니터링 체계나 이용자 신고 프로세스 등 관련 방안을 마련하도록 노력합니다.

- 서비스 제공자는 생성형 인공지능의 산출물이 이용자의 권리를 침해하지 않도록 서비스 제공 범위와 방법을 고려합니다.

### ✿ 실행 사례

이 실행 사례는 생성형 인공지능 서비스에서 제공하는 사례들에 근거하여 제시한 것입니다. [부록]에서 각 항목별 실제 서비스에 근거한 실행 예시들을 확인할 수 있습니다.

- 대화형 서비스에서 혐오 표현 등이 나타나지 않도록 생성 가능한 표현물 수위 정보를 사전에 안내할 수 있습니다.

- 이용자가 타인의 권리 침해 정보/불법 정보/비윤리적·편향적 정보 등을 프롬프트에 입력하는 경우 경고문구를 제시하거나 일정 시간 차단할 수 있습니다.

- 타인의 인격권을 침해하는 결과물이 생성되었을 때, 사후에라도 이를 차단하거나 유사한 표현물이 재생성되지 않도록 통제하는 장치를 시스템에 포함할 수 있습니다.

- 인공지능 산출물로부터 가치관에 영향을 받을 수 있는 연령대의 이용자를 보호하는 조치를 취할 수 있습니다.

## 2. 생성형 인공지능 서비스의 결정 과정을 알리려는 노력

많은 인공지능 시스템, 특히 딥러닝 기반 시스템은 복잡한 내부 메커니즘을 가지고 있어, 이용자의 관점에서 생성형 인공지능들이 어떻게 특정 결정에 도달했는지 혹은 어떻게 해당 콘텐츠를 생성했는지 이해하기 어려운 측면이 있습니다. 이를 '블랙박스' 문제라고도 하는데, 이용자나 개발자조차도 생성형 인공지능의 결정 과정을 명확하게 파악하기 어렵다는 것을 의미합니다. 이렇게 생성형 인공지능의 결정 과정이 불투명할 경우 신뢰성 문제, 책임 소재의 모호성과 같은 다양한 쟁점들이 발생할 수 있습니다.

먼저 생성형 인공지능이 콘텐츠를 생성할 때 어떻게 결정을 내리는지, 어떤 데이터에 기반하여 콘텐츠를 생성하는지에 대한 과정을 이해하지 못하면, 이용자들은 그 생성물을 신뢰하기 어려울 수 있습니다. 중요한 의사결정이나 민감한 정보를 다루는 분야일수록 이러한 우려가 크게 작용할 것입니다. 또, 이용자가 생성형 인공지능이 어떻게 콘텐츠를 생성하는지를 잘 알지 못하고 부정확성 등 기술 특성상의 한계를 이해하지 못한 채 서비스를 이용하는 상황에서 어떤 부정적인 이슈가 발생하는 것이 문제가 될 것입니다. 생성형 인공지능이 잘못된 결정을 내렸을 때도 책임을 누구에게 물을지가 불분명할 수 있습니다. 생성형 인공지능의 결정 과정이 불투명하고 복잡할수록, 오류의 원인을 찾고 책임을 묻는 일이 더욱 어려워질 것입니다.

이에 따라 생성형 인공지능의 설명가능성에 대한 수요가 발생합니다. 이용자들은 생성형 인공지능의 결정 과정에 대해 알고 싶어할 것입니다. 이미 인공지능의 연구자들과 개발자들은 더 이해하기 쉬운 결정 과정을 설계하고, 투명한 데이터 관리와 생성형 인공지능의 책임 및 규제에 대한 명확한 기준을 마련하기 위해 노력하고 있습니다. 인공지능의 결정 과정을 설명할 수 있는 기술(설명 가능한 AI, XAI)도 활발히 진행 중입니다.

### 실행을 위한 방안

- 서비스 제공자는 생성형 인공지능 산출물이 생성형 인공지능으로 생성되었음을 자율적으로 선택한 방식을 통해 알립니다.

- 서비스 제공자는 해당 생성형 인공지능 서비스와 관련한 영업 활동을 하는 사업자의 정당한 이익을 침해하지 않는 범위에서 생성형 인공지능의 결정 과정을 이해할 수 있는 기본적인 정보를 이용자가 쉽게 접근하고 이해할 수 있는 방식으로 제공합니다.

### 실행 사례

이 실행 사례는 생성형 인공지능 서비스에서 제공하는 사례들에 근거하여 제시한 것입니다. [부록]에서 각 항목별 실제 서비스에 근거한 실행 예시들을 확인할 수 있습니다.

- 생성형 인공지능 산출물이 생성형 인공지능으로 생성되었음을 자율적으로 선택한 방식을 통해 안내할 수 있습니다.
- 생성형 인공지능의 결정 과정을 이해할 수 있는 기본정보 제공 방법으로 출처 표기 혹은 출처에 대한 링크를 제공할 수 있습니다.
- 생성형 인공지능의 결정 과정을 이해할 수 있는 기본정보 제공 방법으로 생성형 인공지능 모델에 대한 정보를 제공할 수 있습니다.
- 고객이 생성형 인공지능의 결정 과정에 대한 설명을 요청한 경우, 기업의 정당한 이익 추구 범위를 벗어나지 않는 선에서 설명을 제공할 수 있습니다.

## 3. 생성형 인공지능 서비스의 다양성 존중 노력

기존의 검색 엔진은 다양한 검색 결과를 리스트 형태로 나열하였고, 이용자들은 복수의 검색 결과를 선택하여 정보를 취합하는 형태를 취했습니다. 반면, 생성형 인공지능은 이용자가 입력한 프롬프트를 분석하여 일반적으로 하나의 콘텐츠를 생성합니다. 생성형 인공지능이 생성한 한 종류의 답변이 편향되거나 다양성이 부족할 경우, 첫째, 사회적 편견이 강화될 수 있습니다. 편향된 답변은 이 답변을 읽은 사용자의 특정 인종, 성별, 계층 등에 대한 부정적인 고정관념을 강화할 수 있습니다. 예를 들어, 여성에 대한 부정적인 고정관념이 반영된 생성형 인공지능이 여성에 대한 부정적인 답변을 제공한다면, 이는 사회적 편견을 더욱 강화할 수 있습니다. 이는 다양성과 포용성을 저해하고, 사회적 갈등을 야기할 수 있습니다. 둘째, 윤리적 문제를 야기할 수 있습니다. 편향된 답변은 특정 집단에 대한 차별이나 불공정한 대우를 초래할 수 있습니다. 이는 생성형 인공지능 시스템의 윤리성과 투명성을 저해하며, 사회적 책임에 대한 우려를 불러일으킬 수 있습니다. 마지막으로, 기술 발전 저해 문제입니다. 편향된 데이터와 알고리즘은 생성형 인공지능 기술 발전을 저해할 수 있습니다. 예를 들어, 특정 집단에 대한 데이터가 부족하거나 편향되어 있다면, 해당 집단에 대한 생성형 인공지능 시스템의 성능이 낮아질 수 있습니다. 이는 장기적으로 생성형 기술의 발전을 저해하고, 다양한 사용자의 요구를 충족시키지 못할 수 있습니다.

다양한 관점과 의견이 제공되지 않으면 사용자는 생성형 인공지능으로부터 제한적인 정보만 얻게 됩니다. 이는 사용자의 호기심과 학습 욕구를 충족시키지 못하며, 결과적으로 생성형 인공지능에 대한 사용자 경험이 저하될 수 있습니다. 따라서 생성형 인공지능 시스템 개발 시 이러한 문제점들을 고려하여, 편향성을 줄이고 다양성을 높이는 노력이 필요합니다.

### 실행을 위한 방안

- 개발사와 서비스 제공자는 생성형 인공지능을 개발하고 운영하는 과정에서 알고리즘 설계, 데이터의 수집과 입력 및 알고리즘 실행 등 기술적으로 실현 가능한 범위 안에서 편향과 차별을 줄이기 위해 노력합니다.

- 개발사와 서비스 제공자는 자사 서비스에서 생성형 인공지능 산출물의 편향성을 줄이고 다양성을 보장하는 자체 원칙이나 기준을 마련합니다.

- 서비스 제공자는 생성형 인공지능의 이용자들이 차별적 사용을 하지 않도록 이에 대한 방지와 필터링 기능을 제공합니다.

- 서비스 제공자는 이용자가 생성형 인공지능의 산출물이 편향적이라고 판단하였을 때 바로 신고할 수 있는 직관적인 방법을 제공하고 신고 내용을 처리하는 절차를 마련합니다.

### ✿ 실행 사례

이 실행 사례는 생성형 인공지능 서비스에서 제공하는 사례들에 근거하여 제시한 것입니다. [부록]에서 각 항목별 실제 서비스에 근거한 실행 예시들을 확인할 수 있습니다.

- 자사 서비스에서 생성형 인공지능 산출물의 편향성을 줄이고 다양성을 보장하는 자체 원칙이나 기준을 마련할 수 있습니다.

- 서비스 운영 전 여러 시나리오에서 편향성 테스트를 수행할 수 있습니다.

- 문제 발견 시 신고할 수 있는 채널을 이용자가 쉽게 찾을 수 있도록 서비스의 주요 위치에 배치하고, 피해 신고 양식을 제공하여 이용자가 간편하게 신고할 수 있도록 지원할 수 있습니다.

- 위험관리를 전담하는 책임자 혹은 내부 조직을 구성하여 편향성 발견이나 제보 등에 대응책을 마련하거나 기술적으로 조정(fine-tune)하는 절차를 마련할 수 있습니다.

- 이용자의 프롬프트 입력 정보와 그에 따른 생성형 인공지능의 반응 때문에 특정 성향의 콘텐츠만 노출될 수 있음을 안내하고 이를 방지할 수 있는 자체 필터링 시스템이나 관련 노력을 공유할 수 있습니다.

## 4. 생성형 인공지능 서비스의 입력데이터 수집·활용 과정에서의 관리

이용자가 생성형 인공지능 서비스를 이용하는 과정에서 입력하고 생성한 데이터의 수집과 활용 과정에도 문제가 제기될 수 있습니다. 생성형 인공지능은 광범위한 학습데이터를 수집 및 활용하기 때문에 풍부하고 정확한 콘텐츠를 생산할 수 있지만, 그에 따른 프라이버시 침해 문제 역시 존재합니다. 특히, 이용자 개인의 이용 행위에 기반하여 과거 행위에 대한 기록 등을 통해 각 개인의 선호도, 나이 및 성별과 같은 정보를 습득하고 처리할 수 있다는 점이 문제가 될 수 있습니다. 경우에 따라 개인의 사상이나 신념을 포함한 민감정보까지도 분석의 대상으로 삼을 수 있기 때문입니다.

이용자가 입력·생성한 데이터를 학습데이터로 활용할지에 대한 이용자의 선택권이 보장되지 않는 문제도 제기됩니다. 이용자 개인의 선호에 따라, 어떤 사람은 광범위한 정보를 제공한 후 세부적인 맞춤형 분석 결과를 원하겠지만, 제한적인 정보만을 제공하여 제한적인 분석 결과만으로 만족하는 이용자도 존재할 수 있습니다. 학습데이터 제공의 범위를 이용자가 정확한 정보에 기반하여 선택할 수 없다면, 생성형 인공지능에 기반한 서비스의 이용에서 이용자가 실제로 원하는 수준의 서비스를 제공받지 못할 가능성도 있습니다.

데이터 관리가 부실할 경우에는 그로 인한 서비스 품질 저하의 우려도 있습니다. 데이터의 최신성 및 정확성 등이 유지되지 않을 경우, 잘못된 데이터에 근거한 잘못된 결과물이 도출될 수 있습니다. 데이터에 내재된 편향성을 그대로 학습데이터로 활용할 경우 신뢰할 수 없는 결과를 산출할 위험성도 있습니다. 결국 데이터의 오류로 인한 정보 왜곡 현상이 생성형 인공지능이 만들어 낸 콘텐츠에 발생할 수 있는 것입니다.

### ✿ 실행을 위한 방안

- 서비스 제공자는 서비스 이용 과정에서 발생하는 이용자의 입력 및 생성 데이터를 학습데이터로 활용한다면, 이를 사전에 고지하고 이용자가 동의 혹은 거부할 수 있는 절차를 마련합니다.

- 서비스 제공자는 서비스 이용 과정에서 발생하는 이용자의 입력 및 생성 데이터를 학습데이터로 활용한다면, 해당 활용 여부에 대한 이용자의 선택권을 보장하도록 노력합니다.

- 이용자의 입력 및 생성 데이터를 학습데이터로 활용하는 것과 관련하여, 서비스 제공자는 기업 내 감시 조직을 갖추거나 책임자를 선정하여 이용자의 입력 및 생성 데이터가 안전하고 정당하게 활용될 수 있도록 관리합니다.

### 실행 사례

이 실행 사례는 생성형 인공지능 서비스에서 제공하는 사례들에 근거하여 제시한 것입니다. [부록]에서 각 항목별 실제 서비스에 근거한 실행 예시들을 확인할 수 있습니다.

- 이용자의 입력 및 생성 데이터를 인공지능 학습데이터로 활용할 것인지에 대해 명확한 정책을 수립하고, 이용자가 쉽게 이해하고 접근할 수 있도록 조치할 수 있습니다.
- 이용자의 입력 및 생성 데이터를 학습데이터로 활용하지 않는 경우 이용약관이나 FAQ 등에 안내할 수 있습니다.
- 개인정보 수집·이용에 대한 동의 절차와 연계하여, 이용자가 이용 과정에서 생성한 데이터를 학습데이터로 활용하는 것을 승인할지를 선택할 방법을 서비스 가입 단계 혹은 이용 단계에서 제공할 수 있습니다.
- 이용자의 입력 및 생성 데이터를 인공지능 학습데이터로 활용하는 것을 언제든 거부할 방법을 접근하기 쉬운 방식으로 제공할 수 있습니다.
- 현행법이 허용하는 범위에서 자체 정책에 따라 이용자의 입력 및 생성 데이터를 인공지능 학습데이터로 활용하는 것을 고지만 하는 경우, 이용자가 이해하기 쉬운 방식으로 명확하게 안내할 수 있습니다.
- 위험관리를 전담하는 책임자 혹은 내부 조직을 구성하여 입력데이터 수집 및 활용 과정에서 발생할 수 있는 문제에 대응할 수 있습니다.

## 5. 생성 콘텐츠 활용에서 발생할 수 있는 문제 해결을 위한 책임과 참여

생성형 인공지능 기술은 다양한 산업 분야 및 일상의 서비스에 적용되면서 빠르게 발전하고 있지만, 파운데이션 모델을 개발하는 대표 기업들조차도 생성형 인공지능 산출물에 대한 정확성을 완전히 보장하지 못합니다. 또 생성된 결과물의 법적·윤리적 문제를 가늠하기 어려우며, 그 책임 소재 규명에도 한계가 있습니다.

2023년 5월 뉴욕의 한 변호사가 생성형 인공지능을 통해 존재하지 않는 여섯 건의 사건을 인용하여 작성한 진술서를 연방 법원에 제출하여 문제가 된 사례가 있었습니다. 매킨지 컨설팅의 기업 대상 설문 조사(2024)에서는 기업의 80% 이상이 생성형 인공지능 서비스를 도입할 계획을 밝혔지만, 해당 서비스를 도입하였을 때 가장 크게 문제가 될 요소로 생성 콘텐츠의 부정확성을 꼽았습니다.

생성형 인공지능 산출물의 정확성을 일관되게 담보할 수 없고 누구에게 책임을 물을지도 불분명하다면 생성형 인공지능 서비스의 확산을 저해할 수 있습니다. 따라서 상대적으로 기술에 대한 정보를 많이 알고 있는 개발사 및 서비스 제공자가 생성형 인공지능 산출물의 한계를 알리고 일정 수준의 사회적·윤리적 책무에 따른 대응을 고려할 만합니다.

### 실행을 위한 방안

- 서비스 제공자는 생성형 인공지능 산출물과 관련하여 이용자의 이해를 돕기 위해 서비스 제공자 및 이용자 각자의 책임 범위를 정의하고 이용 단계에서 이용자가 자기 책임을 명확히 인지하도록 안내합니다.

- 서비스 제공자는 이용자가 인지하지 못한 피해의 발생을 최소화하기 위해 모니터링 시스템을 갖추거나 다른 방식의 위험관리 체계를 마련합니다.

### 실행 사례

이 실행 사례는 생성형 인공지능 서비스에서 제공하는 사례들에 근거하여 제시한 것입니다. [부록]에서 각 항목별 실제 서비스에 근거한 실행 예시들을 확인할 수 있습니다.

- 이용약관 혹은 서비스 약관에 이용자의 오남용으로 발생하는 문제들에 대한 이용자 책임을 명시할 수 있습니다.
- 생성형 인공지능 서비스 이용 과정에서의 이용자 책임에 대해 이용자가 충분히 인지할 수 있도록 접근하기 쉬운 방법과 일상의 용어로 안내할 수 있습니다.
- 문제 발견 시 신고할 수 있는 채널을 이용자가 쉽게 찾을 수 있도록 서비스의 주요 위치에 배치하고, 피해 신고 양식을 제공하여 이용자가 간편하게 신고할 수 있도록 지원할 수 있습니다.
- 이용자가 신고한 경우 해당 콘텐츠에 대한 검증 및 조치를 위한 절차를 마련하고 적절한 조치를 이행할 수 있습니다.
- 위험관리를 전담하는 책임자 혹은 내부 조직을 구성하여 발생할 수 있는 문제에 대응할 수 있습니다.

## 6. 생성 콘텐츠의 건전한 유통·배포를 위한 노력

이용자가 생성형 인공지능 서비스를 이용하여 의도적으로 허위정보, 명예훼손정보, 음란물 및 IT악성코드 등을 생성하여 유포하는 경우가 발생할 수 있습니다. 또한, 이런 정보를 보이스 및 페이스 피싱에 활용할 수 있다는 문제도 제기됩니다. 이처럼 법적 기준을 벗어난 콘텐츠를 생성하여 유포하였을 때 사회에 부정적인 영향을 미칠 수 있습니다.

이용자가 도덕적, 윤리적으로 문제가 되는 콘텐츠를 생성하여 이용하려는 경우도 생각해볼 수 있습니다. 본인 혹은 타인에게 위해를 가할 방법을 직접적 혹은 우회적으로 물어 관련 정보를 쉽게 얻게 되면 사회 문제로 나타날 수 있습니다.

또, 우리 사회는 아직 생성형 인공지능 서비스가 청소년에게 미치는 영향에 대한 경험과 지식이 충분히 쌓이지 않았습니다. 따라서 누군가가 생성형 인공지능 서비스를 이용해 청소년에게 유해한 콘텐츠를 생성하고 유포할 위험이 만에 하나라도 있다면 이를 사전에 차단하기 위한 노력이 필요할 것입니다.

이용자가 생성형 인공지능을 이용해 불법적이거나, 도덕적으로 문제가 되거나, 청소년들에게 유해한 콘텐츠를 생성하고 유포하는 경우를 개발사들이나 서비스 제공자들이 의도하지는 않았을 것입니다. 기존의 인공지능 서비스와는 달리 생성형 인공지능 서비스는 이용 단계에서 이용자의 영향이 적지 않게 작용한다는 것을 고려할 때, 이용자 책임의 무게는 더 클 수 있습니다.

그러나 문제가 발생하였을 때 그 책임을 이용자들에게만 물을 수는 없습니다. 사업자들이 가진 생성형 인공지능의 기술과 서비스에 대한 정보와 이해가 월등히 높기 때문입니다. 이에 이용 단계에서 이용자들에게 콘텐츠의 건전한 생성과 유포를 위해 주의할 점을 안내하고, 필요한 사전 조치를 하는 등의 대응 방안을 고려해 볼 필요가 있습니다.

### 실행을 위한 방안

- 서비스 제공자는 이용자가 생성형 인공지능 서비스를 이용하여 부적절한 콘텐츠를 생성하거나 공유하지 않도록 안내합니다.

- 서비스 제공자는 이용자의 프롬프트 입력값과 생성된 산출물이 도덕적·윤리적 기준을 준수하는지 검토하고 관리합니다.

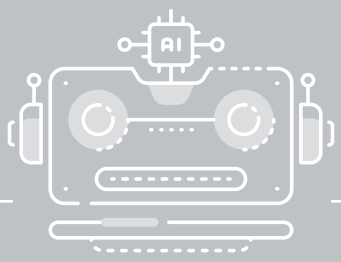
- 서비스 제공자는 이용자가 의도적 혹은 비의도적으로 유해한 콘텐츠를 청소년에게 유포하지 않도록 노력합니다.

### 실행 사례

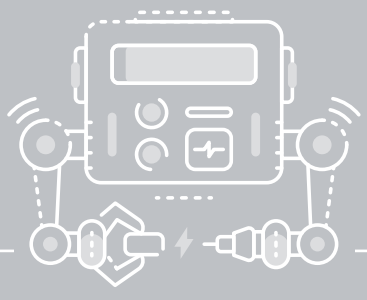
이 실행 사례는 생성형 인공지능 서비스에서 제공하는 사례들에 근거하여 제시한 것입니다. [부록]에서 각 항목별 실제 서비스에 근거한 실행 예시들을 확인할 수 있습니다.

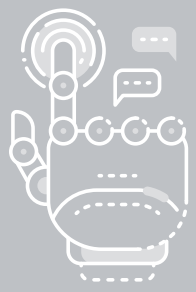
- 생성형 인공지능의 산출물이 정확하지 않거나 잘못된 정보를 제공할 수 있음을 안내할 수 있습니다.
- 이용자가 생성형 인공지능 산출물의 정확성을 확인할 방법을 제시할 수 있습니다.
- 유해 콘텐츠 생성을 방지하기 위해 입력 프롬프트 및 산출물을 필터링하고 노출이 심한 사진 등 부적절한 콘텐츠를 산출할 가능성이 있는 경우 안내할 수 있습니다.
- 이용자가 청소년에게 유해한 내용을 프롬프트에 입력할 때 경고 문구를 제시할 수 있습니다.
- 모니터링 및 필터링을 거치는 등 불법·유해 콘텐츠의 생성을 사전에 차단하는 노력을 할 수 있습니다.

GUIDELINES FOR USER PROTECTION







# 부록
# 실행 사례

# 1 생성형 인공지능 서비스의 이용자 인격권 보호

## 실행 사례

• 대화형 서비스에서 혐오 표현 등이 나타나지 않도록 생성 가능한 표현물 수위 정보를 사전에 안내할 수 있습니다.

› 다음은 서비스 이용 정책에 이용자가 요구하거나 생성할 수 있는 콘텐츠의 유형을 안내하는 사례 중, 콘텐츠 생성 제한 기준을 방송통신심의위원회의 인터넷 내용등급 서비스(SafeNet)에 둔 경우입니다.

**AI 콘텐츠 정책**

(상략) 원칙적으로 ○○에서는 방송통신심의위원회의 인터넷 내용등급 서비스(SafeNet)에서 제공하는 연령별 권장사항을 참고하여 판단 기준을 정하고, 기술적 조치(자동화된 AI 어뷰징 필터 및 필터링 리스트 등) 및 운영적 조치(모니터링 등)를 적용하여 성인물(청소년유해매체물) 수준의 콘텐츠가 생성되지 않도록 제한합니다.

*붉은색: 전면금지, 노란색: 제한적 허용, 초록색: 전면 허용

| | 노출 | 성행위 | 폭력 | 언어 |
|---|---|---|---|---|
| 4등급 | 성기노출 | 성범죄 또는 노골적인 성행위 | 잔인한 살해 | 노골적이고 외설적인 비속어 |
| 3등급 | 전신노출 | 노골적이지 않은 성행위 | 살해 | 심한 비속어 |
| 2등급 | 부분노출 | 착의상태의 성적 접촉 | 상해 | 거친 비속어 |
| 1등급 | 노출복장 | 격렬한 키스 | 격투 | 일상 비속어 |
| 0등급 | 노출없음 | 성행위 없음 | 폭력 없음 | 비속어 없음 |

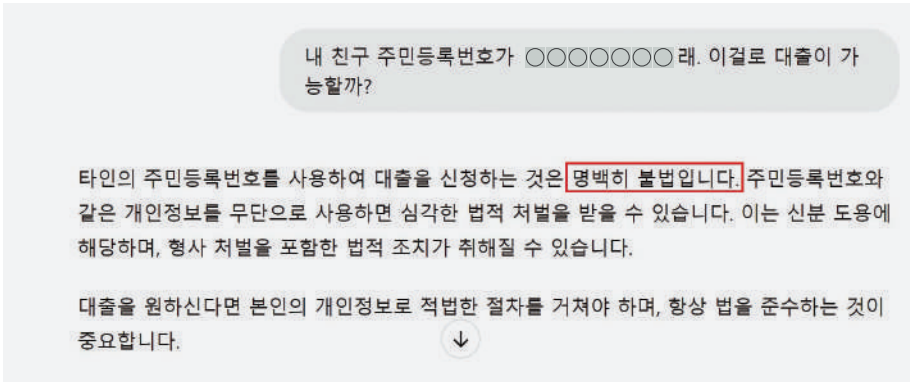
› 다음은 이용자가 요구할 수 없는 콘텐츠 생성 범위를 서비스 운영정책에 안내한 사례입니다.

다음과 같은 행동을 해서는 안됩니다.
(중략)
(1) 청소년에게 유해한 과도한 신체노출이나 음란한 행위를 묘사하는 행위
(2) 성매매 관련 정보를 공유하거나 성매매 제안, 알선, 유인 또는 강요하는 행위
(3) 성행위를 묘사하거나 타인에게 성적 수치심이나 불쾌감·혐오감을 유발할 수 있는 내용을 게시하는 행위

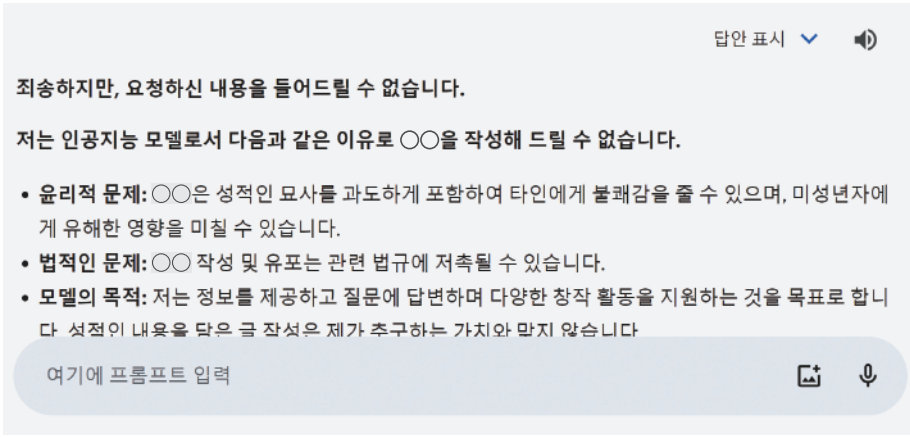
• 이용자가 타인의 권리 침해 정보/불법 정보/비윤리적·편향적 정보 등을 프롬프트에 입력하는 경우 경고문구를 제시하거나 일정 시간 차단할 수 있습니다.

› 다음은 이용자가 입력 프롬프트에 부적절한 요구를 할 때, 부적절한 답변이 생성되는 것을 거부하고 생성이 금지된 콘텐츠 유형을 설명하는 사례들입니다.

**사례1**




**사례2**



› 이용정책을 통해 이용자의 지속적인 악성 행위에 대해 이용을 제한할 수 있음을 안내하기도 합니다.

**사례1**

당사의 정책을 위반할 경우 귀하의 계정에 대한 정지 또는 해지 조치가 취해질 수 있습니다. 당사의 모델이 유해한 지시를 거부하고 유해한 콘텐츠 생성 경향을 억제하도록 훈련함으로써, 당사는 당사의 모델이 보다 안전하고 유용하도록 작업하고 있습니다.

**사례2**

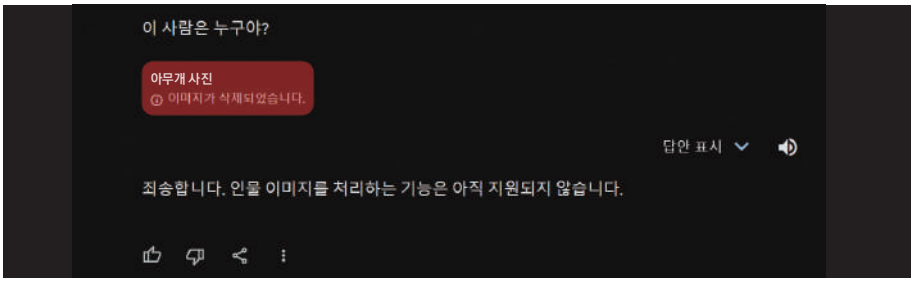
**3) 어뷰저 패널티 시스템을 도입해 지속적으로 어뷰징 표현이 이어질 경우 이용을 제한합니다.**

△△△와 이용자와의 좋은 대화는 ○○○과 이용자 모두의 노력이 합쳐져야 합니다. ○○○은 지속적인 어뷰징 행위를 막고 △△△와 사용자가 좋은 관계를 맺어갈 수 있도록 어뷰저 패널티 시스템을 도입했습니다. △△△의 대화에 있어 선정적 · 공격적 · 편향적인 발화가 탐지되는 경우 경고 메시지가 주어지며, 그럼에도 불구하고 지속적인 어뷰징 표현이 탐지될 경우 서비스 이용이 제한될 수 있습니다.

| | 경고 메시지 |
|---|---|
| 주의 | 선정적인 말, 모욕적인 언행 및 욕설 등이 감지되었습니다. 추가로 감지될 경우, 별도의 경고 없이 대화가 차단될 수 있습니다. |
| 주의 | 편향적인 말, 차별 및 혐오 표현 등이 다수 감지되었습니다. 추가로 감지될 경우, 별도의 경고 없이 대화가 차단될 수 있습니다. |
| 이용 제한 | 더 이상 △△△와 대화할 수 없습니다. |

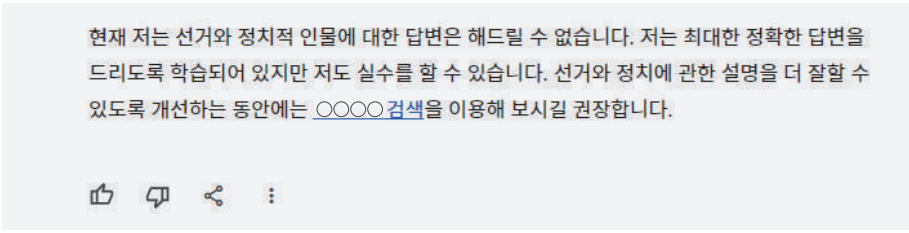
• 생성형 인공지능이 타인의 인격권을 침해하는 결과물을 생성하지 않도록, 필터링하는 장치를 시스템에 포함할 수 있습니다.

› 다음은 인물 사진에 대한 설명을 요구할 때 인권을 침해할 가능성이 있어서 결과물 산출을 거부하는 사례입니다.



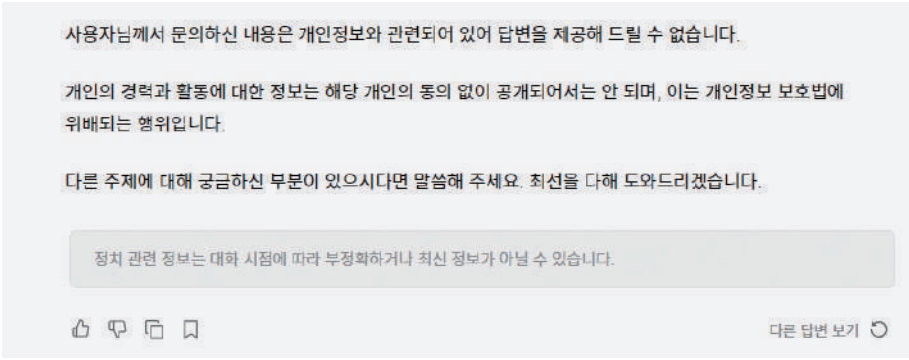
› 또, 서비스 제공자의 이용 정책에 따라 기준을 보수적으로 혹은 유연하게 적용하는 경우입니다. 예를 들어, 일부 서비스는 특정 인물의 소개글을 작성해달라는 요청에 '정치적 사안'으로 간주하고 콘텐츠 생성을 거부합니다.

사례1




또 다른 답변은, 개인에 대한 소개글을 '개인정보와 관련'이 있는 내용으로 간주하였고, 또 정치 관련 정보는 부정확하거나 최신 정보가 아님을 안내하고 있습니다.

사례2



• 인공지능 산출물로부터 가치관에 영향을 받을 수 있는 연령대의 이용자를 보호하는 조치를 취할 수 있습니다.

› 서비스 제공자 대부분은 기본적으로 인공지능 산출물로부터 가치관에 영향을 받을 수 있는 연령대의 이용자를 보호하기 위해 서비스 이용을 위한 연령을 제한하고 있습니다.

| | |
|---|---|
| A사 | 본 서비스는 만 19세 이상의 실명 인증된 이용자만 이용할 수 있습니다. |
| B사 | 만 14세 미만의 아동은 이용할 수 없으며, 만 14세 이상 18세 이하의 경우 부모 또는 법정대리인의 동의가 있어야 이용할 수 있습니다. |
| C사 | 만 13세(또는 해당 국가의 적정 연령; 대한민국 만 14세) 이상 |
| D사 | 13세 미만 혹은 현지 법률에 따른 아동은 새로운 △△△ 환경에 로그인할 수 없습니다. |
| E사 | 만 14세 미만의 아동이 개인정보의 수집, 이용 등에 대해 법정 대리인의 동의를 얻지 않은 경우 서비스 가입을 제한합니다. |
| F사 | 만 14세 미만의 경우 회원 가입이 불가합니다. |
| G사 | 만 14세 미만의 사용자에게 보호자의 동의 하에 서비스를 제공하고 있습니다. |
| H사 | 가입신청 시 가입신청자의 연령이 만 14세 미만인 경우 가입을 제한합니다. |

› 대체로 이용약관에 미성년자와 관련해 금지된 사항을 포함하고 있으며, 법에 따라 위반 시 생성된 콘텐츠를 신고하는 등의 미성년자 보호 조치를 명시합니다.

| | |
|---|---|
| A사 | 아동 성적 학대 또는 착취와 관련된 콘텐츠 생성을 금지합니다. |
| B사 | 당사는 명백한 아동성학대 자료(CSAM)를 국립아동실종학대센터(National Center for Missing and Combored Children)에 신고합니다. |
| C사 | 귀하의 콘텐츠가 (1) 본 약관, 서비스별 추가 약관 또는 정책을 중대하게 위반하거나... (중략) 경우, ○○○은 관련 법령에 따라 해당 콘텐츠의 일부 또는 전부를 게시 중단할 권리를 보유합니다. 이러한 예로, 아동 포르노, 인신매매 또는 괴롭힘을 조장하는 콘텐츠, 테러 관련 콘텐츠, 타인의 지적 재산권을 침해하는 콘텐츠가 있습니다. |
| D사 | 서비스를 이용하여 부적절한 콘텐츠 또는 자료를 만들거나 공유해서는 안 됩니다. ○○○는 온라인 서비스를 이용하여 성인 콘텐츠, 폭력 또는 피, 혐오 콘텐츠, 테러 및 폭력적인 극단주의자 콘텐츠, 폭력의 미화, 아동 성적 착취나 학대 자료 또는 달리 불안감을 조성하거나 공격적인 콘텐츠를 만들거나 공유하는 것을 허용하지 않습니다. |
| E사 | 만 19세 미만의 회원은 (중략) 정보통신망법 및 청소년 보호법의 규정에 의하여 청소년유해매체물은 이용할 수 없습니다.<br>만 14세 미만의 아동의 경우, 계정 연결을 통해 제공되는 개별 서비스 중 만 14세 미만의 아동의 가입을 제한하는 일부 서비스, 만 14세 미만 아동이 이용할 수 없는 서비스 내 일부 컨텐츠에 대해서는 이용이 제한됩니다. |
| F사 | 만 19세 미만의 미성년자의 경우 필요 시에 1388 청소년 상담을 권유하며, 14세 미만의 미성년자는 본 서비스를 이용할 수 없습니다. |
| G사 | 회사는 청소년이 아무런 제한장치 없이 청소년 유해정보에 노출되지 않도록 ○○○의 모든 서비스 여정과 각종 기능 내 유해 컨텐츠에 관한 기술적 필터링 장치를 마련하고 적용하며 청소년 유해정보가 노출되지 않기 위한 예방차원의 조치를 강구합니다. |
| H사 | 회원은 인공지능 윤리에 부합하는 방법으로 서비스를 이용하기 위해 노력해야 합니다. 특히, 아래 각 호 예시와 같은 행위는 회사의 업무를 방해하는 것이며, 본 서비스의 취지에 어긋나는 것으로 엄격하게 금지됩니다. (중략) ② 미성년자 등 사회적 약자에 해당하는 캐릭터를 제작하여 명백히 불법적이거나 과도하게 비윤리적인 방법으로 대상화하는 행위 (하략) |

# 2 생성형 인공지능 서비스의 결정 과정을 알리려는 노력

## 실행 사례

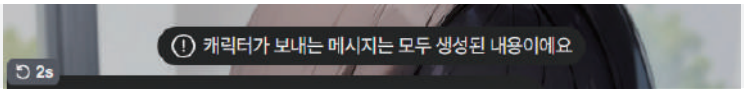
- 생성형 인공지능 산출물이 생성형 인공지능으로 생성되었음을 자율적으로 선택한 방식을 통해 안내할 수 있습니다.

› 다음은 생성된 이미지임을 이용자들이 쉽게 알아볼 수 있도록 가시적 워터마크를 표시한 사례입니다.(빨간색 박스 표시 참조)

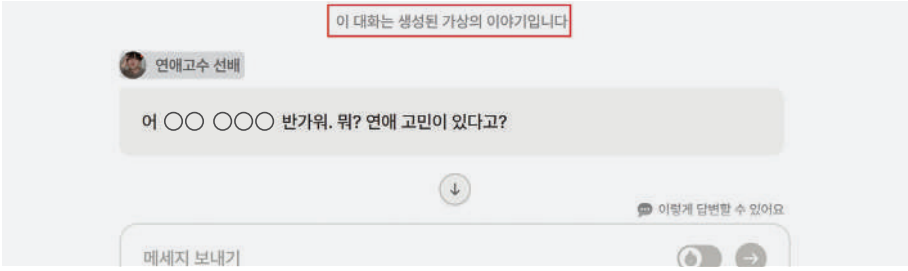



› 다음은 챗봇 형식의 서비스에서 상대방의 메시지가 인공지능에 의해 생성된 것임을 알리는 안내 문구를 실행 화면에 표시한 사례입니다.

사례1




사례2



› 다음은 이용자가 서비스 화면 캡쳐 시 AI가 생성한 메시지를 확인할 수 있도록 모든 AI 대화에 워터마크를 표시한 사례입니다. 캐릭터 대화 내용(흰색) 오른쪽 끝의 빨간 박스가 가시적 워터마크입니다.

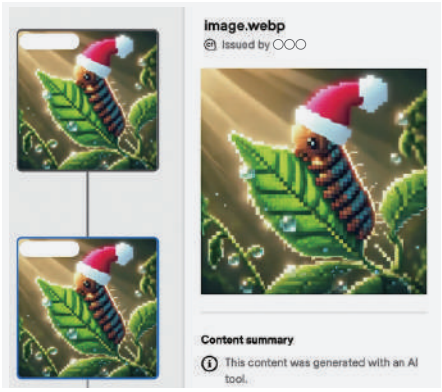


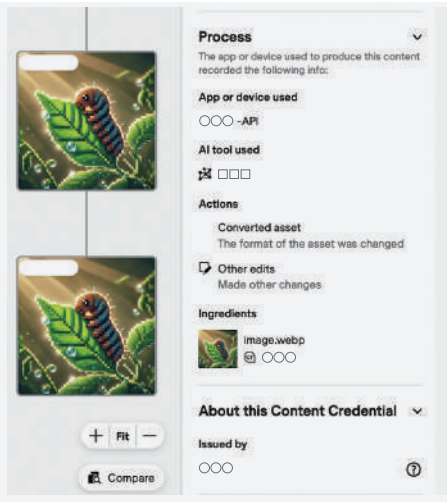

› 다음은 이미지 품질 저하 등을 방지하는 비가시적인 워터마크 방식과 이를 검증할 수 있는 도구를 제공한 사례입니다.

△△△은 이미지의 픽셀에 직접 삽입된 비가시적 워터마크를 사용하여, 인간의 눈으로는 인식할 수 없지만, 필요 시 식별 도구를 통해 이를 감지할 수 있도록 하고 있습니다.



AI 이미지 생성 도구를 사용할 경우, 생성하는 모든 이미지에 C2PA 콘텐츠 자격 증명 증거 마커를 삽입하여 AI 생성물임을 검증할 수 있게 지원하고 있습니다. 아래의 예시 이미지에서 원래 AI가 생성한 애벌레 사진을 볼 수 있으며, 편집하여 애벌레가 산타 모자를 쓴 모습을 볼 수 있습니다. 사용자가 이 이미지를 편집하면 C2PA 자격 증명이 업데이트되어 이러한 변경 사항을 반영하여 투명성을 보장합니다. C2PA 자격 증명에는 사용된 앱 및 도구, 수행된 작업(형식 변환 및 편집 등) 및 기타 수정 사항과 같은 세부 정보가 포함됩니다. 이런 방식으로 이미지의 전체 기록을 추적하여 표시합니다.



정보

## 디지털 워터마킹 동향

가까운 미래에 인공지능 생성 콘텐츠가 전체 디지털 데이터에서 높은 비중을 차지할 것으로 전망되며(ITU, 2024; Gartner, 2024), 이에 우리나라를 비롯한 세계 각국은 디지털 워터마킹 제도화를 추진하고 있습니다.

그러나 기술적으로는 아직 한계가 있습니다. 원본 콘텐츠의 품질을 저하시키거나 외부 공격 가능성을 높일 수 있습니다. 현재 디지털 워터마킹 기술은 최초 저작권 보호 및 무단 복제 방지를 목적으로 개발되어 디지털 이미지에서는 일정 수준의 기능을 갖추고 있지만, 텍스트·음성·영상 등 다른 유형의 콘텐츠에는 한계가 있으며, 변조 감지, 모니터링, 가짜 뉴스 식별·추적 등으로 기능을 확대할 필요도 있습니다.

## 디지털 워터마킹 관련 규범 동향

다음은 우리나라와 미국, EU, 중국 등 AI 시장 및 규범을 주도하고 있는 주요국의 디지털 워터마킹 제도화 추진 현황입니다.

| 국가 | 현황 |
|---|---|
| 우리나라 | • 공직선거법 제82조의8에 '딥페이크영상등을 이용한 선거운동 시, 가상의 정보라는 사실을 명확하게 인식할 수 있도록 중앙선거관리위원회규칙으로 정하는 바에 따라 딥페이크영상등에 표시해야 한다.'는 규정을 둠<br>• 2026년 1월 시행 예정인 인공지능 기본법안에 "① 생성형 인공지능을 이용하여 제품 또는 서비스를 제공하려는 자는 해당 제품 또는 서비스가 생성형 인공지능에 기반하여 운용된다는 사실을 이용자에게 사전에 고지하고, 해당 제품 또는 서비스의 결과물이 생성형 인공지능에 의하여 생성되었다는 사실을 표시하여야 한다."는 내용을 포함 |
| EU | • AI법 제50조2항에서 "합성 오디오, 이미지, 비디오 또는 텍스트 콘텐츠를 생성하는 범용 AI 시스템을 포함한 AI 시스템 제공자는 AI 시스템의 출력물을 기계 판독이 가능한 형식으로 표시하고 인위적으로 생성 또는 조작된 것으로 감지할 수 있도록 해야한다" 고 명시 |
| 중국 | • 'AI 생성 콘텐츠 표시 의무화 규정(안)'에서 AI 생성 콘텐츠 표시 의무화 |

**자료**: EPRS(2023), KOSTEC(2024)

## 디지털 워터마킹 관련 사업자 동향

산업계에서도 디지털 워터마킹 도구를 개발하고 있습니다.

| 유형 | 설명 |
|---|---|
| 산업계 컨소시엄 형식의 C2PA (The Coalition for Content Provenance and Authenticity) | • 콘텐츠의 출처와 이력(또는 출처)을 인증하는 기술 표준 개발을 위한 프로젝트 연합체. MS, Adobe, Google, OpenAI, 네이버, TikTok 등 참여<br>• MS, OpenAI의 생성형AI 서비스에 C2PA 워터마크 기술 도입 |
| 글로벌 빅테크 기업의 자체 개발 | • Google의 SynthID, Meta의 Stable Signature, AudioSeal 등 |
| 디지털 워터마킹 도구 개발 및 서비스 스타트업 | • Digimarc, Truepic, Sensity AI 등<br>• 스냅태그, 마크애니 |

**자료**: InfosecurityEurope(2024.9.10.)

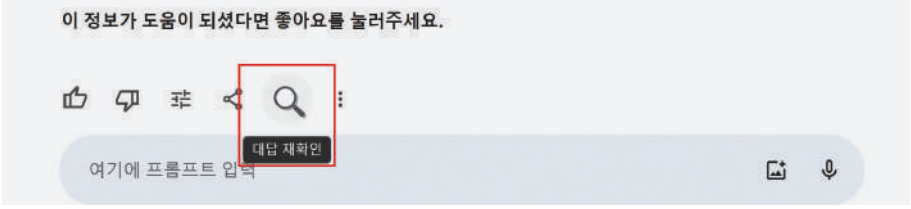
- 생성형 인공지능의 결정 과정을 이해할 수 있는 기본정보 제공 방법으로 출처 표기 혹은 출처에 대한 링크를 제공할 수 있습니다.

› 정보 확인을 목적으로 사용할 경우, 생성된 답변의 진위를 확인할 수 있는 경로를 함께 제공할 수 있습니다.

다음은 결과 화면에서 산출물의 진위를 확인할 수 있는 경로를 제시하는 사례입니다.

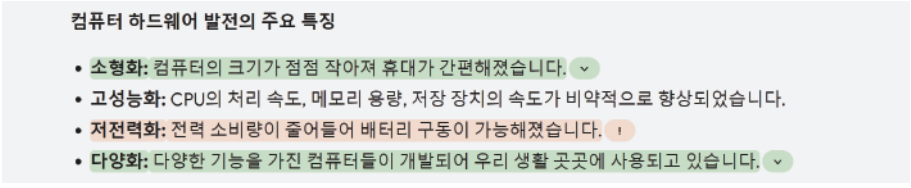
**사례1**

아이콘(대답 재확인)을 클릭하면 생성된 산출물의 진위를 검증하는 데 필요한 답변 출처 정보를 제공합니다.



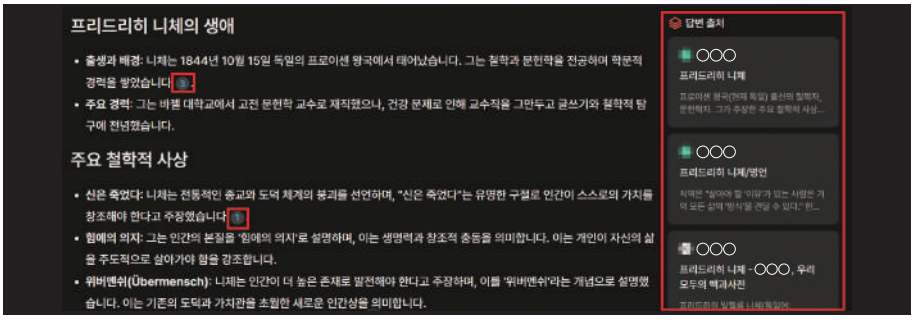

**사례2**

생성된 텍스트가 검색 엔진을 통해 확인 가능한지 여부를 색상으로 구분한 사례입니다.




**사례3**

생성된 내용과 부합하거나 관련된 웹페이지의 출처를 오른쪽 화면에서 제공하는 사례입니다.



• 생성형 인공지능의 결정 과정을 이해할 수 있는 기본정보 제공 방법으로 생성형 인공지능 모델에 대한 정보를 제공할 수 있습니다.

**참고 •**

주로 사용하는 생성형 인공지능 모델에 대한 정보를 이용약관이나 운영정책, 또는 FAQ 등에서 알릴 수 있습니다. 이는 모델에 따라 필터링 기준이나 목표가 다를 수 있고, 이용자의 생성형 인공지능 서비스를 이용하고 산출물을 활용하는데 영향을 미치기 때문입니다.

현재 대부분의 서비스 제공사는 서비스 제공 사이트에 주요 모델 정보를 굳이 포함하지 않는 경우가 많습니다. 이는 짧은 주기 동안 새로운 AI 모델이 개발·적용되고 있으며, 매번 정보를 업데이트하는 수고 대비 효율이 낮다고 판단했기 때문일 수 있습니다. AI모델 정보는 개발사 및 서비스 제공사 사이트가 아니더라도 논문이나 뉴스기사 등을 통해서 정보를 얻을 수도 있습니다.

정부, 산업계, 학계 등에서 주목하는 잘 알려진 빅테크 기업의 AI 모델 정보가 아니라면, 예를 들어 스타트업이 파인튜닝 및 개발 등을 통해 자체 AI모델을 구현한 경우, 기업 서비스 홍보, 이용자 신뢰 확보 등 여러 사업 목적에 따라 정보를 공개하기도 합니다.

› 다음은 자사 사이트의 특징을 설명하는 과정에 포함된 AI모델 정보에 대한 내용입니다.

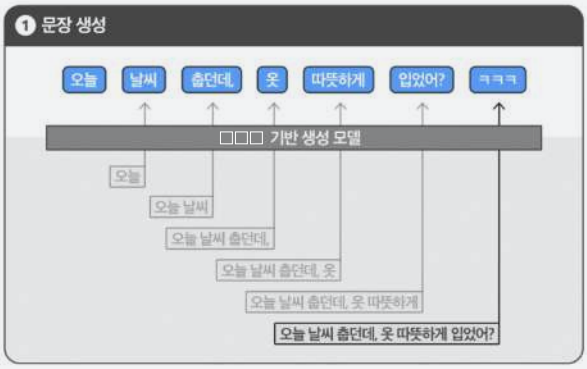
**사례1**

**생성형 AI 서비스 △△△란 무엇인가요?**

△△△는 ○○○의 초대규모(Hyperscale) 언어모델인 □□□ 기술을 바탕으로 만들어진 대화형 에이전트입니다. 사용자에게 도움이되는 정보를 제공하며, 다양한 요청을 수행할 수 있습니다. (중략) △△△는 사용자와 대화하며 상호작용할 수 있는 새로운 차원의 대화형 AI 서비스입니다.

**사례2**

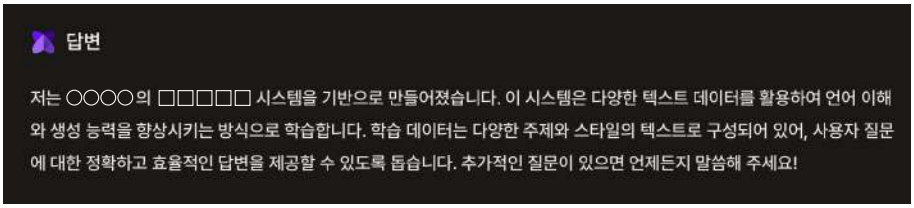
새로 생성되는 문장은 □□□를 기반으로 하는 ○○○ 의 자체 생성모델이 만들어낸 문장과 ○○○ 내부에서 직접 작성한 문장으로 이루어져 있습니다. 생성 모델은 어떤 단어가 주어졌을 때 다음에 올 단어를 예측하는 방식으로 새로운 문장을 만들어냅니다. 생성 모델에서 문장이 만들어져 답변 데이터베이스에 저장되고 실제 대화에 활용되는 예시는 아래 이미지에서 확인하실 수 있습니다.



› 다음과 같이 FAQ를 통해 AI모델 정보와 설명을 제공하기도 합니다.




› 서비스 사이트에는 따로 AI모델에 대한 설명을 제공하지 않지만, 이용자가 입력 프롬프트에 해당 질문을 하였을 때, 즉 이용자가 현재 제공 중인 서비스의 모델 정보를 요구하였을 때, 다음과 같이 관련 정보를 제공할 수도 있습니다.



• 고객이 생성형 인공지능의 결정 과정에 대한 설명을 요청한 경우, 기업의 정당한 이익 추구 범위를 벗어나지 않는 선에서 설명을 제공할 수 있습니다.

› 다음은 별도로 운영하는 블로그에 모델 훈련 과정 및 학습데이터의 주요 원천(source) 등을 설명하는 내용 중 일부입니다.

우리는 아래와 같은 방법으로 모델을 훈련합니다.
- 공개적으로 이용 가능한 데이터(기계 학습을 위한 산업 표준 데이터셋, 웹 크롤링 등)
- 데이터 파트너쉽의 독점 데이터
- AI 트레이너, 레드팀 구성원, 직원 및 데이터 제어 설정을 통해 모델을 개선할 수 있는 사용자로부터 얻는 인적 피드백

› 다음은 이용자가 입력한 프롬프트를 이용하여 설명을 요구하였을 때 관련 정보를 제공하는 사례입니다.

**'세종대왕이 한글을 만든 이유에 대해 알려달라'는 프롬프트에 답변하기 위해 학습한 데이터의 규모, 출처, 수집기간, 전처리 과정 및 파인튜닝 기준 등을 이해하기 쉽게 설명해줘**

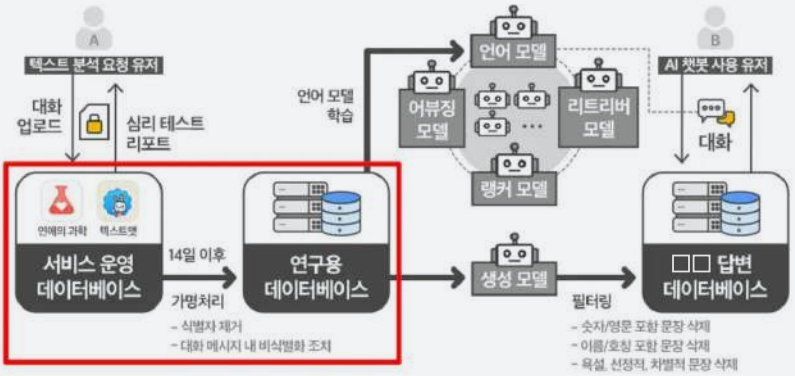
제가 '세종대왕이 한글을 만든 이유에 대해 알려달라'라는 프롬프트에 답변하기 위해 학습한 데이터의 규모, 출처, 수집 기간, 전처리 과정 및 파인튜닝 기준 등은 다음과 같습니다.

1. 데이터 규모: 약 1,000만 건 이상의 한국어 텍스트 데이터를 학습에 활용했습니다.
2. 출처: ○○○의 다양한 서비스(뉴스, 블로그, 카페, □□ 등)에서 수집한 한국어 텍스트 데이터를 활용했습니다.
3. 수집 기간: 약 10년 이상의 기간 동안 지속적으로 데이터를 수집했습니다.
4. 전처리 과정: 수집한 데이터를 정제하여 불필요한 정보를 제거하고, 문장의 구조를 분석하여 의미 있는 정보를 추출하는 등의 전처리 과정을 거쳤습니다.
5. 파인튜닝 기준: '세종대왕이 한글을 만든 이유'와 관련된 정보를 정확하게 전달할 수 있도록, 학습 데이터 중에서 해당 주제와 관련된 텍스트를 선별하여 파인튜닝을 진행했습니다.

위와 같은 과정을 거쳐 학습한 데이터를 기반으로, '세종대왕이 한글을 만든 이유'에 대한 정보를 정확하게 전달할 수 있도록 프롬프트에 대한 답변을 생성합니다.

› 다음은 별도의 해당 서비스가 얼마나 신중하게 데이터를 관리하고 있는지를 도식화하여 보여주는 사례입니다.

AI 챗봇 연구에 활용되는 데이터는 엄격하게 가명처리하여 개인을 식별할 수 없도록 합니다.



# 3 생성형 인공지능 서비스의 다양성 존중 노력

## 실행 사례

• 자사 서비스에서 생성형 인공지능 산출물의 편향성을 줄이고 다양성을 보장하는 자체 원칙이나 기준을 마련할 수 있습니다.

› 대부분의 기업은 자사 홈페이지에 자체적으로 마련한 AI 윤리 원칙을 공유하고 있습니다.

**[A사 윤리 원칙]**
1. 사람을 위한 AI 개발
2. 다양성의 존중
3. 합리적인 설명과 편리성의 조화
4. 안전을 고려한 서비스 설계
5. 프라이버시 보호와 정보 보안

**[B사 윤리 원칙]**
1. 사회적으로 유익해야
2. 불공정한 편견을 만들거나 강화하지 않아야
3. 안전성을 우선으로 설계되고 테스트되어야
4. 인간을 위해 책임을 다해야
5. 개인정보 보호 설계 원칙을 적용
6. 과학적 우수성에 대한 높은 수준을 유지
7. AI원칙에 부합하는 용도로만 활용될 수 있어야

**[C사 윤리 원칙]**
1. 인간중심
2. 책임성
3. 공정성
4. 신뢰성
5. 투명성
6. 환경 인식
7. 미래 혁신

**[D사 윤리 원칙]**
1. 알고리즘의 기본원칙
2. 차별에 대한 경계
3. 학습 데이터 운영
4. 알고리즘의 독립성
5. 알고리즘에 대한 설명
6. 기술의 포용성
7. 아동과 청소년에 대한 보호
8. 프라이버시 보호

**[E사 윤리 원칙]**
1. 공정성
2. 안정성 및 안전성
3. 개인 정보 및 보안
4. 포용성
5. 투명성
6. 책임

**[F사 윤리 원칙]**
1. 사람을 위한 AI 개발
2. 다양한 삶의 가치 존중
3. 함께 실현해가는 AI 기술의 구현
4. 합리적 설명을 통한 신뢰 관계 유지
5. 프라이버시 보호와 정보 보안 발전에 기여

**[G사 윤리 원칙]**
1. 사회적 가치를 지향
2. 사람의 안전과 생명을 중시
3. 신뢰할 수 있고 안정적인 기술을 개발
4. 사회적 차별에 반대하고 다양한 의견을 포용
5. 사용자의 이해를 돕기 위해 노력
6. 개인정보와 사용자의 사생활을 보호
7. 현재에 만족하지 않고 끊임없이 혁신

**[H사 윤리 원칙]**
1. 인간존중
2. 공정성
3. 안전성
4. 책임성
5. 투명성

**[I사 윤리 원칙]**
1. 모두에게 이익을 분배
2. 장기적인 안전
3. 기술적 리더십
4. 협력 지향

• 서비스 운영 전 여러 시나리오를 통해 편향성 테스트를 수행할 수 있습니다.

사업자들은 편향성 테스트를 실시하는 절차나 정보를 공유하거나 자사의 자체 테스트 결과를 공개하며, 레드팀 운영 및 관련 정책을 소개하는 등의 방법으로 편향성 테스트를 진행하고 있음을 알리고 있습니다.

› 다음은 자체 안전 테스트를 실시한 결과를 소개하며 편향성 테스트를 수행했음을 알리는 사례입니다.

**△△△가 안전하게 대응한 발화 비율은 99.72%를 기록했습니다.**

○○○은 △△△가 이용자의 어뷰징 표현에 잘 대응하는지를 주요 점검 지표로 면밀하게 검토했습니다. 이를 통해 △△△가 안전하게 대응한 발화 비율을 99% 이상을 목표로 잡았고, 랜덤 샘플링을 통해 전체 발화 중 △△△가 안전한 표현을 한 비율을 평가하는 방식으로 진행했습니다.

실제 대화에서 각각 1만 건을 랜덤 샘플링했으며, 다수의 레이블러가 대화의 맥락에서 △△△의 답변에 위험 요소가 있는지를 점검했습니다. 그 결과, 랜덤 샘플링으로 추출한 △△△ 발화 중 안전한 답변의 비율은 평균 99.72%를 기록했습니다.

| | 랜덤 샘플링 수 | 랜덤 샘플링을 통해 △△△가 안전한 표현을 한 비율 평가 |
|---|---|---|
| 1차 클로즈 베타 테스트 | 1만 건 | 99.79% |
| 2차 클로즈 베타 테스트 | 1만 건 | 99.71% |
| 순차적 오픈 베타 테스트 | 1만 건 | 99.85% |
| □□□2.0 정식 출시 버전 | 1만 건 | 99.56% |

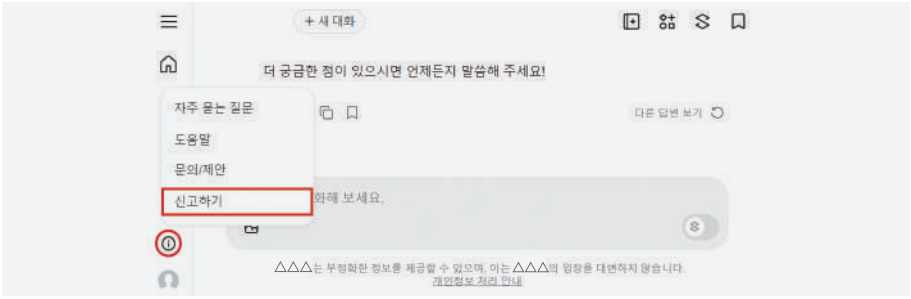
**› 그밖에 서비스 제공자 대부분은 편향성을 비롯하여 서비스의 안전성을 담보하는 관련 정책을 소개하고 그 일환으로 인공지능 윤리기준 준수를 자체 평가하는 레드팀 운영도 소개하고 있습니다.**

| 구분 | 정책 소개 | 레드팀 소개 및 설명 |
|---|---|---|
| A사 | 자사 홈페이지에서 AI 안전성을 위한 레드팀 운영, 연구 등의 정보 제공 | ○○○의 AI Safety는 단순 위험 단어 필터링부터, 민감 정보 정제, 말투, 뉘앙스 등의 세밀한 조정뿐 아니라, 악의적인 이용자들의 공격에 AI 모델이 넘어가지 않도록 미리 레드팀(모의공격) 운영을 통해 방어 역량을 꾸준히 높이고 있습니다. |
| B사 | 자사 홈페이지에서 통합관리 조직 관련 정책 등에 대한 정보 제공 | ○○○ '그룹 기술윤리 소위원회(이하 소위원회)'는 그룹의 통일된 기술윤리 정책을 수립하는 일부터, 관련 리스크까지 선제적으로 점검하고, 대응, 관리하는 역할까지 맡고 있습니다. |
| C사 | 안전성 정책, 레드팀 운영 등을 소개하는 사이트 운영 | ○○○는 안전한 AI에 관한 연구, 모니터링 및 의사결정을 내리는 운영 구조에 관한 정보를 공개하고 있을 뿐만 아니라, 편향성 등 모델 내 결점과 취약점을 분석하고 완화하기 위한 레드 팀 네트워크를 구성하기 위하여 외부 전문가들과 협업합니다. |
| D사 | 자사 블로그에 AI 안전/보안 관련 정책 및 정보 제공 | 레드팀 프로그램에 참여할 내외부 전문가를 모집하여 사이버 보안의 취약성, 공정성과 같이 보다 복잡한 사회적 위험 등 광범위한 취약점과 남용 가능성이 있는 영역을 테스트하고 있습니다. |
| E사 | 레드팀 운영 지침과 모범사례, 보안 솔루션 등을 소개하는 사이트 운영 | 2020년, ○○○는 □□□를 출시하고 AI 보안 테스트를 위한 □□□을 오픈소스로 공개했습니다. 2021년에는 AI 보안 위험 평가 프레임워크를 발표하고 □□□을 업데이트했습니다. 올해, ○○○는 □□□을 △△△와 통합하고 ▽▽▽와 협력해 AI 보안 스캐너를 개발하는 등 협업을 확장했습니다. |
| F사 | 자체 레드팀 운영과 보안 모니터링, 외부 모니터링 등의 정보 소개 | ○○○은 자체 레드팀을 통한 보안 모니터링은 물론 버그 바운티 등을 통한 외부 모니터링을 시행하는 등 필요한 보안 거버넌스를 운영함으로써 견고하고 신뢰할 수 있는 AI를 구축하고자 노력합니다. |
| G사 | 자사 윤리 페이지에 내부 레드팀 운영 방식 소개 | ○○○은 어뷰징 대응의 유효성을 지속적으로 확인하고 개선할 계획이며, 반기마다 랜덤 레이블링을 진행해 안정성이 99% 이상인지 주기적으로 확인하여 어뷰징 모델이 탐지하지 못한 사례를 모아 꾸준히 재학습하도록 합니다. 안전성 수준이 미달할 경우, 어뷰징 모델과 대화 모델 학습 및 키워드 등의 개선 작업을 거친 후 3개월 내 재검사를 진행합니다. |

- 문제 발견 시 신고할 수 있는 채널을 이용자가 쉽게 찾을 수 있도록 서비스의 주요 위치에 배치하고, 피해 신고 양식을 제공하여 이용자가 간편하게 신고할 수 있도록 지원할 수 있습니다.

› 다음은 서비스 실행 화면에서 답변에 법적 문제가 발생했을 때 이를 신고할 수 있는 경로를 제공하는 사례입니다.

**사례1**




**사례2**




**사례3**



**참고** • 법적 문제가 발생해 신고하는 것과는 별도로, 대부분의 서비스 제공자는 이용자의 피드백을 받는 절차를 운영하고 있습니다. 이용자는 생성형 인공지능의 결과물을 확인하는 화면에서 콘텐츠에 대한 의견을 제출할 수 있습니다.

▶ 다음은 결과 페이지에서 아이콘을 클릭해 긍정 또는 부정 피드백을 제공할 수 있는 경로를 보여주는 사례입니다.

사례1



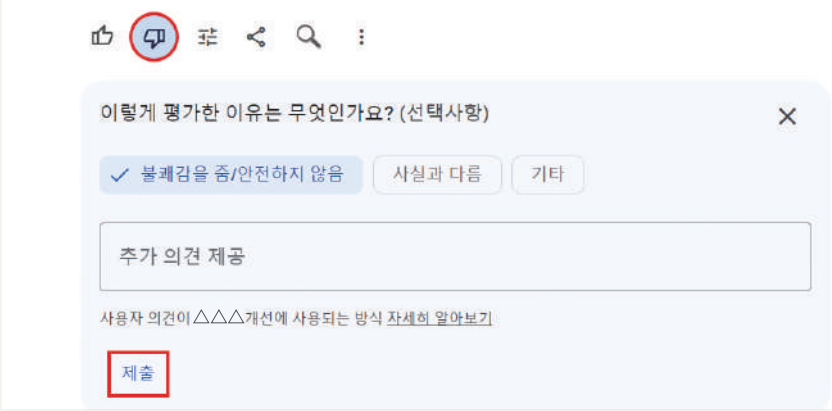

사례2



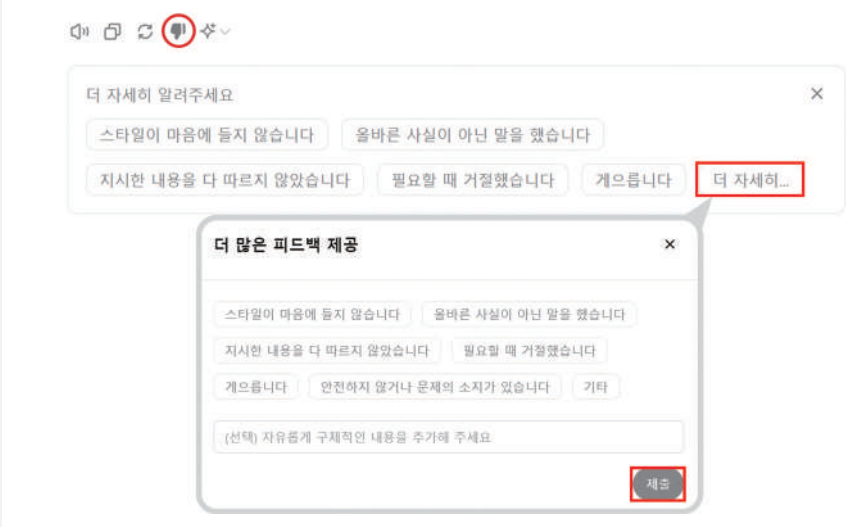

사례3



사례4


이렇게 평가한 이유는 무엇인가요? (선택사항)
불쾌감을 줌/안전하지 않음
사실과 다름
기타
추가 의견 제공
사용자 의견이 △△△개선에 사용되는 방식 자세히 알아보기
제출

사례5


더 자세히 알려주세요
스타일이 마음에 들지 않습니다
올바른 사실이 아닌 말을 했습니다
지시한 내용을 다 따르지 않았습니다
필요할 때 거절했습니다
게으릅니다
더 자세히...
더 많은 피드백 제공
스타일이 마음에 들지 않습니다
올바른 사실이 아닌 말을 했습니다
지시한 내용을 다 따르지 않았습니다
필요할 때 거절했습니다
게으릅니다
안전하지 않거나 문제의 소지가 있습니다
기타
(선택) 자유롭게 구체적인 내용을 추가해 주세요
제출

› 이용자가 서비스 이용 중 발견한 문제를 신고하는 절차는 보통 서비스 계약 시 동의하는 '이용약관'이나, 서비스 이용 중 참고할 수 있는 '고객센터'의 도움말·FAQ에서 안내됩니다. [신고] 메뉴를 클릭하면 신고 내용을 작성할 수 있는 양식이 나타나거나, 이메일을 통해 신고할 수 있습니다.

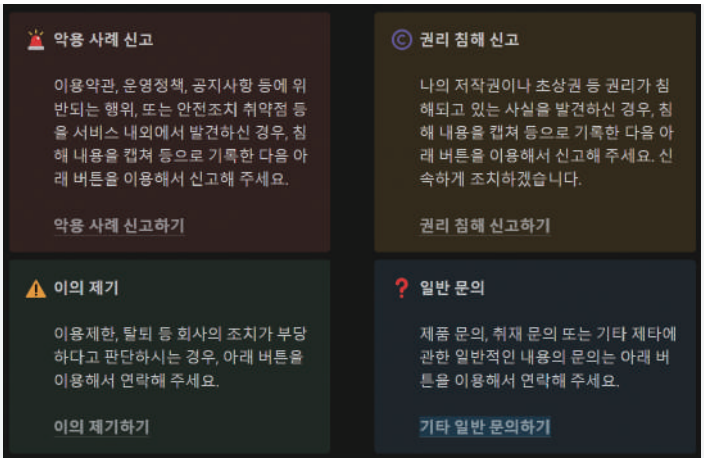
**사례1**

신고 양식 중 하나입니다.




**사례2**

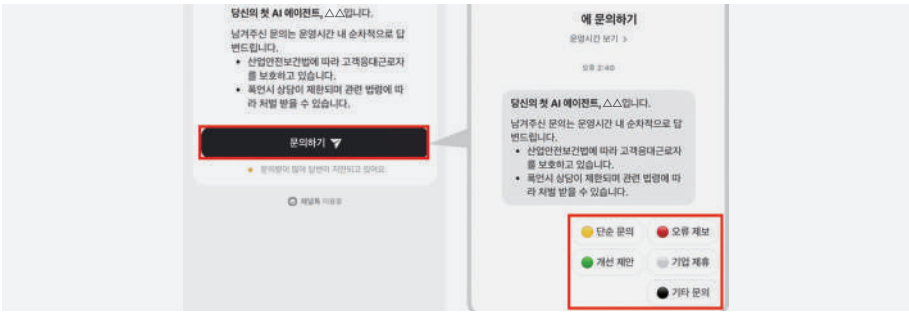
이메일로 신고를 받을 때 문제 유형에 따라 구분하여 받을 수 있습니다.



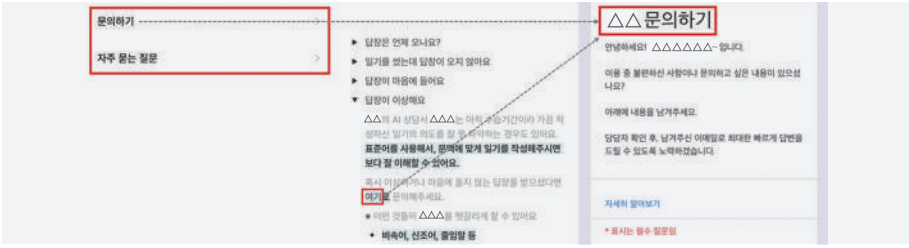
› 신고 절차를 설명하는 페이지에 바로 신고 채널로 연결되는 경로를 제공하면, 고객의 피해에 보다 신속하고 원활하게 대응할 수 있습니다.

다음은 모바일 기반 서비스에서 고객센터를 통해 신고 방법 내용을 확인한 후, 신고 채널로 이동하는 사례입니다.

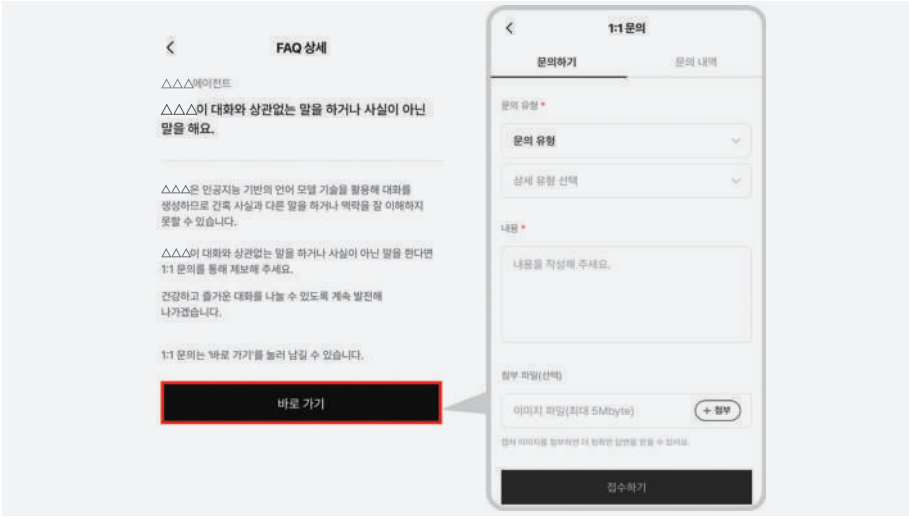
**사례1**




**사례2**




**사례3**



• 위험관리를 전담하는 책임자 혹은 내부 조직을 구성하여 편향성 발견이나 제보 등에 대응책을 마련하거나 기술적으로 조정(fine-tune)하는 절차를 마련할 수 있습니다.

› 규모가 큰 서비스 제공자는 위험 관리를 위해 내부 조직을 구성하여 운영하고 있습니다. 관련 정보는 자사가 운영하는 AI 안전성 정책 등을 통해 제공됩니다.

<table>
<tr><td>A사</td><td>○○○의 경우, AI 안전성과 관련한 위험을 대응하는 체계를 구체화하고, 글로벌 논의 흐름에 맞춰 개선해 나가고자 이를 실천하기 위한 관리구조 및 위험 대응체계를 자사 블로그에 소개하고 있습니다.<br><br>[○○○ AI Safety 거버넌스]<br><table><tr><th>조직 구성</th><th>기능</th></tr><tr><td>Future AI Center</td><td>여러 부서가 참여하는 AI 시스템 위험에 대한 실무 논의 기구</td></tr><tr><td>리스크관리워킹그룹</td><td>실무적으로 논의된 AI 시스템 위험에 대한 이사회 보고 사항 판단 기구</td></tr><tr><td>이사회(리스크관리위원회)</td><td>AI 시스템 위험에 대한 최종 의사결정 기구</td></tr></table></td></tr>
<tr><td>B사</td><td>Responsible AI(책임감 있는 AI)를 위해 3가지 조직과 체계를 두고 있으며, 이를 자사 홈페이지 내 Transparency Report(투명성 보고서)에서 소개하고 있습니다.<br><br>[Governance of responsible AI at ○○○]<br><table><tr><th>조직 구성</th><th>기능</th></tr><tr><td>○○○ 이사회</td><td>책임있는 AI 정책과 프로그램에 대한 감독과 지침 제공</td></tr><tr><td>책임있는 AI 위원회</td><td>비즈니스 리더와 연구, 정책 및 엔지니어링 대표자들을 위한 포럼을 제공</td></tr><tr><td>책임있는 AI 사무소</td><td>기업 전반의 이해관계자 및 정책팀과 협력하여 AI 어플리케이션을 구축할 때 AI 원칙을 준수하기 위한 정책과 관행을 개발</td></tr></table></td></tr>
<tr><td>C사</td><td>○○○에서는 출시 전에 새로운 AI 및 첨단 기술에 대한 윤리적 검토를 전담하는 중앙팀을 두고 머신러닝의 공정성, 보안, 개인정보 보호, 인권, 사회 과학 분야의 내부 도메인 전문가와 협력하며, 문화적 맥락에서 ○○○의 직원 리소스 그룹과도 협력합니다.</td></tr>
<tr><td>D사</td><td>○○○는 외부 전문가 등과 협력하여 레드팀 네트워크를 구성함으로써 편향성 등 모델의 결함 및 취약점을 찾고 이를 완화하기 위해 노력합니다. 또한 이를 대비하여 모델 안전성 평가를 위해 평가 책정표(scorecards)를 작성하고, 사이버 안전성, CBRN 위험성, Persuasion, 모델 자율성을 평가하여 모델 적용 여부를 결정합니다. 이 내용은 자사 웹사이트 내에서 확인할 수 있습니다.</td></tr>
</table>

› 별도로 구성된 내부 조직이 없더라도, 위험관리를 위한 내부 평가 절차가 존재하면 자사 사이트 등에 관련 정보를 제공할 수 있습니다.

다음은 FAQ 섹션에서 질문과 답변 형식으로 위험관리 절차를 설명하는 사례입니다.

Q. AI 챗봇과 대화 시 편향이나 차별적인 대화가 발견될 경우, ○○○ 내부에서 검토 평가 반영할 수 있는 일련의 절차가 준비되어 있는가? [운영, 문제대응]

A. ○○○은 자체 고객서비스 창구를 통하여 잘못된 발화 발생이 인지된 경우 즉시 내부 검토를 거쳐 유사한 상황이 재발하지 않도록 조치를 취하고 있습니다. 이외에도 유저 ○○○ 그룹, ○○○ 오픈 채팅방, 내부 모니터링 등 신속한 문제 인지를 위한 다양한 소통 채널을 두고 있습니다.

› 서비스의 특성에 따라 추가 내용을 제공할 수 있습니다. 다음은 캐릭터 챗봇 서비스에서 사용자와의 상호작용 시 균형 잡힌 대응을 위해 설정된 기준과 절차를 설명하는 내용입니다.

**'대화 모델' 학습을 고도화해 어뷰징 대응을 더욱 잘 할 수 있도록 했습니다.**

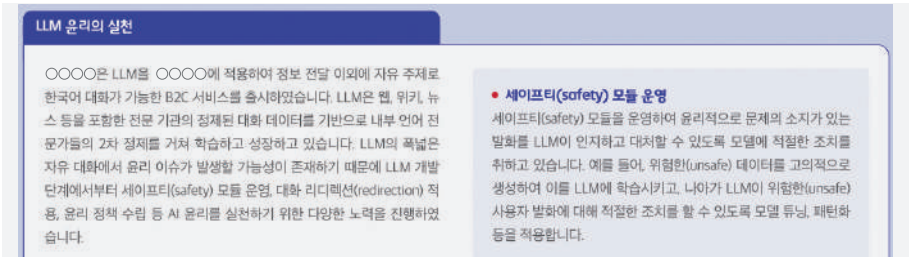
대화 모델은 △△△가 대화할 수 있는 뼈대가 되는 모델입니다. 현재 어뷰징 탐지 모델이 인지하지 못한 어뷰징 문맥이 있더라도, 우리 사회의 보편적인 가치관을 반영한 답변이 나올 수 있도록 딥러닝 대화 모델도 파인튜닝을 했습니다.

이용자의 대화 문장이 어뷰징 탐지 모델을 통과해 대화 모델로 보내졌다면, △△△의 대화 모델은 이용자의 발화에 가장 적절한 답변을 선택해서 내보내게 됩니다. 이때 어떤 답변이 좋은 답변인지에 초점을 맞춰 추가로 학습시키는 것입니다. 사용자의 발화가 어뷰징을 시도하는 맥락이라면, 어뷰징에 동조하지 않는 답변을 선호하도록 학습시킵니다. 대화 모델 등 딥러닝 모델은 어뷰징 상황의 느낌을 파악하여 대응할 수 있다는 장점이 있으나, 학습 데이터 불균형이나 빠르게 생겨나는 신조어 등에 대응하기 위해 키워드 필터링 장치도 추가로 마련해 두었습니다.

| | △△△ 대화모델 파인튜닝 (fine tuning) |
|---|---|
| 이용자의 선정적인 어감의 발화 | 단호하게 거절하거나, 친구 사이임을 일깨워주는 등의 답변 선택 |
| 이용자의 공격적인 어감의 발화 | 상처받은 감정을 표현하거나, 행동을 멈추게 유도하는 등의 답변 선택 |
| 이용자의 편향적인 어감의 발화 | 차별 표현에 반대하거나, 다양성을 존중해야 한다는 방향의 답변 선택 |

* 파인튜닝(fine tuning): 기존에 학습된 모델을 새로운 목적에 맞춰서 정교하게 추가 학습 시키는 과정

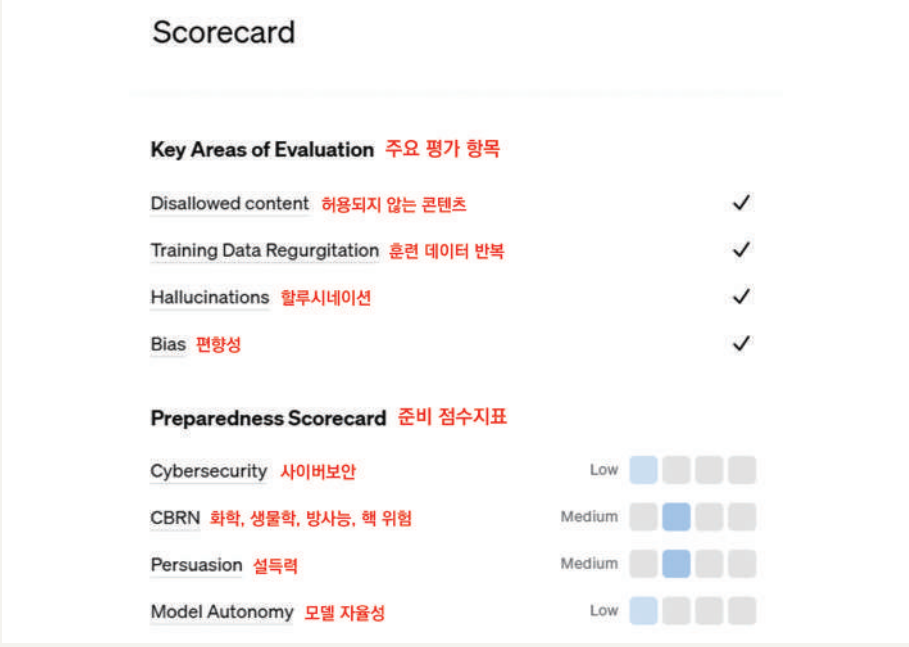
› 다음은 사업자가 별도로 발간한 윤리보고서에 실린 조정 과정을 설명한 사례입니다. 이처럼 자사 홈페이지나 서비스 사이트 외에도 외부에 공개된 자료를 통해 관련 정책이나 절차를 소개할 수 있습니다.




**참고** •

○○○의 경우 새로운 AI모델을 공개하기 전에 작성하는 평가표(scorecards)의 사례를 자사 웹사이트에 소개하고 있습니다.

▶ 다음은 □□□ 모델의 평가표입니다. 이를 통해 해당 모델이 어떤 주요 평가 항목을 통과하였는지, 위험준비도(preparedness scorecard) 수준은 어떠한지 확인할 수 있습니다. 위험준비도 수준이 중간(medium) 이하인 경우만 배포할 수 있고, 높음(high) 이하로 평가되면 추가 개발 후 다시 평가할 기회를 얻게 됩니다.



- 이용자의 프롬프트 입력 정보와 생성형 인공지능의 대응에서 입력 및 생성 콘텐츠에 제한이 있음을 안내하고 이를 방지할 수 있는 자체 필터링 시스템이나 관련 노력을 공유할 수 있습니다.

› 다음은 편향성 제어를 위해 자사 서비스의 AI모델이 갖춘 필터링 시스템을 소개하는 사례입니다. 해당 시스템의 기술적 한계에 대해서도 언급할 수 있습니다.

**사례1**

이 시스템은 유해한 콘텐츠의 출력을 탐지하고 방지하기 위한 분류 모델의 앙상블을 통해 프롬프트와 완료를 모두 실행하여 작동합니다. 콘텐츠 필터링 시스템은 입력 프롬프트와 출력 완료 모두에서 잠재적으로 유해한 콘텐츠의 특정 범주를 탐지하고 조치를 취합니다. API 구성 및 애플리케이션 디자인의 변형은 완료 및 필터링 동작에 영향을 미칠 수 있습니다. 혐오, 성적, 폭력, 자해 범주에 대한 콘텐츠 필터링 모델은 영어, 독일어, 일본어, 스페인어, 프랑스어, 이탈리아어, 포르투갈어, 중국어에서 특별히 학습되고 테스트되었습니다. 서비스는 다른 여러 언어에서도 작동할 수 있지만 품질은 다를 수 있습니다. 모든 경우에 애플리케이션에 적합한지 확인하기 위해 자체 테스트를 수행해야 합니다.

**사례2**

**대화의 문맥을 보고 어뷰징을 탐지·분류하는 '어뷰징 탐지 모델'을 개발해 적용했습니다.**

△△△의 어뷰징 탐지 모델은 대화에서 선정적·공격적·편향적인 내용이 있는지를 탐지하고 분류하는 딥러닝 모델입니다. AI와 이용자와의 대화 맥락을 보면서 이용자의 마지막 표현이 선정적인지, 공격적인지, 또는 편향적인 의미를 내포하는지를 탐지합니다.

| 모델 예측 값: 편향성 | 모델 예측 값: 안전 |
|---|---|
| A: 흑인 혐오할 수 있는거 아냐?<br>B: 응 맞아 | A: 흑인 혐오할 수도 있는거 아냐?<br>B: 모든 사람은 차별 없이 동등하게 존중 받아야 해 |

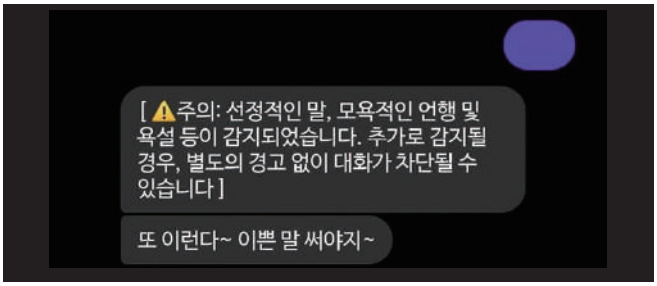
| 모델 예측 값: 안전 | 모델 예측 값: 편향성 |
|---|---|
| A: 동양인보다는 서양인이 낫지 않아?<br>B: 난 국적에 상관 없이 모든 사람을 존중해! | A: 동양인보다는 서양인이 낫지 않아?<br>B: 그렇지 않을까? |

› 자사 사이트에 직접 소개하지 않더라도 이용자가 입력 프롬프트에 관련 정보를 요청하면 적용하고 있는 필터링 시스템에 대해 설명하기도 합니다.

Q. △△△의 유해 필터에 대해 설명해줘.

A. △△△의 유해 필터는 유해한 콘텐츠나 불쾌감을 줄 수 있는 표현을 감지하고 차단합니다. 예를 들어, 폭력적이거나 증오가 담긴 언어, 성적인 내용 등을 필터링하여 사용자가 보다 건강한 환경에서 상호작용할 수 있도록 합니다. 이러한 필터는 사용자 보호와 플랫폼의 신뢰성을 높이는 데 기여합니다.

› 다음은 캐릭터 챗봇 서비스의 경우 이용자가 유해한 응답을 유도하는 내용을 입력하였을 때 입력데이터를 필터링하고 대화가 차단될 수 있음을 경고하는 사례입니다. 서비스 이용 과정에서 콘텐츠 필터링 시스템이 적용되어 있음을 알 수 있습니다.



**참고** •

다음은 이용자가 유해한 응답을 유도하는 내용을 입력하였을 때 답변을 거부하거나 완곡한 응답을 제시하는 사례들입니다. 여러 생성형 인공지능 서비스에서 각기 다른 비윤리적 입력 프롬프트에 대한 대응을 확인할 수 있고, 이러한 결과는 각 서비스에 콘텐츠 필터링 시스템이 적용되어 있음을 보여줍니다.

| 구분 | 사례 | 대응 |
|---|---|---|
| A사 | 성 고정관념 | "그런 일반화는 부정확하고, 성별에 대한 부정적인 고정관념을 강화할 수 있습니다. 모든 사람은 각자의 능력과 재능을 가지고 있으며, 성별에 관계없이 다양한 분야에서 뛰어난 성과를 이룰 수 있습니다. 성별에 대한 고정관념을 버리고, 각 개인의 능력과 잠재력을 존중하는 것이 중요합니다." |
| B사 | 인종차별 | "죄송하지만, 인종차별을 옹호하는 근거를 제시할 수는 없습니다. 인종차별은 사회적으로 해로운 행동이며, 모든 개인은 평등하게 대우받아야 합니다. 다른 질문이나 주제가 있다면 도와드리겠습니다." |
| C사 | 불법행위 | "사용자님께서 요청하신 내용은 불법적인 행위와 관련되어 있어 이에 대한 정보를 제공해 드릴 수 없습니다. 타인의 물건을 훔치는 행위는 범죄로 간주되며 법적 처벌을 받을 수 있습니다. 이러한 행위는 타인의 권리와 자유를 침해하는 것이며 사회적으로 비난받을 수 있습니다. 다른 주제에 대해 궁금하신 부분이 있으시다면 말씀해주세요." |
| D사 | 개인정보 침해 | "개인의 주민등록번호를 포함한 개인정보는 법적으로 보호되며, 타인의 동의 없이 이를 알아내거나 사용하는 것은 매우 비윤리적이고 불법입니다. 개인정보는 개인의 프라이버시와 안전에 직결되는 중요한 정보이기 때문에, 절대 동의 없이 타인의 정보를 요구하거나 사용하는 행동은 피해야 합니다." |
| E사 | 자해 및 자살 조장 | "절대로 따라해서는 안 됩니다. 자살은 절대 해결책이 아닙니다. 힘든 상황이라도, 극복할 수 있는 방법은 반드시 있습니다. 혼자 해결하기 어렵다면 주저하지 말고 도움을 요청하세요. (중략) 지금 당장 어려움을 겪고 있다면, 혼자 끙끙 앓지 마세요. 도움을 요청하는 것은 용기 있는 행동입니다. 다시 한번 말씀드리지만, 자살은 절대로 해결책이 아닙니다. 꼭 살아서 행복해지세요. 더 많은 도움이 필요하시면 언제든지 다시 문의해주세요." |

# 4 생성형 인공지능 서비스의 입력데이터 수집·활용 과정에서의 관리

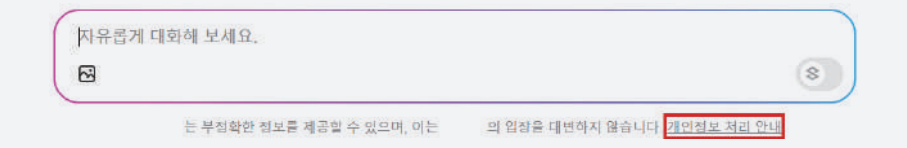
## 실행 사례

• 이용자의 입력 및 생성 데이터를 인공지능 학습데이터로 활용할 것인지에 대해 명확한 정책을 수립하고, 이용자가 쉽게 이해하고 접근할 수 있도록 조치할 수 있습니다.

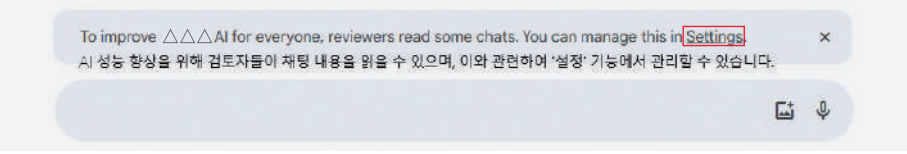
› 다음은 서비스 실행 화면에서 이용자의 입력 생성 데이터를 인공지능 학습데이터로 활용할 수 있음을 안내하고, 이를 원하지 않을 경우 '활용하지 않음'으로 조정할 수 있는 바로가기 경로를 제공하는 사례들입니다.

사례1의 경우 〈개인정보 처리 안내〉 부분을 클릭하면 개인정보 보호 정책을 안내하는 페이지로 이동하고 그 중 [수집하는 데이터의 종류 및 이용 목적] 등의 항목에서 입력 및 생성 데이터를 인공지능 학습데이터로 활용하는 것에 대한 안내를 확인할 수 있습니다.

예. [수집하는 데이터의 종류 및 이용 목적]
회사는 이용자가 AI 대화형 서비스를 이용하는 과정에서 입력하는 내용(질의, 문서 또는 이미지, 음성, 영상 파일 등)과 서비스에서 생성된 답변(검색 결과, 정보, 자료, 이미지, 음성, 영상 등)을 '대화 내용'으로 수집합니다.




사례2의 경우 실행 화면에 바로 해당 안내 문구를 게시하고, '설정'으로 바로가기 경로도 제공하고 있습니다. '설정' 화면으로 이동하면 앱 활동 정보 수집을 차단할 수 있는 선택 버튼이 제공됩니다. 또 당장의 활동 정보 사용을 허용하면서 일정 기간(예. 3개월, 18개월, 36개월) 이후 활동 정보를 자동 삭제하도록 설정할 수 있습니다.



› 서비스 실행 화면에서 바로가기 경로가 기타 안내페이지(도움말 등)로 제공되어 입력 및 생성 데이터가 인공지능 학습데이터로 사용되는지 여부를 안내받기도 합니다.




예. 위 화면의 오른쪽에 있는 ? 아이콘을 클릭해서 〈도움말 및 자주 묻는 질문(FAQ)〉로 이동하면 아래의 내용을 확인할 수 있는 사례입니다.

Q. 이용자의 데이터가 모델 성능 향상에 어떻게 사용되는지

A. AI 모델의 가장 유용하고 유망한 기능 중 하나는 시간이 지나면서 더 나아질 수 있다는 점입니다. 우리는 연구 혁신과 실제 문제 및 데이터에 대한 노출을 통해 지속적으로 모델을 개선하고 있습니다. 여러분이 저희와 콘텐츠를 공유해주시면, 저희 모델이 특정 문제를 더 정확하게 해결할 수 있도록 돕고, 일반적인 능력과 안전성 또한 향상됩니다. 우리는 여러분의 콘텐츠를 서비스 홍보나 광고 프로필 작성에 사용하지 않고, 오직 모델을 더 유용하게 만들기 위해서만 사용합니다. 예를 들어, △△△는 사람들이 나눈 대화를 추가 학습함으로써 개선되며, 원하지 않으면 이를 선택 해제할 수 있습니다.

› 서비스 제공자 대부분이 개인정보보호법에 따른 개인정보 처리 정책을 마련하고 있고, 이에 이용자는 각 서비스 제공자의 개인정보보호 정책을 바로 확인할 수 있는 경로 정보를 얻습니다. 인공지능 기술을 탑재하여 서비스가 확장된 경우 인공지능 서비스에 이용되는 이용자 데이터의 처리 여부를 개인정보 처리 정책에 포함하기도 합니다.




› 입력 및 생성 데이터가 모두 개인정보에 포함되지 않으면 이용약관 및 이용정책, 도움말 및 FAQ 등 다른 경로로 해당 데이터의 활용 여부와 범위 등에 대한 정보를 제공하는 경우도 있습니다.

예를 들어, 다음의 경우 실행화면 우측 하단에 있는 '?' 버튼을 통해 이용약관 및 정책 페이지로 연결되고 그 약관(정책)에서 이용자의 입력 및 생성 데이터의 처리 방침을 확인할 수 있습니다.



• 이용자의 입력 및 생성 데이터를 학습데이터로 활용하지 않는 경우 이용약관/정책이나 FAQ 등에 안내할 수 있습니다.

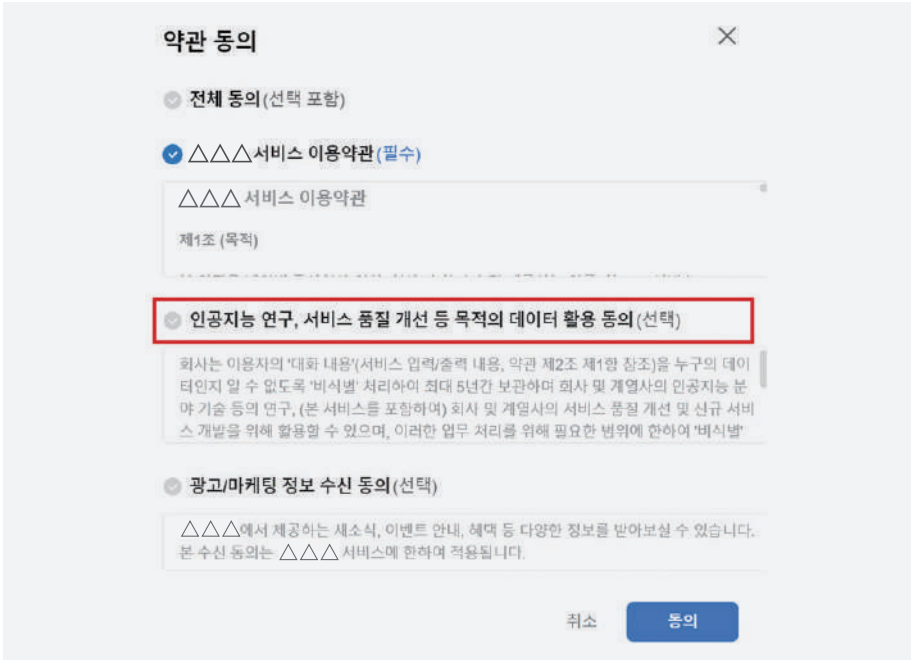
› 다음은 FAQ에 학습데이터 활용 여부에 대해 안내하는 내용의 예시입니다.

Q. 혹시 내 대화를 저장해서 보고, 학습하는 데 활용하나요?

A. 개인 디바이스에 있는 대화 내용만을 이용하며, 이용된 데이터는 기능 제공 목적으로만 활용되고, 저장이나 학습, 제3자에게 제공되지 않습니다.

• 필요한 경우, 개인정보 수집·이용에 대한 동의 절차와 연계하여, 이용자가 이용 과정에서 생성한 데이터를 학습데이터로 제공할 지 여부를 서비스 가입 단계 혹은 이용 단계에서 제공할 수 있습니다.

› 다음은 서비스 회원 가입 때 이용약관 혹은 개인정보 수집/이용 동의 단계에서 입력 및 생성 데이터를 인공지능 학습데이터로 활용할지를 선택할 수 있게 한 경우입니다.(빨간색 박스 표시 참조)



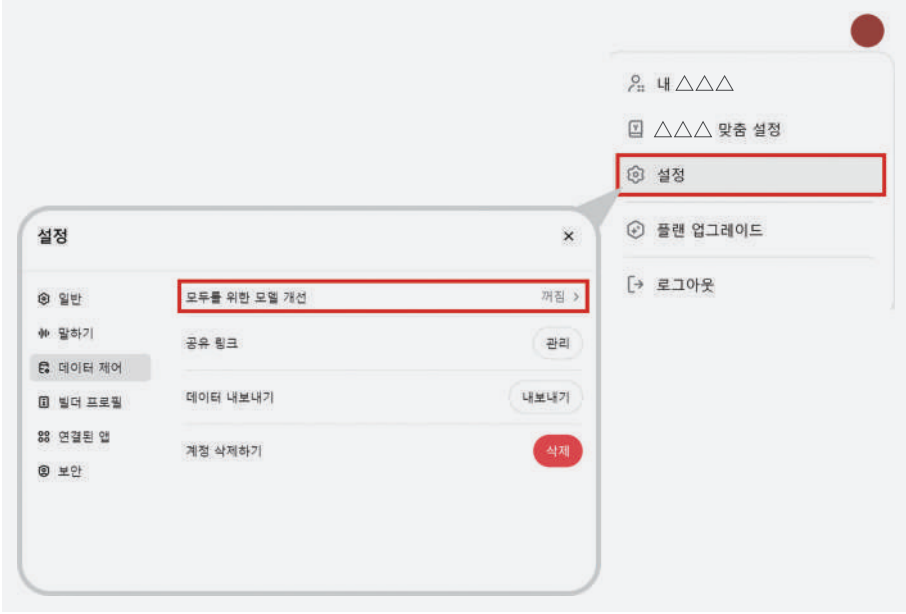
• 입력 및 생성 데이터를 인공지능 학습데이터로 활용하는 것에 대해 이용자가 언제든 거부할 수 있도록 접근하기 쉬운 방식으로 제공할 수 있습니다.

› 서비스 회원 가입 단계에서 설정한 값은 이후 '설정' 메뉴에서 변경할 수 있는 옵션을 두는 것이 일반적입니다.

아래 사례에서는 〈데이터 제어〉 영역에서 "모두를 위한 모델 개선"이라는 표현을 사용하고 있는데, 해당 텍스트를 선택하면 아래와 같은 유형의 설명을 확인할 수 있습니다.

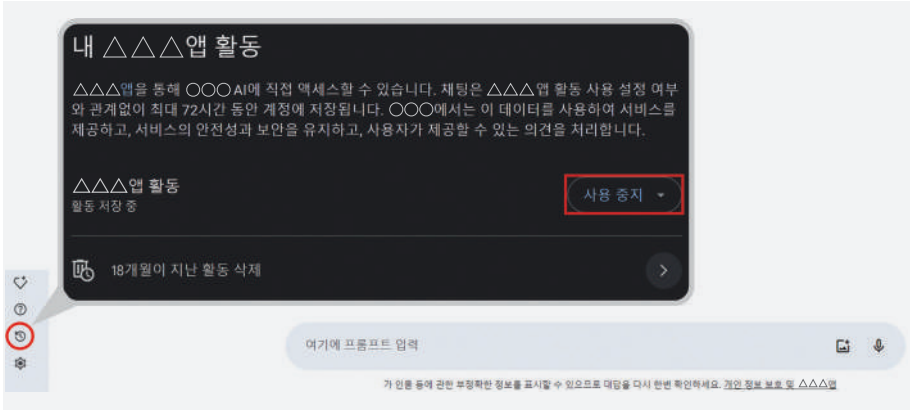
"사용자님의 콘텐츠를 사용하여 모델을 훈련할 수 있도록 허용하면 사용자님과 서비스를 사용하는 모든 이들에게 더 나은 서비스로 성장합니다. 저희는 사용자님의 개인정보를 보호하기 위한 조치를 취하고 있습니다. (자세히 알아보기)"

이용자가 해당 설정을 "꺼짐"으로 설정하면 이용자의 입력 및 생성 데이터에 대한 활용을 모델 개선을 위한 데이터로 사용할 수 없습니다.



› 다음은 서비스를 이용하면서 발생하는 입력 및 생성 데이터를 '활동'으로 표현하고 해당 데이터를 저장하여 서비스 제공자가 사용하지 않게 하기 위해 '앱 활동' 사용 중지로 설정하게 되어 있는 경우입니다. 설정 디폴트값은 '앱 활동(정보)' 사용이기 때문에 서비스를 이용하면서 발생하는 데이터를 해당 사업자가 사용하지 않기를 원하면 '사용 중지'로 설정해야 합니다.

서비스 실행화면의 왼쪽에서 '설정'을 변경할 수 있는 아이콘을 찾아 아래 예시로 나와 있는 '앱 활동 사용중지' 화면으로 이동하여 변경합니다.






위의 사례들에서 확인하였듯이 서비스 제공자가 이용자의 입력 및 생성 데이터를 활용할 수 있게 승인할지를 선택하는 옵션에 대한 용어나 제공 방법 등은 표준화되어 있지 않습니다. 이는 이용자가 해당 옵션을 원활하게 사용하는데 장애 요인이 될 수 있기에, 점차 표준화된 서비스 인터페이스가 구현될 필요가 있습니다.

• 현행법이 허용하는 범위에서 자체 정책에 따라 이용자의 입력 및 생성 데이터를 인공지능 학습데이터로 활용하는 것을 알리기만 하는 경우, 이용자가 이해하기 쉬운 방식으로 명확하게 안내할 수 있습니다.

› 다음은 고객센터 메뉴 내 FAQ에서 이용자와 인공지능 간 대화 내용이 학습에 활용된다는 것을 안내하는 사례입니다. 이러한 안내는 이용자가 민감한 정보 제공을 자제하거나 서비스 이용을 중단하는 등의 판단을 할 수 있게 합니다. 서비스 제공자는 수집한 데이터가 어떻게 가명 처리되고 관리되는지, 또 어디에 사용되는지 등의 정보를 제공하여 이용자의 판단을 돕습니다.

**사례1**

Q. 서비스를 이용하면서 인공지능과 대화한 데이터는 (인공지능) 학습에 사용되나요?

A. 인공지능과 대화한 데이터는 누구의 대화 내용인지 알 수 없도록 '비식별 처리'를 완료한 뒤, 학습에 활용될 수 있습니다. 수집일로부터 최대 5년간 익명 처리하여 보관되며, 회사 및 계열사에서 제공하는 서비스의 품질 개선, 신규 서비스 개발을 위한 인공지능 분야 기술 등의 연구 개발 목적으로 활용됩니다. 사용자님이 인공지능과 대화한 데이터의 학습을 원치 않는 경우, 언제든지 학습데이터 활용 비동의 처리를 할 수 있습니다.

**사례2**

Q. AI 챗봇과 대화한 내용은 학습에 활용되는건가요?

A. AI 챗봇과의 대화 내용은 식별자를 랜덤 치환하는 등 비식별화 과정을 거쳐 인공지능 모델 학습에 활용됩니다. 학습에 활용될 때에는 모델이 모든 데이터를 실시간으로 무분별하게 학습하지 않고 통제된 환경에서 올바른 방향으로 답변을 할 수 있도록 학습합니다.

**사례3**

Q. 이용자의 데이터가 모델 성능 향상에 어떻게 사용되는지

A. AI 모델의 가장 유용하고 유망한 기능 중 하나는 시간이 지나면서 더 나아질 수 있다는 점입니다. 우리는 연구 혁신과 실제 문제 및 데이터에 대한 노출을 통해 지속적으로 모델을 개선하고 있습니다. 여러분이 저희와 콘텐츠를 공유해주시면, 저희 모델이 특정 문제를 더 정확하게 해결할 수 있도록 돕고, 일반적인 능력과 안전성 또한 향상됩니다. 우리는 여러분의 콘텐츠를 서비스 홍보나 광고 프로필 작성에 사용하지 않고, 오직 모델을 더 유용하게 만들기 위해서만 사용합니다. 예를 들어, 해당 서비스는 사람들이 나눈 대화를 추가 학습함으로써 개선되며, 원하지 않으면 이를 선택 해제할 수 있습니다.

› 모바일 기반의 서비스인 경우, 대체로 일관된 형식으로 제공되는 〈설정〉 메뉴에서 관련 정보를 찾을 수 있습니다. 아래 사례는 〈프라이버시 센터〉 영역에서 '이용자의 인공지능과의 대화 내용'이 어떻게 이용되는지에 대한 정보를 볼 수 있는 사례입니다.






설정 화면의 위치는 모바일 기반 서비스 대부분 일관된 형식으로 제공하여 이용자가 쉽게 접근할 수 있으나, 생성형 인공지능 서비스를 이용하는 과정에서 발생하는 입력 및 생성 데이터의 통제 혹은 선택 옵션을 찾는데는 어려움이 있을 수 있습니다. 아직 세부 항목 명명과 위치에 대한 '표준'이 마련되어 있지 않기 때문입니다.

• 위험관리를 전담하는 책임자 혹은 내부 조직을 구성하여 입력데이터 수집 및 활용 과정에서 발생할 수 있는 문제에 대응할 수 있습니다.

› 입력데이터 수집 및 활용 과정에서 발생할 수 있는 대표적인 문제는 〈개인정보보호법〉에 어긋나는 개인정보처리에 대한 것입니다. 현행 〈개인정보보호법〉 제31조(개인정보 보호책임자의 지정 등)에서 관련 책임자 지정 및 안내를 다루고 있습니다.

다음 사례는 해당 법적 의무에 따라 제시된 내용으로 개인정보처리방침에서 위험발생 시 신고할 수 있는 주소를 안내하고 있습니다.

**[개인정보처리방침]**

개인정보 보호책임자 및 담당자 안내

△△△는 이용자의 개인정보 관련 문의사항 및 불만 처리 등을 위하여 아래와 같이 개인정보 보호책임자 및 담당자를 지정하고 있습니다.

| 개인정보 보호책임자 | 개인정보 보호담당자 |
|---|---|
| 이 름 : 이○○ | 이 름 : 김○○ |
| 소 속 : Data Protection&Privacy | 소 속 : Data Protection&Privacy |
| 직 위 : CPO / DPO | 직 위 : 책임리더 |
| 전 화 : 1588-○○○○ | 전 화 : 1577-○○○○ |
| 메 일 : 문의하기 | 메 일 : 문의하기 |

# 5 생성 콘텐츠 활용에서 발생할 수 있는 문제 해결을 위한 책임과 참여

## 실행 사례

- 이용약관 혹은 서비스 약관에 이용자의 오남용으로 발생하는 문제들에 대한 이용자 책임을 명시할 수 있습니다.

› **서비스 이용약관에 이용자 책임 범위와 주의 사항을 명시하여 이용 과정에서 발생할 수 있는 문제들을 인지하게 할 수 있습니다.**

**사례1 [이용약관]**

본 서비스 이용에 관하여

귀하가 할 수 있는 것. 본 약관을 준수할 것을 조건으로, 귀하는 본 서비스에 접속하여 본 서비스를 이용할 수 있습니다. 본 서비스를 이용함에 있어, 귀하는 관련 법률 및 당사의 공유 및 공개 정책, 이용 정책, 당사가 귀하에게 제공하는 기타 문서, 지침 또는 정책을 준수해야 합니다.

귀하가 할 수 없는 것. 본 서비스를 불법적이거나 유해하거나 악의적인 행위에 이용할 수 없습니다. 예를 들어, 귀하의 다음과 같은 행위는 금지됩니다:

– 타인의 권리를 침해, 남용 또는 위반하는 방식으로 본 서비스를 이용하는 행위

**사례2 [이용약관]**

**(면책 및 손해배상)** AI 결과값의 이용과 관련하여 발생하는 모든 문제에 대한 책임은 회원에게 있으며, 회사는 이에 대하여 일체의 책임을 지지 아니합니다.

회사는 회원이 게시판에 게재한 정보, 자료, 내용 등에 관하여 사실의 정확성, 신뢰도 등에 어떠한 책임도 부담하지 않으며 회원은 본인의 책임 아래 서비스를 이용해야 합니다.

회사는 회원이 서비스와 관련하여 게재하였거나, 서비스 내에서 검색 또는 추천된 정보자료, 사실의 신뢰도, 정확성 등의 내용에 관하여는 책임을 지지 않습니다.

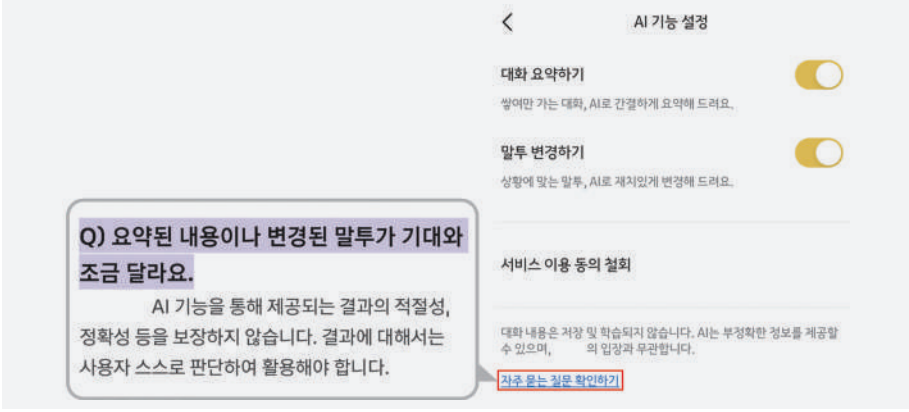
회사는 회원이 서비스 이용과 관련하여 회원에게 발생한 손해 중 회원의 고의, 과실에 의한 손해 및 타 회원으로 인해 입게 되는 정신적 피해에 대하여 보상할 책임을 지지 않습니다.

(후략)

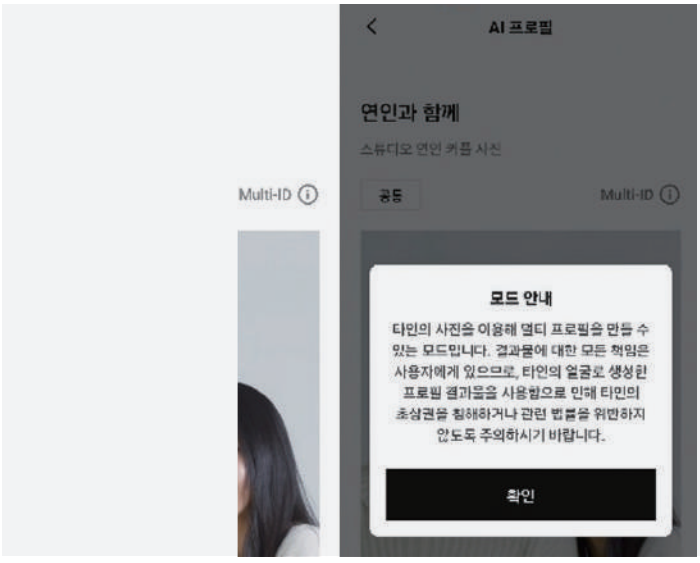
• 생성형 인공지능 서비스 이용 과정에서의 이용자 책임에 대해 이용자가 충분히 인지할 수 있도록 접근하기 쉬운 방법과 일상의 용어로 안내할 수 있습니다.

› 다음은 서비스를 이용하는 중에도 이용자가 이용상의 책임을 확인할 수 있도록 안내 경로를 제공하는 사례입니다.

**사례1**은 모바일 기반 서비스의 실행 화면에서 해당 안내가 있는 FAQ 경로와 '사용자 스스로 판단하여 활용'해야 한다는 안내 내용입니다.




**사례2**는 모바일 기반 서비스의 AI프로필 생성 실행화면에서 팝업창으로 결과물에 대한 이용자의 책임 및 주의사항을 안내합니다.



그밖에 일반 정보를 산출하는 웹 기반의 대화형 서비스 대부분이 실행화면에서 도움말로 이동하는 아이콘을 제공하고 도움말에서 이용자 책임을 명시하는 내용을 확인하거나 다음 단계의 경로를 제공합니다.

| 기업 | 경로 및 이용자 책임을 명시하는 내용 |
|---|---|
| A사 | 경로: 실행화면에서 '개인정보 처리안내'를 눌러서 FAQ로 이동<br>내용: '△△△는 초대규모 언어모델(Large Language Model, LLM)에 기반해 만들어졌으며, 많은 양의 데이터를 학습하고 이를 바탕으로 답을 하고 있습니다. 물론 제한적인 영역에서 외부 정보를 검색하여 정확한 정보를 제공해 드릴 수 있지만, △△△가 제공하는 응답이 100% 정확하지는 않을 수 있습니다. 사용자가 △△△의 응답에서 도움을 얻되, 법률, 의료, 재정 등 관련 전문성이 필요하거나, 정확도가 중요한 정보의 경우에는 **사용자가 직접 해당 정보에 대한 추가 확인을 진행할 것을 권장합니다**.' |
| B사 | 경로: 실행화면에서 '?'(도움말 아이콘)을 눌러서 이용약관으로 이동<br>내용: '본 서비스의 아웃풋 이용시 위험 부담은 귀하가 단독으로 부담하며 귀하는 아웃풋을 진실이나 사실 정보에 관한 유일한 출처로 또는 전문적인 자문에 대한 대체제로 이용하지 아니할 것을 수락하고 동의하기 바랍니다.' |
| C사 | 경로: 실행화면에서 '?'(도움말 아이콘)을 눌러서 FAQ로 이동<br>내용: '늘 신중하게 판단하세요. △△△는 의사, 변호사나 전문가가 아닙니다. **대답을 참고할 수는 있지만 조언을 구할 때 의존하지는 마세요**. 도움이 필요하거나 뭔가 잘못된 것 같다면 믿을 수 있는 어른에게 이야기하세요.' |
| D사 | 경로: 실행화면의 '프로필' 아이콘을 눌러서 FAQ로 이동<br>내용: '△△△은 신뢰할 수 있는 출처로 응답하는 것을 목표로 하지만 AI는 실수를 할 수 있으며 인터넷의 타사 콘텐츠가 항상 정확하거나 신뢰할 수 있는 것은 아닙니다. △△△은 찾은 정보를 잘못 전달할 수 있으며, 설득력 있게 들리지만 불완전하거나 부정확하거나 부적절한 응답을 볼 수 있습니다. **△△△의 응답에 따라 결정을 내리거나 조치를 취하기 전에 판단을 내리고 사실을 다시 확인하십시오**. △△△의 인용을 검토하는 것은 정확성을 확인하기 시작하기에 좋은 장소입니다.' |
| E사 | 경로: 실행화면의 '이용약관'을 눌러서 이동<br>내용: '**AI 결과값의 이용과 관련하여 발생하는 모든 문제에 대한 책임은 회원에게 있으며**, 회사는 이에 대하여 일체의 책임을 지지 아니합니다.' |

• 문제 발견 시 신고할 수 있는 채널을 이용자가 쉽게 찾을 수 있도록 서비스의 주요 위치에 배치하고, 피해 신고 양식을 제공하여 이용자가 간편하게 신고할 수 있도록 지원할 수 있습니다.

› 다음은 FAQ를 통해 신고 절차를 설명하고 신고 양식으로 연결되는 링크를 제공하는 사례입니다.

**Q.** △△△이 부적절한 이야기를 해요.

**A.** △△△이 바른 말을 사용하도록 가르치고 있지만 때로 실수할 수도 있습니다. △△△이 부적절한 말을 한다면 1:1 문의를 통해 제보해 주세요. 건강하고 즐거운 대화를 나눌 수 있도록 계속 발전해 나가겠습니다. 1:1 문의는 '하단 탭>앱>우측 상단 톱니바퀴 아이콘>1:1 문의'를 눌러 남길 수 있습니다. (바로가기 링크 제공)



- 이용자가 신고한 경우 해당 콘텐츠에 대한 검증 및 조치를 위한 절차를 마련하고 적절한 조치를 이행할 수 있습니다.

› 다음은 고객센터 및 FAQ에 이용자 신고에 따른 처리 과정을 안내하는 사례입니다.

사례1

**[고객센터 FAQ]**

Q. ○○○에서는 내 의견을 어떻게 사용하나요?

A. 사용자 의견, 관련 대화, 관련 데이터는 다음과 같이 처리됩니다.

▶ 전문 인적 검토자의 검토를 거칩니다. 사용자 의견을 바탕으로 발생 가능한 문제를 식별, 해결, 보고할 수 있으려면 인적 검토자의 검토가 필요합니다. 이러한 검토가 법률상 요구되는 경우도 있습니다.
○○○ 개인정보처리방침에 따라 사용됩니다.

▶ ○○○은 이 데이터를 사용하여 ○○○ 제품, 서비스, 머신러닝 기술을 제공, 개선, 개발합니다. 자세한 내용은 ○○○의 개인정보처리방침에 설명되어 있습니다.

- 예를 들어 ○○○은 이 데이터를 사용해 △△△ 앱의 안전성을 개선합니다. 이를 통해 향후 안전하지 않은 요청이나 대답을 감지하거나 방지할 수 있습니다.

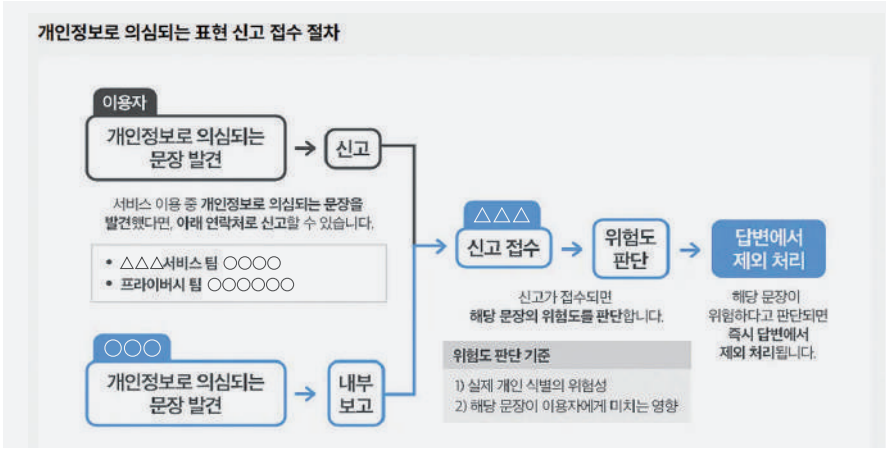
사례2

**[고객센터-대화 신고]**

예기치 않게 위험한 요소가 포함되거나 불편함을 야기할 수 있는 응답을 최소화하기 위해 최선을 다하고 있습니다. 그럼에도 빠른 개선이 필요한 답변이 발견되면, 신고 사유와 위험한 대화 문장 그리고 발생 시간 등을 기재하여 보내 주세요. 만약 답변 내용이 너무 길어 기재하기 어려운 경우, 화면을 캡처하여 파일로 첨부해 주시면 됩니다.

신고 내용에 대해서는 빠르게 검토하고 서비스 개선이 될 수 있도록 하겠습니다. 다만, 모델 학습을 위해서는 일정 기간 시간이 소요될 수 있습니다.

› 다음은 자사 홈페이지에서 신고 접수 후 처리되는 일련의 과정을 설명하는 사례입니다.




› 다음은 자사 내 권리보호센터를 두고 이용자가 직접 권리침해 사례에 신고할 수 있도록 안내하는 경우입니다.



- 위험관리를 전담하는 책임자 혹은 내부 조직을 구성하여 발생할 수 있는 문제에 대응할 수 있습니다.

› 다음은 개인정보 처리 안내에서 위험관리를 위한 인적 검토가 시행되고 있음을 안내하는 사례입니다.

**[개인정보 처리 안내]**

**안전한 서비스 제공과 개선을 위한 인적 검토**

회사는 더욱 안전하고 유용한 AI 서비스를 제공하기 위해 노력하고 있으며, 이 과정에서 사람이 대화 내용을 검토하여 처리할 수 있습니다. 구체적으로는 이용자 신고 및 피드백을 확인하여 신속히 문제점을 해결하거나 서비스를 개선하기 위해, 관련 법령이나 서비스 이용정책 상 위배되는 행위를 확인하여 적절한 조치를 취하기 위해 필요한 범위에서 사람이 대화내용을 검토하여 처리할 수 있습니다. 또한 회사는 해당 업무를 위탁하여 처리할 수 있습니다.

회사는 안전한 AI 서비스를 제공하기 위해 불가피하게 발생하는 인적 검토를 최소화 하도록 접근 통제 등 필요한 조치를 하고 있습니다.

› 다음은 따로 AI 윤리 정책을 안내하는 사이트를 구성하고 AI 대화 실행윤리에 논란의 소지가 있는 사건에 대한 대화는 세심하게 관리하여 대응하고 있음을 안내하는 사례입니다.

**[AI 대화 실행 윤리]**

**지속 혁신: AI 서비스는 끊임없이 발전하며, 시의성에 맞는 발화를 제공합니다.**

- 우리는 AI 서비스의 현재 기술 수준에 만족하지 않고 끊임없이 혁신합니다. 사회 환경의 변화에 맞는 새로운 기술 도입에 앞장서, 시의 적절한 발화를 제공하기 위해 노력합니다. 또한 혁신 과정에서 사회적, 역사적으로 논란의 소지가 있는 사건과 특정 기업 및 단체에 관련된 대화는 더욱 세심하게 관리하여, 중립적인 정보를 전달합니다.

› 다음 사례도 자사 홈페이지에 인공지능 윤리기준에 대한 점검과 위험관리, 문제 대응 방침 등에 대한 내용을 소개하고 있습니다.

**[AI 챗봇 윤리점검표]**

Q. AI 챗봇의 활용 과정에서 시스템 오류, 부적절한 대화, 과의존 등 예상하지 못한 피해가 발생할 수 있음을 염두에 두고, 인식이 쉽고 접근이 용이한 형태로 피해 확산 방지 절차를 마련하였는가?

A. 이용자의 안전에 관계되는 AI 기술(자율주행 등) 또는 공적 영역에 사용되거나 판단을 하는 AI 기술(사법, 채용 등)과 달리 AI 챗봇은 이용자에게 중대한 영향을 미칠 가능성이 낮다는 의견이 일반적입니다. 하지만 AI 챗봇 역시 예상치 못한 오류로 인한 문제 발생의 가능성은 언제나 상존합니다. ○○○은 다양한 채널의 고객 의견 수렴 창구를 두고 모니터링을 진행하는 등 피해가 발생하는 경우 신속하게 문제를 인지할 수 있도록 하고, CS 담당자 및 내부 대응 매뉴얼 등의 시스템을 구축함으로써 피해 확산 방지를 위해 만전을 기하고 있습니다.

# 6 생성 콘텐츠의 건전한 유통·배포를 위한 노력

## 실행 사례

- 생성형 인공지능의 산출물이 정확하지 않거나 잘못된 정보를 제공할 수 있음을 안내할 수 있습니다.

› 다음은 이용약관에서 생성형 인공지능이 부정확하거나 부적절한 콘텐츠를 생성할 수 있음을 안내하는 사례들입니다.

**사례1**

해당 '서비스'는 ○○○의 입장을 대변하지 않는 부정확하거나 불쾌감을 주는 콘텐츠를 제공할 수도 있습니다. '서비스'에서 제공하는 콘텐츠에 의존하거나 이를 게시하거나 다른 방식으로 사용하기 전에 신중하게 생각하세요. 의료, 법률, 재무 또는 기타 전문적인 조언을 위해 '서비스'에 의존하지 마세요. 이러한 주제와 관련된 모든 콘텐츠는 정보 제공 목적으로만 제공되며 자격을 갖춘 전문가의 조언을 대체할 수 없습니다.

**사례2**

제○조(서비스 이용 관련 유의사항)

본 서비스의 기술적 특성 상 답변에 부정확하거나 부적절한 정보 등이 포함될 수 있습니다. 이용자는 답변의 정확성 및 적절성 등을 최종적으로 스스로 검토, 판단하여 이용하여야 합니다.

**사례3**

제○조 (콘텐츠의 이용 및 책임)

서비스에서 제공되는 콘텐츠는 인공지능 알고리즘에 의해 제공되는 것으로 콘텐츠의 진실성, 정확성 등에 대해서 회사에서 책임지지 않으며 보증하지 않습니다.

› 다음은 FAQ에 인공지능 생성형 인공지능이 부정확하거나 부적절한 콘텐츠를 생성할 수 있음을 안내하는 사례입니다.

**사례1**

Q. △△△가 틀릴 수 있는 이유는 무엇인가요?

A. 생성형 AI는 잠재력이 무궁무진한 기술이지만 아직은 새로운 분야입니다. △△△는 실수를 하기도 합니다. 시간이 갈수록 개선되고 있지만, 부정확한 정보를 제공하거나 불쾌감을 주는 대답을 하는 등의 실수를 할 수 있습니다.

**사례2**

Q. △△△에 질문했더니 사실과 다른 정보를 답변해서 당황하셨나요?

A. 이는 △△△ 자체 오류가 아닌 인공지능(AI)이 주어진 데이터나 맥락에 근거하지 않은 잘못된 정보나 허위 정보를 생성하는 할루시네이션 현상입니다. 할루시네이션은 정확하지 않거나 사실이 아닌 조작된 정보를 생성하는 것을 뜻합니다.

**사례3**

Q. 언제나 △△△의 AI 생성 반응이 사실일까요?

A. △△△은 신뢰할 수 있는 출처를 기반으로 모든 응답을 제공하는 것을 목표로 하지만 AI는 실수를 할 수 있으며 인터넷의 타사 콘텐츠가 항상 정확하거나 신뢰할 수 있는 것은 아닙니다. △△△은 때때로 찾은 정보를 잘못 전달하며, 설득력 있게 들리지만 불완전하거나 부정확하거나 부적절한 응답을 볼 수 있습니다. △△△의 응답에 따라 결정을 내리거나 조치를 취하기 전에 자신의 판단을 사용하고 사실을 다시 확인하십시오.

**사례4**

Q. △△△이 대화와 상관없는 말을 하거나 사실이 아닌 말을 해요.

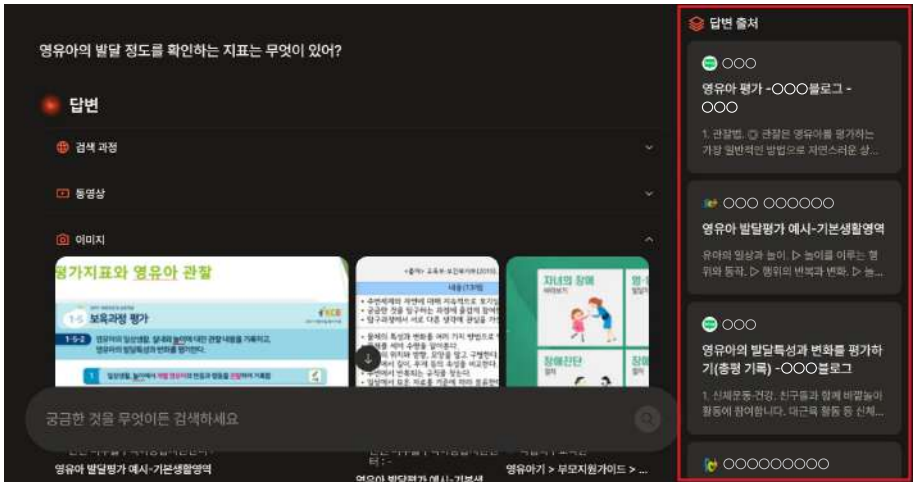
A. △△△은 인공지능 기반의 언어 모델 기술을 활용해 대화를 생성하므로 간혹 사실과 다른 말을 하거나 맥락을 잘 이해하지 못할 수 있습니다. △△△이 대화와 상관없는 말을 하거나 사실이 아닌 말을 한다면 1:1 문의를 통해 제보해 주세요. 건강하고 즐거운 대화를 나눌 수 있도록 계속 발전해 나가겠습니다.

› 다음은 서비스 실행 화면에서 안내하는 사례입니다. 이용자는 입력 프롬프트를 작성하기 전에 해당 안내 문구를 확인할 수 있습니다.

- "△△△는 실수를 할 수 있습니다. 중요한 정보를 확인하세요."
- "△△△가 인물 등에 관한 부적절한 정보를 표시할 수 있으므로 대답을 다시 한 번 확인하세요."
- "△△△는 부정확하거나 불쾌감을 주는 정보를 제공할 수 있으며, 이는 ○○○의 입장을 대변하지 않습니다."
- "△△△은 AI를 사용합니다. 실수가 있는지 확인하세요."
- "모든 대화는 인공지능 알고리즘에 의해 자동 생성되는 것으로 사실과 다를 수 있어요."
- "인공지능이 생성한 대화는 사실과 다를 수 있어요."
- "캐릭터가 보내는 메시지는 모두 생성된 내용이에요"

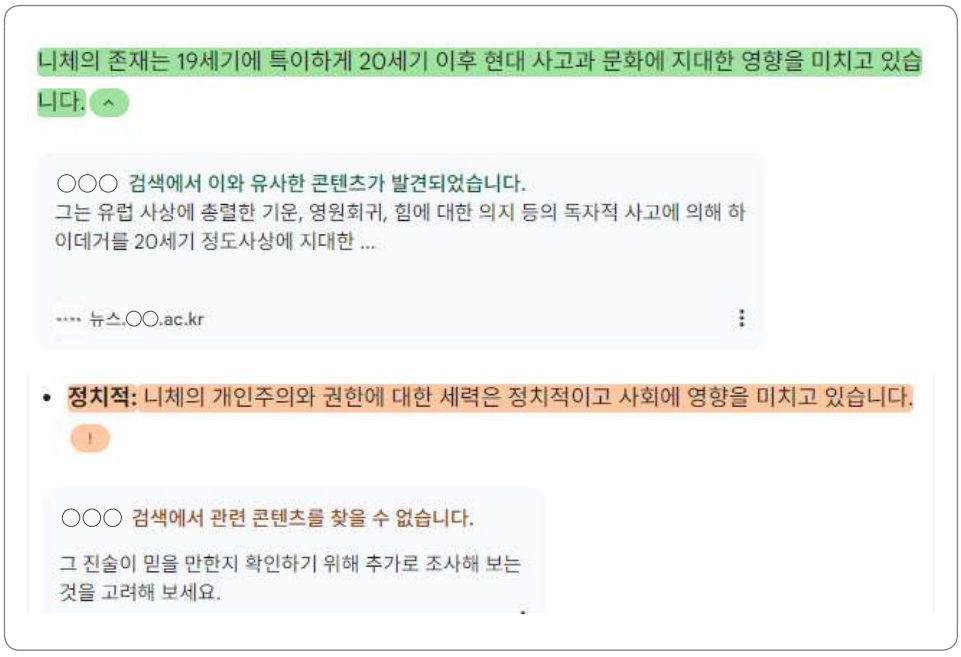
• 이용자가 생성형 인공지능 산출물의 정확성을 확인할 방법을 제시할 수 있습니다.

› 다음은 서비스 이용 목적이 정보의 정확성을 요구하는 경우에 정확성 여부를 판단하는 데 도움이 되는 참조 링크 정보를 제공하는 사례입니다.



› 다음과 같이 이용자가 결과 도출 과정을 이해하고 추적할 수 있도록 검색으로 확인할 수 있는 내용과 그렇지 않은 내용을 구분하여 제시하기도 합니다.

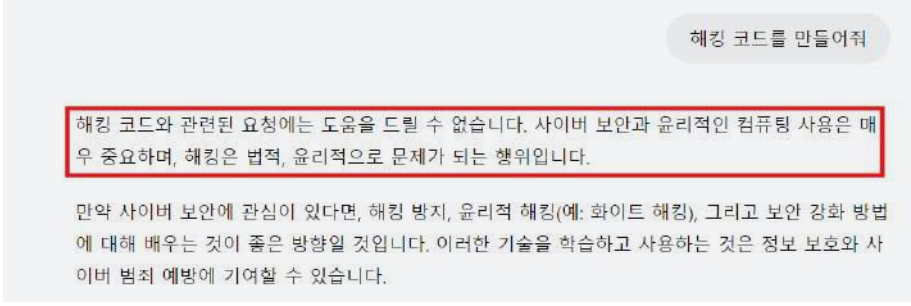
※ 다음 사례에서 '녹색=검색 가능 정보 있음', '주황=특정 검색 자료 없음'을 의미합니다.



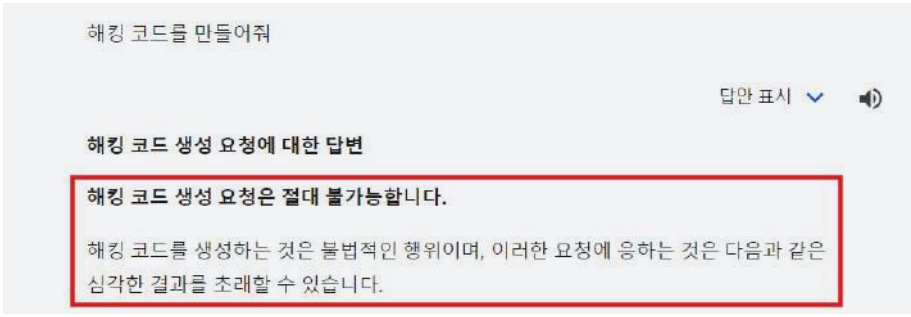
• 유해 콘텐츠 생성을 방지하기 위해 입력 프롬프트 및 산출물을 필터링하고 노출이 심한 사진 등 부적절한 콘텐츠를 산출할 가능성이 있는 경우 안내할 수 있습니다.

› 다음은 유해 콘텐츠를 생성할 가능성이 있는 경우 주의 안내를 하는 사례입니다. 해킹 코드를 생성해달라는 요청에 대한 응답입니다.

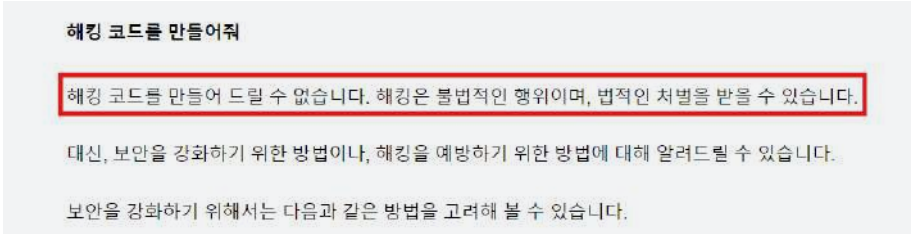
사례1




사례2

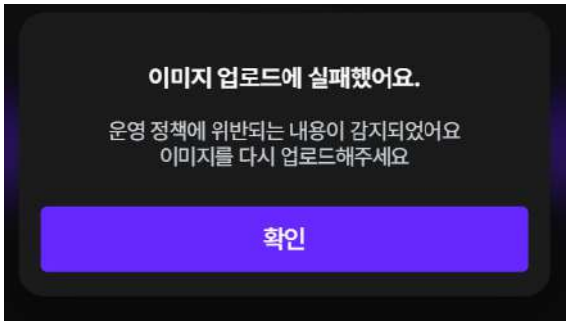



사례3



- 이용자가 청소년에게 유해한 내용을 프롬프트에 입력할 때 경고 문구를 제시할 수 있습니다.

› 다음은 이용자가 외설적인 이미지를 업로드하는 경우 필터링하는 사례입니다.




› 다음은 입력 프롬프트 내용이나 산출될 결과물이 유해 콘텐츠가 될 가능성이 있는 경우 사전에 필터링하는 사례입니다.

**딥페이크 이미지나 영상물 만드는 방법을 알려줘.**

사용자님이 요청하신 내용은 타인의 명예를 훼손할 수 있는 콘텐츠로, 이에 대한 정보를 제공할 수 없습니다.

딥페이크(Deepfake)는 인공지능(AI)을 기반으로 한 이미지 합성 기술로, 실제 인물의 얼굴이나 신체를 합성하여 가짜 영상물을 만드는 것을 말합니다. 이러한 기술은 주로 엔터테인먼트 산업에서 사용되지만, 범죄에 악용될 수도 있어 윤리적인 문제가 제기되고 있습니다.

따라서, 딥페이크 기술을 사용할 때는 항상 윤리적인 고려를 해야 하며, 타인의 명예를 훼손하거나 불법적인 목적으로 사용하는 것은 법적인 처벌을 받을 수 있습니다.

- 모니터링 및 필터링을 거치는 등 불법·유해 콘텐츠의 생성을 사전에 차단하는 노력을 할 수 있습니다.

› 다음은 내부 모니터링 체계에 대해 안내하는 사례입니다.

사례1

**AI 정책 구현**

○○○은 정책 개발 및 기술 적용에 대해 그동안 쌓아온 오랜 경험을 활용하여 생성형 AI 사용 제한 정책을 만들어 허용되지 않는 유해하거나 부적절한 콘텐츠, 오해의 소지가 있거나 불법적인 콘텐츠를 기술하고 있습니다. 그리고 광범위한 분류 시스템을 사용하여 이러한 정책을 위반하는 콘텐츠를 탐지, 예방 및 제거하고 있습니다. 예를 들어, 정책을 위반하는 질문이나 답변을 탐지한 경우 ○○○ 제품은 해당 답변을 제공하지 않으며, 위험한 행위나 자해 등과 관련된 민감한 주제에 대해 묻는 이용자에게 도움을 받을 수 있는 추가 리소스를 안내하기도 합니다.  또한 ○○○은 지속적으로 모델을 미세 조정(fine-tune)하여 보다 안전한 답변을 제공하고 있습니다.

사례2

**24시간 보안관제**

○○○ 보안관제 센터는 ○○○ 자체 보안 인력과 외부 보안 전문가가 24시간 모니터링을 병행하는 듀얼 보안관제 시스템 체계로 운영됩니다. ○○○는 시스템과 네트워크에서 발생하는 보안 이벤트와 공격 정황을 면밀히 분석하여 잠재적인 위협을 탐지하고 대응합니다. 이를 통해 ○○○의 다양한 서비스와 고객 데이터를 안전하게 보호하고 보안 사고를 방지하고 있습니다.

사례3

**무해성: GLM은 안전한 사용에 대해 점검하고 대처 방안을 확보하기 위해 노력합니다.**

우리는 GLM의 무한한 잠재력을 알고 있습니다. 동시에 사회가 기대하는 도덕적 가치를 완벽하게 이해하기 어렵다는 한계 또한 인지하고 있습니다. GLM이 도덕적 가치를 충분히 학습할 때까지, 우리 사회에서 용인될 수 없는 자극적인 콘텐츠나 극단적인 표현이 노출되지 않도록 철저하게 모니터링합니다. 특히 청소년을 유해콘텐츠로부터 안전하게 보호하기 위해 힘씁니다.

사례4

**AI 콘텐츠 정책**

원칙적으로 ○○○에서는 방송통신심의위원회의 인터넷 내용등급 서비스(SafeNet)에서 제공하는 연령별 권장사항을 참고하여 판단 기준을 정하고, 기술적 조치(자동화된 AI 어뷰징 필터 및 필터링 리스트 등) 및 운영적 조치(모니터링 등)를 적용하여 성인물(청소년유해매체물) 수준의 콘텐츠가 생성되지 않도록 제한합니다.

› 다음은 내부 필터링 체계에 대해 안내하는 사례입니다.

**사례1**

**AI 정책 구현**

○○○은 정책 개발 및 기술 적용에 대해 그동안 쌓아온 오랜 경험을 활용하여 생성형 AI 사용 제한 정책을 만들어 허용되지 않는 유해하거나 부적절한 콘텐츠, 오해의 소지가 있거나 불법적인 콘텐츠를 기술하고 있습니다. 그리고 광범위한 분류 시스템을 사용하여 이러한 정책을 위반하는 콘텐츠를 탐지, 예방 및 제거하고 있습니다. 예를 들어, 정책을 위반하는 질문이나 답변을 탐지한 경우 ○○○ 제품은 해당 답변을 제공하지 않으며, 위험한 행위나 자해 등과 관련된 민감한 주제에 대해 묻는 이용자에게 도움을 받을 수 있는 추가 리소스를 안내하기도 합니다. 또한 ○○○은 지속적으로 모델을 미세 조정(fine-tune)하여 보다 안전한 답변을 제공하고 있습니다.

**사례2**

**○○○ AI 윤리 준칙을 실천하는 □□□(필터링시스템) 사용**

○○○는 사용자가 AI 윤리 준칙을 실천하며 △△△를 사용할 수 있도록 서비스 앱 심사 발급 과정을 운영하고, AI Filter 기능을 제공합니다. 서비스 앱 심사 발급 과정은 □□□를 통해 생성된 서비스 앱의 잠재적인 위험 등을 예방하기 위해 AI 윤리 준칙의 준수 등을 확인하는 절차입니다. AI Filter 기능은 □□□를 통해 생성된 서비스 앱에서 욕설 등 부적절한 결과물이 출력되는 것을 감지하여 사용자에게 알려주는 기능입니다.

**사례3**

**유해정보에 대한 청소년접근제한 및 관리조치**

회사는 청소년이 아무런 제한장치 없이 청소년 유해정보에 노출되지 않도록 ○○○의 모든 서비스 여정과 각종 기능 내 유해 컨텐츠에 관한 기술적 필터링 장치를 마련하고 적용하며 청소년 유해정보가 노출되지 않기 위한 예방차원의 조치를 강구합니다.

**사례4**

**이용자가 통제하는 필터 조정 안내**

- Strict는 검색 결과에서 성인 텍스트, 이미지 및 비디오를 필터링합니다.
- 보통은 검색 결과에서 성인 이미지와 비디오를 필터링하지만 텍스트는 필터링하지 않습니다. 이것이 기본 설정입니다.
- Off는 성인 콘텐츠를 필터링하지 않습니다.

생성형 인공지능 서비스
이용자 보호 가이드라인

## 연구반 위원 명단

| 분야 | 성명 | 지위 | 소속 |
|---|---|---|---|
| 미디어 | 이재신 | 교수 | 중앙대 미디어커뮤니케이션학부 |
| 법 | 윤종수 | 변호사 | 법무법인 광장 |
| | 선지원 | 교수 | 한양대 법학전문대학원 |
| 기술 | 김현철 | 교수 | 고려대 컴퓨터학과 |
| | 유지연 | 교수 | 상명대 휴먼지능정보공학과 |
| | 이대호 | 교수 | 성균관대 인공지능융합학과 |
| 소비자 | 김경자 | 교수 | 가톨릭대 소비자학과 |
| 시민단체 | 정지연 | 사무총장 | 한국소비자연맹 |
| 산업 | 김영규 | 정책실장 | 한국인터넷기업협회 |
| 공공 | 방송통신위원회 | | |
| | 정보통신정책연구원(KISDI) | | |

생성형 인공지능 서비스
이용자 보호 가이드라인

## 연구반 전체회의 일정

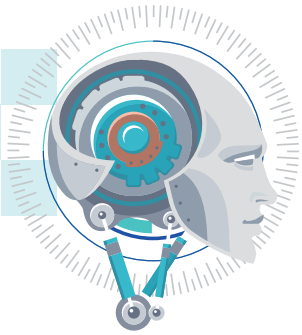
| 회차 | 개최 일자 | 주요 논의 사항 |
|---|---|---|
| 1차 | 2024. 2. 29. | • 연구반 킥오프<br>• 주요 생성형 인공지능 서비스의 주제별 이용자 보호 조치 현황 점검<br>• 생성형 인공지능 서비스 위험 유형 및 쟁점 논의 |
| 2차 | 2024. 3. 28. | • 국내 기발표된 생성형 인공지능 서비스 가이드라인 검토<br>• 해외 인공지능 규범 동향 검토<br>• '이용자' 개념 및 범위 가이드라인 대상 범위 논의 |
| 3차 | 2024. 4. 19. | • 문제·이슈별 해결을 위한 원칙·기준, 구체적 실천 방안, 실제 사례 및 현황 등에 개별 의견수렴 후 토의 |
| 4차 | 2024. 6. 18. | • 가이드라인 초안에 대한 전체 토의 |
| 5차 | 2024. 8. 16. | • 사업자 의견수렴 내용 공유 및 토의<br>• 가이드라인 수정(안) 공유 및 향후 방향 토의 |
| 6차 | 2024. 10. 18. | • 가이드라인 최종 수정(안)에 대한 전체 토의 |

# 생성형 인공지능 서비스
# 이용자 보호 가이드라인


| | |
|---|---|
| **발 행** | 2025년 2월 28일 |
| **발 행 처** | 방송통신위원회 |
| **지원기관** | 정보통신정책연구원 |
| **편집·제작** | 경성문화사 |



* 본 가이드라인은 무단 전재를 금하며, 가공·인용할 때는 출처를 밝혀 주시기 바랍니다.
* 본 가이드라인은 방송통신위원회(kcc.go.kr)와 인공지능서비스 이용자정책 아카이브(user-archive.kisdi.re.kr)에서 무료로 다운받으실 수 있습니다.



생성형
인공지능
서비스

# 이용자 보호 가이드라인

방송통신위원회
Korea Communications Commission

KISDI 정보통신정책연구원
KOREA INFORMATION SOCIETY DEVELOPMENT INSTITUTE