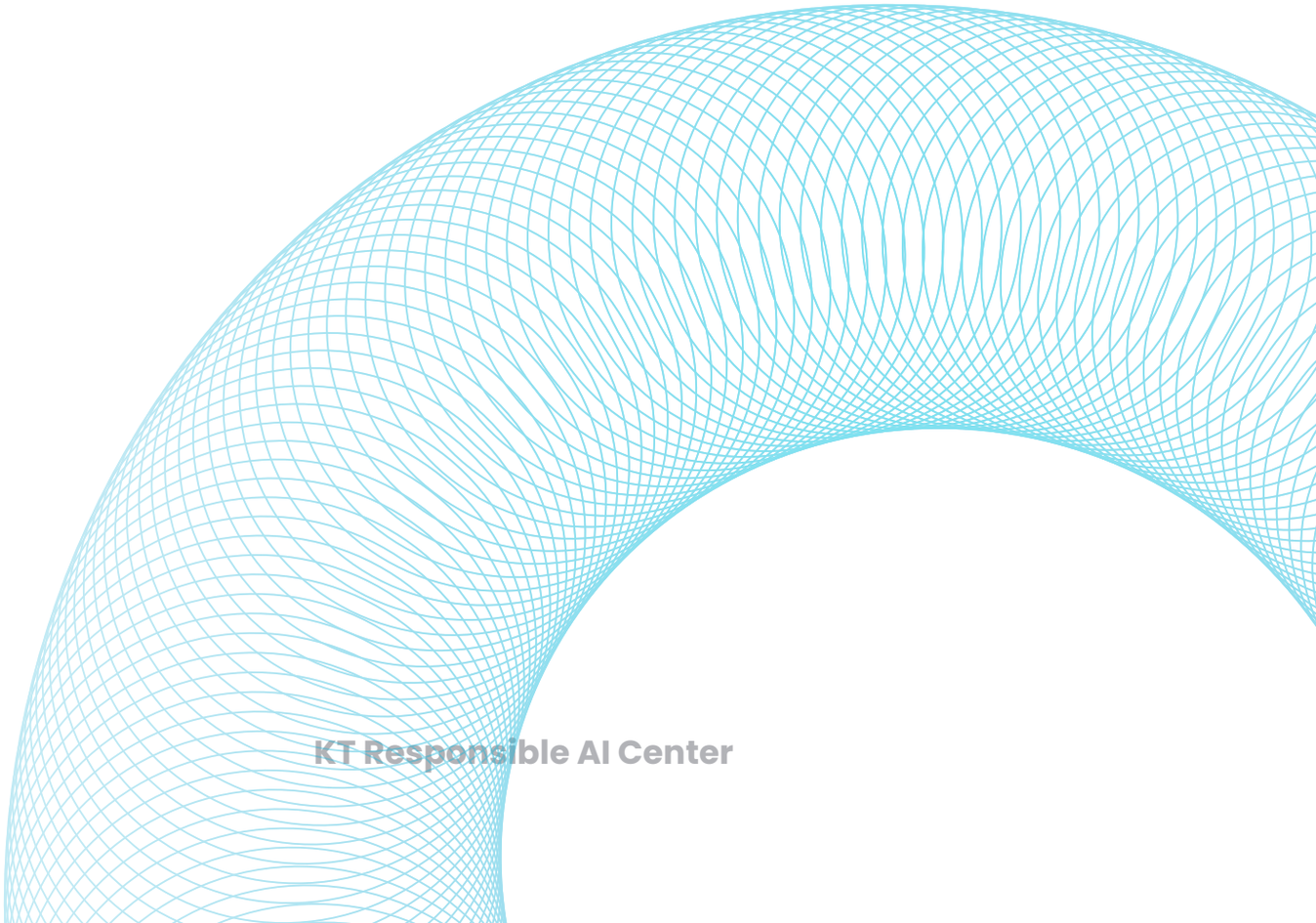
KT

RESPONSIBLE AI

REPORT



KT Responsible AI Center

CONTENTS

Responsible AI

Responsible AI 여정 4

Responsible AI Center (RAIC) 5

Responsible AI 프레임 워크

Responsible AI 거버넌스 7

Responsible AI 윤리원칙 8

Responsible AI 프로세스 9

Responsible AI 미래

Responsible AI 여정으로의 초대 11

Responsible AI

| Responsible AI 여정

KT는 AI대전환 시대에 발맞춰 AICT 중심 경영으로 도약 중입니다.
'AICT 혁신 파트너' 비전 달성을 위해서 무엇보다 중요한 것은 AI가 사회의 윤리 및 공공의 가치에 부합하고, 유익하고 바람직한 방향으로 사용될 수 있도록 responsible AI 체계를 확립하는 것입니다.

이러한 기조는 그간 KT가 걸어왔던 발자취를 통해서도 엿볼 수 있습니다.
KT는 자체 초거대 언어모델(LLM) 믿:음을 보유한 AI 선도기업으로 responsible AI에 대한 세계적인 물결에 동참하고자 많은 고민과 노력을 기울였습니다.

KT는 2024년 인간 존엄성과 공공성 증진이라는 기본가치를 기반으로 AI 윤리원칙을 제정('24.03)하고, responsible AI를 위한 본격적인 이행을 위해 2024년 4월 'Responsible AI Center (책임감 있는 인공지능 센터, RAIC)'를 신설하였습니다.

이후 글로벌 responsible AI 움직임에 적극적으로 참여해왔으며 최근에 있었던 제2회 AI 안전 정상회의(AI Safety Summit)인 AI Seoul Summit 2024 ('24.05)에서 Microsoft, 삼성전자, Anthropic 등 국내외 주요 기업과 함께 AI의 책임·발전·혜택 등 기업이 추구할 방향을 담은 자발적 약속인 '서울 AI 기업 서약'에도 동참하였습니다.

2017.01
국내 최초 AI 스피커 'GiGA Genie' 출시

2023.10
SCFA*와 통신 AI 산업 개발 백서 출간

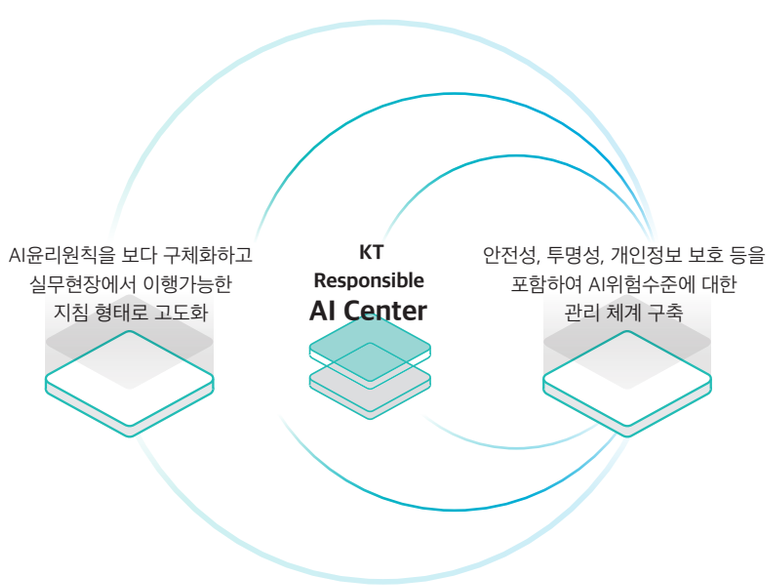
2024.03
KT AI 윤리원칙 제정

2024.04
KT Responsible AI Center 신설

2024.05
AI 서울 기업 서약 동참

2024.05
KT 초거대 AI 믿:음 1.0 출시

안전한 AI서비스로 고객에게 유익한 가치 제공



출처: KT ESG Report 2024

* SCFA (Strategic Cooperation Framework Agreement): 한·중·일 통신사업자간 전략적 협의체

Responsible AI

| Responsible AI Center (RAIC)

AI가 사회적 가치와 목표에 부합하여 활용될 수 있도록 하는데 있어 중추적인 역할을 수행하는 KT Responsible AI Center (이하 RAIC)는 KT가 윤리성과 신뢰성을 갖춘 AI를 제공할 수 있도록 KT만의 responsible AI 프레임워크를 연구개발하고 수립합니다.

KT RAIC는 KT AI가 인간의 가치와 존엄 및 사회의 지속가능성을 향상시키는데 앞장서야 한다고 생각합니다. 더불어, KT AI가 안전하고 신뢰성 있는 서비스로 제공되기를 바랍니다. 또한, 국내외 AI 윤리성/신뢰성 관련 정책 및 논의를 선도하고자 합니다. 궁극적으로는 모든 고객이 안심하고 AI를 활용할 수 있는 AI 혁신 파트너가 되고자 노력합니다.

KT RAIC는 AI 기술이 사용자에게 유익한 가치를 제공할 수 있도록 관련된 위험을 최소화하기 위한 연구를 수행하며, AI 시스템의 취약점을 분석하여 위험 수준에 대한 관리 체계를 구축하는데 노력합니다.

VISION
모든 고객이 안심하고 활용할 수 있는 AI 혁신 파트너

MISSION
국내외 책임 있는 AI 윤리 정책 및 agenda 선도

AI 위험성 최소화

AI 기술이 사용자에게
유익한 가치를 제공할 수 있도록
잠재적 위험을 최소화하는 연구 수행

AI 관리 체계 구축

AI 시스템의 취약점을 분석하여
위험 수준에 대한
관리 체계 구축

실무 이행 지침 제작

AI 윤리원칙을
실무에서 이행할 수 있는
지침으로 제작 및 배포

Responsible AI
프레임워크

| Responsible AI 프레임워크

안전하고 신뢰할 수 있는 AI 서비스 제공의 근간이 되는 responsible AI 프레임워크는 매우 중요합니다. AI 기술은 지속적으로 더욱 빠르게 발전하고 있으며 사람의 지능을 넘어서는 범용 AI인 AGI*의 등장도 예상됩니다.
한편, AI의 발전으로 윤리적, 사회적 부작용에 대한 우려의 목소리도 커지고 있습니다. KT는 AI 기술의 복잡성과 예측 불가능성으로 인한 우려에 깊이 공감하고 있으며, 미래의 AI 시대에 대비하고자 responsible AI 프레임워크를 개발했습니다.

Responsible AI 거버넌스

KT는 responsible AI 실행을 위한 내부 의사결정 체계를 만들고 responsible AI 원칙의 내재화 및 임직원 인식 제고·확산을 위해 노력하며, 국내외 AI윤리 연구소 및 공공 · 규제 관련 기관들과의 협력을 추구합니다.
더 나아가서는 KT파트너사에게 responsible AI 철학을 전파하고 지원하며 지속가능한 responsible AI 생태계를 만들고자 합니다.

Responsible AI 윤리원칙

KT의 responsible AI 윤리원칙은 KT의 핵심가치 및 윤리경영원칙과 국내외 규제를 참고로 하여 만들어 졌습니다.
모든 responsible AI 논의는 이 원칙에서부터 시작하고자 합니다.

Responsible AI 프로세스

국내외 AI 윤리 정책 및 법규 등을 반영한 responsible AI 지침서를 만들어 내부에 적용합니다. Responsible AI 지침서는 AI 모델 및 서비스에 대한 개발 프로세스, 리스크 평가 방법 및 기준, 리스크 분석 및 완화 프로세스 등으로 구성되어 있습니다.

* AGI (Artificial General Intelligence): 인간과 유사한 지능과 스스로 학습할 수 있는 능력을 갖춘 소프트웨어를 만들려는 이론적 AI 연구 분야

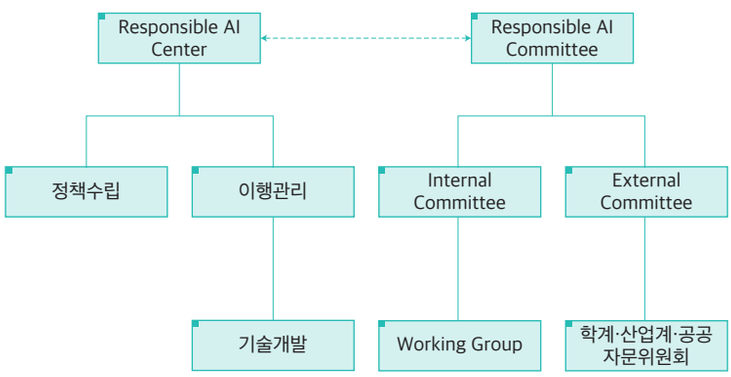
Responsible AI
프레임워크

Responsible AI 거버넌스

| Responsible AI 거버넌스

Responsible AI 추진 거버넌스

RAIC는 responsible AI구현을 위한 주요 역할들을 수행하는 한편, 전사의 responsible AI 거버넌스를 구성하는 구심점으로서, 사내부서와의 협업 및 외부 이해관계자와의 협업을 수행합니다. 또한, AI서비스가 이뤄지는 전 과정에 대해 responsible AI가 구현될 수 있도록 responsible AI 관련 정책과 지침 등을 전파하는 한편 평가 및 관리체계를 운영해 평가와 피드백을 계속해 나갑니다.



외부 협업 체계

KT의 responsible AI 프레임워크는 AI 기술의 발전과 법 · 규제의 변화에 따라 지속적인 개선과 보완이 필요합니다. 이를 위해 외부 협업 체계를 구축하고, 주기적인 워크샵과 세미나를 통해 responsible AI 분야의 글로벌 동향을 파악하고 규제를 반영하여 responsible AI 프레임워크를 견고하고 안정적으로 갖춰 나가고자 합니다. 이를 통해 우리는 responsible AI 분야에서의 리더십을 가지고 지속적인 발전을 이루겠습니다.

글로벌 파트너로는 Microsoft와 AI 기술 및 정책 분야 협력을 진행하고 있습니다. Microsoft는 2016년부터 responsible AI에 대한 논의를 진행하여 세계적인 responsible AI 선도 기업으로 인정받고 있습니다. KT는 Microsoft와 responsible AI 프레임워크 구축 노하우를 공유하고, AI 기술 협업 체계를 구축하고 있으며, 이를 통해 responsible AI 분야에서 선도적인 역할을 수행해 나가고자 합니다.

또한, 학계 파트너와 함께 AICT 응용기술 공동 연구 개발 및 AI 기술 협업을 진행해 왔습니다. KT는 responsible AI를 실천하기 위해 AI 기술의 빠른 변화를 주도하고 학계와 긴밀히 협력하겠습니다.

Responsible AI
프레임워크

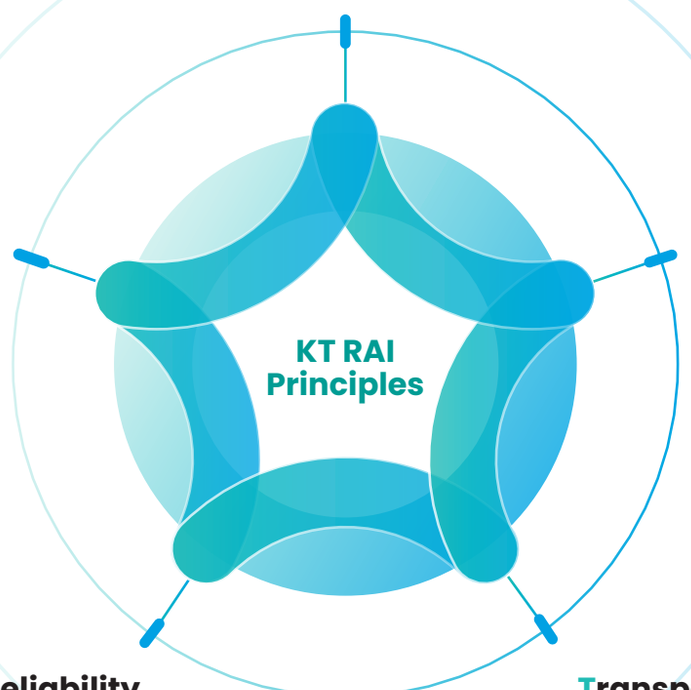
Responsible AI 윤리원칙

| Responsible AI 윤리원칙

KT는 responsible AI를 확보하기 위해 윤리 및 기술 측면에서 **5가지 핵심 원칙(ASTRI)**을 정립하였습니다.
이 원칙들은 북극성 길잡이별(ASTRI)처럼 KT의 모든 responsible AI 논의의 방향을 가리키는 이정표 역할을 수행합니다. 이 원칙들은 글로벌 responsible AI 흐름 뿐 아니라 우리나라에서 통용되는 사회·문화적 가치들을 반영하여 선정했습니다. 또한 한 · 중 · 일 주요 통신사가 함께 선언한 AI 윤리 및 신뢰성 원칙 선언 등 기존의 KT AI 방향성을 고려하였습니다.

Accountability

AI 기술이 세상에 미치는 영향을
깊이 생각하고 책임을 다하고
인간의 삶에 영향을 주는 문제에서
AI가 최종 결정을 하지 않는 것을 보장합니다.



Inclusivity

AI의 편향을 최소화하며 AI를 통해
모든 사회 구성원을 이롭게 합니다.
다양성을 존중하여 포용적이고
개방적인 환경을 구축합니다.

Sustainability

지속가능한 환경적 요소와
사회 경제적 요소를 고려합니다.
탄소 배출을 최소화하는
AI 생태계를 조성합니다.

Reliability

AI가 도출한 결과의 신뢰성을 보장합니다.
시스템을 안전하고 견고하게 유지하며,
잠재적 위험 요소는
철저히 분석하고 대비합니다.

Transparency

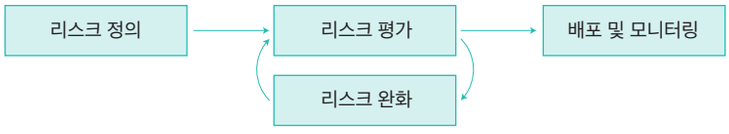
데이터 수집부터 모델 배포, 활용에 이르는
AI Lifecycle 전 과정을 투명하게 관리합니다.
시스템의 운영절차 및 결과에 대해
설명 가능한 환경을 지향합니다.

Responsible AI 프레임워크

Responsible AI 프로세스

| Responsible AI 프로세스

KT는 AI를 기획하고 검증하는 단계에서 KT의 모든 AI 제품 및 서비스가 KT의 responsible AI 윤리원칙을 준수할 수 있도록 내부 평가 절차를 수립하고 있습니다. 각 AI 개발 단계에서 발생할 수 있는 리스크를 정의하고, 이를 완화하기 위한 다양한 활동을 시행하고 있습니다.
KT는 AI 리스크 분석 및 완화를 위해 다음과 같은 프로세스를 운영합니다.



리스크 정의

AI 시스템의 위험 요소와 잠재적인 문제 사항을 검증하는 단계입니다. 해당 단계에서는 영향 평가서 작성을 통해 AI 모델이나 서비스의 목적, 제공하는 기능, 사용자 분석, 잠재적인 피해 등을 단계적으로 도출하게 됩니다. 이를 통해 사용자들이 AI 서비스나 모델을 이용하면서 발생할 수 있는 리스크를 사전에 정의하고, 문제 상황에서 피해를 최소화하기 위해 안전장치를 마련하도록 가이드하고 있습니다.

Responsible AI 지침서

개발자와 사용자가 responsible AI를 더 잘 이해하고 실천할 수 있도록 국내외 법률, 규제, 정책과 더불어 연구기관들의 윤리원칙 및 가이드라인을 기반으로 responsible AI 지침서를 개발하였으며 자문 위원회의 검증을 통해 지침서를 고도화합니다.

리스크 평가

앞서 정의한 리스크들의 위험도에 대해 평가하는 단계입니다. AI 모델이나 서비스의 위험도가 낮거나 악용될 소지가 낮다고 판단되는 경우에는 출시 후 지속적인 모니터링을 통해 그 위험성을 관리해가지만, 위험도가 높거나 악용될 소지가 높다고 판단되는 경우에는 출시를 보류하게 됩니다. 보류가 결정되면 리스크 완화를 통해 그 위험성을 감소시켜야 하며, 충분히 안전하다고 판단될 때 출시가 가능합니다. KT에서 리스크를 평가하는 방법 중 하나로는 레드티밍이 있습니다.

레드티밍

레드티밍은 다양한 적대적 기법을 활용하여 AI 모델이나 시스템의 취약점을 찾고 안전성을 평가하는 활동을 말합니다. KT는 사외의 독립적인 AI red team과 협력하여 보다 다양하고 객관적인 진단을 받고자 노력하고 있습니다.

Responsible AI 프레임워크

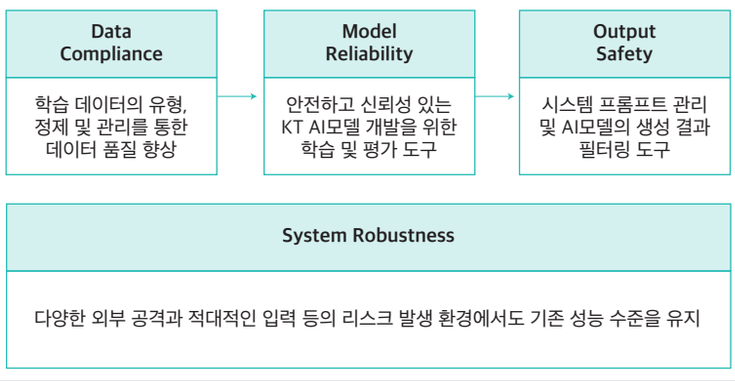
Responsible AI 프로세스

리스크 완화

다양한 도구와 방법을 통해 발생가능한 피해를 완화하는 단계 입니다. AI 개발은 사전 학습, 미세조정, 테스트, 배포로 이루어집니다. 각 단계에서 발생할 수 있는 잠재적인 리스크를 최소화하기 위한 기술적 조치로 KT는 AI 원칙과 정책, 그리고 기술을 수용하는 리스크 완화 파이프라인을 적용하고 있습니다.

리스크 완화 파이프라인

학습 데이터에 대한 내부 관리 통제, 모델 리스크 식별 및 설명가능성을 고려한 모델 설계 프로세스, AI 모델의 생성 결과물에 대한 관리 기술 등으로 구성되어 있습니다.



배포 및 모니터링

문제 해결이 완료된 AI 시스템을 사용자들에게 배포하는 단계입니다. KT는 AI 모델 배포 시 정보제공을 위해 모델카드를 배포합니다. 배포 후에는 문제가 다시 발생하지 않는지 집중적으로 모니터링을 수행하며, 발생하지 않는 것이 확인되면 평상시 수준으로 모니터링을 진행합니다.

모델카드

모델카드를 통해 사용자들에게 모델의 정보를 제공하여 적절한 사용을 돕고, 잠재적 위험과 한계에 대한 이해를 도우면서 투명성과 책임성을 강화하고 있습니다. AI 모델 카드에는 사용 목적, 부적절한 사용 사례, 예상되는 리스크와 완화하기 위한 방안, 모델의 한계 등이 기재되어 있습니다. KT는 이를 통해 모델에 내재된 편향, 오류 및 시스템의 보안 취약점을 확인하고 잠재적인 위험에 대응하고 있습니다.

Responsible AI
미래

최근 빠르게 발전하는 AI 기술에 대한 위험성과 부작용에 대한 우려가 커지고 있으며, 이를 안전하고 편리하게 활용하기 위해 많은 논의가 있습니다.
KT는 모든 고객이 믿고 안전하게 사용할 수 있는 AI를 제공하기 위하여 responsible AI 프레임워크를 만들었습니다.
이를 통해 사전에 발생 가능한 다양한 문제를 끊임없이 고민하고, 다양한 계층의 이해관계자들과 지속적인 논의로 리스크를 예방하고자 합니다.

KT의 AI 윤리원칙은 AI를 개발하고 운영하는데 있어서 의사결정과 행동의 기준이 될 것이며, responsible AI 거버넌스와 프로세스는 윤리원칙을 지키기 위한 관리 체계로서 동작할 것 입니다. 이를 통해 안전하고 편리한 AI를 만들어 고객에게 제공하고, 더 발전된 미래를 만드는데 기여하겠습니다.

| 디지털 혁신파트너 KT의 responsible AI 여정으로의 초대

KT의 responsible AI를 향한 여정에 귀 기울여 주신 모든 분들께 감사드립니다.
Responsible AI를 만들기 위해 중요한 사항으로 꼽히는 것 중 하나는 다양한 사람들이 모여서 만들어야 한다는 것입니다. 다양한 계층의 사람들이 모여서 사회 기술적 관점에서 responsible AI를 논의하고 반영해야 합니다.

KT는 기업, 기관, 고객 누구나 KT의 responsible AI를 만들어가는 여정에 초대 드립니다.

KT는 AI 디지털 혁신 파트너로서, 고객의 혁신을 돕고 산업 발전에 이바지해 왔으며, responsible AI를 구현하기 위해 외부 협력과 오픈 생태계를 지향하고 다양한 파트너들과 함께 혁신을 이루어 나가고 있습니다.
다양한 이해관계자들의 참여와 협력은 오픈 생태계를 더욱 풍요롭게 만들 것이며, 그 결실은 오픈 생태계 참여자들의 공동의 노력이라고 생각합니다.
KT의 responsible AI와 함께 성장하며, 지속 가능한 디지털 혁신을 이루어 나가길 기대합니다.

KT Responsible AI Report

KT Responsible AI Center

배순민 센터장
박완진 담 당
노희진 팀 장
장지환 수 석
김수영 책 임
오승우 책 임
나 현 선 임
정유진 전 임

발행일

2024년 10월

발행처

Responsible AI Center

kt, 당신과 미래 사이에