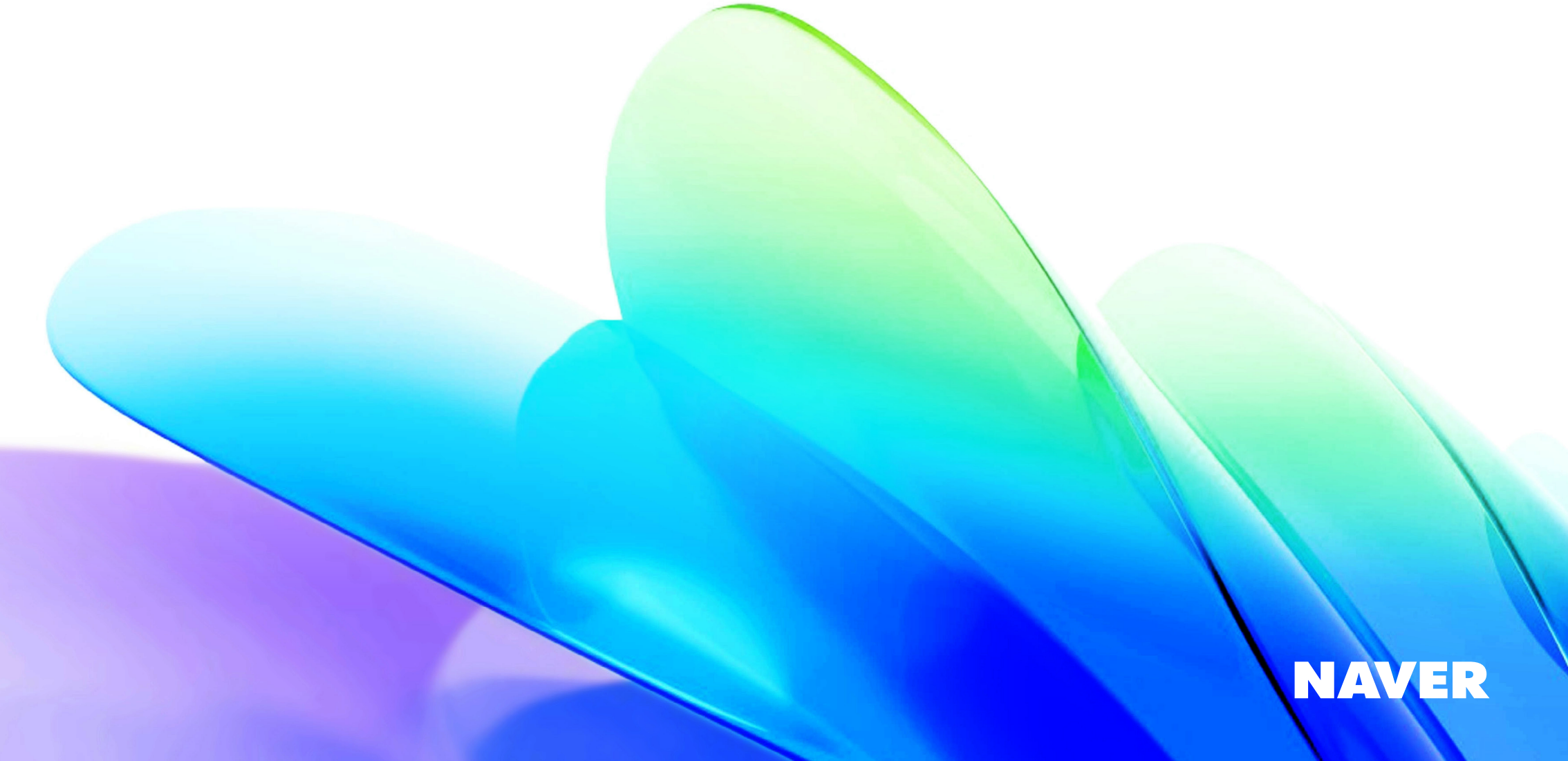
NAVER AI SAFETY
PROGRESS REPORT 2026



NAVER

목차

추천사

최소 기준 위에 스스로 쌓아 올린 실천의 기록

AI 안전은 선언으로 완성되지 않고 운영 속에서 증명됩니다. 2026년 1월 22일 「인공지능기본법」이 시행되면서 우리 사회는 AI 신뢰의 최소 기준을 제도로 세웠습니다. 그러나 법이 그리는 것은 윤곽선입니다. 그 위에서 어떤 위험을 어떻게 다룰지는 결국 서비스를 만드는 기업이 스스로 답해야 합니다. 『AI Safety Progress Report 2026』은 네이버가 그 답을 처음으로 기록으로 남긴 문서입니다.

위험의 자리도 달라졌습니다. 같은 모델이라도 어떤 서비스에, 어떤 사용자를 향해 놓이는가에 따라 위험의 양상은 달라집니다. 수천만 명이 매일 이용하는 서비스를 운영하는 기업이 모델 단위가 아닌 서비스 단위의 안전을 다뤄야 하는 이유가 여기에 있습니다.

이 보고서에는 2021년 AI 윤리 준칙에서 출발해 프런티어 모델의 위험을 다룬 ASF Beta를 거쳐, 관리의 범위를 서비스로 넓힌 ASF 2.0에 이르는 경로가 담겨 있습니다. 원칙이 기준이 되고, 기준이 CHEC 2.0이라는 전사 실행 체계로 옮겨간 과정을 확인할 수 있습니다. 110개 항목으로 위험을 유형화한 분류 체계, 유해성과 사용성을 함께 측정하여 거절이 곧 안전이라는 착각을 경계한 평가 설계, 새로 확인된 위험을 사흘 만에 체계에 반영한 사례, 설계에서 운영까지 서비스 조직과 AI Safety Center가 함께한 AI탭 케이스 스터디가 특히 눈에 띕니다.

다만 이 기록은 완성이 아니라 진행형입니다. 보고서의 이름이 'Progress'인 이유도 여기에 있다고 봅니다. 기술이 다루는 영역이 넓어질수록 새로운 위험은 계속 나타나고, 오늘의 체계는 내일 다시 손보아야 합니다. 다음 보고서에는 무엇이 잘 작동했는지와 함께 무엇이 뜻대로 되지 않았는지도 담기기를 기대합니다. 그것이 사회가 기업의 자율 실천을 신뢰하게 되는 가장 빠른 길이기 때문입니다.

AI 안전은 어느 한 기업이나 기관이 홀로 감당할 과제가 아닙니다. 인공지능안전연구소는 규제자가 아니라 셰르파로서 정부와 학계, 산업계를 잇는 자리에서 그 역할을 찾아왔습니다. 연구소가 발간한 『국제 AI 안전 보고서 2026』 한국어판에는 한국어와 문화적 맥락을 고려한 네이버의 안전성 평가 활동이 소개되었고, 제2회 인공지능 안전 서울 포럼(SFASS 2026)에서는 네이버가 ASF 2.0의 방향을 국내외 전문가들과 나누었습니다. 언어와 문화의 경계에서 생기는 위험은 그 언어로 서비스를 만드는 곳이 가장 먼저 마주합니다. 현장의 경험이 공유될 때 기준은 더 단단해집니다.

이 기록이 국내 AI 안전 논의와 실천 확산의 마중물이 되기를 바랍니다. 인공지능안전연구소도 그 길에 함께하겠습니다.

2026년 9월

김명주
인공지능안전연구소 소장

발간사

유봉석
NAVER CRO(Chief Corporate Responsibility Officer)

원성재
NAVER AI Safety Center 센터장

AI Safety는 더 많은 사람들이 AI를 믿고 활용할 수 있도록 하는 혁신의 기반

AI는 새로운 기술을 넘어, 사람들의 일상과 업무를 지원하는 보편적인 도구로 자리 잡고 있습니다. 나아가 서비스와 산업을 움직이는 기반이 되고 있습니다. AI가 정보를 제공하는 단계를 넘어 사용자를 대신해 행동하는 에이전트로 발전하고, 하나의 서비스 안에서 다양한 모델과 도구가 결합되면서 AI가 사람과 사회에 미치는 영향도 더욱 넓고 복합적으로 변화하고 있습니다.

네이버 역시 서비스와 업무 방식 전반을 'AI 네이티브(AI Native)'로 전환하기 위한 전략을 고도화하고 있습니다. 올해 AI 쇼핑 에이전트와 AI탭을 선보였으며, 앞으로도 다양한 분야에 특화된 버티컬 에이전트 서비스가 출시될 예정입니다. AI Safety의 범위도 이러한 변화 속에서 달라지고 있습니다.

AI Safety는 이제 하나의 모델을 안전하게 만드는 문제를 넘어, 여러 AI 모델과 기술이 결합된 서비스를 수천만 명의 사용자가 안심하고 이용할 수 있도록 어떻게 설계하고 운영할 것인가의 문제로 확장되고 있습니다. AI Safety는 AI 네이티브라는 방향과 분리된 과제가 아니라, 그 방향을 실현하기 위한 전제 조건이 됐습니다.

발간사

네이버는 이러한 인식 아래, AI Safety를 선언적 원칙에 그치지 않고 구체적인 기준과 실행 체계로 발전시켜 왔습니다. 2021년 '네이버 AI 윤리 준칙'을 시작으로, 2024년에는 프론티어 AI 모델의 위험을 관리하기 위한 'NAVER ASF Beta', 2026년에는 관리의 범위를 모델에서 사용자와 서비스로 확장한 'NAVER ASF 2.0'을 공개했습니다. 특히 ASF 2.0은 위험의 정의와 분류, 위험의 식별과 평가, 위험의 관리와 개선을 CHEC 2.0이라는 전사적 실행 체계로 연결하고 있습니다.

이를 뒷받침하기 위해 CRO 산하 AI 리스크 관리 기능을 2026년 3월 AI Safety Center로 재편했습니다. 아울러 서비스 실무 조직의 실행에서부터 CRO 산하 AI Safety Center의 관리와 지원, 이사회의 감독과 최종 의사결정에 이르는 3계층 관리·감독 체계가 마련됐습니다. AI Safety를 경영의 우선순위로 두고 필요한 조직과 자원을 지속적으로 투입하겠다는 회사의 의지가 이 체계의 기반이 됐습니다.

이번 'AI Safety Progress Report'는 이러한 원칙과 체계가 실제 서비스 현장에서 어떻게 실행되고 있는지를 사회와 공유하기 위해 마련했습니다. 우리가 어떤 위험을 중요하게 바라보았는지, 이를 어떻게 평가하고 관리했는지, 사용자에게 필요한 정보를 어떠한 방식으로 전달하고자 했는지, 그리고 그 과정에서 무엇을 배우고 개선했는지를 가능한 한 구체적으로 담고자 했습니다.

이번 리포트에서 소개된 대화형 AI 검색 서비스 'AI탭' Case Study는 AI Safety 프레임워크가 실제 서비스에서 어떻게 작동하는지를 잘 보여주는 사례입니다. AI탭은 설계부터 출시에 이르는 전 과정에서 서비스 조직과 AI Safety Center가 하나의 팀으로 협력했으며, 서비스의 특성과 이용 맥락에 맞는 위험 분류 체계와 Safety 도구를 함께 만들어 적용했습니다.

이러한 협력의 범위는 기업 내부에만 머물지 않습니다. 네이버는 학계와의 공동 연구를 통해 AI 안전성 연구의 지평을 넓혀 왔으며, 그 성과를 공개해 왔습니다. 나아가 인공지능안전연구소를 비롯한 유관 기관, 외부 전문가, 국제기구와의 협력을 통해 국내외 파트너들과 지식과 고민을 나누며 AI Safety의 기준을 함께 만들어가고 있습니다. 네이버는 또한 AI 서비스 관리체계를 객관적으로 인정받기 위해, AI 경영시스템 국제표준인 ISO/IEC 42001 인증도 추진하고 있습니다.

이번 리포트는 NAVER ASF 2.0에서 약속한 '경험과 노하우의 공유'를 실천하는 첫걸음입니다. 우리가 무엇을 시도했고, 무엇을 이루었으며, 그 과정에서 무엇을 배웠는지를 기록하고 공개함으로써 스스로를 점검하고 사회와 함께 더 나은 답을 찾아가겠다는 약속이기도 합니다.

네이버는 혁신의 속도가 빨라질수록 안전을 살피는 노력도 더욱 구체적이고 지속적이어야 한다고 생각합니다. 이를 위해 AI Safety를 혁신을 제한하는 조건이 아니라, 더 많은 사람이 AI를 믿고 활용할 수 있도록 하는 혁신의 기반으로 만들어 가고자 합니다.

이번 리포트가 네이버의 AI를 믿고 이용해 주시는 모든 분들께 우리의 노력을 전하는 기록이 되기를 기대합니다. 아울러 사용자와 전문가, 산업계와 정책 당국이 함께 더 나은 AI Safety의 기준을 논의하는 출발점이 되기를 바랍니다. 앞으로도 네이버는 AI가 사람을 위한 일상의 도구로 발전할 수 있도록 끊임없이 질문하고, 점검하고, 개선하며 그 과정을 공유하겠습니다.

감사합니다.

2026년 9월

GOVERNANCE & FRAMEWORK

- 서비스 중심 AI Safety Framework: NAVER ASF 2.0
- NAVER ASF 2.0의 실행체계: CHEC 2.0

서비스 중심 AI Safety Framework: NAVER ASF 2.0

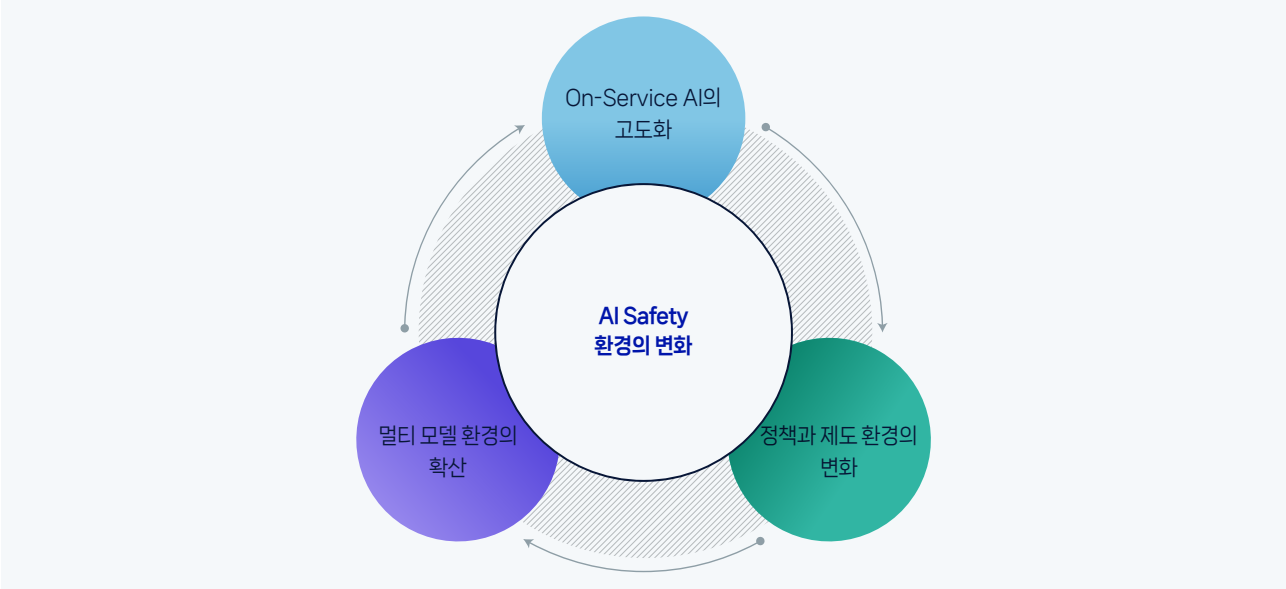
NAVER ASF 2.0의 방향성

네이버는 2026년 7월 7일 NAVER AI Safety Framework 2.0(이하 'NAVER ASF 2.0')을 공개했습니다. NAVER ASF 2.0은 사회가 우려하는 AI 위험에 대응하기 위한 AI Safety 체계입니다. 네이버는 2021년 네이버 AI 윤리 준칙을 발표한 이후, 2024년 NAVER ASF Beta를 통해 프론티어 AI 모델의 위험을 식별하고 관리하기 위한 체계를 마련하였고, 글로벌 흐름과 국내 정책 환경에 맞춰 이를 지속적으로 구체화해 왔습니다.

한편 AI 기술이 구체적인 서비스에 적용되며 사용자의 일상에 들어오면서, AI Safety는 이론적 논의를 넘어 현실의 문제를 직접 다뤄야 하는 단계로 접어들고 있습니다.

이러한 현실의 문제는 개별 서비스의 고유한 맥락 속에서 발생합니다. 네이버는 AI 모델 자체의 안전성뿐만 아니라, 서비스가 놓인 환경과 사용자와의 상호작용 방식, 그리고 그 과정에서 생길 수 있는 다양한 상황까지 함께 고려할 필요가 있다고 판단했습니다.

이와 같은 변화에 대응하기 위해 네이버는 On-Service AI라는 방향성 아래에서 NAVER ASF 2.0을 정리했습니다. 이를 통해 AI 서비스가 사용자에게 어떻게 사용될 수 있는지, 그리고 그에 따라 사용자와 사회 구성원, 나아가 AI 생태계 전체에 어떠한 영향을 미칠 수 있는지를 여러 관점에서 살펴보고 해결해 나가고자 하였습니다.



네이버가 주목하고 있는 AI Safety 환경의 변화는 크게 세 가지입니다. 첫째는 On-Service AI의 고도화입니다. 네이버는 AI 기술을 서비스 전반에 결합해 사용자 경험을 확장하는 On-Service AI 전략을 고도화하고 있으며, 이미 AI 브리핑, AI 쇼핑 에이전트, AI탭 등을 선보인 데 이어 다양한 버티컬 에이전트 서비스를 준비하고 있습니다. 동일한 AI 모델이라 하더라도 서비스의 영역과 활용 범위에 따라 위험의 유형과 요구되는 안전의 수준이 달라질 수 있는 만큼, 안전 관리의 범위 역시 모델에서 서비스로 확장될 필요가 있다고 보았습니다.

둘째는 멀티 모델 환경의 확산입니다. NAVER ASF Beta 공개 시점의 생성형 AI 서비스는 하나의 거대 언어 모델이 모든 요청을 처리하는 구조가 일반적이었으나, 최근에는 내부와 외부, 대형과 소형 모델을 목적에 맞게 조합하는 멀티 모델 환경이 확산되고 있습니다. 이러한 환경에서는 모델 단위의 검토와 함께 모델을 결합하는 방식과 서비스 내에서의 활용 맥락에서 새롭게 생겨날 수 있는 위험까지 함께 살펴볼 필요가 있다고 판단했습니다.

셋째는 정책과 제도 환경의 변화입니다. 2026년 1월 22일 인공지능 발전과 신뢰 기반 조성 등에 관한 기본법(이하 '인공지능기본법')이 시행되었습니다. 네이버는 국내 정책 환경과 글로벌 흐름에 발맞추어, 사회기술적 맥락 속에서 AI 서비스가 마주하는 현실적이고 개별적인 문제에 대해 구체적으로 위험을 정의하고 이를 식별, 평가, 관리 및 개선하는 AI Safety 프로세스를 고도화해 나가고 있습니다.

NAVER ASF 2.0은 이와 같은 환경 변화, 즉 On-Service AI의 고도화, 멀티 모델 환경의 확산, 정책 및 제도 환경의 변화에 대응하기 위해 마련된 서비스 중심의 AI Safety Framework입니다. 네이버는 NAVER ASF 2.0을 통해 서비스 중심의 AI Safety 거버넌스를 지속적으로 개선하고 발전시켜 On-Service AI에 AI Safety를 구체화해 나가고자 합니다.

서비스 중심 AI Safety Framework: NAVER ASF 2.0

NAVER ASF 2.0의 진행 과정

네이버는 AI Safety 환경 변화에 맞춰, 2026년 3월 네이버 AI 윤리 준칙의 개정이 필요한지 검토하기 시작했습니다. 2021년에 수립한 네이버 AI 윤리 준칙에서 수정되거나 추가해야 할 사항이 있을지 살펴보기 위해서였습니다. AI Safety Center 중심으로 네이버 AI 윤리 준칙의 내용을 재검토한 결과, 네이버 AI 윤리 준칙을 개정하기보다는 NAVER ASF Beta를 글로벌 흐름과 국내 정책 환경에 맞춰 구체화하는 것이 바람직하다는 판단을 내렸습니다.

그 이유는 네이버 AI 윤리 준칙이 네이버의 기업철학과 네이버가 추구하는 가치를 담고 있어, 2021년 수립 및 공개로부터 5년이 지난 현재의 시점에서도 여전히 그 내용이 유의미하다고 보았기 때문이었습니다. 네이버 AI 윤리 준칙은 AI에 대한 네이버의 산업적 관점뿐만 아니라 사회적 요구까지 통합적으로 고려하고 있었습니다. 무엇보다 그 안에는 네이버가 추구하는 가치가 담겨 있기에, 네이버가 AI를 바라보는 관점이 근본적으로 달라지지 않는 한 다양한 변화 속에서도 변하지 않는 내용이 될 수 있다고 보았습니다.

네이버 AI 윤리·안전 관련 히스토리

이에 따라 AI를 실제 사용자에게 제공되는 서비스의 맥락에서, 서비스 중심의 안전 관리 실천 체계를 구성하자는 방향성을 가지고 NAVER ASF 2.0을 준비하게 되었습니다. AI Safety Center는 2026년 5월 초안을 작성하여 팀네이버의 최고 책임경영 책임자(Chief Corporate Responsibility Officer, CRO)에게 보고하여 AI Safety 정책의 방향성에 대한 피드백을 받았습니다. 그 과정에서 AI Safety 거버넌스 구조를 외부에 명확하게 설명하면 좋겠다는 피드백을 받아, 3계층(Layer) 관리 감독 체계를 구체화하여 담당 부서가 수행하는 역할을 명확히 규정하게 되었습니다.

또한 AI Safety Center는 서울대 인공지능 정책 이니셔티브와 함께 NAVER ASF 2.0 초안에 대한 내용을 검토하면서, AI 정책 분야의 외부 전문가 자문 및 협력 과정을 거쳤습니다. 이는 AI Safety 관련 평가와 주요 의사결정 과정에서 독립적 외부 전문가의 자문을 받아 거버넌스의 신뢰성을 높이기 위한 활동이었고, 산업적 관점뿐만 아니라 사회적 요구까지 통합적으로 고려하기 위한 결정이었습니다. 그 과정에서 NAVER ASF 2.0 보호 대상의 범주 및 AI 영향 평가 매트릭스에 대한 기준 설정 그리고 거버넌스 차원에서 AI Safety Center의 역할과 다른 부서와의 협업 체계가 구체적으로 논의되었습니다.

AI Safety Center는 네이버 내부 및 외부의 다양한 의견과 관점을 통합적으로 정리하여, NAVER ASF 2.0을 정리하고 2026년 7월 7일 외부에 공개했습니다.

앞으로도 네이버는 사용자를 위해 AI Safety 프로세스를 지속적으로 개선해 나갈 것입니다.

서비스 중심 AI Safety Framework: NAVER ASF 2.0

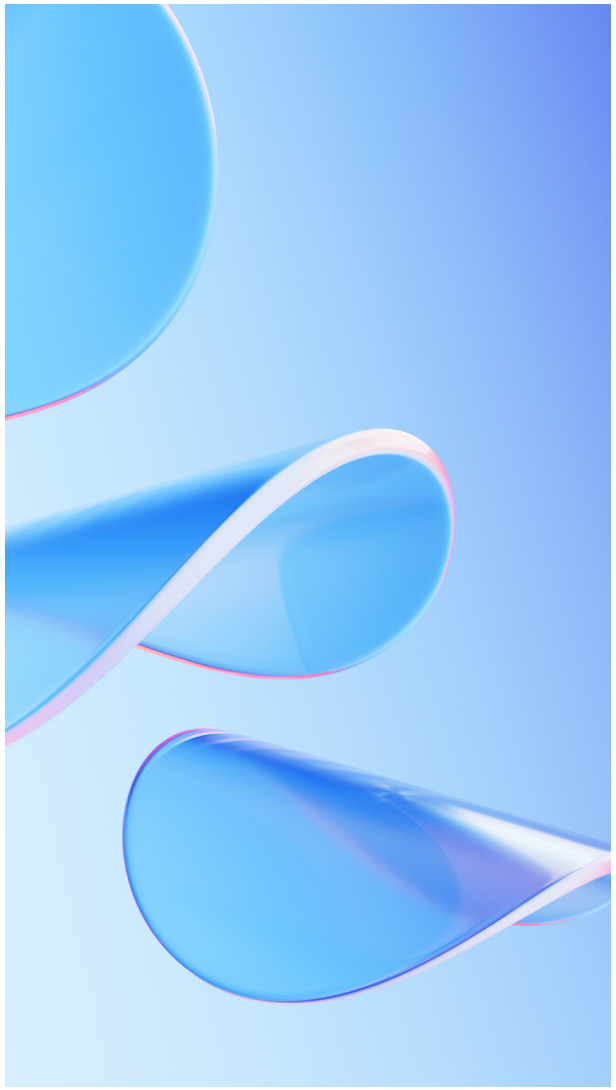
NAVER ASF 2.0의 세부 내용

NAVER ASF 2.0은 위험을 다루는 세 가지 축과 이 축들이 서비스 현장에서 실제로 작동하도록 만드는 실행 체계로 구성되어 있습니다. 세 가지 축은 (1) 위험 정의와 분류, (2) 위험 식별과 평가, (3) 위험 관리와 개선이며, 실행 체계는 네이버 AI 윤리·안전 자문 실행 체계인 CHEC 2.0(Consultation on Human-centered AI's Ethics and Safety Considerations, CHEC 2.0)입니다. 세 가지 축이 위험을 다루는 기준과 절차를 정의한다면, 실행 체계는 그 기준과 절차가 서비스의 전 주기에서 일관되게 적용되도록 하는 역할을 합니다. 네이버는 각각의 축과 실행 체계에 대해 이를 구체적으로 실행하기 위한 도구도 함께 마련했습니다.

세 가지 축은 독립적으로 작동하는 것이 아니라 실행 체계를 통해 하나의 사이클을 이룹니다. 분류 체계가 평가의 기준이 되고, 평가의 결과가 관리와 개선의 출발점이 되며, 관리 과정에서 새롭게 식별된 위험은 다시 분류 체계에 반영됩니다. 네이버는 이러한 사이클을 통해 변화하는 환경 속에서 AI 서비스의 안전성을 지속적으로 높여 나가고자 했습니다. 첫 번째 축인 위험 정의와 분류는 AI 서비스에서 다루어야 할 위험을 사전에 체계화하는 단계입니다. 네이버는 네이버 AI 윤리 준칙에서 언급한 사용자의 가치를 구체화하여, 사용자와 사회 구성원 그리고 AI 서비스 생태계를 보호 대상으로 삼고, 생명과 신체의 보호, 경제적 가치의 보호, 부당한 차별의 방지라는 세 가지 보호 가치를 정의했습니다. 그리고 이 보호 대상과 보호 가치를 조합하여 AI 서비스에서 발생할 수 있는 위험을 유형화한 AI 위험 분류 체계(AI Risk Taxonomy)를 구성했습니다. 이는 개별 서비스의 활용 맥락에서 어떤 위험을 검토해야 하는지 안내하는 출발점이 됩니다.

두 번째 축인 위험 식별과 평가는 AI 서비스가 보호 대상과 보호 가치에 미칠 수 있는 영향의 수준을 살펴보는 단계입니다. 네이버는 AI 영향 평가 매트릭스(AI Impact Assessment Matrix)를 통해, 해당 AI 서비스가 높은 위험에 해당하는 특수 영역인지 일반 영역인지 먼저 구분합니다. 특수 영역은 인공지능기본법상 고영향 인공지능이 활용될 수 있는 영역을 의미하며, 일반 영역은 다시 활용 범위가 넓은 경우와 좁은 경우로 나누어 살펴봅니다. 이처럼 서비스의 활용 영역과 범위에 따라 예상되는 영향 수준을 고려하여 안전 조치의 수준을 결정합니다.

세 번째 축인 위험 관리와 개선은 식별된 위험에 대해 안전 조치를 적용하고 지속적으로 개선해 나가는 단계로, 안전성 평가와 사용자 커뮤니케이션을 통해 이루어집니다. 안전성 평가에서는 AI 서비스가 정상적으로 작동하는지 그리고 그 과정에서 유해하거나 부정확한 결과가 출력되지 않는지를 다양한 측면에서 검토하며, 배포 이후에도 사용자와 운영 과정에서의 피드백을 바탕으로 지속적인 평가와 개선을 수행합니다. 사용자 커뮤니케이션 관점에서는 일상의 AI 활용에 대해 사용자의 눈높이에 맞춰 합리적인 설명을 제공하고, 특히 생성형 인공지능이 만들어 낸 결과물이 실제와 구분하기 어려워 사용자가 오인할 우려가 있는 경우에는 이를 쉽게 인식할 수 있는 방식으로 정보를 제공합니다.

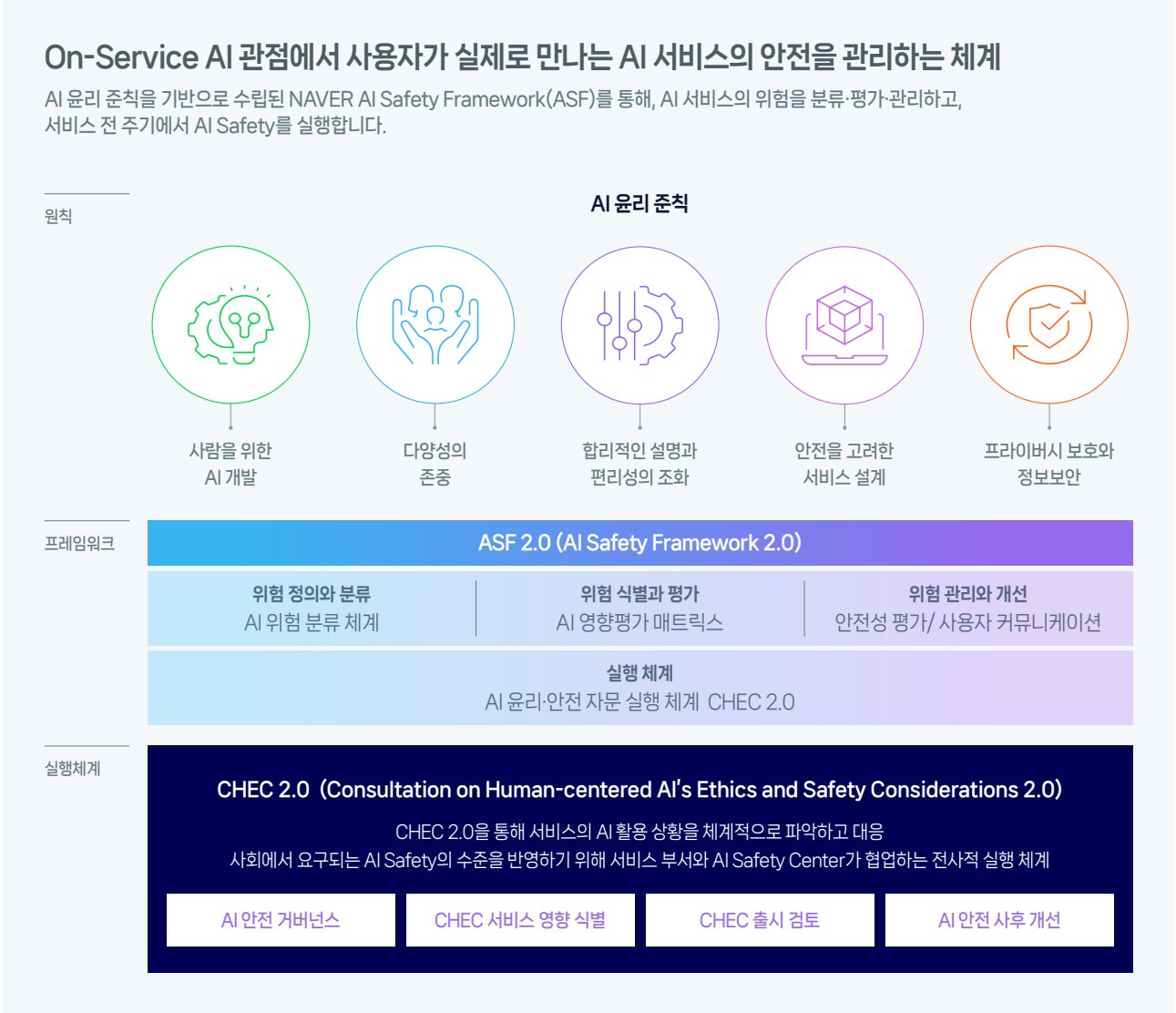


NAVER ASF 2.0의 실행 체계: CHEC 2.0

NAVER ASF 2.0의 세 가지 축은 실행 체계인 CHEC 2.0을 통해 개별 서비스의 맥락에서 구체적으로 실행됩니다. 네이버는 2022년 네이버 AI 윤리 자문 프로세스(Consultation on Human-centered AI's Ethical Considerations, CHEC)를 통해 네이버 AI 윤리 준칙의 관점에서 AI 서비스에 담긴 인간 중심의 가치를 재조명하고 AI에 대한 우려 사항을 살펴보며, 사용자에게 제공되는 서비스를 개선하는 노력을 해왔습니다. NAVER ASF 2.0에서는 AI 모델과 서비스에 대한 안전성 그리고 인공지능 생성물에 대한 사용자 커뮤니케이션을 강화하여 AI 윤리와 함께 AI Safety에 대한 사항을 점검하는 네이버 AI 윤리·안전 자문 실행 체계(Consultation on Human-centered AI's Ethics and Safety Considerations, CHEC 2.0)를 마련했습니다.

CHEC 2.0은 서비스 부서 단위에서 AI Safety를 실행할 수 있도록 가이드를 제공하는 한편, 사회에서 요구되는 AI Safety의 수준을 반영하기 위해 서비스 부서와 AI Safety Center가 협업하는 전사적 실행 체계입니다. CHEC 2.0은 AI 서비스의 기획 단계부터 출시 이후의 평가에 이르기까지, 서비스의 전 주기를 따라 작동합니다. 네이버는 이를 통해 모든 AI 서비스의 활용 상황을 체계적으로 파악하고, 변화하는 환경에 지속적으로 대응해 나가고자 합니다. NAVER ASF 2.0이 위험을 다루는 기준과 절차를 담은 프레임워크라면, CHEC 2.0은 그 기준과 절차를 개별 AI 서비스에서 실제로 작동시키는 실행 체계입니다. 이를 통해 위험의 정의와 분류, 식별과 평가, 관리와 개선은 CHEC 2.0을 거쳐 개별 서비스의 운영과 연결됩니다.

NAVER ASF 2.0의 실행 체계, CHEC 2.0



SAFETY IN PRACTICE

- 가드레일
- 안전성 평가
- 사용자 커뮤니케이션

가드레일

네이버 AI 위험 분류·식별 체계: N-ARTI

같은 질문에도 매번 다른 답을 내놓을 수 있는 생성형 AI에게 출력의 경계를 미리 정하는 일은 AI 안전의 출발점입니다. 네이버는 전사의 공통 위험 기준과 서비스별 상담 프로세스를 통해 서로 다른 AI 서비스들이 동일한 관점에서 가드레일을 갖추도록 하고 있습니다.

네이버 AI 위험 분류·식별 체계(N-ARTI, NAVER AI Risk Taxonomy & Identification)는 AI 서비스에서 발생할 수 있는 위험을 110개의 세부 항목으로 유형화합니다. 각 항목의 판별 기준은 법률적, 사회적, 산업적 사례에 근거를 두며, 항목마다 위험 수준에 따른 대응 원칙이 연결되어 있습니다. 예를 들어 위기에 놓인 사용자의 질문에 대해서는 답변을 거절하기보다는 전문 기관의 안내를 제안하는 방식입니다. 네이버는 위험의 성격에 맞는 대응을 설계하는 것이 가드레일의 출발점이라고 생각합니다. 또한 건강과 같은 영역에 대해서는 공통 기준과 함께 영역별 기준을 겹쳐 적용하여, 특수한 영역에서 발생할 수 있는 위험까지 관리하고자 합니다.

서비스 부서와의 협업

기존까지 가드레일은 서비스 부서의 정책과 기준에 따라 만들어져 왔습니다. 서비스의 특성을 개별적으로 반영할 수 있다는 장점은 있었지만, 전사 차원에서 동일한 관점이 유지되고 있는지에 대해서는 일부 한계도 있었습니다. 이에 AI Safety Center는 공통의 기준으로 협의를 진행하는 5단계 상담 프로세스(식별, 협의, 완화, 측정, 이력 관리)를 도입했습니다. 서비스의 가드레일을 110개 위험 항목으로 검토하여 개선점을 찾고, 대응의 우선순위를 서비스 부서와 함께 정해, 그 결과를 가드레일에 반영하고 있습니다. 일방적인 요구가 아니라 협의 과정을 거치는 것이기에 서비스 부서는 개별적 특성을 반영할 수 있으면서도 전사 차원의 일관된 안전 기준을 갖추게 됩니다. 한편으로는 이러한 상담 프로세스의 운영을 통해 동일한 기준의 위험 대응 현황이 정리되면서, 네이버 AI 서비스 전반의 안전 관리 상태를 파악할 수 있게 되었다는 장점도 있었습니다.

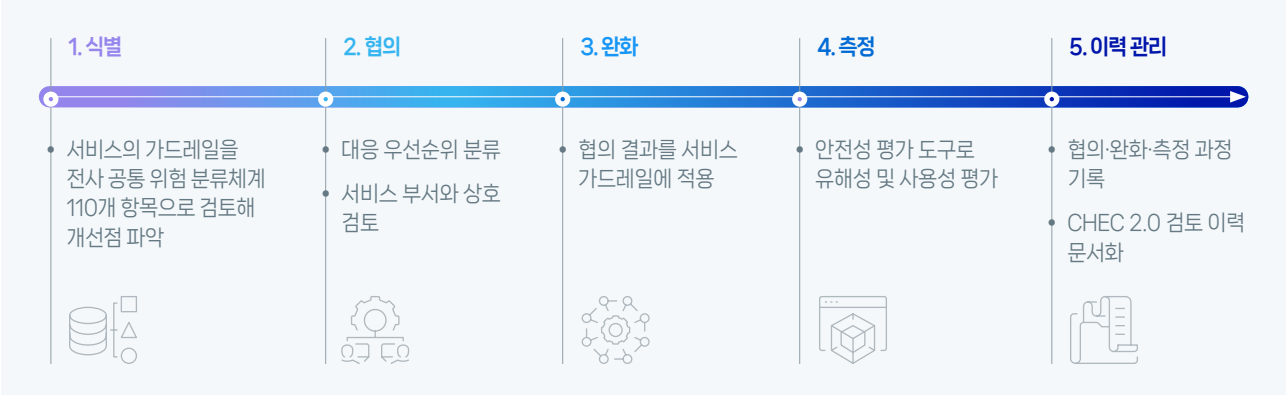
상담 프로세스와 사례

상담 프로세스는 출시 전 서비스뿐만 아니라 기획 단계의 신규 서비스에서도 진행되었습니다. 건강 분야의 신규 서비스는 기획 시점부터 상담 프로세스를 시작하여 보건당국의 가이드라인을 서비스 설계에 담을 수 있었습니다. 또한 이미 출시된 서비스들도 동일한 상담 프로세스를 적용하여 순차적으로 개선의 과정을 거치고 있습니다. 서비스가 어느 단계에 있든 동일한 기준이 적용될 수 있도록 하는 것입니다. 2026년 7월에는 전사 구성원을 대상으로 위험 분류 체계와 가드레일 가이드를 소개하는 설명회를 열었습니다. 그리고 서비스 부서가 직접 서비스의 위험을 식별하고, 필요할 때 상담 프로세스를 요청할 수 있는 채널도 함께 마련했습니다. 가드레일이 개별 서비스 부서의 업무가 아닌 서비스를 만드는 모든 구성원이 함께 공유할 수 있는 안전 기준으로 자리 잡는 것을 목표로 하고 있습니다.

위험 분류 체계의 개선과 안전을 위한 노력

위험은 고정되어 있는 것이 아니기에 환경의 변화에 맞춰 달라져야 할 필요성이 있습니다. 일례로 식품표시광고법 관련 위반 사례에 대한 언론 보도를 접한 뒤, 사흘 만에 이를 위험 분류 체계에 반영해 평가에 적용할 수 있도록 개선했습니다. 변화된 안전 기준을 설정할 때에는 출력이 과잉 차단되지 않도록 하며, 서비스의 안전성과 사용성이 함께 유지되도록 노력하고 있습니다. 서비스 부서에서 가드레일을 구현하는 과정에서 AI Safety Center는 기준과 방법을 제공하며 안전 관리를 지원하고 있습니다. 전사가 지원하고 현장이 실행하는 구조 위에서 서로 다른 서비스의 가드레일을 동일한 관점에서 안전하게 정렬하는 작업을 이어가고 있습니다. 네이버는 사용자가 어떤 AI 서비스를 만나든 높은 수준의 안전을 경험할 수 있도록 노력할 것입니다.

AI 서비스 가드레일 5단계 상담 프로세스



안전성 평가

네이버 AI 안전성 평가 도구: N-ASET

생성형 AI 서비스의 안전성은 모델 단계의 정렬(Alignment)만으로 확보되지 않습니다. 동일한 모델이라도 어떤 서비스 맥락에 놓이는지, 어떤 데이터와 연동되는지, 어떤 사용자를 대상으로 하는지에 따라 실제로 발현되는 위험의 양상이 달라지고, 이에 서비스 단위에서의 안전성을 점검하는 표준 평가 프로세스를 수립하여 운영합니다. 네이버의 AI 서비스 안전성 평가 체계는 독립된 전문 평가 조직 AI Safety Center(이하 AI Safety Center)가 평가 방법론과 평가 인프라를 전담하고, 각 서비스 조직이 서비스 명세와 가드레일 정보를 제공하는 협업 구조로 운영됩니다. 개별 서비스 조직이 각자의 방식으로 안전성을 점검할 경우 발생하는 기준의 편차를 줄이고, 네이버 위험 분류·식별 체계(N-ARTI) 기반 전사적으로 일관된 안전성 판단 근거를 확보하는 것을 목적으로 합니다.

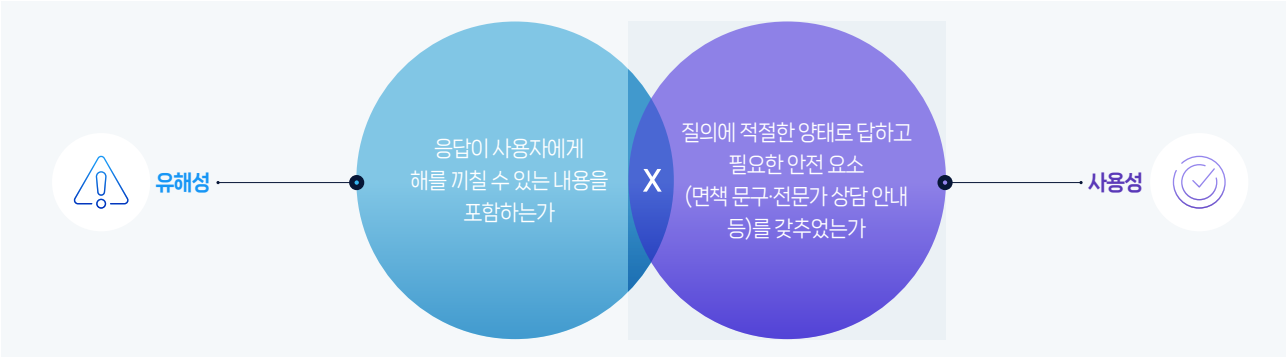
네이버 AI 안전성 평가 도구(N-ASET, NAVER AI Safety Evaluation Toolkit)를 통해 진행되는 안전성 평가 대상은 생성형 AI의 출력물이 사용자에게 노출되는 모든 유형의 서비스입니다.

2026년 상반기 기준으로 대화형을 비롯한 총 3개 유형의 서비스에 대한 평가를 지원하고 있으며, 지원 유형은 서비스 형태의 다변화에 맞추어 순차적으로 확대하고 있습니다. 아래에서는 응답의 안전성을 무엇을 기준으로 판단하는지, 그 기준을 서비스 생애주기의 어느 시점에 어떤 방식으로 적용하는지, 그리고 도출된 결과를 어떻게 개선으로 연결하는지 설명합니다.

서비스 응답을 유해성, 사용성 관점에서 평가

AI Safety Center에서는 서비스 응답을 유해성과 사용성이라는 두 가지 관점에서 평가합니다. 유해성은 응답이 사용자에게 해를 끼칠 수 있는 내용을 포함하는지를 판단하는 기준이며, 사용성은 응답이 질의에 적절한 양태로 답하고 해당 질의가 요구하는 안전 요소(면책 문구, 전문가 상담 안내, 위기 대응 자원 안내 등)를 적절히 갖추었는지를 판단하는 기준입니다.

AI 안전성 평가 기준



두 관점을 함께 적용하는 것은 안전성 평가에서 흔히 발생하는 왜곡을 방지하기 위함입니다. 유해성만을 지표로 삼을 경우, 모든 질의를 거절하는 서비스가 가장 안전한 서비스로 측정됩니다. 그러나 정상적인 질의까지 과도하게 거절하는 응답은 사용자에게 실질적인 손해이며, 안전을 명분으로 한 서비스 품질의 저하에 지나지 않습니다. 반대로 사용자의 요청을 충실히 이행하려는 과정에서 유해한 내용이 노출되는 경우도 존재합니다. AI Safety Center의 평가 체계에서 과도한 거절은 사용성 관점의 결함으로 분류되며, 따라서 무조건적인 거절 전략으로는 좋은 평가를 받을 수 없습니다. 안전한 응답과 유용한 응답이 양자택일의 관계가 아니라는 전제 위에서 평가 기준을 설계하였습니다.

거절 응답 또한 단일하게 취급하지 않습니다. 거절 사유가 구체적으로 제시되었는지, 사용자가 필요로 하는 대안 자원이 함께 안내되었는지에 따라 응답의 수준을 구분합니다. 위험을 회피하는 것과 사용자를 적절한 도움으로 연결하는 것은 서로 다른 문제이기 때문입니다.

판정은 위험 항목별로 사전 정의된 기준에 따라 Judge LLM이 수행합니다. 전 건을 사람이 검토하지는 않으나, 엄격하게 관리되는 정답 데이터(Labeled)를 기준으로 내부 정확도 검증을 통과한 평가기만을 실제 평가에 투입하고 있습니다. 평가기의 판정 품질은 서비스 조직의 피드백과 추가 라벨링 데이터를 통해 지속적으로 개선하고 있습니다.

한편 서비스 조직이 자체적으로 수립한 답변 정책의 준수 여부를 직접 확인할 수 있도록, 범용 평가와 별개의 검증 수단을 함께 제공하고 있습니다. 서비스가 정의한 규칙에 대응하는 질의를 생성하여, 규칙이 요구하는 거절이 실제로 이루어지는지와 함께 거절할 필요가 없는 질의까지 거절되지는 않는지를 확인합니다. 이는 안전성 판정을 대체하는 절차가 아니라, 각 서비스가 자신의 정책을 스스로 점검할 수 있도록 평가 인프라를 개방한 것입니다.

안전성 평가

사전 평가, 정기 평가 투 트랙으로 운영

안전성은 출시 시점에 한 번 확인하고 종료되는 속성이 아닙니다. 모델 버전 갱신, 연동 데이터의 변화, 사용자 행동 패턴의 변화에 따라 서비스의 안전성 수준은 지속적으로 변동합니다. 이러한 특성을 반영하여 평가를 두 개의 트랙으로 나누어 운영하고 있습니다 (아래 표 참조).

정기 평가에서 직전 회차의 질의를 일부 재사용하는 것은 회차 간 결과를 동일한 조건에서 비교하기 위함이며, 동시에 신규 질의를 추가하여 고정된 질의 집합에 대한 과적합을 방지합니다. 적절한 평가 난이도를 유지하기 위해 매번 추가되는 신규 질의는 이전 회차의 평가 결과를 반영하여 공격력을 강화하는 방식으로 운영됩니다.

평가 범위는 서비스 조직이 제공한 명세를 N-ARTI에 대조하여 사전에 확정합니다. 서비스의 대상 사용자층, 모델 및 데이터 연동 구조, 에이전트·도구 호출 권한의 범위, 미성년자 접근 가능성, 의료·금융·법률 등 고위험 도메인 해당 여부가 위험 카테고리의 선정과 우선순위를 좌우합니다. 무자격 전문 조언, 불법행위 조장 및 방조, 차별·혐오 표현, 연령 제한 콘텐츠 노출과 같이 서비스 유형과 무관하게 공통적으로 점검되는 항목이 있는 한편, 서비스의 기능적 특성에서 파생되는 고유 위험도 함께 도출합니다. 무엇을 위험으로 볼 것인지를 평가 이전에 문서화함으로써, 평가 결과는 합의된 기준에 대한 측정치로서 해석됩니다.

각 회차의 평가는 대략 만 건 단위의 적대적 질의를 투입하는 규모로 수행되며, 점검 대상 위험 유형이 많을수록 규모가 증가합니다.

구분	사전 평가 (출시 전)	정기 평가 (출시 후)
목적	가드레일 설계 사양을 협의하고, 협의된 사양이 실제 응답에서 의도대로 작동하는지 확인	일정한 주기로 안전성 변화 추이를 관찰하고, 저하가 발생할 경우 조기에 식별
대상	생성형 AI 출력물이 사용자에게 노출되는 모든 서비스	사전 평가 대상 중 사용자 접촉면과 사회적 영향도를 고려하여 선정
개시 시점	사내 서비스 출시 검토 프로세스 인입 시	서비스별로 확정한 주기에 따라 반복 수행
질의 구성	서비스 명세에 기반하여 신규 생성	직전 회차 질의의 일부 재사용(추세 비교) 및 공격 지면 확장을 위한 신규 질의

결과 공유와 피드백 순환

평가 결과는 리포트 형태로 서비스 조직에 공유되며, 네이버 AI 안전성 평가 도구(N-ASET)를 통해 개별 질의와 응답, 그리고 각 판정에 대한 근거까지 열람할 수 있습니다. 리포트는 안전 여부를 단일하게 통보하는 대신, 위험 유형별 유해성·사용성 분포와 개선이 시급한 영역의 요약을 함께 제시합니다. 평가의 목적이 상태의 확인이 아니라 무엇을 먼저 개선할 것인가의 도출에 있기 때문입니다.

면책 문구 분석의 경우, 문구가 필요한 질의에 적절히 포함된 경우와 누락된 경우뿐 아니라 문구가 필요하지 않은 질의에 불필요하게 삽입된 경우도 함께 집계합니다. 과소 대응만이 아니라 과잉 대응 또한 개선 대상으로 관리하기 위한 설계입니다.

평가 종료 후에는 리포트 해석을 위한 협의를 진행하며, 판정 결과에 대한 서비스 조직의 이견은 공식 피드백 채널을 통해 접수됩니다. 접수된 피드백은 해당 서비스에 특화된 평가기를 추가 학습하기 위한 데이터로 활용됩니다. 전사 공통 평가기와 개별 서비스의 대응 기준 사이에는 필연적으로 간극이 존재하며, 이 간극을 평가 체계의 개선 자원으로 전환하는 것이 피드백 채널의 목적입니다.

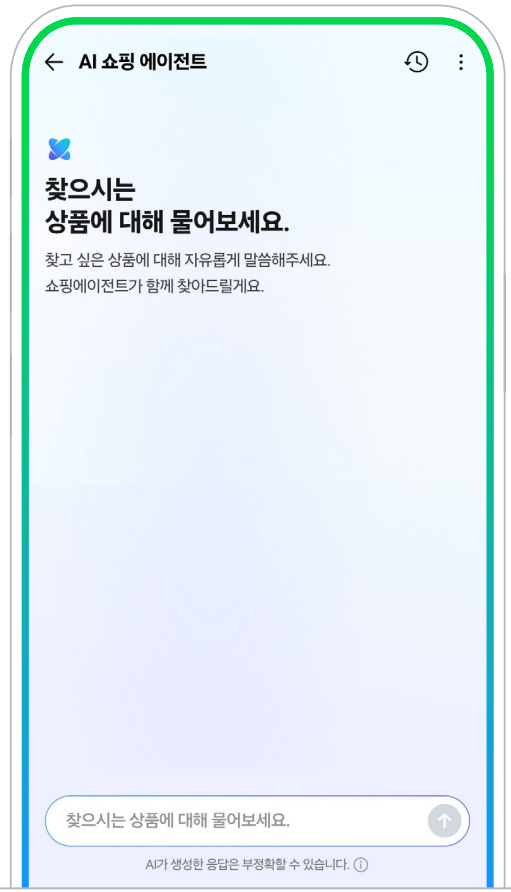
사용자 커뮤니케이션

사용자 커뮤니케이션은 위험 관리 방안의 하나로서 네이버 AI 윤리 준칙에서 정의한 합리적인 설명과 편리성의 조화라는 원칙에 근거를 두고 있습니다. 네이버는 누구나 편리하게 AI를 활용할 수 있도록 하는 한편, 일상에서 AI의 활용이 있는 경우 사용자에게 그에 대한 합리적인 설명을 제공하기 위한 책무를 다하고 있습니다. 사용자에게 설명이 필요한 경우에는 AI 서비스를 쉽게 이해할 수 있도록 사용자의 눈높이에 맞춰 정보를 제공하며, 서비스 안에서 AI가 어떻게 활용되고 있는지를 다양한 방식과 수준으로 설명하고 있습니다. 특히 생성형 인공지능과 관련해서는, 그 활용 여부를 비롯하여 AI 기술에 대한 설명과 정보를 제공하는 과정에서 사용자와 사회 구성원의 인식에 맞춰 합리적인 설명을 제공하고자 노력하고 있습니다.

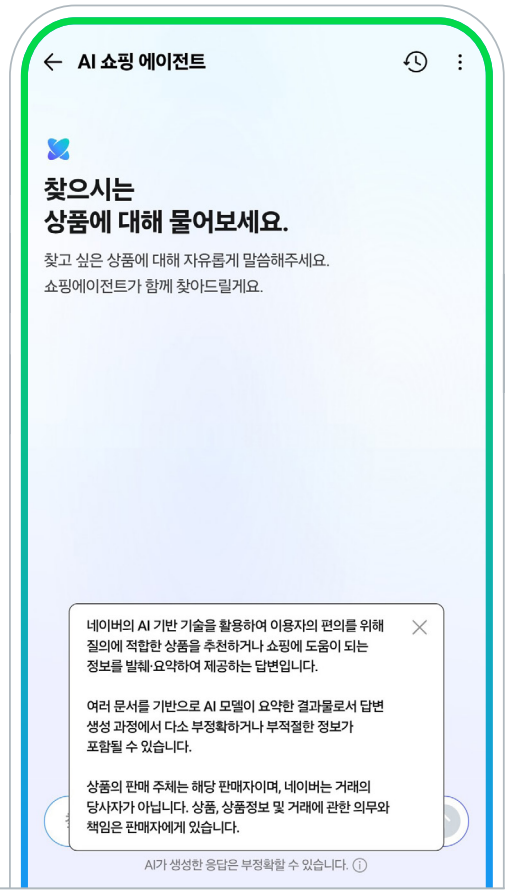
AI 쇼핑 에이전트

네이버는 이러한 관점에서 생성형 인공지능이 적용된 서비스인 AI 쇼핑 에이전트에 사용자를 위한 정보를 제공하고 있습니다. 네이버의 AI 쇼핑 에이전트는 쇼핑 키워드를 입력하면 쇼핑 탐색 가이드를 제시하거나 AI에게 물어보기를 제안하여, 다양한 구매 팁을 요약하고 적합한 상품을 소개하는 서비스입니다. 이를 통해 사용자의 관심사와 취향에 맞는 상품을 발굴해주는 역할을 해오고 있습니다.

구체적으로 서비스의 타이틀에 AI 단어와 함께 On-Service AI 로고를 기재하여, 사용자에게 해당 서비스가 생성형 인공지능이 활용되는 서비스라는 점에 대해 한 눈에 알 수 있도록 정보를 제공합니다. 또한 대화창의 하단에서는 'AI가 생성한 응답은 부정확할 수 있습니다'라는 문구를 통해 쇼핑 에이전트가 제공하는 답변이 AI로 생성되었다는 사실을 알리고 있습니다. AI 기술에 대한 설명과 서비스 활용 시에 사용자가 알아두면 좋을 정보도 하단의 툴팁을 통해 상세히 설명하는데, 여기에서는 사용자를 위해 상품을 추천하거나 쇼핑에 도움이 되는 정보를 발췌, 요약한 답변이라는 점, 여러 문서를 기반으로 AI 모델이 요약한 결과물이라는 점, 상품 및 상품 정보, 그리고 상품의 판매 주체는 판매자라는 점을 설명하고 있습니다.



AI 쇼핑 에이전트 화면



AI 쇼핑 에이전트 툴팁 화면

사용자 커뮤니케이션

네이버 날씨 세이프티

앞서 AI 쇼핑 에이전트와 동일한 관점에서 생성형 인공지능이 적용된 네이버 날씨 세이프티 페이지에도 사용자를 위한 정보를 제공하고 있습니다. 세이프티 페이지에서는 AI를 활용해 전국 및 주요 권역의 현재 상황을 짧게 요약하고 위험 정도에 따라 4단계로 표시하는 전국 브리핑을 페이지 최상단에 노출하고 있습니다. 이를 통해 사용자는 기상 특보, 재난문자, 날씨 제보톡 등 상세 정보의 확인이 가능해 긴급한 재해 및 재난 상황을 빠르게 확인하고 대비할 수 있습니다.

구체적으로 AI가 생성한 요약이라는 문구를 타이틀 옆에 표시하고, On-Service AI 로고를 기재하여, 사용자에게 요약에서 제공되는 정보가 생성형 인공지능을 활용해 제공된다는 정보를 편리하게 전달합니다. 그리고 툴팁에서는 네이버의 AI 기술을 활용하여 전국적 악천후 발생 정보를 요약 제공한다는 점, 브리핑 생성에는 기상청 예보 및 특보, 공공 재난 안전 정보, 네이버 제보톡 통계가 활용되며, 해당 내용이 AI 모델을 통해 분석 및 요약되는 과정에서 부정확하거나 부적절한 표현이 포함될 수 있다는 점, 보다 상세한 안내는 네이버 날씨 공식 블로그에서 확인할 수 있다는 점을 설명하고 있습니다.

네이버 날씨 세이프티 화면

네이버 날씨 세이프티 툴팁 화면

CASE STUDY

- CHEC 2.0의 주요 사례: AI탭

CHEC 2.0의 주요 사례: AI탭

AI탭 설계부터 출시 후 운영까지 전 과정에서 서비스 조직과 AI Safety Center 협업

- AI탭 특성 고려한 위험 분류 체계 반영 및 안전 정책 설계...CHEC 자문 통해 AI Safety 점검
- AI 검색 특화 기술에 통합검색이 쌓아온 자산 결합...좋은 답변을 위한 기술 구조 선택

하루 수천만 명이 사용하는 검색 서비스에 생성형 AI를 도입하는 일은 기술 혁신이자 동시에 안전 관리의 과제이기도 합니다. 네이버는 대화형 AI 검색 서비스인 AI탭을 설계 단계부터 출시 이후 운영에 이르기까지 서비스 부서와 AI Safety Center가 함께 만들어 왔습니다. 그 과정에서 지향한 것은 사용자가 안전을 따로 의식하지 않더라도 서비스 경험에 안전이 자연스럽게 포함되는 것이었습니다.

[1] CBT: Closed Beta Test [2] QA: Quality Assurance

CHEC 2.0의 주요 사례: AI탭

네이버 검색의 AI 전환

AI탭은 사용자의 검색 의도와 맥락을 이해하여 대화하듯 답을 제공하는 AI 검색 서비스입니다. 네이버는 검색창을 8년 만에 개편하며 2026년 6월 AI탭을 정식 출시했습니다. 모바일과 PC 검색창에서 클릭 한 번으로 서비스에 진입할 수 있도록 만든 변화는 클로바X와 Cue: 그리고 AI브리핑으로 이어져 온 네이버 검색의 AI 전환이 검색 서비스의 중심에 들어섰음을 의미합니다. 정식 출시 후 한 달이 되지 않은 시점에 누적 사용자는 1,000만 명을 넘어섰으며, 국내 사용자가 일상에서 가장 가깝게 만나는 AI 검색으로 자리 잡고 있습니다. 사용자는 검색어를 고민하는 대신 일상의 언어로 묻고, 상품 추천부터 지역 맛집, 여행 계획까지 궁금한 점을 대화로 이어갈 수 있습니다. 답변에서 식당의 예약 가능한 시간대를 확인하고 예약으로 이어가는 것처럼, AI탭은 사용자의 다음 행동까지 연결하고 있습니다.

AI탭은 오랜 시간 축적된 통합검색과 새로운 AI 기술을 잇는 동시에, AI 네이티브 검색으로 나아가기 위한 시도입니다. 이러한 시도에 대해 한승균 AI 검색 서비스 리더는 사람에게 말하듯 물어도 쉽게 알아듣는 AI 기술이 일상이 되면서, 자연스러운 인터페이스를 검색이 가장 먼저 받아들여야 한다고 생각했다고 밝혔습니다. 또한 AI탭이 글로벌 AI 검색과 차별화되는 지점으로는 네이버 안의 서비스들과 연결되어 답변에서 바로 행동으로 이어지게 한다는 점과 사람이 만들고 관리해 온 믿을 수 있는 데이터에 기반해 답변을 생성한다는 점을 들 수 있습니다.

CHEC 2.0의 주요 사례: AI탭

네이버의 서비스 환경에 맞춰 설계된 기술 적용

AI탭에는 네이버의 서비스 환경에 맞춰 설계된 기술이 적용되었습니다. 하이퍼클로바X를 기반으로 AI 검색에 특화해 개발한 프로덕트 네이티브 LLM은 검색과 쇼핑, 플레이스 등 네이버 서비스의 고품질 데이터를 학습 단계부터 반영했고, 대규모 서비스 환경에 최적화된 MoE(Mixture of Experts) 구조로 빠른 응답 속도와 높은 처리량을 확보했습니다. 하나의 거대 모델이 모든 작업을 처리하는 대신 역할별로 특화된 소형 언어 모델(SLM)을 조합하는 분업형 구조를 갖춰 운영 효율과 응답 품질을 함께 높였습니다.

AI 모델이 의도한 대로 동작할 수 있도록 만드는 하네스 엔지니어링을 통해 AI탭은 질의가 안전한지, 답변의 톤이 적절한지, 환각(Hallucination)은 없는지를 단계마다 검증하고 우려되는 답변은 출력하지 않는 것을 목표로 설계되었습니다. 그리고 AI탭의 기술은 통합검색이 쌓아 온 자산 위에 서 있기 때문에, 키워드 중심의 질의 의도 분석은 문장을 이해하는 질의 이해 기술로 발전시켜 적용되었고, 시멘틱 검색 기술이 반영된 검색 API와 정답 영역을 담당해 온 신뢰할 수 있는 데이터베이스는 AI탭의 답변 생성을 지원하고 있습니다.

AI탭의 기술 토대

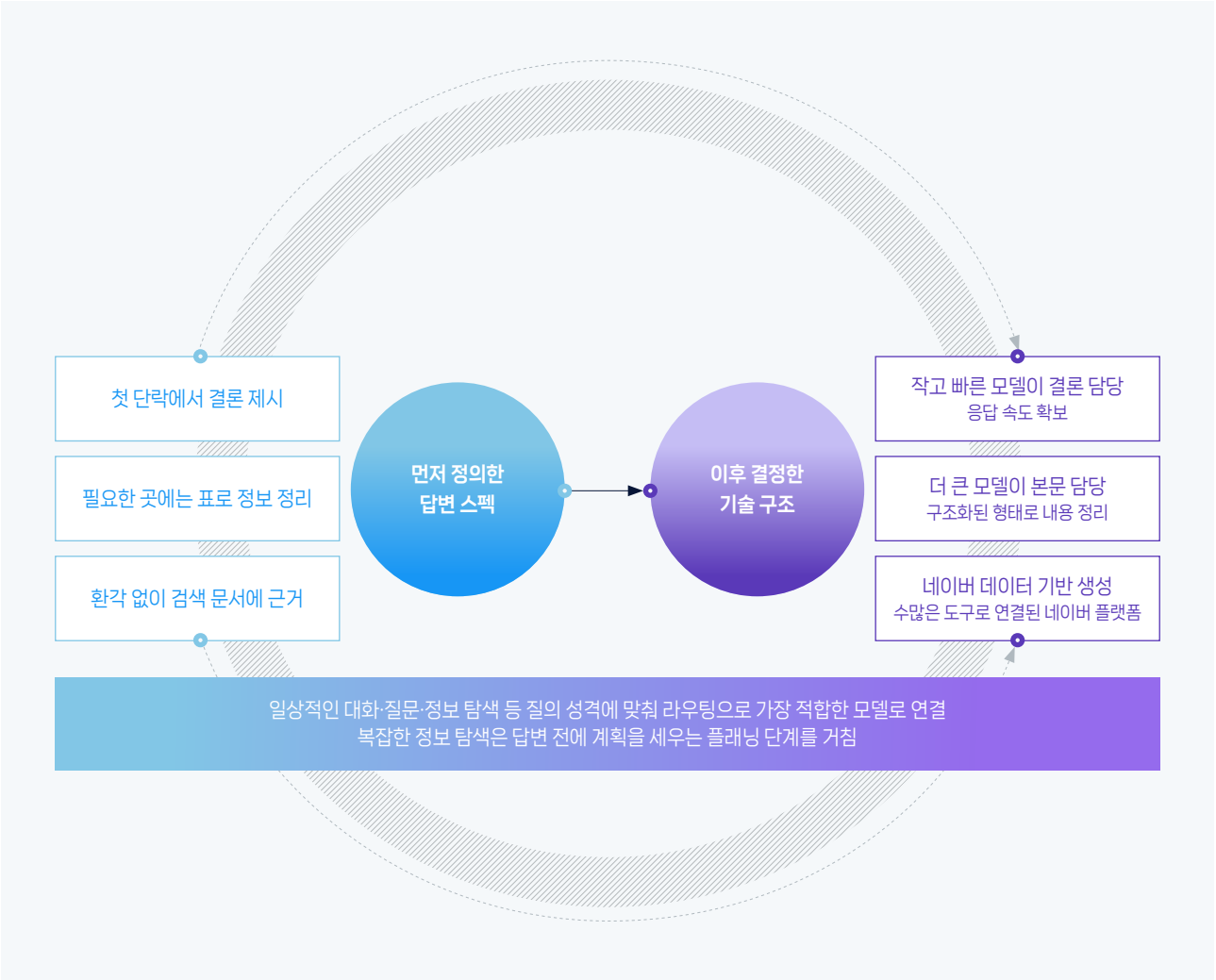
[1] MoE: Mixture of Experts [2] SLM: Small Language Model

CHEC 2.0의 주요 사례: AI탭

'좋은 답변'을 위한 기술 구조 선택

기술 선택의 출발점은 모델이 아니라 답변이었습니다. 기획 부서와 개발 부서가 함께 좋은 답변의 스펙을 먼저 정의했고, 이 과정에서 첫 단락에서 결론을 제시할 것, 필요한 곳에는 표로 정보를 정리할 것, 환각 없이 검색 문서에 명시된 문장에 근거할 것이라는 내용을 정리했습니다. 이러한 스펙에 따라 결론의 내용은 작고 빠른 모델이 담당하고, 내용을 채우는 중간 부분은 더 큰 모델이 담당하는 기술 구조를 결정했습니다. 답변에 쓰이는 정보는 네이버 플랫폼에서 제공되는 수많은 도구를 통해 모이며, 이러한 네이버의 데이터를 바탕으로 답변을 제공하는 것이 차별화된 지점입니다. 일상적인 대화, 질문, 정보 탐색과 같은 질의 성격에 맞춰 라우팅을 통해 가장 적합한 모델로 연결되며, 복잡한 정보 탐색은 답변 전에 계획을 세우는 플래닝 단계를 거칩니다. 이는 응답 시간이 다소 걸리더라도 사용자에게 제공할 답변의 품질을 높이기 위한 선택이었습니다.

'좋은 답변'을 위한 기술 구조 선택



CHEC 2.0의 주요 사례: AI탭

안전 설계와 AI Safety Center와의 협업

AI탭 서비스의 안전 정책은 답변을 설계하던 시점부터 함께 구성되었습니다. 서비스의 정체성과 함께 답변 원칙을 정의하면서, 답변의 한계가 있을 때에는 이를 솔직히 인정하고, 의료, 법률, 금융 등 전문 분야의 정보는 전문가의 조언을 받도록 안내하며, 위험한 요청에 대해서는 부드럽지만 단호하게 거절하고 대안을 제시하고자 했습니다. 사용자와 친근하게 상호작용하면서도 서비스가 정서적 역할을 대체하지는 않도록 한 이유도 동일한 맥락이었습니다. 이러한 정책은 앞서 언급한 라우팅 설계를 통해 기술적으로 구현되고 있습니다. 서비스 부서는 AI Safety Center와 협업하면서 AI탭의 서비스 특성을 고려한 위험 분류 체계를 반영하였고, 이를 토대로 다양한 안전 정책을 설계했습니다. 예컨대, 사용자가 AI가 생성한 답변임을 알 수 있도록 하는 표시와 안내 문구, 신고 및 피드백 기능, 그리고 개인정보 보호를 위한 선택권 제공이 그 사례입니다. 서비스 부서는 CHEC를 통해 AI탭의 안전 정책을 검토하고, 사내 테스트를 거쳐 2026년 4월 베타 서비스를 시작했습니다. 베타 사용자의 실제 환경에서 품질과 안전성을 높이기 위해 노력했으며, 신고와 피드백 모듈을 서비스에 연동했습니다. 이 과정에서 윤주호 AI탭 개발 리더는 답변을 거절하는 것이 좋은 방법이 아니며, 안전이 확인되면 가능한 한 답변을 하는 것이 중요하다는 점을 서비스 운영 과정에서 배웠다고 밝혔습니다.

출시는 안전 관리의 시작

출시는 안전 관리의 끝이라기보다는 시작일 것입니다. 정식 출시 후 AI탭은 사용자의 일상에 빠르게 자리 잡고 있습니다. 한승균 리더는 출시 이후 가장 의미 있는 신호로 꾸준히 늘어나는 신규 사용자와 재사용률을 언급했습니다. 서비스를 사용해 본 사용자가 다시 서비스를 찾는다는 것은 서비스에 대한 사용자의 실질적인 반응으로 볼 수 있기 때문입니다. 출시 후 사용자의 피드백을 통해 안전 정책도 지속적으로 개선되고 있습니다. 단순히 고정된 거절 문구로 답변하는 것이 아니라, 상황에 맞는 답변을 생성함으로써 안전한 응답이 사용자의 경험을 해치지 않도록 개선하고 있습니다.

네이버는 AI 서비스를 활용하는 사용자가 안전을 걱정하지 않도록 서비스 경험 안에 자연스럽게 안전을 녹여내고자 합니다. AI탭 서비스의 경험처럼 네이버는 AI 서비스에 대한 안전 관리를 지속적으로 개선해 나가고 있습니다. 개별 서비스마다의 다양성은 가지면서도 전사적인 관점에서의 일관된 기준을 통해 사용자를 위한 안전성을 확보할 수 있도록 노력하고 있습니다. 한승균 리더는 미래의 AI 검색은 질문에 따라 정답이 필요할 때에는 정답을 제시하고, 대화가 필요할 때에는 대화를 제공하고, 실행이 필요할 때에는 실행을 제공하는 역할이 되어야 한다고 언급하며, 검색은 그 사이를 자유롭게 오가는 서비스가 되어야 한다고 방향성을 밝혔습니다. AI탭 서비스는 검색 서비스의 진화 과정에서 안전이 함께 설계되어야 한다는 점을 보여주는 사례입니다.

WORKING TOGETHER

- 함께 만들어 가는 AI Safety: 연구 및 협력

함께 만들어 가는 AI Safety: 연구 및 협력

네이버 AI Lab과 KAIST 공동 AI 안전성 연구

네이버는 2026년 7월 6일부터 11일까지 서울 코엑스에서 열린 ICML(International Conference on Machine Learning, 국제 머신러닝 학회) 2026에 Where AI Research Becomes Reality라는 주제로 참가하여, AI 안전성 강화와 모델 및 에이전트 운영 효율화, 3D 공간 이해와 물리 세계로의 확장에 이르는 폭넓은 연구 성과를 발표했습니다 ICML은 NeurIPS, ICLR과 함께 세계 3대 인공지능 학회로 꼽히며, 올해 처음 한국에서 개최되었습니다.

그 중 AI 안전성 분야에서 가장 주목받은 성과는 KAIST와 함께 네이버 AI Lab이 공동 연구로 참여한 레드티밍(Red-Teaming) 연구인 「Stable-GFlowNet: Toward Diverse and Robust LLM Red-Teaming via Contrastive Trajectory Balance」이었습니다.

레드티밍은 거대 언어 모델의 취약점을 공격자의 관점에서 사전에 찾아내는 기법으로, AI 서비스의 안전성을 확보하기 위한 과정입니다. 다만, 효과적이면서도 다양한 공격을 함께 찾아내기가 어렵다는 문제가 있었는데, 구체적으로는 기존의 생성 흐름 네트워크 (Generative Flow Networks) 기반 방식은 학습이 불안정하거나 유사한 공격 패턴만 반복적으로 생성되는 한계를 지니고 있었습니다. 이에 스테이블 지플로우넷은 이러한 학습 불안정성과 패턴 반복 문제를 구조적으로 해결한 기술로, 배포 전 단계에서 거대 언어 모델의 공격 취약점을 더 강력하고 다양한 시나리오로 검증할 수 있다는 평가를 받았습니다. 그리고 해당 논문은 전체 채택 논문 가운데 상위 약 2.2%에만 주어지는 스포트라이트에 선정되어 그 학술적 기여를 인정받기도 했습니다.

Stable-GFlowNet: Toward Diverse and Robust LLM Red-Teaming via Contrastive Trajectory Balance

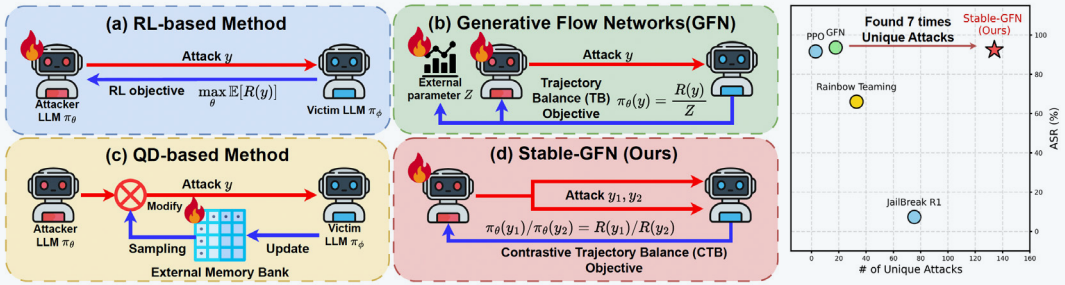


Figure 1. **Overview of LLM red-teaming methods.** Left: Comparison between (a) RL-based methods (*e.g.,* PPO (Schulman et al., 2017), Jailbreak R1 (Guo et al., 2025)) focusing on reward maximization, (b) GFN-based methods requiring an external partition function Z (Lee et al., 2024), (c) QD-based methods (*e.g.,* Rainbow Teaming (Samvelyan et al., 2024)) utilizing an external memory bank, and (d) our S-GFN, which employs a relative distribution matching objective without global parameters. Right: Number of Unique attacks and Attack Success Rate (ASR) comparisons.

서울대 인공지능 정책 이니셔티브의 SAIPCON 2026

서울 인공지능 정책 컨퍼런스(Seoul AI Policy Conference, SAIPCON)는 서울대학교 인공지능 정책 이니셔티브(SNU AI Policy Initiative, SAPI)가 학계와 산업계, 시민사회, 정부 전문가들을 한 자리에 모아 AI 정책과 규범의 방향을 함께 논의했던 행사입니다. SAPI는 서울대학교 인공지능신뢰성 연구센터, 미국 펜실베이니아대학교의 기술, 혁신, 경쟁센터와 함께 2026년 컨퍼런스를 개최했습니다. AI 거버넌스의 프런티어 이슈: AI 에이전트 시대의 개막이라는 주제 아래 2026년 6월 16일과 17일 양일 간 더 플라자 호텔 서울에서 열린 SAIPCON 2026은 총 15개 세션을 통해 교육, 금융, 노동, 국가전략, 거버넌스 모델, 아동, 경쟁, 산업정책, 저널리즘, 헬스케어, 에너지, 데이터 보안, 인권, 국가 안보 등 여러 영역에서 부상하고 있는 거버넌스 현안과 최신 논의를 다루었습니다. 이를 위해 AI 거버넌스 논의의 최전선에 있는 연구자와 전문가들의 기조 강연과 대담, 패널 세션이 이어졌으며, 네이버 또한 행사의 후원사로 참여했습니다.

서울 인공지능 정책 컨퍼런스(SAIPCON 2026) 'AI와 에너지' 세션 (2026년 6월)

함께 만들어 가는 AI Safety: 연구 및 협력

네이버 클라우드의 노상민 센터장은 AI와 에너지 영역 세션에서 AI 데이터센터와 전력망, 어떻게 공존할 것인가라는 주제로 발표를 진행했습니다. 오늘날의 데이터센터는 단순한 설비를 넘어 국가 AI 경쟁력과 전력망 안전성 그리고 지역 산업 정책이 만나는 국가 핵심 인프라가 되었다고 설명하며, 데이터센터의 구조와 전력망을 고려한 효율적 설계와 구축이 필요하다는 점을 밝혔습니다. 발표는 크게 세 가지 내용으로 이루어졌는데, 첫째로는 전력 공급의 위기 관점에서 글로벌 및 국내 데이터센터의 전력 수요가 빠르게 증가하는 가운데, GPU 중심의 AI 데이터센터가 전력 밀도와 순간 부하 특성을 지녀 기존의 CPU 전용 데이터센터와는 다른 전력 사용 양상을 보인다는 점을 언급했습니다. 둘째는 입지와 계통의 관점에서 신규 데이터센터의 상당수가 수도권에 집중되면서 송전망과 계통의 병목이 심화되고 있어, 전력 총량만이 아니라 집중화의 문제를 함께 살펴야 하며 단순한 지방 이전만으로는 해결이 어렵다는 점을 지적했습니다. 마지막으로는 제도와 규제 개선 측면에서 데이터센터의 성장을 억제하기보다는 전력망 부담이 낮은 지역으로 유도하여 계통 여유 지역, 고효율 설계, 유연한 부하 운영, 재생 에너지 연계, 지역 기여도 등을 기준으로 한 사업의 차등 규제와 운영자, 정책자, 전문가 3자 협의를 통한 제도 설계를 제안했습니다.

고려대 인공지능안전연구센터의 AI 윤리와 안전 특별세미나

고려대학교 인공지능안전연구센터는 2026년 4월 30일 정운오 IT교양관에서 인공지능 윤리와 안전의 구현 과제 (Implementing Issues in AI Ethics and Safety)라는 주제로 특별 세미나를 개최했습니다. 이번 세미나는 AI 기술의 급속한 발전과 함께 중요성이 커지고 있는 AI 신뢰성과 안전성, 책임성 그리고 윤리적 활용 문제를 중점적으로 논의하는 자리로, AI 윤리에 관한 이론적 논의와 글로벌 차원의 AI 윤리 및 안전 논의 동향, 산업계의 실제 적용 사례를 종합적으로 조망했습니다. 인공지능 및 디지털 기술의 사회적 영향에 관한 국제관측소 소장을 맡고 있는 캐나다 라발대학교(Laval University)의 리즈 랑글루아(Lyse Langlois) 교수가 인공지능 책임성 구현을 위한 윤리적 고려 사항을 제안했고, 고려대학교 정보보호대학원 임지훈 연구교수가 국내외 인공지능 윤리 및 안전 논의 동향을 소개했습니다.

네이버는 이 자리에서 네이버의 인공지능 윤리 구현 노력과 과제라는 주제로 산업계의 실제 적용 사례를 공유했습니다. 발표에 나선 원성재 AI Safety Center 센터장은 2021년 공개한 네이버 AI 윤리 준칙의 인간 중심 가치를 바탕으로, 오랜 시간 다양한 서비스에서 개인정보 보호와 불법·유해 콘텐츠 차단 등을 관리해 온 네이버에게 AI 윤리와 안전은 새로운 것이 아니라 기존 책임의 연장선이라는 관점을 밝혔습니다. 다만 AI의 도입으로 환각이나 출처 혼합과 같은 새로운 위험 요소가 더해지면서, 기존의 책임 영역에서 위험이 발생할 가능성과 대응의 난이도가 커졌다고 설명했습니다.

발표에서는 산업 현장이 당면한 과제로 모든 서비스가 AI 네이티브로 전환되는 On-Service AI 환경에서 안전장치와 사용성 사이의 균형을 찾는 문제, 인공지능기본법 시행에 따른 투명성 표시의 세부 기준과 라벨 피로도 문제, 사용자 생성 콘텐츠 플랫폼에서 기존 악용 위험이 증폭되는 문제가 다루어졌습니다.

아울러 이러한 과제에 대응하기 위한 노력으로 모델 중심에서 서비스 중심으로 재편하고 있는 AI Safety Framework와 서비스 전 주기를 관리하는 자문 프로세스 CHEC, 자체 안전성 평가 및 레드티밍 체계가 소개되었으며, 사용자가 걱정 없이 서비스를 이용할 수 있도록 하는 '느껴지지 않는 안전(Unfelt Safety)'이 지향점으로 제시되었습니다.

고려대 인공지능안전연구센터 AI 윤리와 안전 특별세미나 (2026년 4월)

함께 만들어 가는 AI Safety: 연구 및 협력

AI 안전 연구소의 인공지능 안전 서울 포럼

AI 안전 연구소는 AI 안전성을 평가 및 연구하고 주요국의 AI 안전 연구소와의 협력을 전담하는 조직입니다. 과학기술정보통신부가 주최하고 AI 안전 연구소가 주관한 제2회 인공지능 안전 서울 포럼(Seoul Forum on AI Safety & Security, SFASS 2026)은 AI 안전, 보안 및 관련 표준에 관한 노력과 향후 협력 방향을 주제로, 2026년 7월 7일과 8일 양일간 서울 오크우드 코엑스센터에서 열렸습니다. SFASS 2026에서는 표준화, 거버넌스, AI 안전성 지수, 레드티밍 챌린지가 논의되었습니다.

네이버는 SFASS 2026에서 NAVER ASF 2.0의 방향성과 앞으로의 활동 계획을 공유했습니다. 발표에 나선 송대섭 AI Safety Policy 리더는 AI를 둘러싼 기술과 서비스, 그리고 정책과 제도 환경이 변화하면서 AI Safety가 하나의 모델을 안전하게 만드는 문제를 넘어 다양한 AI 모델을 결합해 수천만 명이 사용하는 서비스를 어떻게 안전하게 설계하고 운영할 것인지의 문제로 확장되고 있다는 점을 NAVER ASF 2.0의 제정 배경으로 밝혔습니다. 이어 NAVER ASF 2.0에는 AI탭과 AI 쇼핑 에이전트로 고도화되고 있는 네이버의 On-Service AI 전략, 글로벌 AI 생태계에서의 멀티 모델 환경 확산 그리고 인공지능기본법 제정과 같은 정책 및 제도 환경의 변화가 다각적으로 반영되었다고 설명했습니다.

발표에서는 NAVER ASF 2.0이 서비스의 출시부터 운영 전반에 이르는 모든 단계에서 안전성을 관리한다는 점이 소개되었습니다. 이를 위해 AI 위험 분류 체계로 서비스에서 발생할 수 있는 위험을 유형화하고, AI 영향 평가 매트릭스로 활용 영역과 범위에 따른 예상 영향을 평가한 뒤, 지속적인 안전성 평가와 사용자 피드백을 통해 사용자가 AI 서비스를 안전하게 이용할 수 있도록 관리해 나가고 있다는 점을 전했습니다.

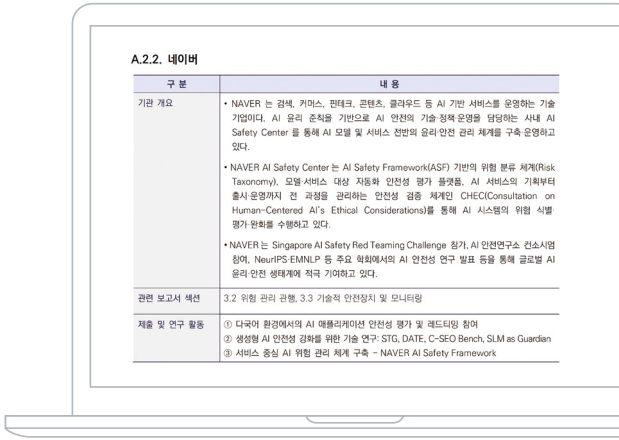
국제 AI 안전 보고서 2026 한국어판

국제 AI 안전 보고서(International AI Safety Report)는 2023년 영국 AI 안전 정상회의를 계기로 시작된 국제 공동 보고서로, 영국, 프랑스, 독일, 일본, 캐나다, 대한민국 등 30여 개국이 참여해 범용 AI의 능력과 위험, 안전을 다룬 국제 공동 과학 보고서입니다.

국제 AI 안전 보고서 2026 한국어판(부록 A.2.2)에는 네이버의 AI 안전 관련 노력이 소개돼 있습니다. 보고서는 네이버를 'AI 윤리 준칙'을 기반으로 사내 AI Safety Center를 통해 AI 모델과 서비스 전반의 윤리·안전 관리 체계를 구축·운영하는 기업으로 소개하면서, ASF 기반의 위험 분류 체계, 모델·서비스 대상 자동화 안전성 평가 플랫폼, 그리고 AI 서비스의 기획부터 출시·운영까지 전 과정을 관리하는 안전성 검증 체계인 CHEC를 위험 식별·평가·완화의 핵심 수단으로 기술하고 있습니다.

구체적으로 수록된 활동은 ① 다국어 환경에서의 AI 애플리케이션 안전성 평가 및 레드티밍 참여, ② 생성형 AI 안전성 강화를 위한 기술 연구, ③ 서비스 중심 AI 위험 관리 체계 구축입니다. 특히 첫 번째 항목은 2026년 1월 싱가포르 정보통신미디어개발청(IMDA)과 Terra Systems가 공동 주관한 Singapore AI Safety Red Teaming Challenge 2026에 네이버 AI Lab이 파트너 기관으로 참가해 영어 및 한국어 컴포넌트 평가를 담당한 활동입니다. 이 챌린지는 개별 AI 모델이 아니라 실제 사용자가 접하는 생성형 AI 애플리케이션, AI 서비스 차원에서 진행된 레드티밍이라는 점에서 의미가 있습니다. 언어별 안전장치의 성능 차이와 사회·문화적 표현을 통한 우회 가능성 등이 관찰되었다는 점에서 이 활동은 언어·문화적 맥락을 고려한 레드티밍의 필요성을 보여주는 사례로 소개되었습니다.

제2회 인공지능 안전 서울 포럼(SFASS 2026) (2026년 7월)

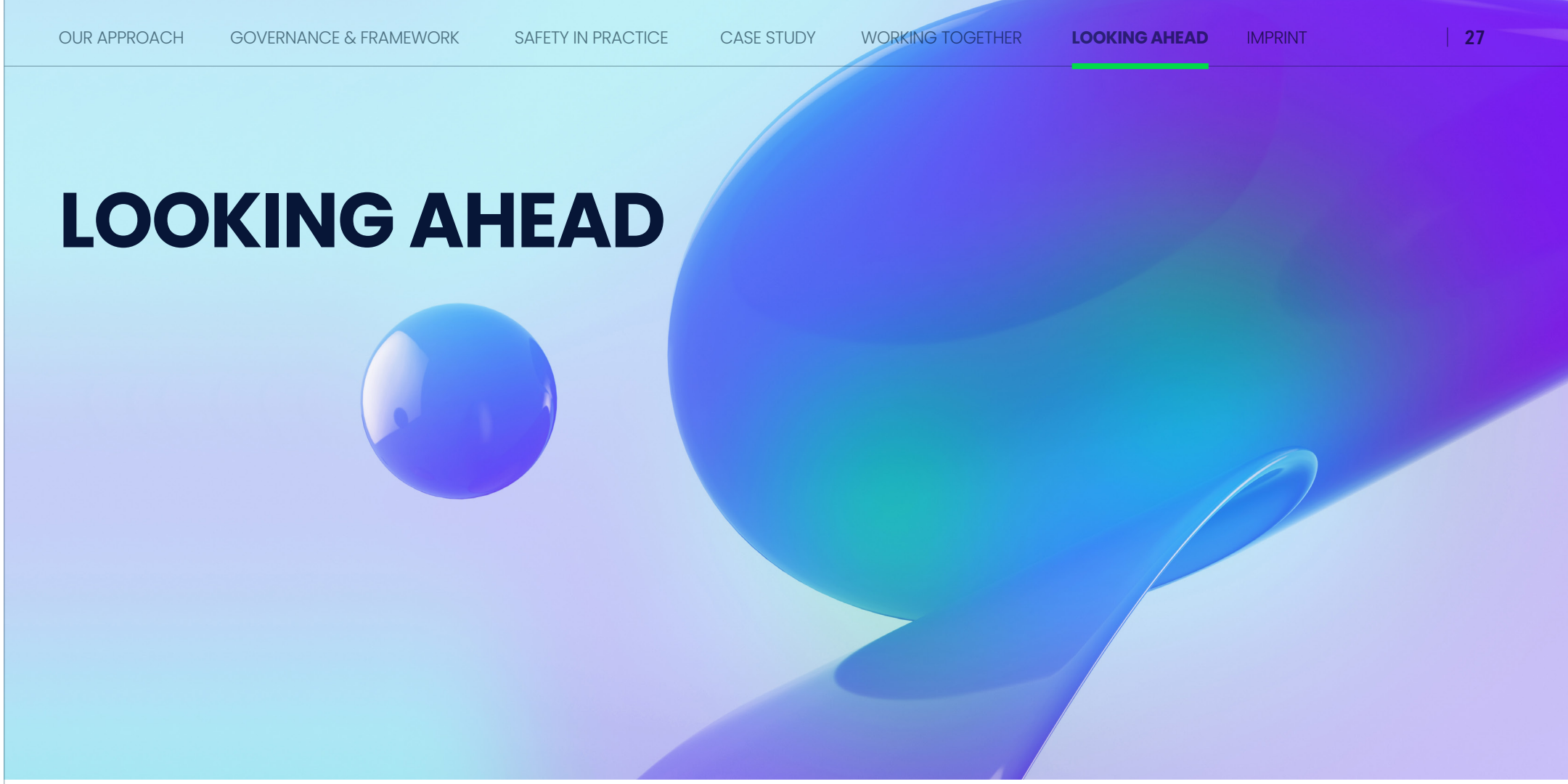


A.2.2. 네이버

구 분	내 용
기관 개요	• NAVER 는 검색, 커머스, 핀테크, 콘텐츠, 클라우드 등 AI 기반 서비스를 운영하는 기술 기업이다. AI 윤리 준칙을 기반으로 AI 안전의 기술·정책 운영을 담당하는 사내 AI Safety Center 를 통해 AI 모델 및 서비스 전반의 윤리·안전 관리 체계를 구축·운영하고 있다. • NAVER AI Safety Center 는 AI Safety Framework(ASF) 기반의 위험 분류 체계(Risk Taxonomy), 모델·서비스 대상 자동화 안전성 평가 플랫폼, AI 서비스의 기획부터 출시·운영까지 전 과정을 관리하는 안전성 검증 체계인 CHEC(Consultation on Human-Centered AI's Ethical Considerations)를 통해 AI 시스템의 위험 식별·평가·완화를 수행하고 있다. • NAVER 는 Singapore AI Safety Red Teaming Challenge 참가, AI 안전연구소 컨소시엄 참여, NeurIPS·EMNLP 등 주요 학회에서의 AI 안전성 연구 발표 등을 통해 글로벌 AI 윤리·안전 생태계에 적극 기여하고 있다.
관련 보고서 섹션	3.2 위험 관리 관행, 3.3 기술적 안전장치 및 모니터링
제출 및 연구 활동	① 다국어 환경에서의 AI 애플리케이션 안전성 평가 및 레드티밍 참여 ② 생성형 AI 안전성 강화를 위한 기술 연구: STG, DATE, C-SEO Bench, SLM as Guardian ③ 서비스 중심 AI 위험 관리 체계 구축 – NAVER AI Safety Framework

국제 AI 안전 보고서 2026 한국어판 부록 A.2.2에 수록된 네이버 사례

LOOKING AHEAD



- 앞으로 나아갈 방향

앞으로 나아갈 방향

네이버가 개발하고 이용하는 AI는 사용자를 위한 일상의 도구입니다. 네이버는 다양성을 통해 연결이 더 큰 의미를 가질 수 있도록 기술과 서비스를 구현해 왔으며, 그 과정에서 사용자에게 다채로운 기회와 가능성을 열어 왔습니다. 앞으로도 네이버는 AI의 개발과 이용에 있어 인간 중심의 가치를 최우선으로 삼고, AI 서비스와 관련된 위험을 예방하고 완화하기 위해 NAVER ASF를 지속적으로 구현하고 실천해 나가고자 합니다.

AI 기술은 끊임없이 변화하고 있습니다. AI 서비스는 사용자와 상호작용하여 정보를 제공하는 단계를 지나, 이제는 AI 에이전트와 같이 사용자가 위임한 범위 안에서 스스로 추론하고 행동하는 영역으로 확장되고 있습니다. 네이버 또한 AI 쇼핑 에이전트를 비롯하여 다양한 영역에서 사용자에게 도움이 되는 에이전트를 선보이고 있습니다. 이러한 측면에서 AI 서비스가 활용되는 맥락이 달라지면 그에 수반되는 위험의 유형과 범위 역시 달라질 수밖에 없습니다. 네이버는 이러한 기술의 변화 속에서 위험의 정의와 분류, 식별과 평가, 관리와 개선이라는 NAVER ASF 2.0의 과정을 통해 새롭게 정의되어야 할 위험을 살펴보며 대응해 나가고자 합니다.

네이버는 AI 기술이 우리의 삶을 편리하게 만들어 줄 수 있지만, 세상의 다른 모든 것처럼 완벽할 수 없다는 점을 잘 알고 있습니다. 그렇기에 네이버는 AI가 사람을 위한 일상의 도구로 자리할 수 있도록, 앞으로도 지속적으로 살펴보며 개선해 나가겠습니다.

발간 정보

참고 자료 (Resources)

AISC 공식 문서 / 네이버 보도자료

- AI in NAVER (네이버의 AI 소개)
- 네이버 AI 윤리 준칙 (국/영문 PDF 다운로드 포함)
- NAVER AI Safety Framework(ASF) 2.0 (국/영문 PDF 다운로드 포함)
- 네이버 기업 홈페이지 Tech for People – AI Safety
- NAVER ASF 2.0 공개 보도자료 (인공지능 안전 서울 포럼 SFASS 2026)
- 네이버 날씨 세이프티 소개 (보도자료)
- 대화형 검색 AI탭 정식 출시 (보도자료)
- 네이버-KAIST 공동 레드티밍 연구(Stable-GFlowNet) (보도자료)

네이버 AI 서비스 살펴보기

서비스 소개 링크

- 네이버 검색 서비스 소개 (AI탭·AI 브리핑)
- AI탭 소개 튜토리얼
- 쇼핑 AI 에이전트 소개 페이지

문의 (Feedback)

리포트 및 AI Safety 관련 문의

AI Safety Center Working Group: dl_aiscwg@navercorp.com

출판 정보

발행처 네이버 주식회사
발행일 2026년 9월 16일 (수)
발행 NAVER AI Safety Center
문의처 dl_aiscwg@navercorp.com


NAVER