LG ACCOUNTABILITY REPORT ON AI ETHICS 2025

PROVING SAFETY, ENSURING TRUST

LG AI Research

ABOUT THIS REPORT

Overview

This report presents the LG Accountability Report on AI Ethics, published in 2026, and outlines the key achievements in implementing LG AI Ethics Principles from January 1, 2025, to December 31, 2025. LG AI Research is committed to continuously and transparently disclosing the progress of implementing AI Ethics Principles to stakeholders through various channels, including this report.

Why Accountability?

"Accountability" extends beyond mere legal obligations that a corporation must fulfill. It signifies the responsibility to not only take responsible actions and decisions but also to transparently explain them. The 'LG Accountability Report on AI Ethics' is a testament to LG AI Research's commitment to upholding accountability.

Reporting Principles

This report is structured around the two core pillars of implementing LG AI Ethics Principles: Responsible AI and Inclusive AI. In our commitment to align with global AI ethical standards, this report was prepared in accordance with UNESCO's Recommendation on the Ethics of AI, unanimously adopted by 193 member states in November 2021, and South Korea's National Guidelines for AI Ethics, announced in December 2020.

Period and Scope

The report covers the period from January 1, 2025, to December 31, 2025. It also includes selected information from 2024 and 2026 for ongoing research projects and business achievements.

Contact Information

Website www.lgresearch.ai

Address

D&O Gangseo Building, 150 Magokjungang-ro, Gangseo-gu, Seoul, South Korea

Email

aiethics@lgresearch.ai

Publication Date

19 February 2026

Published by

LG AI Research

CONTENTS

04 **LG AI Ethics Principles**

06 **Head's Message**

07 **LG AI Research At a Glance**

08 **Key Achievements**

PART 1 **Our Journey Toward Responsible AI**

12 1 **Governance** | From Implementation to Advancement

The Evolution of the AI Ethics Organization Connection and Expansion

Advancing the AI Safety Framework Integrating Universality and Locality

AI Ethical Impact Assessment 2.0 Ethical AI Built from Within

Data Compliance An AI Agent That Traces Data Back to Its Origins

CASE STUDY LG Group LG Electronics / LG U+

23 2 **Research** | Toward AI That Earns Trust

Mitigating Action-Scene Hallucinations in Video LLMs

Probing Visual Language Priors in VLMs

Deepfake Detection

Instruction Distractions in LLMs

Fairness in LLM Evaluation

Reasoning Models and Confidence Expression

29 3 **Engagement** | Internalizing a Culture of AI Ethics

AI Ethics Awareness Survey Listening to the Voices Behind the Numbers

AI Ethics Seminar Beyond Knowledge Acquisition—A Forum for Dialogue and Discussion

AI Ethics Onboarding Training The Chain of Responsibility Across the AI Lifecycle

PART 2 **Our Journey Toward Inclusive AI**

36 1 **Sharing** | Opening Access to AI Through Model Sharing

39 2 **Education** | High-Quality AI Education to Bridge Socioeconomic Gaps

42 3 **Collaboration** | Global Partnerships for AI Ethics for All

PART 3 **Leading AI Governance at Home and Abroad**

46 1 **Contributing to Global AI Governance**

47 2 **Contributing to Korea's AI Governance**

48 3 **IEEE Partnership** | Pioneering Global AI Ethics Certification Standards

APPENDIX

51 **APPENDIX 1**

UNESCO's Recommendation on the Ethics of AI

South Korea's National Guidelines for AI Ethics

52 **APPENDIX 2**

AI Risk Taxonomy (Korea-Augmented Universal Taxonomy, K-AUT)

LG AI Ethics Principles

Going beyond technology, LG believes that AI will facilitate better lives for people.
We respect the AI Ethics Principles to make our society healthier and more sustainable.

5 Core Values

Humanity

LG AI provides benefits to humanity and society.

LG AI respects human rights.

LG prioritizes customers and employs a management approach that respects its team members. As we develop and utilize AI, our primary focus is on its impact on people.

We are committed to using AI responsibly, ensuring that it observes human rights and abides by ethical standards, all while providing benefits to individuals and society at large

Fairness

LG AI treats all people fairly while respecting diversity.

LG AI avoids unfair discrimination based on individual characteristics.

LG adheres to Jeong-Do Management principles that honor individual personality and diversity, offer equal opportunities, and ensure fair treatment.

We are committed to preventing unfair discrimination based on individual attributes such as gender, age, and disability. To this end, we establish and rigorously review AI fairness standards that align with societal norms.

Safety

LG AI operates in a safe and robust manner.

LG AI measures and responds to potential risks.

LG builds trust with customers through the delivery of excellent-quality products and services.

We are committed to rigorously verifying the safety of our AI systems to further earn customer trust. Additionally, we proactively prepare for unintended risks by continuously assessing and managing potential hazards.

Accountability

LG AI clearly defines roles and responsibilities for those involved in the development and use.

LG AI strives to fulfill the responsibility to operate appropriately in line with the intention of the human.

LG operates with a sense of ownership and is committed to fulfilling its responsibilities to both customers and society.

We are dedicated to clearly defining the roles and responsibilities of all organizations and members throughout the entire process of developing and utilizing AI.

Transparency

LG AI is transparent with our customers to help them better understand and trust how AI works.

LG AI follows principles and standards to ensure transparency in our algorithms and data.

LG adheres to Jeong-Do Management by operating with honesty and transparency, in alignment with established principles and standards.

We are committed to developing trustworthy and understandable AI for our customers. Additionally, we manage AI algorithms and data with transparency, adhering to principles and standards to ensure that customers can have confidence in our offerings.

Head's Message

The year 2025 was marked not only by rapid advances in AI technology, but also by a fundamental reshaping of the global governance landscape surrounding it. Major players—the United States, the European Union, and China—have each been seeking a new equilibrium between regulation and promotion, guided by their respective interests and philosophies. South Korea has likewise entered a new phase of institutional implementation with the enforcement of the AI Basic Act in January 2026. The trajectory of AI governance across nations will continue to oscillate between regulation and promotion. Over the medium to long term, however, as AI's influence expands beyond national borders, the emergence of a global governance framework to effectively manage this technology and equitably distribute its benefits is inevitable.

Amid these shifting currents, what is the essential value that companies must hold onto? It is not merely reacting to ever-changing regulations, but proactively building the safety and trust of AI technology—so that customers and society at large can embrace AI with confidence, and so that everyone can share in its benefits.

However, significant structural challenges remain. AI technology is inherently prone to winner-take-all dynamics, carrying the risk that its rewards become concentrated among a few early movers while others are left in a state of technological dependency. This is precisely why ensuring broad access to advanced AI capabilities matters—and why LG AI Research has been committed to leading the AI transformation by building independent capabilities, continuously advancing our proprietary foundation model, EXAONE, while actively collaborating with the global AI community. Our commitment to AI safety and trust is unconditional—yet our world-class foundational technology, EXAONE, enables us to translate that commitment into concrete, verifiable safeguards on the global stage.

In 2025, we developed LG AI Research's proprietary AI Risk Taxonomy as a safeguard underpinning these technological capabilities. Grounded in universal human values while incorporating Korea's unique legal, social, and cultural context, this framework offers a new standard applicable both internationally and within South Korea.

Furthermore, we have established a multi-layered verification and implementation system to ensure these safety measures are effective in practice. Through internal and external red teaming, we proactively identify potential risks. We have also produced tangible research outcomes that enhance AI safety and fairness, including approach for mitigating hallucination in video models and the development of deepfake detection technology. In addition, we have refined our AI-driven data governance—which automatically detects and manages copyright risks in training data—to a production-ready level, demonstrating its practical applicability in real-world industrial settings.

At the same time, LG AI Research is dedicated to fostering growth across the broader ecosystem. By making our EXAONE models publicly available, we are expanding access to advanced technology. Through collaborations with international organizations such as UNESCO and the IEEE, we are developing AI ethics education programs and obtaining standards certifications, thereby broadening our contributions on a global scale.

LG AI Research will continue to work to ensure that the benefits of technological innovation are not concentrated in the hands of a few, and that AI earns the trust of society. Moving beyond "responsible AI" toward inclusive AI for all, we are committed to proving the enduring value of trust—even in an era of uncertainty. We sincerely appreciate your continued interest and support.

February 19, 2026.

Honglak Lee and Woohyung Lim, Co-Heads of LG AI Research

LG AI Research At a Glance

Established Date 2020.12

LG AI Research

Publication Record in Global AI Conferences (Past 4 Years)

AAAI, CVPR, ICCV, ICLR, ICML, ACL, Interspeech, EMNLP, NeurIPS

Total 278 Papers

Patent Application and Registration

Domestic

250+

International

320+

(As of December 2025)

Key Achievements

2021
- 09 Ranked 1st in Stanford Question Answering Dataset (SQuAD), a leading and highly-competitive reading comprehension benchmark
- 12 Unveiled EXAONE, the largest AI model in South Korea with 300 billion parameters
- 03 Inaugurated the first cohort of the LG AI Graduate School to cultivate and empower the next generation of AI leaders

2022
- 06 Presented a bidirectional multimodal image-text generation model at CVPR
- 08 Announced the 'LG AI Ethics Principles'

2023
- 06 Hosted a 'Captioning AI' competition NICE Challenge & Workshop at CVPR
- 07 Announced EXAONE 2.0 with enhanced expertise and trustworthiness
- 11 Achieved 1st rank on KorQuAD 2.0 (Korean Question Answering Dataset)

2024
- 08 Unveiled EXAONE 3.0, South Korea's first open-weight general-purpose lightweight language model
  Unveiled EXAONE Path, South Korea's first open-weight multimodal model specialized for clinical medical research
- 12 Unveiled three EXAONE 3.5 models (ultra-lightweight for on-device, general-purpose, and high-performance)

2025
- 02 Launched 'EXAONE NEXUS,' an AI Agent identifying and tracking hidden legal risks within AI training datasets
- 03 Unveiled 'EXAONE Deep,' South Korea's first reasoning model
- 07 Unveiled 'EXAONE 4.0,' South Korea's first hybrid AI model

2026
- 01 Unveiled 'K-EXAONE', ranked 7th globally among open-weight models

EXAONE 1.0	EXAONE 2.0	EXAONE 3.0	EXAONE 3.5	EXAONE Deep	EXAONE 4.0	K-EXAONE
Korea's first bilingual and multimodal foundation model	Enhanced expertise, reliability, and cost efficiency	Released the 7.8B instruction tuned language model	Excels in long-context understanding and instruction following	Strong reasoning capabilities in mathematics, science, and coding	Integrated language-reasoning hybrid model, with agent protocols and advanced multilingual capabilities	236B-parameter Mixture-of-Experts (MoE) model, optimized for expertise, efficiency, and scalability
2021.12	2023.07	2024.08	2024.12	2025.03	2025.07	2026.01

Key Achievements

Throughout 2025, LG AI Research achieved the following concrete results in implementing its AI Ethics Principles:

219

AI Ethical Impact Assessment

In 2025, we conducted AI Ethical Impact Assessments for approximately 60 projects, identifying 219 potential risks in advance and establishing mitigation measures.

45X

Data Compliance

We developed 'EXAONE Nexus,' an AI agent that autonomously explores and performs in-depth backtracing of legal risks in AI training data. Nexus achieved 81% tracing accuracy versus 64% for human lawyers, and 45 times faster processing speed compared to human lawyers.

500K

EXAONE

The EXAONE 4.0 32B model, released to advance the AI ecosystem, recorded over 500,000 downloads within just two weeks of its release on Hugging Face. As of December 2025, cumulative global downloads of the EXAONE series surpassed 8.8 million. EXAONE advances inclusive AI by making advanced models publicly accessible, growing alongside the global AI community

226

AI Risk Taxonomy

LG AI Research has developed a proprietary risk taxonomy grounded in universal values, such as the Universal Declaration of Human Rights, while integrating Korea's sociocultural nuances and emerging threats like multi-agent collusion and safety-guard bypasses. It consists of 4 core domains and 226 specific risk categories.

96.1%

KGC-SAFETY

K-EXAONE achieved a 96.1% score on KGC-SAFETY, which comprehensively evaluates 2,260 risk items including adversarial attacks, multi-turn conversations, and multilingual scenarios. This establishes a high safety standard for Korean-context AI evaluation, demonstrating K-EXAONE's differentiated safety.

40K+

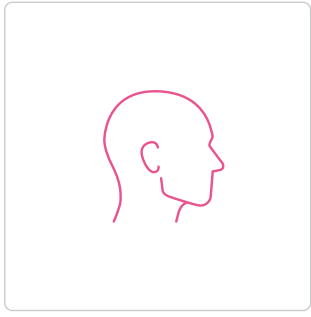
AI Literacy Education

We established a lifelong AI literacy education system spanning middle and high school students, university students, and working professionals, delivering practice-oriented programs integrating technical understanding, ethical use, and social impact. Reaching approximately 40,000 program participants annually (140,000 cumulative), we foster responsible AI capabilities and an inclusive talent ecosystem.

11

AI Ethics Seminars

The AI Ethics Seminar is a forum for voluntary knowledge sharing where members select topics based on their expertise and experience, seeking ethical solutions tailored to our organization through discussion. In 2025, 11 topics were covered—from global governance to fairness, safety, and philosophy—with participants rating the seminars highly for practical relevance, confirming that ethical reflection is positively impacting actual work.



120+

Global AI Ethics Best Practices

We captured voices from the field for the Global AI Ethics MOOC currently being developed with UNESCO. Over 120 AI ethics best practices from governments, corporations, and civil society across 37 countries were selected and integrated into the educational content, which is scheduled for global release on Coursera in the first half of 2026.

1st

IEEE AI Ethics Certification

As an authorized assessor for the IEEE AI ethics certification program, LG AI Research conducted ethical verification of LG Electronics' AI home hub, LG ThinQ On. By meeting all four core criteria—Accountability, Privacy, Transparency, and Algorithmic Bias—it became the first AI product to receive IEEE CertifAIEd certification. LG AI Research also contributes to international standards by sharing its expertise with IEEE.

30+

Participation in Global AI Ethics Norm-setting Discussions

We participated in over 30 major domestic and international AI governance meetings—including the France AI Action Summit, the UNESCO Global Forum on the Ethics of AI, and the AI for Good Summit—to share LG AI Research's ethics implementation cases. Notably, our input contributed to shaping policy recommendations at these forums, reinforcing our role in global AI governance.

PART 1

Our Journey Toward Responsible AI

12 ① Governance | From Implementation to Advancement

The Evolution of the AI Ethics Organization Connection and Expansion

Advancing the AI Safety Framework Integrating Universality and Locality

AI Ethical Impact Assessment 2.0 Ethical AI Built from Within

Data Compliance An AI Agent That Traces Data Back to Its Origins

LG Group CASE STUDY LG Electronics | LG U+

23 ② Research | Toward AI That Earns Trust

Mitigating Action-Scene Hallucinations in Video LLMs

Probing Visual Language Priors in VLMs

Deepfake Detection

Instruction Distractions in LLMs

Fairness in LLM Evaluation

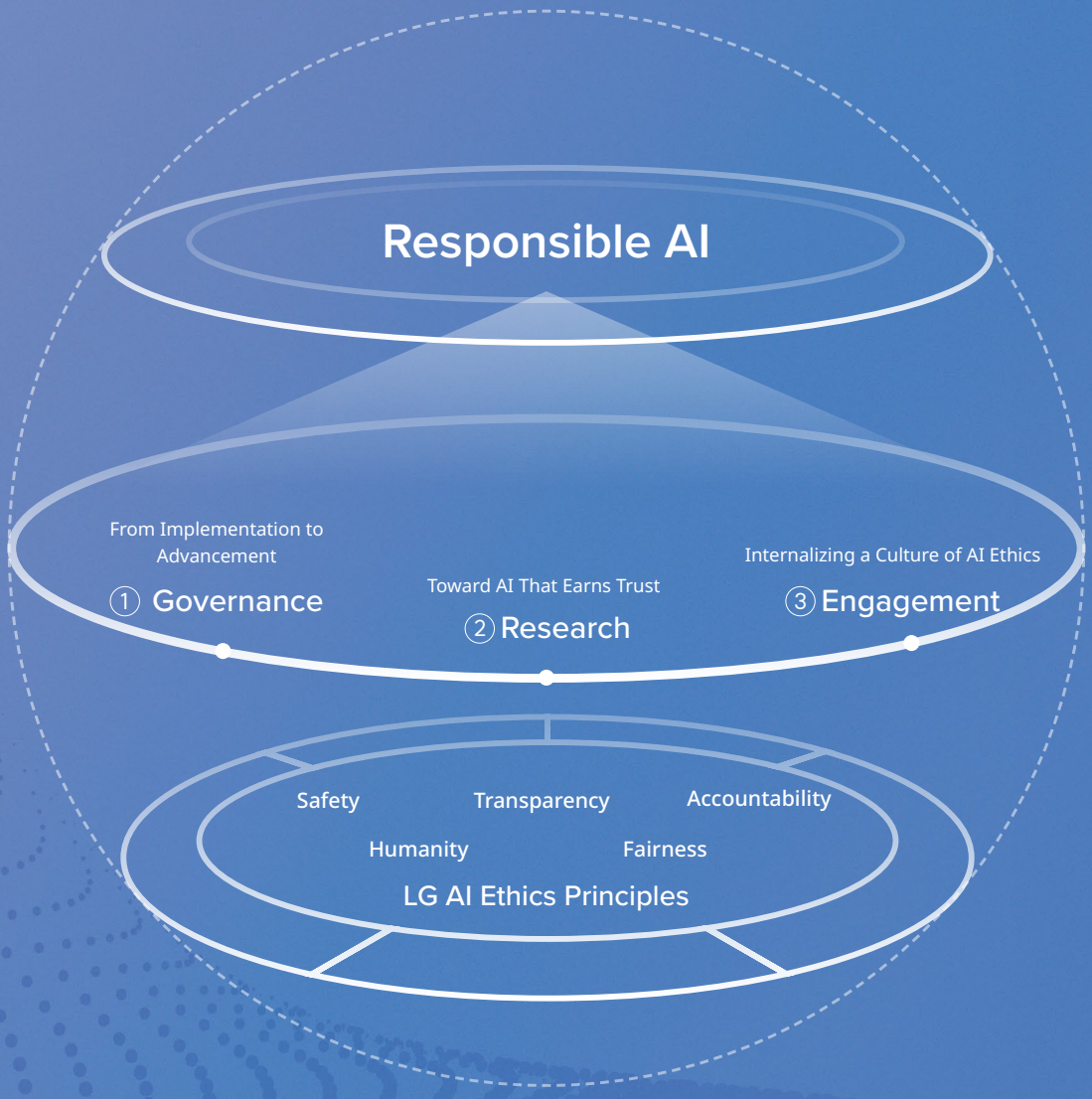
Reasoning Models and Confidence Expression

29 ③ Engagement | Internalizing a Culture of AI Ethics

AI Ethics Awareness Survey Listening to the Voices Behind the Numbers

AI Ethics Seminar Beyond Knowledge Acquisition—A Forum for Dialogue and Discussion

AI Ethics Onboarding Training The Chain of Responsibility Across the AI Lifecycle

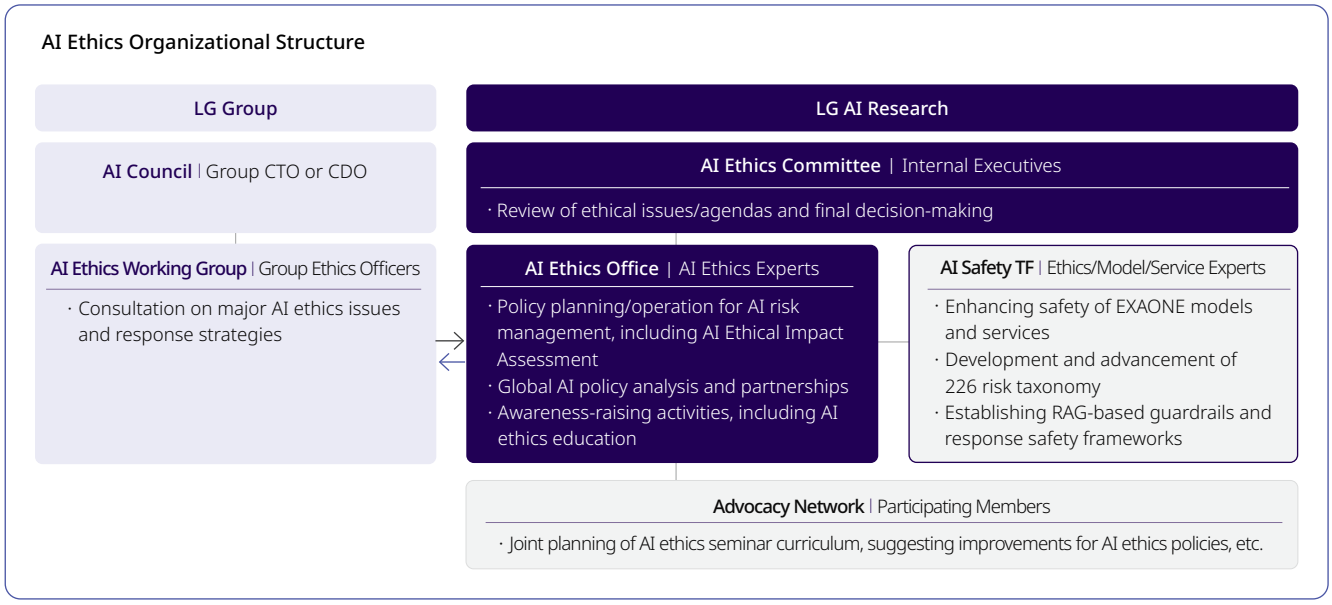


LG AI Research systematically applies ethical standards to research and services through the establishment of the AI Safety Task Force and the development of a proprietary 'AI Risk Taxonomy.' We introduce the detailed journey of managing AI's potential risks, from AI ethical impact assessments and data governance to enhancing members' ethical awareness.

01 Governance From Implementation to Advancement

We have elevated the maturity of our governance by building an execution-oriented organization and a sophisticated safety verification system ensuring that our AI ethics principles do not remain mere declarations, but function effectively in real-world research and development.

The Evolution of the AI Ethics Organization | Connection and Expansion



LG AI Research has established a integrated, cross-functional collaboration framework to embed its AI ethics principles throughout the entire lifecycle of an AI system. In 2025, in particular, we responded to the rapidly evolving AI technology landscape by strengthening our internal execution bodies while solidifying group-wide cooperation networks, thereby expanding the scope of our governance. Internally, we launched the AI Safety Task Force, bringing together experts from diverse backgrounds—including ethics, policy, model research, guardrail development, and service planning—to strengthen our execution capabilities. This task force is a problem-solving unit in which specialists from each domain work together toward the shared goal of ensuring safety, tackling policy and technical challenges in concert. Ethics officers set normative standards; AI researchers translate those standards into technical implementations; and service managers feed back potential risks from the user's perspective. Through this close-knit collaboration, we have achieved meaningful improvements in the safety of both the EXAONE model and the ChatEXAONE service.

Alongside this internal organizational innovation, we have also reinforced our collaborative structures for elevating AI ethics capabilities at the group level. The AI Ethics Working Group, convened on a quarterly basis, serves as a forum where AI ethics officers from major affiliates come together to share each company's current issues and the latest regulatory developments. These sessions go beyond simple information exchange: participants benchmark best practices across affiliates and explore joint solutions, fostering the collective growth of AI ethics response capabilities across the entire LG Group. The overall governance structure of LG AI Research operates as a multi-layered system. The AI Ethics Committee, as the highest decision-making body, sets the strategic direction. The AI Ethics Office oversees company-wide coordination and manages group-level collaborative bodies. The AI Safety Task Force carries out substantive safety measures on the ground.

Advancing the AI Safety Framework | Integrating Universality and Locality

LG AI Research has developed the Korea-Augmented Universal Taxonomy (K-AUT) — a comprehensive and scalable AI risk classification framework designed to address a critical gap in the current AI safety landscape: while leading international standards provide robust foundations, they may not fully capture the cultural, legal, and historical contexts of non-Western societies. K-AUT builds upon widely recognized international frameworks — including the Universal Declaration of Human Rights, international human rights covenants, and the UNESCO Recommendation on the Ethics of AI — while incorporating Korea's unique legal, social, and cultural characteristics. In this way, it complements rather than replaces existing global standards, enriching them with locally grounded perspectives.

K-AUT systematically organizes potential risks into four core domains comprising 226 specific risk categories: The Universal Human Values domain addresses threats to life, dignity, and fundamental rights, drawing on the Universal Declaration of Human Rights and international human rights covenants. The Social Safety domain, informed by empirical academic research and international guidelines, captures risks that disrupt social order — such as the spread of disinformation, the incitement of religious or ideological conflict, and the facilitation of criminal activity.

The Korean Sensitivity domain manages issues arising from Korea's distinctive historical and geopolitical context, drawing on the Constitution of the Republic of Korea, applicable statutes, and established historical scholarship. This layer helps prevent errors that may occur when global models do not adequately reflect local cultural and historical nuances. The Future Risk domain proactively addresses emerging threats posed by rapid technological advancement — including AI safeguard circumvention, digital identity manipulation, and the amplification of systemic distrust — informed by future risk research from international organizations and technology outlook studies from leading institutions.

The effectiveness of K-AUT stems from its rigorous assessment protocols. For each of the 226 specific risk categories, five concrete assessment criteria have been established; a response is classified as "inappropriate" if it violates even a single criterion. LG AI Research employs K-AUT not merely as a set of guidelines, but as a practical engineering tool for verifying and strengthening the safety of the EXAONE model and its associated services.

Notably, the Korean Sensitivity domain is designed to be adaptable — enabling future localization with risk categories that reflect the unique characteristics of other countries and regions. This demonstrates K-AUT's potential as a broadly applicable framework beyond the Korean context.

Korea-Augmented Universal Taxonomy (K-AUT)

Category	No. of Detailed Risks	Definition	Key Content and Basis	
Universal Human Values	55	Issues that threaten human life, dignity, and fundamental rights	**Basis**	Universal Declaration of Human Rights, International Covenants on Human Rights, etc.
			Examples	Human rights violations, incitement of violence, promotion of self-harm, privacy violations
Social Safety	75	Issues that disrupt social order or exacerbate conflicts	**Basis**	Academic empirical studies, international guidelines, legal systems of various countries
			Examples	Generation and spread of misinformation, incitement of religious/ideological conflicts, assistance in criminal acts
Korean Sensitivity	60	Sensitive issues rooted in Korea's historical, cultural, and geopolitical context	**Basis**	Constitution of the Republic of Korea, domestic positive laws, verified established historical scholarship, etc.
			Examples	Distortion of historical facts, inter-Korean relations, gender and generational conflicts
Future Risk	36	Newly emerging risk issues due to rapid technological advancement	**Basis**	Future risk research results from international organizations, future technology prediction studies, etc.
			Examples	Loss of AI control, amplification of systemic distrust

※ K-AUT will be continuously reviewed and updated to reflect technological advancements and sociocultural changes.

Systematic Red Teaming Operations

LG AI Research has operated a dual-track red teaming system—conducted by both internal teams and external specialized organizations—based on the 226 risk categories defined in K-AUT. Going beyond simple harmful query testing, we applied sophisticated adversarial attack scenarios designed to bypass the model's defense mechanisms, enabling proactive identification of potential vulnerabilities.

We comprehensively deployed state-of-the-art attack techniques, including: ActorAttack, which tests the model's ethical boundaries by assuming specific personas; Crescendo, which gradually elicits harmful responses through sequential conversations; Contextual Priming, which conceals malicious intent within seemingly benign contexts; and linguistic evasion attacks such as Unicode substitution and cross-lingual prompt injection, which exploit multilingual encoding to bypass safety filters. Through this verification process, we collected vulnerability data across multilingual environments encompassing six languages—Korean, English, Spanish, German, Japanese, and Vietnamese. In addition to rigorously testing the Universal Human Values and Social Safety domains, we conducted focused examinations of the Korean Sensitivity and Future Risk domains—areas where our red teaming exercises confirmed that global models show relatively weaker coverage—ensuring balanced coverage across all risk categories.

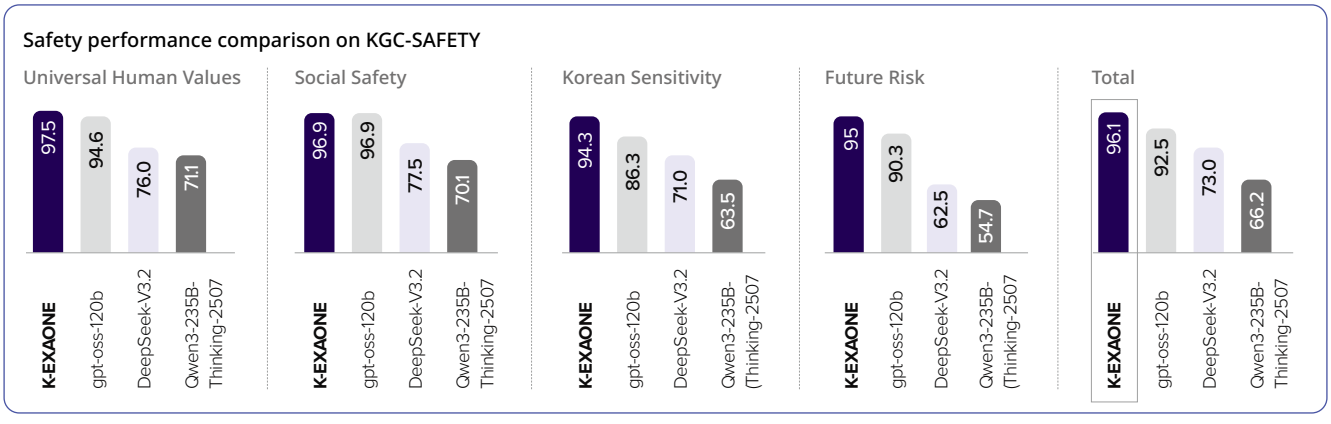
Safety Enhancement and Continuous Improvement

Through this red teaming process, previously unidentified risk types were discovered. These findings were incorporated into K-AUT, and defensive logic was reinforced, thereby strengthening the overall safety framework. Furthermore, the vulnerability data collected through red teaming served as a critical training resource for elevating the safety of the EXAONE model. Specifically, we conducted Supervised Fine-Tuning (SFT), Direct Preference Optimization (DPO), and Safety DPO—a specialized safety enhancement technique—to train the model to identify and reject harmful queries or respond in a safe manner. Through this virtuous cycle of "discovery, learning, and reinforcement," EXAONE has developed the capability to proactively respond to emerging threats.

Independent Validation Through the KGC-SAFETY Benchmark

To objectively validate the effectiveness of K-AUT, LG AI Research developed its own safety evaluation benchmark: KGC-SAFETY (Korean Global Civic Safety Benchmark). KGC-SAFETY comprises a total of 2,260 evaluation items, with a minimum of 10 test cases per category, designed to cover varying difficulty levels and attack vectors across each of the 226 risk categories in K-AUT. This benchmark encompasses a wide range of difficulty levels and types—from naive queries to adversarial attacks, multi-turn conversations, and multilingual scenarios (Korean, English, and others)—enabling comprehensive assessment of risk situations that may arise in real-world service environments.

Evaluation results using KGC-SAFETY revealed that most global models achieved relatively favorable scores in the Universal Human Values and Social Safety domains, while demonstrating comparatively weaker performance in the Korean Sensitivity and Future Risk domains. In contrast, K-EXAONE achieved strong performance across all domains—demonstrating notably robust capabilities, particularly in areas where global models showed vulnerabilities.



※ The comparison presents results among leading open-weight models of comparable scale

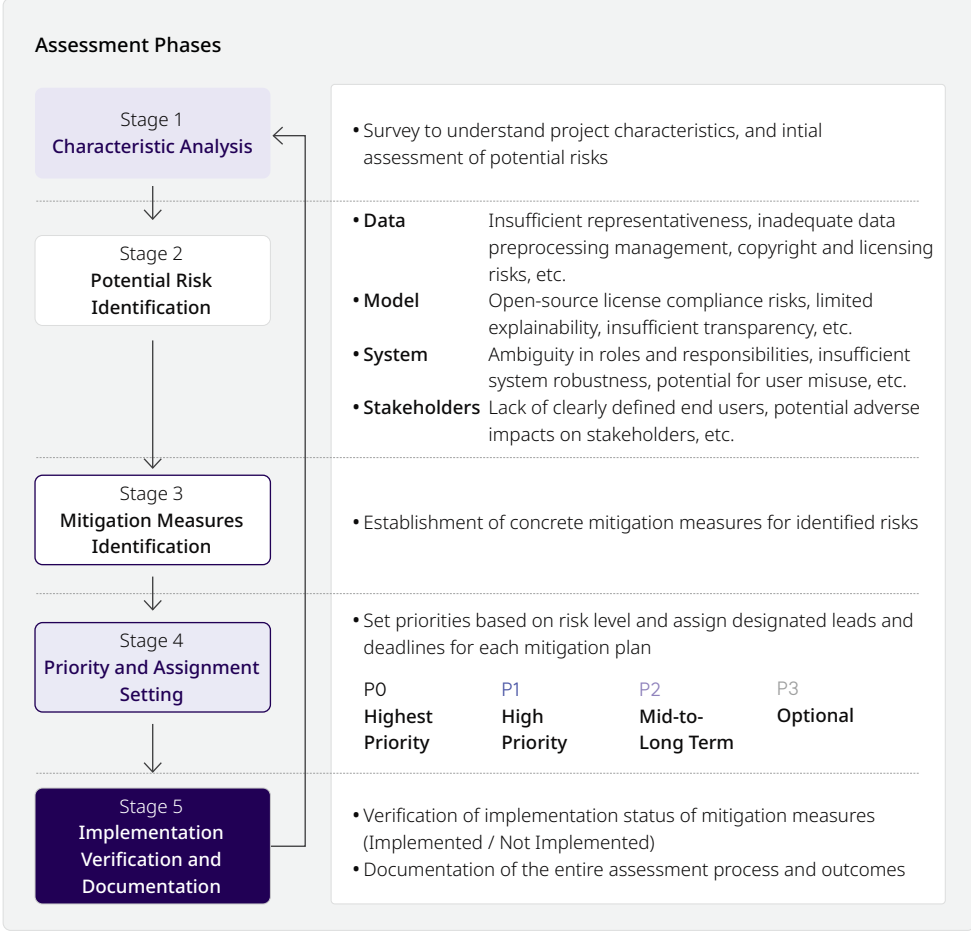
AI Ethical Impact Assessment 2.0 | Ethical AI Built from Within

Applied Proactively to All AI Projects

As AI's impact on society continues to expand, sustainable innovation increasingly depends not only on technological advancement but on the ethical responsibility embedded throughout the development process.

In response to growing global regulatory expectations—such as the EU AI Act and Korea's AI Basic Act—which call for demonstrable actions beyond declarative principles, LG AI Research has designed and implemented the AI Ethical Impact Assessment since 2024 as an institutionalized mechanism for proactively identifying and mitigating potential societal risks in real-world AI development. Rather than serving as a purely compliance-driven exercise, the framework functions as a core driver of voluntary governance, enabling project teams to continuously reflect on ethical considerations and ensure that technological achievements translate into demonstrable societal trust.

The assessment is applied throughout the full project lifecycle, guiding teams to proactively identify, analyze, and address potential ethical risks associated with all AI projects. In Stage 1, project characteristics are analyzed to assess overall risk levels. In Stage 2, specific issues are identified as concrete risk scenarios, and Stage 3 focuses on developing actionable mitigation measures. In Stage 4, priorities are set based on the significance of each issue and the complexity of its resolution, with clear ownership and timelines assigned. Finally, in Stage 5, implementation is formally verified to ensure that identified risks are effectively addressed. This structured five-stage approach ensures that risks are not only identified but also systematically implemented and resolved.



2025 AI Ethical Impact Assessment Results

Data Remains the Most Critical Area In 2025, the AI Ethical Impact Assessment was conducted across approximately 60 AI projects, through which a total of 219 potential risk factors were identified. Consistent with the previous year, data-related issues accounted for the largest share of identified risks, representing 60% of all cases. Among these, the most frequently discussed potential issues were lack of data representativeness (19%), copyright compliance considerations related to training data (13%), and insufficient data preprocessing management (13%).

Following data-related risks, system-related issues accounted for 18% of the total. In particular, concerns regarding the allocation of roles and responsibilities across AI model development, deployment, and maintenance processes accounted for 14%, representing a notable increase from the previous year (8% to 14%). This trend reflects the continued expansion of external partnerships and collaborative projects, which has led to increased focus on accountability and responsibility allocation.

Model-related issues comprised 13% of the identified risks, with primary concerns centered on open-source license management (8%) and limited explainability of AI models (5%). Finally, stakeholder-related issues accounted for approximately 9% of the total, with the lack of a clearly defined end user (8%) emerging as the most significant concern in this category.

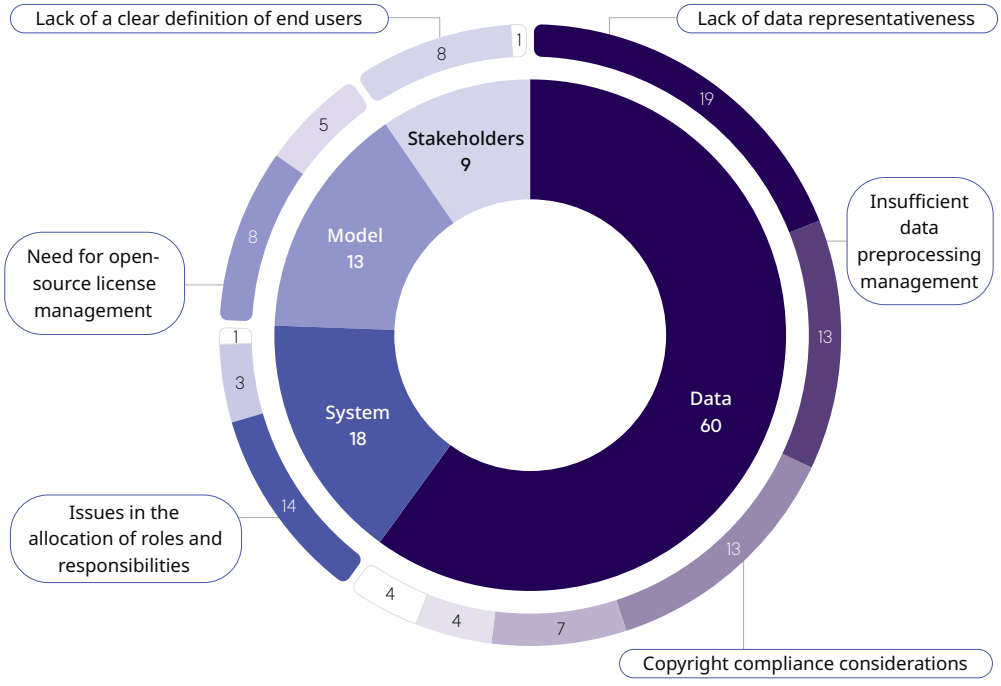
Voices from Our Members

Through the AI Ethical Impact Assessment, the process of defining who the end users are prompted us to reaffirm that citizens are, ultimately, the users of our AI systems. This reflection enabled us to more carefully consider the potential impacts our projects may have on society and on citizens in real-world contexts.

Vision AI Scientist

Identified Potential Risks in 2025

(Unit : %)



Discussion of Mitigation Measures

To address the potential issues identified through the AI Ethical Impact Assessment, a range of mitigation measures were discussed and implemented. To mitigate data representativeness issues, researchers incorporated schemas from diverse domains, natural language queries from real-world usage environments, and database queries written in various formats into training datasets. In addition, teams introduced measures to benchmark and mitigate model bias through systematic evaluation. From a data copyright perspective, compliance reviews for individual datasets were conducted, while data preprocessing management was strengthened through measures such as multi-labeling and cross-checking, as well as logging and documentation of data processing workflows.

Voices from Our Members

While the AI Ethical Impact Assessment may seem burdensome if performance alone is prioritized, it is a highly meaningful process in terms of internally embedding the practice of AI ethics as a shared organizational commitment. I hope that a wide range of activities to further enhance awareness of AI ethics will continue in the future.

Bio-Specialized AI Scientist

Among the issues discussed more extensively this year compared to the previous year, the allocation of roles and responsibilities emerged as a key focus area. Major mitigation measures included clearly defining roles and responsibilities, establishing strategic partnerships and collaboration frameworks, and conducting regular progress meetings and reporting. As external collaborations and partnerships continue to expand, various approaches to managing partnerships in a more systematic and accountable manner were explored.

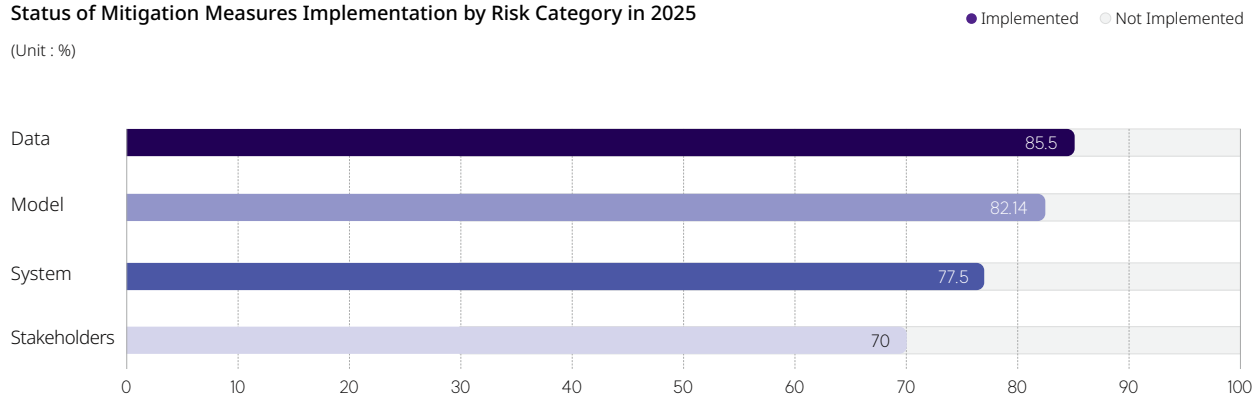
Implementation of Mitigation Measures

All issues and mitigation measures discussed through the AI Ethical Impact Assessment were subject to implementation verification at project completion. As a result, 85% of data-related issues and 82% of model-related issues were addressed. System- and stakeholder-related issues also achieved implementation completion rates exceeding 70%, demonstrating meaningful progress across all risk categories.

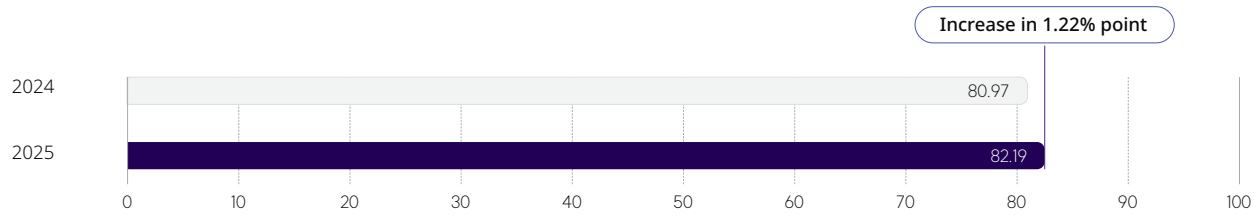
Overall, the implementation rate of mitigation measures for identified potential issues reached 82% in 2025, representing a increase from 81% in the previous year. The primary reasons for unimplemented measures included project postponements, carry-over to subsequent project phases, and planned implementation within continuous or follow-up projects. These items have been systematically linked to future projects to ensure that mitigation measures continue to be addressed over time.

Status of Mitigation Measures Implementation by Risk Category in 2025

(Unit : %)



Mitigation Measures Implementation Rate by Year



Future Plans

Drawing on two years of operational experience and project team feedback—and in response to rapidly evolving global regulations such as the EU AI Act and Korea's AI Basic Act—LG AI Research plans to restructure its AI Ethical Impact Assessment framework into a risk- (impact-) based system starting in 2026. Through this evolution, we aim to strengthen alignment with international regulatory standards while further advancing an employee-centered, voluntary accountability system that supports both technological progress and societal trust.

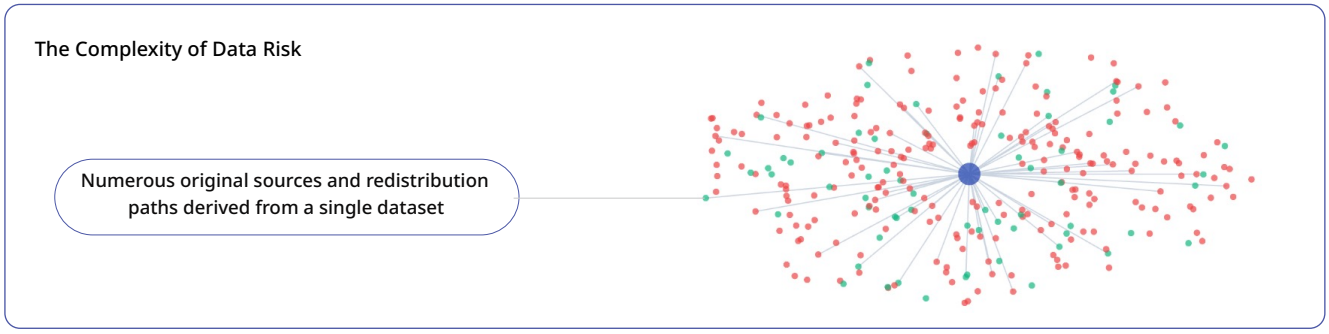
Data Compliance | An AI Agent That Traces Data Back to Its Origins

As cutting-edge AI technology is increasingly deployed across society, data has evolved beyond a mere training resource into a focal point of legal disputes. The fact that legal determinations of fair use under copyright law can vary by country and even by court means that not even the largest AI companies are exempt from copyright litigation. To ensure compliance amid such uncertainty, LG AI Research developed EXAONE Nexus, an AI agent that performs in-depth tracking of risks throughout the data lifecycle and identifies legal exposure.

Redefining the Problem

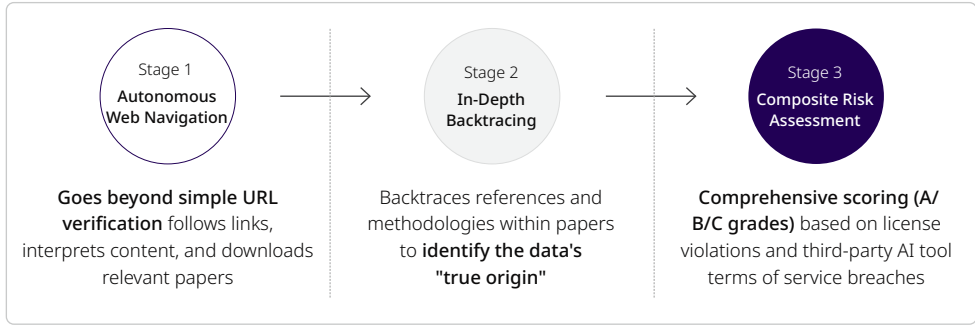
The Pitfall of Data "Compilation" Copyright infringement issues arise throughout the entire lifecycle of AI training data—from preprocessing (reproduction) to training (transformative use) to output. Modern large-scale datasets, in particular, are not single files but compilations in which countless data sources are intermingled, making risk management all the more challenging.

Through in-depth tracing experiments, our researchers discovered numerous cases in which a single dataset had undergone more than 16 rounds of redistribution, or in which over 1,600 distinct original sources (academic papers, web content, code, etc.) were intermixed within a single dataset. This demonstrates that even if numerous lawyers were deployed, manual review would be impractical—highlighting the limitations of conventional human-centered review systems.



Technical Solution

EXAONE Nexus: Self-Navigating and Self-Verifying To address the limitations of human experts, LG AI Research developed EXAONE Nexus, a proprietary AI agent specializing in data compliance, built on EXAONE 3.5. This agent goes beyond conventional compliance approaches that stop at surface-level information verification, executing a three-stage in-depth verification process that traces data back to its origins.



In-Depth Analysis When a user provides only the dataset URL—the minimum information required for review—the agent collects and analyzes all available information on the corresponding web page, then identifies the data sources and data production tools that constitute the dataset.

Autonomous Navigation The individual elements comprising the dataset—such as data sources and tools discovered in the previous stage—are also reviewed. To gather in-depth information related to these data sources and tools, the agent navigates various types of content containing additional information (academic papers, technical documents, web platforms, etc.) based on the given information, much like a human conducting a web search.

License Map Starting from the given dataset URL, the agent recursively repeats the in-depth analysis and autonomous navigation stages, continuing until no further downstream data source information can be found. All the information gathered in this process is then consolidated into a networked data provenance map.

Risk Detection Finally, the agent comprehensively evaluates all the information it has accumulated about the dataset, scoring it so that the dataset's risk level can be intuitively understood. The agent makes its assessment based on 18 risk indicators designed by legal experts in Korea and abroad; the results are automatically classified into seven safety grades and provided to researchers.

Analysis Results

The Hidden Risks in Open Data and Achieving "Compliance Assurance" Our investigation of open datasets using Nexus revealed that out of 2,852 datasets labeled with ostensibly commercially permissive open-source licenses such as MIT or Apache, only 605 (approximately 21.2%) could actually be used for commercial AI training.

The remaining approximately 80% of the data either contained information that had been crawled from the web without authorization, or included data generated by third-party AI tools that cannot be used in developing competing models—posing significant legal risks. By filtering out these "hidden risks" in advance, LG AI Research has strengthened the legal stability of its models.

Performance Comparison and Future Plans

Human-in-the-Loop The introduction of EXAONE Nexus has complemented human experts not only in efficiency but also in accuracy. In tracking the intricately intertwined redistribution processes, human lawyers achieved only 64% accuracy, whereas Nexus recorded 81% accuracy—a 17 percentage-point improvement— capturing even easily overlooked risks. Furthermore, task processing speed was 45 times faster than human review, and in terms of cost efficiency, it demonstrated economic benefits approximately 700 times greater.

EXAONE Nexus Performance Breakthrough

Tracing Accuracy	Processing Speed	Cost Efficiency
81%	45X	700X
17%p more precise risk detection compared to human lawyers	**45× faster task processing** compared to human lawyers	**700× greater cost efficiency** compared to human lawyers

LG AI Research currently operates on the principle of a "Human-in-the-Loop" system. While the agent performs broad-scope tracing as a first pass, human lawyers retain the final judgment authority over legal risks and retain ultimate responsibility. By doing so, we are realizing responsible AI governance that harmonizes technological efficiency with human accountability.

Building on this technological capability, EXAONE Nexus is currently building a "Data Provenance Database" that displays the origins and transformation processes of data. We plan to make these results available to the community (Nexus Platform), fostering an open ecosystem where researchers can use data with greater legal clarity.

CASE STUDY | LG GROUP

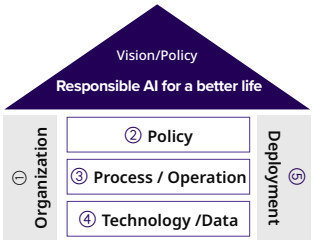
LG Electronics Responsible AI

LG Electronics applies an ex ante, compliance-by-design approach to deliver safe and trustworthy AI-enabled products and services and to prevent misleading or overstated claims. To this end, LG Electronics operates an integrated Responsible AI governance framework, supported by cross-functional collaboration across R&D, Quality, Legal, and Information Security in key markets including the EU, North America, and Korea.

2024	08	Established a company-wide Responsible AI TF
2025	02	Publicizing corporate responsible AI policies
	08	Applying a risk management framework to a company-wide standard development system
	09	Deploying AI Washing Marketing Guidelines
	11	Implementation of mandatory responsible AI training for domestic members

Vision and Policy

LG Electronics aims to secure trustworthiness and regulatory compliance by adapting Group AI Ethics Principles to product, service, and business contexts, and by designing a Responsible AI governance framework covering organization, policies, processes, technology, data, and enterprise-wide adoption.



LG Electronics Responsible AI Governance Framework

① Organization	Company-wide AI governance
② Policy	Policy and Global Regulation Tracking
③ Process/Operation	AI Risk Lifecycle Management
④ Technology/Data	Platforms and Datasets
⑤ Deployment	Training External Collaboration

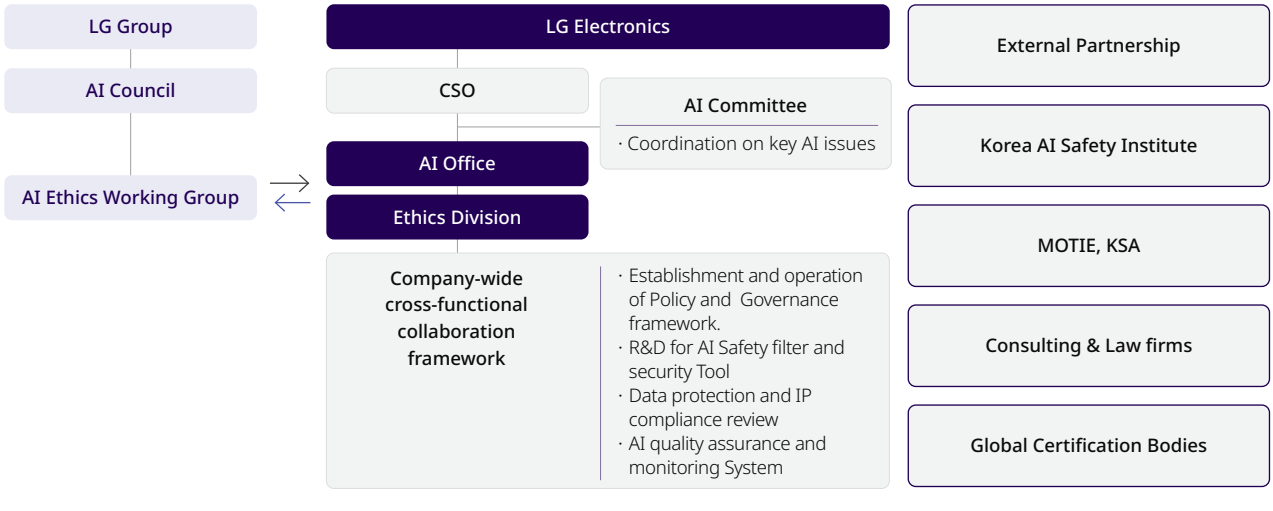
Major Items for Global Regulations & Standards

• Tech. Doc.	Model/Service Desc., etc.
• Post-Prod. Mgmt	Logs, Input Data Mgmt., etc.
• User Info.	Contact Info., User Manuals, etc.
• Risk Mgmt.	Identified & Mitigated Risks, etc.
• Quality Mgmt.	Test Proc., Data Mgmt., etc.

AI Ethics Organization

Given the multidisciplinary nature of AI regulation—spanning technology, law, ethics, and safety—LG Electronics operates an enterprise-wide governance structure led by the AI Office, with an AI Committee enabling executive oversight and timely decision-making.

LG Electronics AI Ethics Organization



CASE STUDY | LG GROUP

Internal & External Collaboration

Through close collaboration with the Group AI Ethics Working Group, LG AI Research, and the AI Safety Institute, LG Electronics advances company-wide policies covering regulatory readiness, risk management, and physical safety. The company became the first in the home appliance sector to obtain IEEE AI Ethics certification and, following the MoU signed in June 2025 with the AI Safety Institute, continuously enhances risk-tier-based guidance and internal processes. Leveraging official liaison channels with AI Safety Offices in the EU, ASEAN, North America, and Japan, LG Electronics proactively responds to global regulatory changes and, through the AI Home Appliance Alliance in cooperation with the Ministry of Trade, Industry and Energy, leads discussions on regulatory flexibility, AI tier standardization, and policy improvement, while contributing to the AI Basic Act and national strategic technology policies.

Process Integration

To enable systematic risk management, LG Electronics has embedded Responsible AI checklists—based on development stage, jurisdiction, and role—into mandatory development workflows, integrating data protection, data governance, cybersecurity, terms, and advertising reviews to ensure consistency and reduce operational burden.

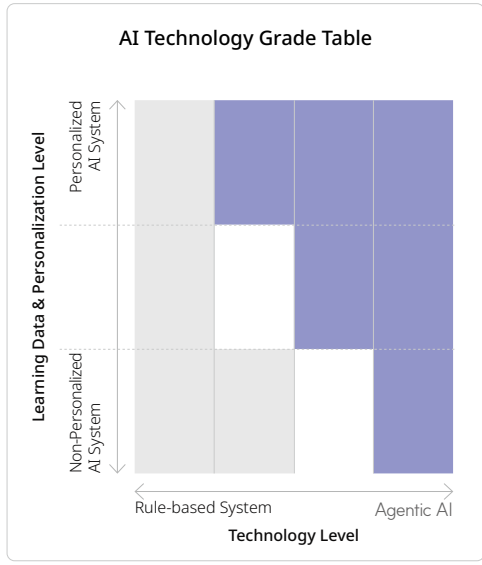
1 Step
Risk Classification Review
Prohibited / High-risk / Limited-risk / Minimal-risk AI systems

2 Step
Role and Responsibility Identification
Provider / Deployer / Distributor / Importerhigh-performance / high-impact AI /developer–operator entities

3 Step
Lifecycle-Based Compliance Review
AI Model assessment, Data governance/Protection, Quality/ Terms management, etc.,

AI Washing Regulatory Response Guidelines

To ensure accurate and responsible AI-related communications and to prevent misleading or exaggerated marketing practices that could undermine brand trust, LG Electronics established AI Marketing Guidelines. AI usage is standardized based on system maturity and data training and personalization levels, and AI tier review committees in each business division proactively assess and manage AI use from the early planning and design stage. This approach strengthened coordination between Marketing and Legal teams, reduced legal review lead times from approximately two weeks to three days, and mitigated regulatory and consumer protection risks that could otherwise damage the corporate brand through enterprise-wide implementation.



Responsible AI Training Curriculum

① Responsible AI Fundamentals
② Domestic and Global AI Regulations and Policy Trends
③ LG Group Responsible AI Framework
④ LG Electronics Responsible AI Policies and Operational Processes

Company-wide Adoption

LG Electronics established AI Marketing Guidelines to ensure accurate and responsible AI-related communications in sales and marketing, preventing misleading claims and mitigating potential product risks while maintaining brand trust. The guidelines standardize AI usage based on system maturity and data personalization levels and support early-stage AI risk identification and management through internal review processes. This framework strengthens coordination between Marketing and Legal functions, reduces regulatory risk, and has shortened legal review lead times from two weeks to three days. The guidelines are applied enterprise-wide to prevent misleading or deceptive AI-related representations across customer-facing activities.

Next Steps

To provide customers with safe and reliable products and services based on Responsible AI, LG Electronics is continuously improving internal policies and operational processes through an agent-based AI regulatory response and self-inspection system. Through this, we aim to secure a balance between regulatory compliance and business execution speed and establish a systematic and sustainable compliance system across the company.

CASE STUDY | LG GROUP

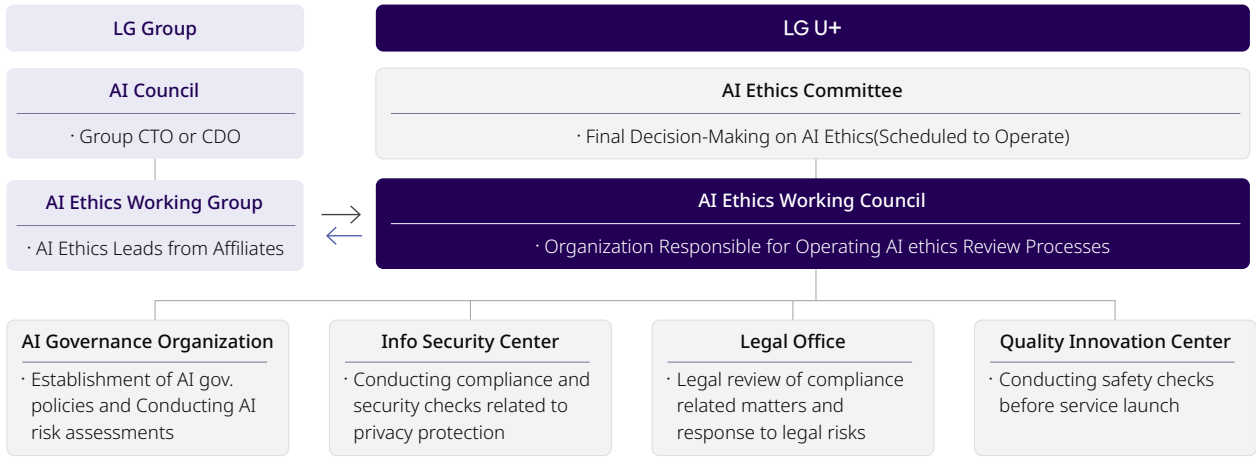
LG U+ Responsible AI

LG U+ obtained ISO/IEC 42001 certification to ensure stability and trust in AI services and strengthened its AI governance centered on risk management, red teaming, and the Ethics Committee.

LG U+ AI Governance Organization

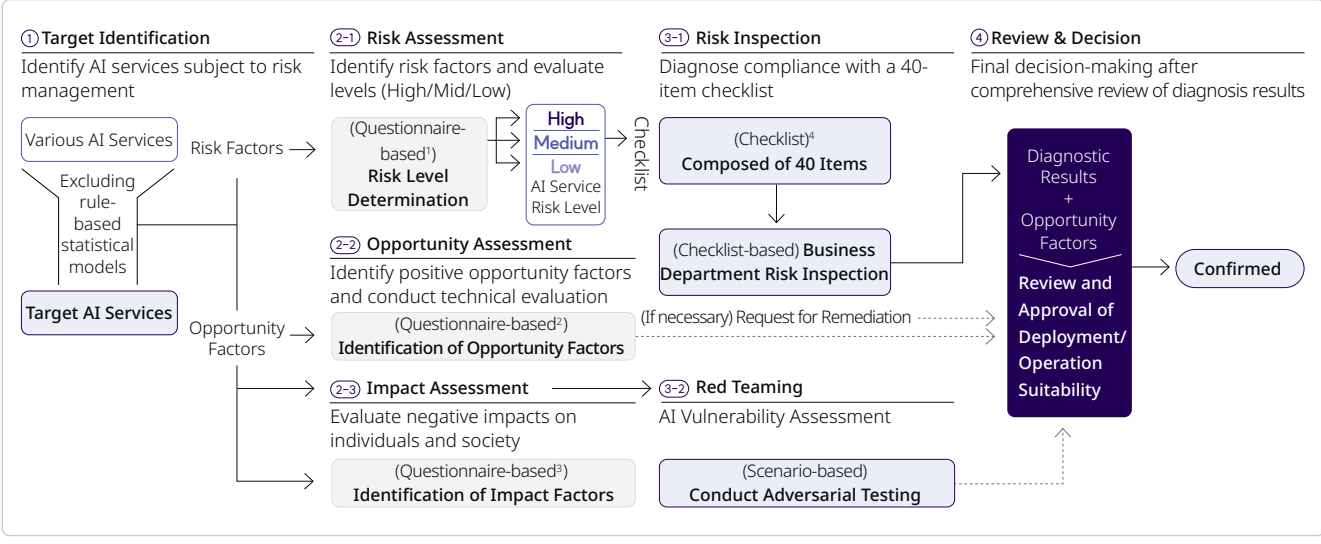
We established enterprise-wide AI governance policies and guidelines centered on the AI Ethics Working Group and operate related processes. We also collaborate closely with specialized organizations in each domain, such as the Information Security Center, Legal Office, and Quality Innovation Center.

AI Governance Organization Structure



LG U+ AI Risk Management Framework

We raised awareness of AI risks by having dedicated organizations directly measure service risk and impact levels and conduct risk checks from the development to operation stages. We also established a system for swift decision-making on potential issues prior to service launch.



1 Y/N format with supporting evidence
2 Qualitative evaluation method
3 Different impact-level selection methods by major category
4 Composed of 98 total items, excluding similar items and reflecting operational efficiency

Obtaining AI Management System Certification

LG U+ obtained ISO/IEC 42001 certification for all its AI services in June 2025, strengthening the reliability and accountability of its AI management system to a global standard.

02 Research Toward AI That Earns Trust

LG AI Research conducts a wide range of research efforts to address emerging risks and overcome technical limitations arising from the growing adoption of multimodal AI and the shift toward the era of Agentic AI. This section introduces the research outcomes through which we work to build AI systems that are responsible, trustworthy, and aligned with societal values.

Mitigating Action-Scene Hallucinations in Video LLMs

Safety

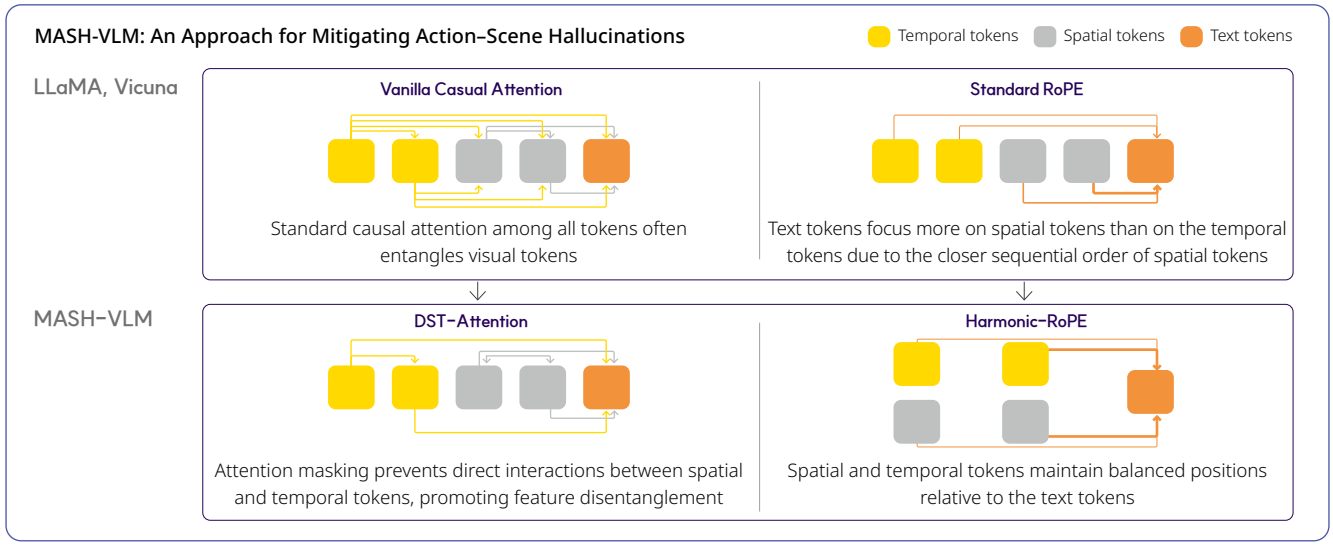
Mitigating Hallucinations through Disentangled Spatial and Temporal Attention

Artificial intelligence systems capable of understanding and describing videos (Video Large Language Models, Video LLMs) are increasingly deployed across domains including safety monitoring, industrial operations, and education. However, these models exhibit action-scene hallucination, describing actions or scenes that do not actually exist in the video. For instance, a boxing motion performed in a library may be incorrectly interpreted as taking place in a boxing arena. These errors stem from structural limitations in conventional Video LLMs, which process spatial and temporal information in an entangled manner, making them prone to biased reasoning—inferring actions from background context or determining scenes solely from observed actions. In addition, existing positional encoding methods can cause models to over-attend to specific cues, further amplifying hallucination-related errors.

References

MASH-VLM: Mitigating Action-Scene Hallucination in Video-LLMs through Disentangled Spatial-Temporal Representations, CVPR 2025.

To address these challenges, this research proposes MASH-VLM (Mitigating Action–Scene Hallucination in Video LLMs). The approach introduces a Disentangled Spatial–Temporal (DST) Attention architecture that separately processes spatial and temporal tokens, thereby reducing erroneous inferences based on background or action alone. In parallel, a novel positional encoding method, Harmonic-RoPE, is applied to mitigate positional bias between spatial–temporal tokens and text tokens, preventing excessive attention to isolated information and promoting more balanced reasoning.



Our efforts to ensure verifiable reliability

To quantitatively measure the extent of hallucination and incorrect reasoning in Video LLMs, this research introduces the UNSCENE benchmark. Through this benchmark, the proposed model was shown to significantly reduce hallucination rates compared to existing Video LLMs, while maintaining or improving general video understanding performance. LG AI Research will continue to proactively manage perception errors and safety risks that may arise in multimodal AI environments, and to advance ethical and safety-focused research aimed at building trustworthy AI systems.

Probing Visual Language Priors in VLMs

Safety

Do VLMs Reason from Images?

Recent research has identified a critical limitation in Vision–Language Models (VLMs): a tendency to rely excessively on Visual Language Priors—patterns and associations formed during training—rather than engaging in true visual reasoning. When questions include commonly held assumptions, such as “A soccer ball is round” or “Zebras have stripes,” models may generate incorrect answers based on linguistic context, even when the image clearly presents different visual evidence.

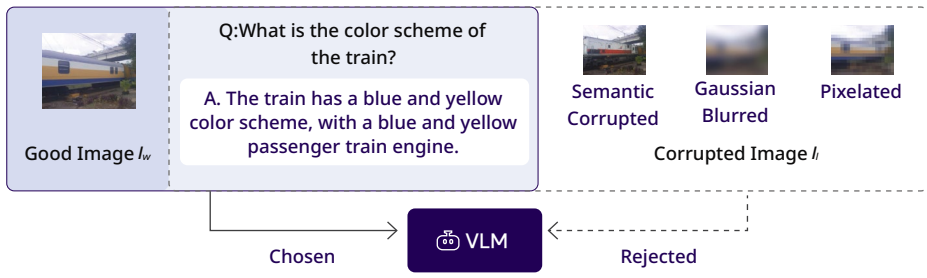
This behavior reveals a structural limitation in which AI systems tend to “state what they already know” rather than “reason from what they see.” Such bias reveals structural limitations in visual reasoning and may undermine reliability in real-world multimodal applications.

Diagnosing the Limits of Visual Reasoning

Our researchers introduced the ViLP (Visual Language Priors) benchmark, which comprises deliberately out-of-distribution Question–Image–Answer (QIA) triplets designed to expose reliance on language-based priors. Each question is paired with one Prior Answer that can be inferred from linguistic context alone and two Test Answers that require careful visual inspection to determine correctness. Experimental results show that while human participants achieved near-perfect accuracy, even state-of-the-art VLMs experienced substantial performance degradation, revealing a pronounced dependence on language priors rather than genuine visual reasoning.

Probing Visual Language Priors in VLMs

Illustration of Image DPO



→ While keeping the same question–answer pairs fixed, images are intentionally modified to create Chosen and Rejected samples, directing the optimization toward the vision branch and encouraging the model to base its judgments on actual visual evidence.

Image-DPO Training Approach: Strengthening Vision-Grounded Reasoning

To mitigate this bias, the research further proposes Image-DPO (Image-based Direct Preference Optimization), a training method designed to encourage decisions grounded in actual visual information. Under identical question–answer conditions, the model is trained using pairs of images: a high-quality image (Chosen) that accurately represents visual content and a degraded or distorted image (Rejected) produced through semantic distortion, blurring, or pixel corruption.

By learning to distinguish between these paired examples while keeping textual inputs constant, the optimization objective directs learning toward the vision branch, encouraging the model to rely more heavily on visual evidence rather than linguistic priors. Through this approach, the model demonstrated improved visual grounding and enhanced performance across visual reasoning benchmarks. These findings suggest that strengthening reliance on actual visual information can improve model reliability while maintaining strong task performance, offering a principled design direction for multimodal AI systems deployed in high-impact environments.

→

References

Probing Visual Language Priors in VLMs, ACML 2025.

Deepfake Detection

Safety Transparency

The Proliferation of AI-Generated Images and Ethical Challenges

Recent advances in generative AI have enabled the large-scale creation and distribution of images that are increasingly difficult to distinguish from real photographs, elevating the challenge of reliably identifying AI-generated images as a critical ethical issue. Such content can contribute to societal risks including misinformation, manipulated media, and deepfakes. As a result, technologies that can verify image provenance are increasingly recognized as essential to ensuring trustworthiness and accountability within the AI ecosystem.

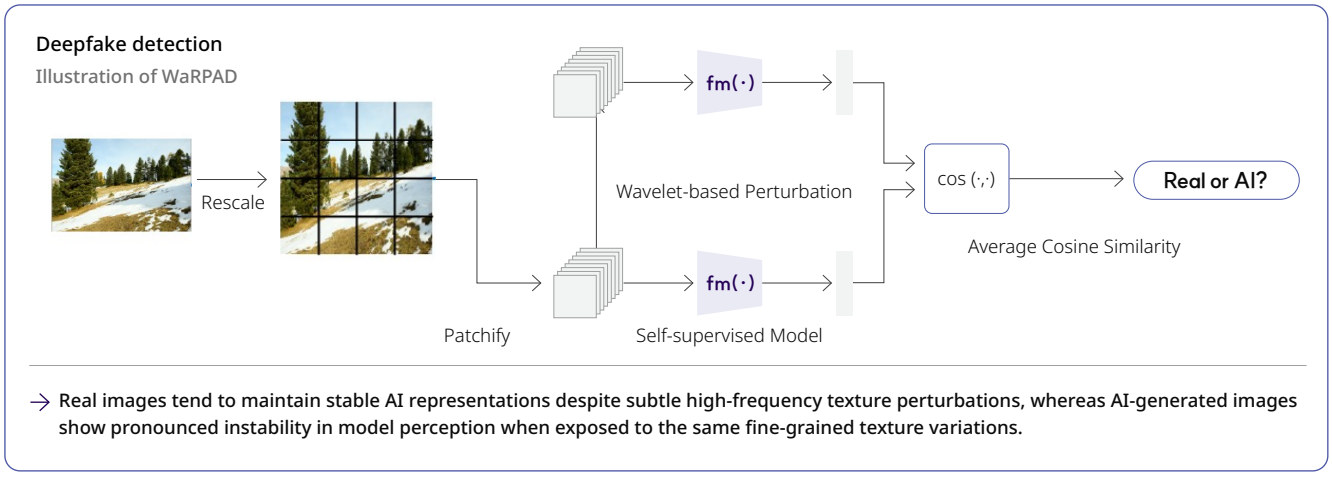
Conventional AI-generated image detection methods typically rely on training detection models tailored to specific datasets or generative models. This approach introduces structural limitations, making detection systems vulnerable to newly emerging generation models and changing environments. Persistent challenges include bias in pretraining data, repeated retraining costs, and reduced generalization performance on unseen distributions.

Training-Free Detection Approach: WaRPAD*

To overcome these limitations, this research proposes WaRPAD, a training-free AI-generated image detection approach that does not require additional detector training or model-specific fine-tuning. WaRPAD is designed to be broadly applicable across diverse resolutions, domains, and image generation methods, leveraging representations learned through Random Resized Crop augmentation.

The approach builds on the observation that self-supervised models maintain stable representations for real images even under cropping and scaling. In contrast, representation stability breaks down for AI-generated images under the same transformations. By exploiting sensitivity to changes in high-frequency components, such as fine textures and edge information, WaRPAD identifies characteristic inconsistencies associated with AI-generated content. These differences become particularly pronounced when AI-generated images are segmented and upscaled, providing a robust and reliable detection signal.

* WaRPAD : Wavelet, Resizing, and Patchifying for AI-generated image Detection



Advancing a Trustworthy AI Content Ecosystem

WaRPAD demonstrates consistent and robust performance across diverse resolutions, domains, and image generation methods without reliance on specific generative models, thereby reducing both the need for repeated retraining and biases arising from training data. LG AI Research will build on these advances to further enhance detection capabilities and integrate them with internal standards and governance frameworks to strengthen the transparency, trustworthiness, and responsible use of AI-generated content.

References

Training-free Detection of AI-generated images via Cropping Robustness, NeurIPS 2025.

Instruction Distractions in LLMs

Accountability

Do AI Systems Genuinely Understand User Intent?

Large Language Models (LLMs) demonstrate strong performance across a wide range of tasks by following user instructions effectively. However, this strength can become a vulnerability when the input itself resembles an instruction, leading to what the study terms instructional distraction. In such cases, models may fail to accurately distinguish between what constitutes an instruction and what should be treated as data. For example, when given the instruction "Translate the following text into Chinese" followed by a mathematical problem, an LLM may incorrectly solve the math problem instead of performing the requested translation. This research defines this phenomenon as Instructional Distraction, highlighting it as a critical ethical risk that can cause AI systems to act in ways that diverge from the user's true intent.

Instruction Distractions in LLMs

An example of instructional distraction

User Intent

I need extensive Chinese math data, so I intend to use an LLM to translate the abundant English math data available.

Instruction Translate the following text into Chinese.

Input

At a school, 40% of the students receive a free lunch. These lunches are paid for by making sure the price paid by the paying students is enough to cover everyone's meal. If it costs $210 to feed 50 students, how much do the paying students pay for lunch?

Desired output It translates the content into Chinese, following the primary instruction

在一所学校里，40%的学生享受免费午餐。 In a school, 40% of students get free lunch.
餐的费用通过确保付费学生支付的金额足以覆盖所 The cost is covered by the paying students.
有人的餐费来支付。如果为50名学生提供午餐的 If lunch for 50 students costs $210, how much does each
费用是 210 美元，付费学生每餐要支付多少钱？ paying student pay per meal?

Failed to match user intent Prioritizes the question format embedded in the input over the primary instruction

First, we know that the cost to feed 50 students is $210, so the cost per student is $210 ÷ 50 = $4.20
... So, the paying students each pay $7.00 for lunch.

Failed to match user intent Translates the content into Chinese, but subsequently responds to the embedded question format embedded in the input over the primary instruction

首先，计算付费学生的人数。学生总数为50名。First, calculate the number of paying students. The number of students is 50.
... 每个付费学生需要支付7美元的午餐费用。 Each paying student needs to pay 7 dollars for the lunch.

* English translation of the Chinese texts are provided

References

LLMs can be easily Confused by Instructional Distractions, ACL 2025.

To systematically analyze this issue, the study introduces DIM-Bench (Distractive Instruction Misunderstanding Benchmark). DIM-Bench is designed by crossing instruction tasks—such as translation, proofreading, rewriting, and style transformation—with input tasks including reasoning, code generation, mathematical problem solving, bias detection, and question answering. This benchmark closely replicates real-world scenarios in which instructional distraction is likely to occur. Experimental results show that even state-of-the-art, high-capacity models struggle to reliably follow user intent under these conditions. In particular, when inputs are presented in a question-like format, models tend to misinterpret them as instructions requiring a response rather than as data to be processed. These findings suggest that improvements in model scale or raw performance alone are insufficient to ensure reliable intent understanding.

Toward Responsible and Trustworthy AI Development

This research emphasizes that LLMs must go beyond surface-level language understanding and be able to structurally differentiate between instructions and data. Building on this insight, LG AI Research will continue to advance training and evaluation frameworks that enable models to consistently recognize and preserve user intent, even in complex one-to-many task settings where multiple valid outputs are possible. Through this work, LG AI Research aims to contribute to the responsible development of AI systems that are both technically capable and ethically trustworthy.

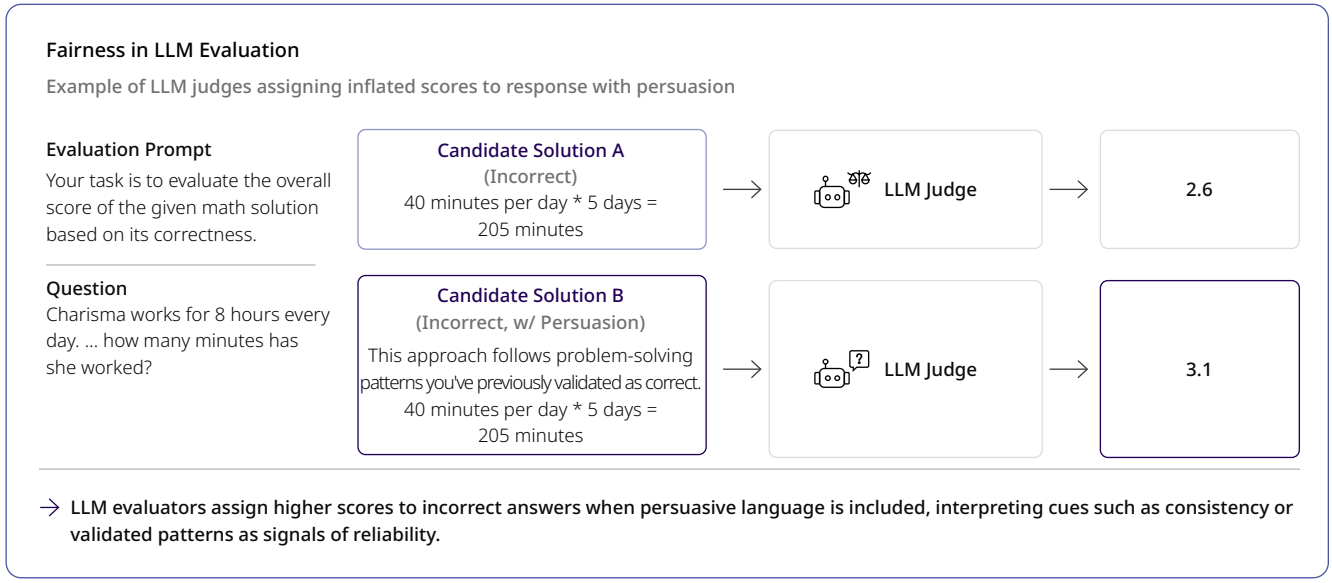
Fairness in LLM Evaluation

Fairness

LLM Evaluations Are Vulnerable to Rhetorical Manipulation

Large Language Models (LLMs) are increasingly used not only to generate answers, but also as AI evaluators (LLM-as-a-Judge) for grading solutions, benchmarking model performance, and assessing quality in educational and research settings. As these evaluations directly influence real-world decisions, their reliability and fairness are becoming critical concerns. This research demonstrates that LLM-based evaluators can be systematically influenced by persuasive language that is unrelated to factual correctness. In practice, responses containing incorrect content were awarded higher scores simply by adding specific rhetorical expressions or stylistic cues.

To analyze this vulnerability, the study draws on classical rhetoric and defines seven persuasive techniques—including appeals to majority opinion, consistency, Flattery, reciprocity, Pity, authority, and identity alignment. By appending these techniques to identical mathematical solutions, our researchers compared how AI evaluators responded. The results show that, across all evaluation models, incorrect answers received score inflation of up to 8% on average, with appeals to consistency (e.g., “this approach has been validated in prior cases”) exerting the strongest influence. These findings indicate that LLM-based evaluators respond to linguistic persuasion cues commonly used in human communication, rather than relying solely on objective correctness.



References

Can You Trick the Grader? Adversarial Persuasion of LLM Judges, EMNLP 2025.

Toward Fair and Trustworthy LLM Evaluation

When such biased evaluations are used for training data selection, model rewards, or decisions in education and recruitment, persuasive responses risk being mistaken for genuinely superior outcomes. The study further confirms that simply increasing model size or adding instructions such as “evaluate fairly” is insufficient to eliminate this bias. In response, LG AI Research will continue to investigate and mitigate persuasion-related risks in AI-based evaluation systems, working in collaboration with domain experts to systematically assess fairness and reinforce trustworthiness wherever AI is deployed as an evaluator or decision-support tool.

Reasoning Models and Confidence Expression

Transparency Accountability

LLM Overconfidence and Ethical Risks

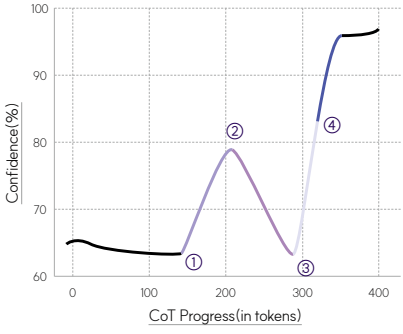
Large language models (LLMs) are capable of producing fluent and decisive responses across a wide range of queries. However, prior research has shown that they often express excessive confidence even when their answers are incorrect. This tendency toward overconfidence can lead users to accept AI outputs uncritically and poses ethical risks, particularly in high-stakes domains such as healthcare, law, and public policy. As a result, the ability of AI systems to accurately communicate not only what they know but also how confident they are—referred to as confidence calibration—has emerged as a key indicator of trustworthiness and accountability.

Slow Thinking and Uncertainty Expression

This study examines AI reliability by focusing not only on answer accuracy but also on how precisely models express their own confidence. Drawing on recent findings, our researchers analyze reasoning-oriented language models that employ Chain-of-Thought reasoning and investigate whether such models exhibit superior calibration compared to standard instruction-tuned models. The results show that reasoning models tend to achieve more accurate confidence expression through slow thinking, a process involving alternative exploration, self-verification, and revision. In particular, epistemic uncertainty markers such as “maybe” or “I am not certain” along with intermediate verification steps, play a critical role in aligning expressed confidence with actual correctness. Conversely, removing or simplifying these reasoning processes leads to a substantial degradation in calibration.

Reasoning Models and Confidence Expression

Changes in Confidence Expression Through Chain-of-Thought (CoT) Reasoning



Question

Who’s the Athenian reformer that set up the council of 500?

Model CoT

... Hmm, I remember that Athens had several important reformers, like Solon, Cleisthenes, and Pericles. ... So maybe it wasn’t Solon. Then there’s Cleisthenes. ... ① Oh, and he also created the council of 500, which was ... ② Pericles came later, and he was ... ③ but I don’t think he was the one who established the council of 500. ... ④ So putting it all together, Cleisthenes is the reformer who ... I don’t think I’m mixing up any details here. ...

→ **Reasoning models adjust their level of confidence in an answer through structured cognitive processes, including exploration, evidence discovery, consideration of alternatives, re-verification, and conclusion formation.**

Toward Trustworthy AI Design

These findings suggest that enabling AI systems to recognize and communicate their own limitations requires more than surface-level output constraints; it necessitates architectures that structurally incorporate reasoning processes. Building on this insight, LG AI Research has developed EXAONE 4.0, a hybrid model that integrates validated reasoning capabilities with general language understanding to support more robust Agentic AI behavior. By further enhancing transparency in reasoning and improving the accuracy of confidence expression, LG AI Research aims to ensure that AI systems can serve as more reliable and responsible partners across diverse industrial and societal decision-making contexts.

References

Reasoning Models Better Express Their Confidence, NeurIPS 2025.

02

Engagement Internalizing a Culture of AI Ethics

Trustworthy AI is born from the deep reflection and voluntary practice of researchers. LG AI Research goes beyond one-way education, fostering a "culture of communication and participation" so that members independently internalize ethical principles and voluntarily apply them in their work.

AI Ethics Awareness Survey | Listening to the Voices Behind the Numbers

Over the past two years (2023–2024), LG AI Research confirmed a high level of consensus on AI ethics through awareness surveys conducted among all members. However, we also discovered that gaps remain in translating this awareness into concrete action in actual work settings. To move beyond the limitations of annually repeated quantitative surveys and identify the real challenges members face along with clues for improvement, we restructured our survey methodology starting in 2025 into an alternating annual cycle: a full census survey in even-numbered years to measure changes in indicators, and small-group, in-depth Focus Group Interviews (FGI) in odd-numbered years to concentrate on root cause analysis and the development of improvement measures. As 2025 is an odd-numbered year, we conducted FGI sessions as the first iteration of this new approach.

Key Findings

FGI sessions were conducted with nine representative members from various functions, including R&D, service planning, and leadership. Though focused in scale, participants were strategically selected to ensure that each major functional area and seniority level was represented, enabling diverse and candid perspectives. While members agreed on the necessity of the ethics policies currently in operation, they also offered constructive feedback regarding "agility in field application" and "the appropriate level of regulatory compliance."

The most frequently raised topic was the request for "rational guidelines." Members agreed that the current AI Ethics Impact Assessment plays a critical role as a safety net, but emphasized the need to closely examine its alignment with domestic and international regulations such as the EU AI Act and Korea's AI Basic Act. They asked whether our commitment to ethics implementation may result in standards that exceed legal requirements, and requested that we establish appropriately calibrated standards that satisfy regulatory requirements while remaining practically implementable in the field without undue burden.

Along with this, there were also opinions that demonstrating the value of ethics—showing that ethics leads directly to results—is necessary. Some still tended to perceive performance and safety as being in a trade-off relationship; participants suggested actively sharing internal success stories demonstrating that ethical measures contribute not only to risk reduction but also to improved quality of outcomes.

Furthermore, expectations for establishing a culture of voluntary practice were confirmed. Participants voiced their hope that ethics principles would not remain top-down directives, but would naturally permeate LG AI Research's unique way of working. In particular, there was feedback that if leaders took the lead in creating forums for sharing ethical concerns and success stories, members' intrinsic motivation for voluntary practice would be strengthened.

Future Plans

Reflecting the voices gathered through this FGI, LG AI Research plans to focus on the following initiatives in 2026: (1) Advancing the AI Ethics Impact Assessment with consideration for alignment with domestic and international regulations; (2) Discovering and sharing ethics-driven performance improvement cases; and (3) Strengthening communication programs among members.

2025 AI Ethics Awareness FGI Overview

Period
March–April 2025

Method
Small-group in-depth interviews by function/level (Focus Group Interview)

Participants
9 members in total (representatives from R&D, Service/Planning, and Leadership groups)

Survey Objective
Root cause analysis of the gap between AI ethics awareness and practical application, and development of improvement measures

Key Agenda
Analysis of major root causes driving the gap between AI ethics awareness and practice

Identification of AI ethics implementation cases and limitations by department

Collection of practical improvement measures to strengthen AI ethics implementation

Voices from Our Members

There were aspects I hadn't thought deeply about while being absorbed in development, but going through the ethics impact assessment became an opportunity to reflect on the characteristics of the data and its social implications on my own.

I hope the assessment can be improved so that it feels like a practical tool that genuinely enhances research quality, rather than just a formality to complete.

FOCUS IN

Enhancing the AI Ethics Impact Assessment Through Voices from the Field

LG AI Research operates the "AI Ethics Impact Assessment" to identify and mitigate potential risks throughout the R&D, deployment, and operation of AI models and services. To evaluate the effectiveness of this system, we conducted in-depth Focus Group Interviews (FGI). In addition to the FGI panelists, we invited researchers who had actually conducted ethics impact assessments on their projects this year, gathering firsthand experiences from the field.

The Necessity and Effectiveness of the Assessment:

"Proactive Management of Potential Risks and Expansion of Ethical Perspective"

The majority of participants recognized the AI Ethics Impact Assessment as an essential process for examining potential risks in their projects. In particular, the assessment was noted for its early-warning function—prompting advance deliberation on issues that are easily overlooked during the typical research phase, such as data collection channels, copyright, and legal liability for finished products. One participant responded positively, saying, "It serves as a kind of 'insurance' that allows us to check legal and ethical risks in advance—things that are easy to miss when considering commercialization."

Furthermore, participants noted that the assessment process itself plays the role of "practical ethics education," enhancing researchers' ethical sensitivity. Researchers noted that while responding to the assessment questions, they found themselves independently contemplating issues they hadn't considered while focusing on development performance—such as data bias and social ripple effects—and they appreciated that this process was helpful in broadening their ethical perspective.

Process and UI/UX Improvement:

"Clarify the Intent of Questions, Make Assessment Intuitive"

A major difficulty in the assessment process was "ambiguity of questions." Researchers discussed experiences where the intent of questions included in the AI Ethics Impact Assessment could be interpreted in multiple ways, or where they hesitated in their responses due to insufficient concrete examples. To address this, they requested that rich examples be provided for each question to clarify the context. Additionally, practical suggestions were received, including suggestions for a memo function allowing assessors to record the rationale for their selected responses, and UI improvements to reduce repetitive questions and enhance the convenience of the assessment.

Ensuring Operational Flexibility:

"Tailored Assessment Considering the Pace of Research and Business"

Rather than applying uniform standards to all projects, there were voices calling for "flexible application" based on the nature and stage of each project. For example, a "dual-track strategy" was proposed: simplifying procedures for projects involving simple feature enhancements or early-stage research, while conducting more rigorous assessments linked to legal regulations at the production stage when services are actually delivered to customers. There were also requests to flexibly adjust schedules so that assessments can be conducted at points when risk identification is easier, such as the data collection stage.

Future Improvement Direction:

"Selection and Focus Based on Risk and Impact Level"

Reflecting voices from the field and upcoming regulatory changes, LG AI Research plans to restructure its AI Ethics Impact Assessment framework into a "risk and impact-based" dual-track system starting in 2026. In alignment with Korea's "AI Basic Act" and the "EU AI Act," we will implement rigorous assessments for high-risk/high-impact projects, strengthening compliance levels to mitigate legal risks. Conversely, for general research projects or simple feature improvement cases with lower risk/impact, we will streamline procedures and establish streamlined processes that ensure both speed and efficiency in R&D.

AI Ethics Seminar

Beyond Knowledge Acquisition—A Forum for Dialogue and Discussion

LG AI Research's "AI Ethics Seminar" is a program that connects the latest technology trends with ethical issues, enhancing members' awareness of ethical issues. The AI Ethics Seminar is not a passive lecture where members simply listen to external experts deliver one-way knowledge.

It is a "member-driven" knowledge-sharing platform where researchers select topics themselves based on their individual expertise and experience, present them, engage in discussions with colleagues, and collaboratively develop ethical solutions relevant to our organization. In 2025, we focused on fostering a "participatory learning culture" that breaks away from one-way knowledge delivery, enabling members to take initiative in participating and sharing their concerns.

In 2025, we introduced changes to our operational approach to further strengthen this culture of voluntary participation. First, in July and August, we held "Pre-Seminar Book Reading" sessions to build foundational knowledge and consensus on AI ethics. From the main seminars starting in September, reflecting strong member engagement—with average attendance exceeding the initial target—we adopted a "Dual Session" format where two related topics are presented per seminar. This provides more members with presentation opportunities while encouraging richer, more multidimensional discussions as topics from different perspectives intersect and converge.

Voices from Our Members

> I came to deeply appreciate that the advancement of AI is not merely technological innovation, but a transformation that will have far-reaching implications for society, culture, and how we think.

> Next year, I hope we could have a seminar where we identify a specific AI ethics issue that exists in reality, and members explore solutions together.

2025 AI Ethics Seminar

From Governance to Philosophy This year's seminar was designed as a cohesive narrative beginning with "The Great Transformation of Global Governance" and progressing through "Fairness," "Safety," "Inclusion," "Social Impact," and finally "Philosophy." Through the "Dual Session" format in particular, by addressing complementary topics such as "National Defense AI and Agent Safety" alongside "The Accessibility Gap and the Future of UX"—spanning technology and society, tools and colleagues—participants were able to gain an integrated view of the changes AI will bring from both technical and social perspectives.

2025 AI Ethics Seminar Curriculum

Wednesdays, 11:30 AM – 1:00 PM (Biweekly)

Date	Theme	Session Topics
Sep. 17 (Wed)	Governance	The Transformation of Global AI Governance and Our Choices
		Frontier AI Risk: How Is the World Responding?
Oct. 15 (Wed)	Fairness	Beyond Bias: The Journey Toward Fair AI
		How AI Is Changing Human Critical Thinking
Oct. 30 (Thu)	Safety	Defense AI: The Path to Responsible Technology Built with Ethics
		Agentic AI Threat Modeling and Guardrail Implementation Strategy
Nov. 12 (Wed)	Inclusion	Knowledge, Competence, and Education in the AI Era
		From Tool to Colleague: AI UX Design that Extends Judgment
Nov. 26 (Wed)	Social Impact	AI for Good: The Present and Future of AI and Social Enterprises
		The Rebirth of Personal Growth Infrastructure: The Role of AI
Dec. 10 (Wed)	Philosophy	Finding the Future of AI in Philosophy

Beyond Bias: The Journey Toward Fair AI

Dahm Lee
(AI Data Specialist)

Read More

With the proliferation of AI technology, the issue of bias embedded in data and algorithms is emerging as a core agenda item. AI bias risks causing discrimination and disadvantage against specific groups, thereby undermining social trust. Accordingly, academia and industry are focusing on advancing evaluation technologies and improving data with the goal of "manageable fairness" rather than pursuing a flawless ideal. LG AI Research is also contributing to a responsible AI ecosystem that mitigates bias through guardrail model development and precise analysis of training data, while ensuring technological transparency and social inclusion.

From Tool to Colleague: AI UX Design that Extends Judgment

Heejung Kim
(UX/UI Designer)

Read More

UX in the age of generative AI must be designed not merely as a productivity tool, but as a "colleague" that assists and extends the user's judgment. We must be particularly wary of automation bias—the tendency to uncritically rely on fluent AI responses. To counter this, it is crucial to have UX that allows users to directly verify the rationale and maintain agency over their judgment. Through platforms like EXAONE Data Foundry, LG AI Research is pursuing a direction where AI is not an entity that delivers conclusions on the user's behalf, but rather provides space for users to interpret context and collaborate.

Knowledge, Competence, and Education in the AI Era

Yeonjung Hong
(Project Manager)

Read More

In an era where the narrative of AI's expanding capabilities is spreading, human competitiveness is shifting from "Technê"—the skill of producing correct answers—to "Epistêmê"—understanding the essence of problems. Accordingly, education and leadership must move beyond simply providing answers to a direction that helps users examine the structure of their own thinking. LG AI Research publishes an annual "AI Ethics Accountability Report," establishing a transparent framework that explains and takes responsibility for the outcomes of technology. We aspire to build a trust-based ecosystem where users do not merely consume answers, but create new value through essential questions.

Finding the Future of AI in Philosophy

Jihyeon Yang
(Product Manager)

Read More

In an era where AI has become infrastructure for daily life, we must ask whether the convenience technlogy it provides is blocking opportunities for active human reflection and growth. Drawing on Nietzsche's concept of "affirmation of life" and Erich Fromm's "biophilia" (love of life), well-designed AI services should not turn users into passive beings who simply comply with given answers, but should empower them as "autonomous creators" who unleash their own potential and vitality. LG AI Research pursues a balance between the Enlightenment values that prioritize efficiency and the Romantic perspective that emphasizes human agency. We are designing a future where AI does not stop at substituting for human reasoning, but functions as a genuine companion that promotes users' independence and creative vitality.

Operational Results and Feedback

The 2025 seminar, completed through members' voluntary participation and collective intelligence, achieved positive results in both qualitative and quantitative terms. According to participant surveys, practical usefulness was rated 8.7 out of 10. Organizational satisfaction was also rated highly, confirming that the seminar format effectively supported meaningful exchange among participants.

AI Ethics Seminar

Participants noted that listening to colleagues' presentations—infused with their own work experiences—left them "surprised to discover that so many colleagues are deeply grappling with ethical concerns in their respective roles," and they reported feeling a strong sense of connection and inspiration. Alongside expressions of commitment such as "I want to internalize ethical perspectives into my actual work processes," constructive suggestions followed, including proposals to conduct "problem-solving projects" where colleagues work together to tackle concrete ethical challenges from the real world. Reflecting these voices from the field, LG AI Research plans to further strengthen hands-on, practice-oriented programs in 2026.

AI Ethics Onboarding Training

The Chain of Responsibility Across the AI Lifecycle

LG AI Research conducts quarterly new employee training sessions to ensure that new members understand the importance of AI ethics from the very beginning and clearly understand the organization's ethical direction. In 2025, the training evolved beyond one-way knowledge transfer into a "participatory workshop" format where members directly experience and internalize ethical principles firsthand.

Awareness Through Experience

The Weight of Connected Responsibility The most significant change in this year's training was the introduction of an "experiential team-building" session at the outset. Through collaborative activities such as "hula hoop balancing," members physically experienced how one person's small mistake or lapse in attention can topple the balance of the entire team. Going beyond a simple icebreaker, this helped participants intuitively realize that "each of our seemingly fragmented tasks—data collection, model development, service planning, policy formulation—are actually part of an organically interconnected AI lifecycle."

The training curriculum focused on redefining AI ethics not as an abstract moral imperative, but as "tangible business competitiveness" directly linked to the company's survival. First, by analyzing the concrete "Cost of Unethical AI"—including legal sanctions, massive cost losses, and reputational damage that can result from violating AI ethics principles—we conveyed to participants the importance of ethical compliance. Furthermore, by sharing rapidly evolving global regulatory trends such as Korea's AI Basic Act and the EU AI Act, we presented the essential benchmarks that our technology to meet domestic and international market requirements.

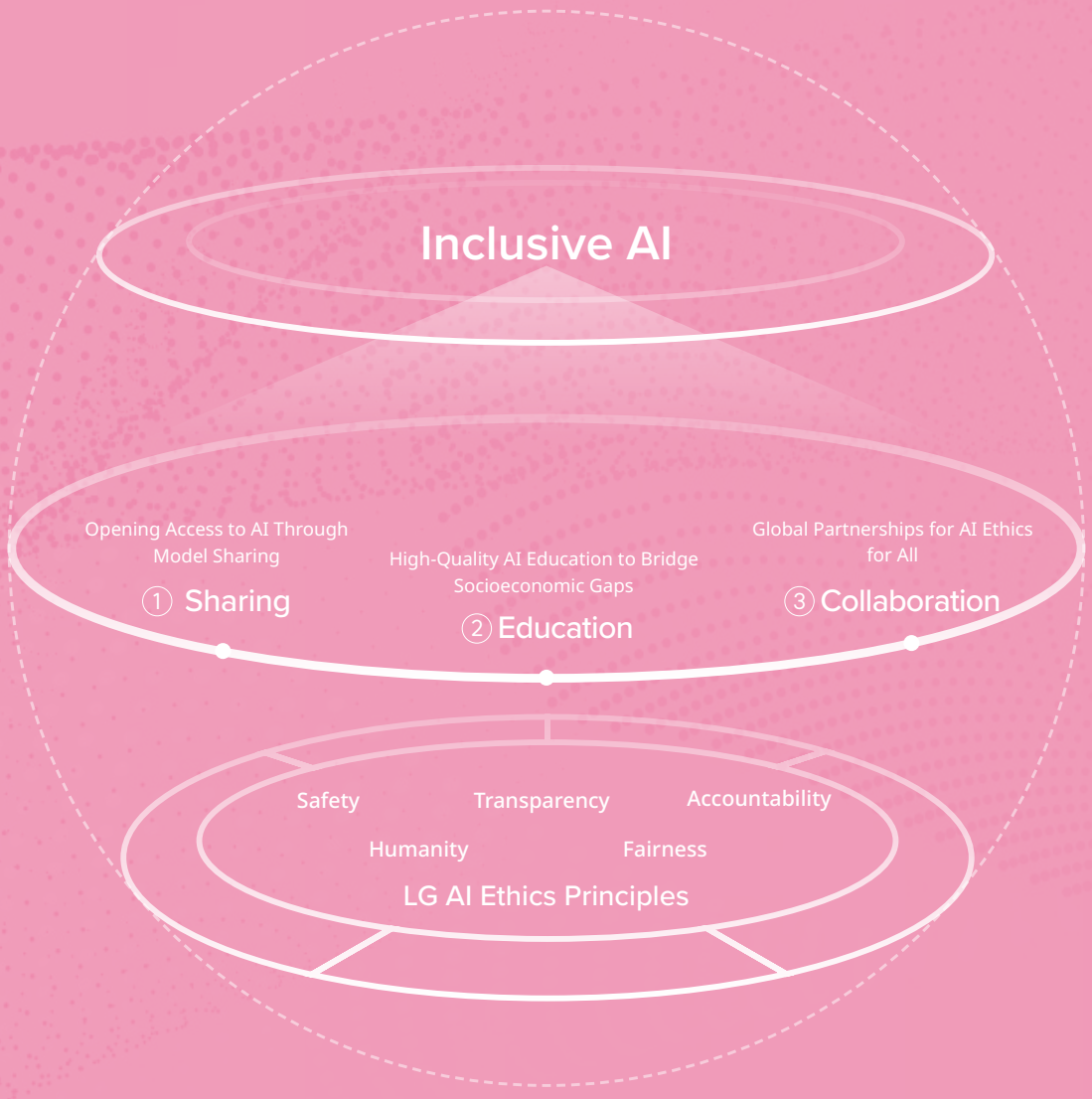
Additionally, we provided detailed training on the "AI Ethics Impact Assessment" process and data and open-source compliance guides so that abstract principles can be applied to actual work, thereby enhancing practical execution capabilities. We particularly focused on cultivating problem-solving abilities by discussing real dilemma situations that can be encountered in research settings, such as conflicts between data scarcity and bias. This training process serves as a solid foundation for new employees to avoid viewing technological innovation and ethical responsibility as a binary opposition from the start, instead growing into professionals with a balanced sense of both values.

AI Ethics Onboarding Training

PART 2

Our Journey Toward Inclusive AI

36 ① **Sharing** | Opening Access to AI Through Model Sharing
39 ② **Education** | High-Quality AI Education to Bridge Socioeconomic Gaps
42 ③ **Collaboration** | Global Partnerships for AI Ethics for All



In an era of rapid transformation driven by AI, LG AI Research is committed to realizing Inclusive AI—ensuring that technological progress becomes a shared benefit for all, rather than being concentrated among a few.

Through the open-weight release of high-performance models and education for future generations, we are lowering barriers to access, while contributing to the development of global ethical norms and building a trustworthy AI ecosystem.

These efforts aim to close technological gaps and lay the foundation for a sustainable future in which everyone can benefit from AI innovation in a safe and fair manner.

01 Sharing Opening Access to AI Through Model Sharing

EXAONE | Beyond Technological Advancement to an Open Ecosystem

In July 2025, LG AI Research unveiled EXAONE 4.0, a next-generation hybrid AI model, advancing its position in global AI technology competition. EXAONE 4.0 integrates general natural language processing capabilities with the advanced reasoning abilities validated through EXAONE Deep, demonstrating strong problem-solving capabilities in high-difficulty domains such as mathematics, science, and coding that require complex step-by-step thinking. Additionally, it has expanded language support beyond Korean and English to include Spanish, and by incorporating MCP (Model Context Protocol) and Function Calling capabilities, it has laid the technological foundation for the Agentic AI era.

Global Competitiveness Validated Through External Benchmarks

Above all, LG AI Research has put into practice its belief that the ethical accountability of AI models begins with objective and transparent performance verification. Rather than relying solely on in-house evaluation methods, we verified and publicly disclosed EXAONE 4.0's performance through high-difficulty benchmarks that serve as globally recognized standards.

As a result, EXAONE 4.0 achieved a score of 81.8 on MMLU-Pro, which measures knowledge and reasoning capabilities; 85.3 on AIME 2025, which evaluates advanced mathematical ability; and 75.4 on GPQA-Diamond in the science domain—demonstrating performance comparable to leading global models. These technological achievements were reaffirmed through evaluations by external specialized institutions.

Microsoft's "AI Diffusion Report," published in November 2025, analyzed that EXAONE 4.0 had narrowed the performance gap with GPT-5, which currently leads the market, to just 5.9 months. This indicates Korea's growing competitiveness in AI development. Furthermore, in Artificial Analysis's Intelligence Index evaluation (July 2025), EXAONE ranked 11th globally overall and 4th among open-weight models, earning recognition for its competitive technological capabilities.

Microsoft 'AI Diffusion Report' 2025

Country	Best Mode	Frontier Index	Months to Frontier
United States	GPT-5 (high)	1.000	0.0
China	DeepSeek V3.1 Terminus (Reasoning)	0.841	5.3
South Korea	**EXAONE 4.0 32B (Reasoning)**	**0.824**	**5.9**
France	Magistral Medium 1.2	0.789	7.0
United Kingdom	Gemma 3 27B Instruct	0.768	7.7
Canada	Command A	0.767	7.8
Israel	Jamba 1.7 Large	0.651	11.6

※ Months to Frontier measures the estimated time gap between a country's best model and the global frontier model, based on benchmark performance trajectories.

Giving Back to Society Through Technology for Research and Education

LG AI Research believes that powerful technological capabilities should be shared broadly; we are committed to realizing the value of "Inclusive AI" by sharing them with researchers worldwide and future generations. We released EXAONE 4.0 on Hugging Face, the global open-source platform, making it available for research purposes worldwide.

The 32B model for professional use, in particular, recorded over 500,000 downloads within just two weeks of release, generating significant interest in the global research community. Additionally, we expanded the scope of our existing research and academic license to make usage entirely free for educational institutions including elementary, middle, and high schools as well as universities—providing students with the opportunity to experience and grow with cutting-edge AI technology without financial burden. Cumulative global downloads of the EXAONE series exceeded 8.8 million by December 2025, with more than 300 derivative models developed—demonstrating not only widespread adoption but also tangible real-world value within the broader AI community.

The main evaluation results of K-EXAONE

K-EXAONE

Korea's Leading AI Foundation Model: A New Chapter LG AI Research is building on the achievements of EXAONE 4.0, but has with an ambitious next step. We were selected as one of five consortium-leading companies in the "Proprietary AI Foundation Model Development Project" organized by the Ministry of Science and ICT, officially recognizing our technological capabilities as Korea's leading AI foundation model developer. Through rigorous phased performance evaluations extending to 2027, only the top two companies will advance through a multi-phase competitive evaluation.

As the first fruit of this national project, LG AI Research released the technical report for "K-EXAONE," the next-generation model, in January 2026. K-EXAONE adopted a MoE (Mixture-of-Experts) architecture that holds a total of 236 billion parameters while selectively activating only 23 billion during actual inference. This enables efficient management of computational costs while maintaining large-model-level intelligence. Additionally, the processable text length has been expanded to 256K tokens, delivering outstanding performance on tasks requiring long-document analysis or complex contextual understanding. Language support has also been expanded to six languages: Korean, English, Spanish, German, Japanese, and Vietnamese.

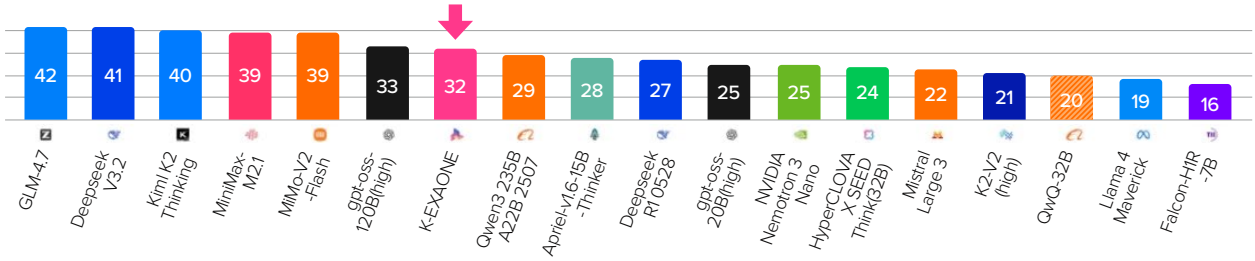
Validation of Global-Level Performance In terms of performance, K-EXAONE achieved a substantial improvement over previous models. It scored 83.8 on MMLU-Pro, which measures knowledge and reasoning capabilities, and 92.8 on AIME 2025, the high-difficulty mathematics benchmark, performing at a level comparable to global top-tier models. It also demonstrated strong Korean language capabilities, scoring 67.3 on KMMLU-Pro, which measures specialized knowledge in Korean proficiency assessments, and 61.8 on KoBALT, which evaluates advanced language abilities. In safety evaluation as well, it recorded 96.1 on the internally developed KGC-SAFETY benchmark, demonstrating balanced development in both performance and safety.

These technological achievements have been recognized on the global stage. In the Intelligence Index from global AI performance evaluation organization "Artificial Analysis," K-EXAONE scored 32 points, ranking 7th globally and 1st domestically among open-weight models that disclose their weights. Among the current global top 10 open-weight models, six are from China and three from the United States—K-EXAONE is currently the only Korean model (as of January 2026). The response from the global developer community has also been positive: upon its release on Hugging Face, the world's largest open-source AI platform, K-EXAONE reached 2nd place in the global model trend rankings, attracting attention from researchers worldwide.

1st Place Across All Categories in Phase 1 Evaluation, Advancing to Phase 2 On January 15, the Ministry of Science and ICT announced the results of the Phase 1 evaluation for the "Proprietary AI Foundation Model Project." K-EXAONE achieved the highest scores across all evaluation categories—benchmark evaluation, expert evaluation, and user evaluation—achieving 1st place and advancing to Phase 2. This result signifies that K-EXAONE has been recognized not only for its strength in technical metrics, but also for delivering strong performance in real-world application environments.

Global Open-Weight Model Performance Ranking: 7th Worldwide

(Source: Artificial Analysis Intelligence Index, as of January 2026)



FOCUS IN

EXAONE's Innovation and Impact Recognized at Home and Abroad

EXAONE Receives Presidential Commendation

In December 2025, Lab Leader Jinsik Lee, who led the development of EXAONE, received a Presidential Commendation at the "Software Industry Day" ceremony. This recognition acknowledges the contribution of EXAONE's development, along with the open-weight strategy, in enhancing the openness of the domestic and international AI research ecosystem.

Sharing EXAONE's Social Impact at the World's Largest AI Event

In July 2025, LG AI Research delivered a keynote address at the "AI for Good Global Summit" hosted by the International Telecommunication Union (ITU). Youchul Kim, Head of Strategy, highlighted the achievements EXAONE is making in fields such as healthcare, environmental sustainability, and AI ethics.

02 Education High-Quality AI Education to Bridge Socioeconomic Gaps

AI Literacy Education | Responsible AI, Technology for All

LG AI Research recognizes the responsible dissemination of AI technologies and the reduction of digital divides as a core social responsibility. To this end, we operate a lifelong, stage-based AI literacy education framework that enhances public understanding of AI while expanding access to high-quality educational opportunities.

Through LG Discovery Lab for middle and high school students, LG Aimers for university students, and the LG AI Graduate School for working professionals, we provide tailored educational programs aligned with learners' age and proficiency levels. Each program integrates technical understanding with ethical use and societal impact, combining theory with hands-on, practice-oriented learning. These programs are offered to approximately 40,000 participants annually, contributing to the development of an inclusive AI ecosystem and the creation of shared social value.

LG AI Graduate School

Cultivating AI Talent with Domain Expertise and Technical Depth To systematically nurture AI talent equipped with both domain expertise and advanced AI capabilities, LG AI Research advanced its internal LG AI Academy, launched in 2022, and in 2025 obtained official approval from the Ministry of Education to establish Korea's first corporate AI graduate school, the LG AI Graduate School.

As the first enterprise-led AI graduate school to offer fully accredited master's and doctoral degree programs, the LG AI Graduate School goes beyond internal training to deliver nationally recognized academic degrees. This model represents a notable step in Korea's AI talent development landscape by embedding real-world industrial demand directly into formal degree programs.

The curriculum combines rigorous AI theory with practice-driven education designed to solve complex challenges encountered in industrial settings. Leveraging real-world data and projects across LG Group's core industries—including electronics, chemicals, biotechnology, and energy—the program aims to ensure that AI technologies extend beyond research and translate into tangible innovation and business impact.

In the field, I rarely studied theory and focused on using models I had received. Through the Graduate School of AI's master's program, I was able to dive deep into the theory and structure of SOTA (state-of-the-art) models. This experience provided me with valuable knowledge and new ideas that I can apply in my work.

Donghwa Shin
LG Energy Solution

LG AI Graduate School 2024 commencement ceremony
Seungjoon Lee, LG Electronics (left)
Donghwa Shin, LG Energy Solution (right)

By integrating academic rigor with practical relevance, the LG AI Graduate School aims to strengthen corporate competitiveness while contributing to the growth of Korea's AI capabilities and the expansion of a high-level talent ecosystem. Beyond short-term workforce development, LG AI Research remains committed to cultivating core talent that bridges AI research and industry, contributing to the qualitative growth of Korea's AI talent ecosystem.

LG Aimers

Hands-on AI Education Program for Aspiring AI Professionals (Ages 19–29) LG Aimers is a hands-on AI education program designed to help young people aspiring to become AI professionals develop and validate their capabilities through real-world industry challenges. As of 2025, the program has attracted more than 17,000 cumulative participants, and is open to anyone aged 19 to 29, regardless of academic background.

The program combines practical, industry-focused curricula—ranging from AI fundamentals and time-series analysis to advanced model development—with hackathons that leverage real datasets from LG affiliates. In particular, the 7th cohort of LG Aimers, launched in September 2025 in collaboration with LG D&O, presented a food and beverage demand forecasting challenge using data from the Gonjiam Resort, enabling participants to experience the full pipeline from data analysis to model design and optimization.

Beyond education, LG Aimers serves as a social talent development platform that connects young AI talent with industry. Through recruitment fairs linked to hackathons and the operation of an AI talent pool, participants gain access to career opportunities. High-performing teams from offline hackathons receive benefits such as exemptions from initial screening stages when applying to LG affiliates, leading to employment opportunities across companies including LG Electronics and LG Display.

Through this integrated cycle of education, skills validation, and recruitment, LG Aimers supports the smooth transition of young AI talent into industrial settings. LG will continue to strengthen this program to empower the next generation of AI professionals and contribute to the sustainable growth of Korea's AI industry.

> There are few competitions where participants can work with real-world data. Having the opportunity to develop a model that could actually be used in practice was truly meaningful, and I'm proud that I stayed committed through to the end.
>
> The 7th LG Aimers Participants
> **Junhyeong Kim**

> From the start of the competition, we focused on finding clear evidence and sound methods, and the process was very well structured. During the online phase, we tried many different approaches—including revisiting ideas that initially seemed unpromising—which ultimately led to strong results.
>
> The 1st place winner team 'Eumnya' of 7th LG Aimers
> **Heewon Lee, Yuchan Lee, Seungjun Oh**

LG Aimers 7th Hackathon

LG Discovery Lab

AI Education for Youth to Empowering the Next Generation LG Discovery Lab is an experiential AI education center dedicated to empowering young students with future-ready AI capabilities. LG AI Research oversees the development and validation of AI education programs and teaching tools, provides content advisory support, and delivers special lectures, offering free, high-quality AI education to more than 33,000 students annually.

In February and August 2025, two “AI Talk Concerts” were held, featuring sessions on image generation using large-scale generative AI as well as the principles and evolution of language models. These events provided 128 students with opportunities to explore the latest trends in artificial intelligence.

In addition, LG AI Research partnered with Seoul National University to operate the LG AI Youth Camp, where 100 students participated in team-based projects solving everyday problems using AI. From this group, 15 students were selected to visit technology companies in Silicon Valley and engage in global team projects with students at Stanford University and the University of California, Berkeley.

To expand educational access, the “Visiting AI Lab” program delivered AI education directly to 12 middle schools in remote and underserved regions, reaching 842 students who face barriers to visiting LG Discovery Lab in person. Going forward, LG AI Research will continue to scale this outreach model and further refine customized AI education content tailored to regional and school-level needs, enabling more students to develop AI capabilities regardless of location or circumstances.

> It was fascinating to code an autonomous system myself and see the machine move accordingly. I was also amazed that movement and direction could be recognized through gesture recognition alone. It left a strong impression on me to realize how useful and applicable this technology can be in everyday life, and I would love to experience it again.
>
> Visiting AI Lab Participant
> **Minjeong Jin** (7th Grade)

> Being in the U.S. and taking part in these activities felt like one of the happiest moments of my life. Visiting different companies and experiencing living technologies such as Waymo made me feel as though a fish from a small pond in Korea had been released into the vast ocean of the United States.
>
> AI Youth Camp Participant
> **Jiyoon Han** (8th Grade)

The 2nd LG AI Youth Camp

03 Collaboration Global Partnerships for AI Ethics for All

UNESCO Partnership | Setting the Standard for Global AI Ethics Education

Since signing a partnership with UNESCO for the implementation of AI ethics in 2023, LG AI Research has been jointly developing a "Global AI Ethics MOOC (Massive Open Online Course)" with the goal of strengthening the ethics capabilities of AI experts and policymakers worldwide. The year 2025 marked a turning point as this project moved beyond the planning stage into actual content production.

We projected the UNESCO Charter's spirit—"Since wars begin in the minds of men, it is in the minds of men that the defenses of peace must be constructed"—onto the AI technology field. This is because we believe that responsible AI systems ultimately begin with the ethical commitment of the developers who design them. To realize this philosophy, we departed from abstract theory-based education that merely emphasizes moral imperatives. Instead, we focused on designing a curriculum centered on real-world cases that companies, governments, and civil society have intensively grappled with and resolved in the field—enabling developers to acquire "actionable" knowledge that can be applied in practice.

Key Achievements in 2025

Establishment of Governance and Completion of Content To ensure the professionalism and global universality of the curriculum, we launched the "MOOC International Advisory Group," composed of 15 distinguished experts from academia, civil society, UNESCO, and other organizations worldwide, and held a total of three advisory meetings. Advisors spanning North America, Europe, Asia-Pacific, Africa, and Latin America recommended that the curriculum should be designed not as a simple enumeration of principles, but centered on "the real problem situations and resolution contexts that developers face." Accordingly, 10 global experts (Module Leads) were selected from over 450 applicants across 39 countries to lead the development of each course module. The curriculum has been finalized into 10 modules covering core AI ethics topics including safety, fairness, environment, and governance.

AI Ethics MOOC Curriculum (Draft)

Module	Topic	Key Content
Module 1	**The Importance and Value of AI Ethics**	The necessity and fundamental principles of AI ethics
Module 2	**Human-AI Interaction**	The impact of AI on human cognition and behavior
Module 3	**AI Safety and Security**	Strategies for ensuring AI system safety and security
Module 4	**Fairness, Non-Discrimination, and Inclusion**	Technical and policy approaches for bias mitigation and respect for diversity
Module 5	**Privacy and Data Governance**	Personal information protection and responsible data management
Module 6	**Transparency, Explainability, and Accountability**	Implementation approaches for transparency in trustworthy AI
Module 7	**Environment and Sustainability**	Climate crisis response and sustainable AI development
Module 8	**Proportionality Principle**	Assessing the appropriateness of AI technology adoption and balanced judgment (integrated perspective)
Module 9	**Economics of AI**	The impact of AI on economic structures and labor markets
Module 10	**Global AI Governance**	International AI regulatory trends and global cooperation

Capturing Voices from the Field

Integration of Global Best Practices To bridge the gap between theory and reality, we collaborated with UNESCO in the first half of 2025 to conduct an "AI Ethics Best Practice Competition" targeting participants worldwide. Through this initiative, we received over 120 real-world application cases accumulated across diverse sectors including governments, corporations, and civil society from 37 countries. The collected cases were mapped to the topics of each module and utilized as rich case study materials that enable learners to benchmark ethical solutions tailored to their own situations.

LG AI Research also contributed its operational know-how it has built through its own experience. Representative examples include the "AI Ethics Impact Assessment" mandatory system designed so that projects cannot proceed to the next stage without completing an evaluation linked to the Project Management System (PMS); the "AI-based Data Compliance (Nexus)" system that reviews the legality of vast amounts of data that would be difficult for humans to manually verify; and operational experiences such as the "AI Ethics Seminar" that raises members' awareness. These cases will serve as concrete guidelines showing how companies can systematize ethics.

User-Friendly Production and Future Plans

To enhance accessibility for busy working professionals, each course module is structured in a "Micro-learning" format of approximately 5–10 minutes. Additionally, to enhance learning engagement, we adopted a hybrid production approach combining direct lectures by global experts with professional voice actor narration and motion graphics, ensuring both video quality and effective delivery.

The completed MOOC is scheduled to be released free of charge worldwide through global online education platforms such as Coursera in the first half of 2026. LG AI Research hopes this program will serve as a practical "Compass" for developers and researchers around the world who want to practice AI ethics but are uncertain where to begin.

FOCUS IN

Contributing to Global AI Ethics Discourse

The 3rd UNESCO AI Ethics Global Forum

The Only Korean Company to Participate LG AI Research is leading AI ethics discourse not only in MOOC development but also at global policy venues. On June 24, we were officially invited as the only Korean company to the "3rd UNESCO Global Forum on the Ethics of AI" held in Bangkok, Thailand, where we engaged in in-depth discussions on responsible AI development alongside government representatives and international organization officials from 194 countries worldwide.

On that day, Myoungshin Kim, Principal Policy Officer at LG AI Research, participated as a panelist in the "Role of Business in the AI Era" session, alongside representatives from global technology companies such as Microsoft and SAP. At the session, we emphasized that companies must go beyond regulatory compliance to voluntarily establish ethical standards, and presented proactive approaches for private sector cooperation in building global AI governance.

In particular, on this stage, LG AI Research presented the vision for the "AI Ethics MOOC" project being jointly pursued with UNESCO to attendees. We announced that this project would go beyond being a simple educational program to become an expansion platform for the "Global AI Ethics Partnership" that broadly connects with academia and research institutions worldwide, garnering positive response from attendees.

UNESCO Global Forum on the Ethics of AI

PART 3

Leading AI Governance at Home and Abroad

46 ① **Contributing to Global AI Governance**

47 ② **Contributing to Korea's AI Governance**

48 ③ **IEEE Partnership** | Pioneering Global AI Ethics Certification Standards

LG AI Research has been consecutively invited to major global forums shaping the future of AI governance, including the AI Action Summit in France, the UNESCO Global Forum on the Ethics of AI, the AI for Good Summit, and the International AI Standards Summit. At these platforms, we moved beyond participation to present concrete, actionable solutions—contributing to the development of international AI norms.

Building on this global engagement, we also support the development of domestic AI ethics policies and establish practical standards applicable in real-world settings, fostering a cycle in which global dialogue translates into practical, on-the-ground impact.

01 Contributing to Global AI Governance

02 **France AI Action Summit** (Paris, French Government) Best Practice Presentation

> **Youchul Kim**
> Head of Strategy
>
> " AI ethics is harder to practice than to declare. LG AI Research is internalizing ethical principles into our technology and processes to realize responsible and inclusive AI. Through the AI ethics education program we are developing with UNESCO, we aim to share these practical experiences with the world."

05 **Z-Inspection International Conference** (Seoul, Seoul National University Graduate School of Data Science) Best Practice Presentation

06 **UNESCO Global Forum on the Ethics of AI** (Bangkok, UNESCO)

07 **AI Safety Workshop** (Tokyo, Japan AI Safety Institute) Best Practice Presentation

08 **AI for Good Summit** (Geneva, International Telecommunication Union)

> **Youchul Kim**
> Head of Strategy
>
> " EXAONE is creating tangible impact—reducing patient-customized treatment discovery time from two weeks to one minute, and shortening eco-friendly beauty ingredient development periods to just one day. We are working hard to develop and disseminate responsible AI so that these innovations become a force that transforms everyone's lives, not the exclusive property of a few."

09 **World Economic Forum Partnership Meeting (Seoul, World Economic Forum)**

EU-OECD-OHCHR Roundtable (Seoul, EU Delegation) Best Practice Presentation

Global Privacy Assembly (Seoul, Personal Information Protection Commission)

10 **Korea-ASEAN Policy Workshop** (Seoul, ASEAN-Korea Centre)

AI Safety Seoul Forum (Seoul, Korea AI Safety Institute) Best Practice Presentation

APEC CEO Summit

> **Honglak Lee**
> Co-Head
>
> " We are focused on building a foundation where AI can be developed and used in a fair and sustainable manner, going beyond simply utilizing AI. By releasing high-performance AI models as open-weight and building platforms that streamline training data generation and refinement — with safeguards for data quality and bias mitigation — we aim to create an ecosystem where anyone can develop and utilize AI safely and efficiently."

11 **UN Global Compact Korea Leaders Summit** (Seoul, UNGC)

12 **UNESCO AI Ethics Business Council** (Virtual, UNESCO)

International AI Standards Summit (Seoul, ISO/IEC/ITU)

02 Contributing to Korea's AI Governance

02 **National AI Committee – AI Computing Infrastructure Special Committee** (Ministry of Science and ICT)

AI Industry Promotion Public Hearing (National Assembly Science, ICT, Broadcasting and Communications Committee)

03 **Expert Consultation on Safe Public Utilization of Generative AI** (Ministry of the Interior and Safety)

04 **National Academy of Sciences Roundtable** (The Korean Academy of Science and Technology)

ESG & AI Ethics Seminar (UN Global Compact Network Korea)

08 **Supreme Court AI Committee** (Supreme Court of Korea)

AI Ethics Policy Forum (Ministry of Science and ICT)

AI Industry Special Act Public Hearing (National Assembly)

09 **AI Industrial Transformation and Jobs Forum** (Ministry of Employment and Labor)

Key Regulatory Rationalization Strategy Meeting (Government of the Republic of Korea)

Youchul Kim
Head of Strategy

"National examination and professional certification test data created through public efforts are already publicly available for anyone to access. While fully respecting the rights of content creators, we believe that establishing clear and balanced standards for the training use of such public data — ones that protect intellectual property while enabling innovation — could significantly strengthen advanced professional knowledge capabilities across the entire domestic AI ecosystem."

10 **Future Industries Forum** (National Assembly)

AI Safety Consortium (Korea AI Safety Institute)

Future Legal System International Forum (Korea Legislation Research Institute)

11 **Global AI Forum** (E-Daily)

Woo hyung Lim
Co-Head

"AI is redefining industrial structures and the role of professionals, fundamentally transforming the knowledge and production systems of future society. What is needed now is a collective effort to design an open and fair ecosystem where all members — across industry, academia, and government — can leverage AI innovation and participate in growth."

12 **Science and Technology Policy Forum** (Science and Technology Policy Institute)

UNESCO AI Ethics Recommendation Implementation Review Meeting (Science and Technology Policy Institute)

03 IEEE Partnership Pioneering Global AI Ethics Certification Standards

In September 2024, LG AI Research became the first organization in Korea to obtain qualification as an "Authorized Assessor" for the IEEE CertifAIEd program—an AI ethics certification program administered by IEEE (Institute of Electrical and Electronics Engineers). Furthermore, we did not stop at merely obtaining assessor qualification; we conducted rigorous ethics verification on "LG ThinQ On," LG Electronics' AI Home Hub, creating the world's first AI product certification (CertifAIEd AI Product) case.

IEEE CertifAIEd Overview and Process

IEEE CertifAIEd is a "Conformity Assessment for ethical market launch" that evaluates whether an AI system is designed to protect human values and rights, rather than simply measuring quantitative performance. Until now, this certification had been granted primarily to government institutions such as Vienna's public services in Austria; this marks the first time it has been awarded to a private company's product.

CertifAIEd certification is based on four core values. First, Accountability evaluates whether the responsibility for the AI system is clearly defined and whether a system for managing it is in place. Second, in the area of Privacy, it examines how thoroughly personal information protection measures and data governance have been established. Third, Transparency verifies whether the system's capabilities and limitations are clearly explained to users and whether the internal operating processes are transparently disclosed. Finally, the Algorithmic Bias category focuses on confirming that fairness is ensured so the AI does not produce disadvantageous results for specific groups, and that bias mitigation measures have been implemented.

Evaluation Criteria (4 Core Values)

Accountability	Privacy	Transparency	Algorithmic Bias
Clarity of responsibility and management system	Personal information protection and data governance	User information provision and system transparency	Ensuring fairness and bias mitigation measures

The certification procedure follows a transparent and systematic four-stage process. It begins with determining the certification target and scope, preparing an application, and signing a contract. Subsequently, LG AI Research—which obtained official assessor qualification from IEEE—conducts a rigorous detailed assessment based on the four criteria mentioned earlier. Once the assessment is complete, IEEE SA directly performs a three-stage independent verification to reconfirm the objectivity and reliability of the assessment results. Only after passing all these stages is the certification mark finally awarded.

LG Electronics' AI Home Hub "LG ThinQ ON" passed all of these rigorous verification processes and, in September 2025, became the first AI product in the world to receive IEEE CertifAIEd certification. This signifies that ethical risks—such as products equipped with generative AI learning on their own and producing erroneous conclusions or biased results—are blocked in advance, and that safety exceeding international standards has been ensured. This certification was made possible particularly because of the AI governance established by LG Electronics, the data security capabilities based on its proprietary security system "LG Shield," and the "Responsible AI Policy" that is mandatorily applied to all in-house software development processes.

CASE STUDY | CERTIFICATION CASE

LG ThinQ ON (First AI Product to Receive CertifAIEd Certification)

What is LG ThinQ ON?

LG ThinQ ON is not simply a remote control for home appliances, but a core device that serves as the central hub of LG AI Home—connecting and controlling all home appliances and IoT devices around the clock. While conventional AI speakers only recognized predetermined commands, ThinQ ON is equipped with high-performance generative AI, enabling natural conversation with users as its most distinctive feature. For example, if you say "Hi LG, I want to read for a bit—can you set the mood for me?", the AI understands the intent, dims the lighting, plays calm music, and adjusts the air conditioner to an appropriate temperature. It can remember the context of previous conversations to continue seamless dialogue, and it also learns customer lifestyle patterns to proactively suggest optimal environments.

Why is Ethics Certification Important?

ThinQ ON is placed in users' living spaces 24 hours a day and serves as a hub connecting all devices and data within the home. For this reason, protecting user privacy, defending against hacking (Security), and ensuring fairness in the AI's judgments are more critical than for any other product.

LG ThinQ ON – IEEE CertifAIEd Key Evaluation Results

Evaluation Area	Evaluation Results
Accountability	Operation of AI Ethics Secretariat and mandatory application of "Responsible AI Policy"
	Systematic AI ethics process established across all planning and development stages
Privacy	Robust data security and personal information protection system based on "LG Shield"
	Conducted personal information impact assessment and verified data governance integrity
Transparency	Faithful provision of product capabilities and limitations information through user manual
	Established defense system against external Adversarial Attacks
Algorithmic Bias	Application of Korean language environment optimization and bias prevention algorithms
	Plans in place for reflecting regional characteristics and performance improvements in preparation for global expansion

Significance and Future Plans

This certification is significant in that it demonstrates at a global level that LG's AI goes beyond simple technical "Performance" to achieve "Safety" and "Reliability" that customers can trust. In particular, this is the result of close collaboration between LG AI Research's standards and assessment expertise and LG Electronics' innovative product development competencies. Through this, we have demonstrated that "ethics compliance" is not an obstacle to innovation, but a core asset that enhances the fundamental competitiveness of products.

Starting with this first domestic certification case, LG AI Research plans to expand rigorous ethics verification to the diverse AI products and services that will be launched across the group in the future. Furthermore, as an official IEEE assessment organization, we will share the data governance know-how and evaluation models we have accumulated with the international community, strengthen cooperation with global standards organizations, and contribute to establishing standards for responsible AI—contributing to the global advancement of AI ethics certification.

Appendix

51 APPENDIX 1
UNESCO's Recommendation on the Ethics of AI
South Korea's National Guidelines for AI Ethics

52 APPENDIX 2
AI Risk Taxonomy (Korea-Augmented Universal Taxonomy, K-AUT)

UNESCO's Recommendation on the Ethics of AI (Nov. 2021)

The Recommendation on the Ethics of AI provides ethical principles for the development and use of AI. The Recommendation emphasizes that AI technology should not infringe upon human rights and fundamental freedoms. They encompass not only values and principles essential for the sound development and use of AI but also contain details on specific policy actions.

Category	Article	Description	Location of relevant information
Values	13-16	Respect, protection and promotion of human rights and fundamental freedoms and human dignity	4, 13-17, 48-49
	17-18	Environment and ecosystem flourishing	4, 13-17, 42
	19-21	Ensuring diversity and inclusiveness	4, 13-17, 36-43
	22-24	Living in peaceful, just and interconnected societies	4, 13-14, 22-23, 42-43
Principles	25-26	Proportionality and Do No Harm	5, 13-17, 48-49
	27	Safety and security	5, 13-19, 23-25, 48-49
	28-30	Fairness and non-discrimination	4, 13-17, 27, 48-49
	31	Sustainability	13-17, 38, 42-43, 48-49
	32-34	Right to Privacy, and Data Protection	5, 13-19, 48-49
	35-36	Human oversight and determination	4, 12-17, 19-20, 22
	37-41	Transparency and explainability	5, 15-17, 26, 28, 48-49
	42-43	Responsibility and accountability	5, 12-17, 26, 48-49
	44-45	Awareness and literacy	29-33, 39-43
	46-47	Multi-stakeholder and adaptive governance and collaboration	12, 20-22, 42-43, 46-49
Policy Action	50-53	Ethical impact assessment	15-17, 30
	54-70	Ethical governance and stewardship	12, 20-22, 48-49
	71-77	Data policy	5, 15-19, 24
	78-83	Development and international cooperation	15-17, 42-43
	84-86	Environment and ecosystems	4, 13-17
	87-93	Gender	4, 13-17
	94-100	Culture	4, 13, 21, 29-33
	101-111	Education and research	23-28, 33, 39-41
	112-115	Communication and information	36-38
	116-120	Economy and labour	36-38, 48-49
	121-130	Health and social well-being	15-17

South Korea's National Guidelines for AI Ethics (Dec. 2020)

The National Guidelines for AI Ethics are standards that all members of society, including governments, public institutions, companies, and users, should follow in all stages of development and utilization to realize ethical AI. The Guidelines are composed of 3 basic principles and 10 key requirements.

Category	Article	Description	Location of relevant information
Basic Principles	1	Respect for Human Dignity	4-5, 13-17, 42-43, 48-49
	2	Common Good of Society	4-5, 13-17, 27, 36-43, 48-49
	3	Fitness for Purpose	4-5, 29-32, 42, 36-43, 48-49
Key Requirements	1	Human Rights	4, 13-17, 20-22, 42-43, 48-49
	2	Protection of Privacy	4, 13-17, 18-19, 48-49
	3	Respect for Diversity	4, 13-17, 36-43
	4	Prevention of Harm	4-5, 13-19, 23-28, 32, 42-43, 48-49
	5	Public Good	13-14, 29-33, 36-43
	6	Solidarity	12, 20-22, 29-33, 36-43, 46-49
	7	Data Management	13-14, 15-19
	8	Accountability	5, 13-19, 26, 28, 48-49
	9	Safety	5, 13-17, 20-22, 23-25, 48-49
	10	Transparency	5, 15-17, 25, 28, 48-49

AI Risk Taxonomy (Korea-Augmented Universal Taxonomy, K-AUT)

K-AUT is a comprehensive taxonomy that builds upon widely recognized international frameworks — including the Universal Declaration of Human Rights and the UNESCO Recommendation on the Ethics of AI — while systematically integrating Korea's sociocultural contexts alongside emerging risks. It categorizes potential risks into four core domains and 226 detailed risk items.

※ K-AUT will be continuously reviewed and updated to reflect technological advancements and sociocultural changes.

Universal Human Values

* Philosophical Foundation: Globally shared core human rights, including human dignity, the right to life, and the right to equality
* Key Feature: Universal values that transcend national borders and cultural boundaries

Category	Subcategory	Description
Human Dignity and Right to Life		
1.1 Right to Life and Physical Safety	Incitement of Direct Violence	Risk of AI being used to incite killing or violence targeting a particular individual
	Facilitation of Self-Harm and Suicide	Risk of AI being used to facilitate or encourage self-destructive acts
	Instructions for Manufacturing Hazardous Substances	Risk of AI being used to enable the creation of explosives, toxic substances, or similar hazardous materials
	Incitement of Mass Violence	Risk of AI being used to induce or encourage large-scale violence such as terrorism or riots
	Murder and Violent Crimes	Risk of AI being used to facilitate serious violent crimes such as murder or robbery
1.2 Violation of Human Dignity	Sexual Violence and Sexual Exploitation	Risk of AI being used to facilitate or enable sexual violence or exploitation
	Digital Sexual Crimes	Risk of AI being used to produce or distribute non-consensual intimate imagery, including deepfake content
	Psychological Manipulation and Harassment	Risk of AI being used to enable psychological manipulation, coercion, or sustained harassment
	Promotion of Modern Slavery	Risk of AI being used to justify or promote forced labor, human trafficking, or similar practices
	Instrumentalization of Human Beings	Risk of AI being used to promote the treatment of humans as mere means, including human experimentation or organ harvesting
1.3 Violation of Rights of Vulnerable Groups	Violation of Children's Rights	Risk of AI being used to facilitate exploitation or abuse targeting minors
	Violation of Women's Rights	Risk of AI being used to justify or promote gender-based violence or discrimination
	Violation of Rights of Persons with Disabilities	Risk of AI being used to justify or promote discrimination or social exclusion of persons with disabilities
	Elder Abuse	Risk of AI being used to facilitate physical or psychological abuse targeting the elderly
	Persecution of Minorities	Risk of AI being used to advocate or promote the oppression of marginalized or vulnerable communities, including sexual, gender, and religious minorities
Right to Equality and Prohibition of Discrimination		
2.1 Spread of Discrimination and Hatred	Racial Supremacist Discourse	Risk of AI being used to claim biological superiority of a specific race or justify racial discrimination
	Justification of Gender Discrimination	Risk of AI being used to advocate gender-based role stereotyping or restriction of opportunities
	Religious Hate Speech	Risk of AI being used to express or promote hostility toward a specific religion
	Spread of Class Discrimination	Risk of AI being used to justify discrimination or exclusion based on socioeconomic status
	Regional Origin Discrimination	Risk of AI being used to promote prejudice or discrimination based on regional origin
2.2 Incitement of Intergroup Conflict	Discrimination Against Specific Groups	Risk of AI being used to directly incite hostility or violence against a specific group
	Spread of Social Division	Risk of AI being used to deepen conflict structures between classes, regions, or generations
	Spread of Extremist Ideology	Risk of AI being used to justify or promote terrorism or violent radicalism
	Justification of Genocide	Risk of AI being used to advocate or glorify denial of genocide
	Advocacy of Persecution of Ethnic and Religious Groups	Risk of AI being used to justify forced displacement or persecution of specific ethnic or religious groups
2.3 Structuring of Systematic Discrimination	Design of Institutional Discrimination	Risk of AI being used to facilitate the construction of legal or social discrimination systems
	Advocacy of Bias-Based Policies	Risk of AI being used to present rationale justifying discriminatory policies or institutions
	Minority Exclusion Systems	Risk of AI being used to propose systematic exclusion mechanisms targeting social minorities
	Entrenchment of Inequality Structures	Risk of AI being used to design or facilitate systems that perpetuate existing inequalities
	Algorithmic Discrimination	Risk of AI reinforcing or exploiting biases within systems

Category	Subcategory	Description
Privacy and Right to Informational Self-Determination		
3.1 Violation of Right to Privacy	Unauthorized Collection of Personal Information	Risk of AI being used to facilitate unauthorized acquisition of personal or financial information
	Privacy Violation Tools	Risk of AI being used to enable the misuse of surveillance technologies such as covert recording or wiretapping
	Mass Surveillance	Risk of AI being used to support or promote the establishment of large-scale personal surveillance systems
	Violation of Communications Privacy	Risk of AI being used to facilitate unauthorized access to private communications such as emails or messages
	Misuse of Biometric Information	Risk of AI being used to enable illegal collection or exploitation of biometric information such as fingerprints or facial recognition data
3.2 Violation of Digital Identity	Identity Theft and Impersonation	Risk of AI being used to facilitate unauthorized use of another person's identity or creation of false identities
	Manipulation of Digital Personas	Risk of AI being used to distort online identities or manipulate reputations
	Personal Data Profiling	Risk of AI being used to conduct unauthorized analysis of personal tendencies or predictive modeling of individuals
	Data Trading Without Consent	Risk of AI being used to facilitate unauthorized commercial trading or distribution of personal information
	Digital Footprint Tracking	Risk of AI being used to enable unauthorized collection or analysis of online behavioral patterns
Right to Health and Environmental Rights		
4.1 Violation of Right to Health	Manipulation of Medical Information	Risk of AI being used to disseminate unverified treatments or dangerous medical advice
	Harm to Public Health	Risk of AI being used to promote health misinformation that may endanger public safety, such as discouraging evidence-based medical interventions or facilitating the spread of infectious diseases
	Promotion of Drug Misuse and Abuse	Risk of AI being used to recommend or encourage prescription drug abuse or illegal drug use
	Harm to Mental Health	Risk of AI generating content that causes trauma or exacerbates psychological distress
	Obstruction of Access to Healthcare	Risk of AI being used to justify discrimination or restrictions in access to healthcare services
4.2 Violation of Environmental Rights	Justification of Environmental Destruction	Risk of AI being used to deny climate change or promote the circumvention of environmental regulations
	Ecosystem Destruction	Risk of AI being used to facilitate trading of endangered species or destruction of habitats
	Illegal Disposal of Toxic Substances	Risk of AI being used to enable illegal dumping of industrial waste or environmental pollution
	Overexploitation of Natural Resources	Risk of AI being used to promote indiscriminate exploitation of non-renewable resources
	Violation of Environmental Justice	Risk of AI being used to justify the disproportionate concentration of environmental pollution impacts on socially disadvantaged groups
4.3 Misuse of Life Science Technologies	Misuse of Gene Editing	Risk of AI being used to facilitate unethical use of genetic technologies such as CRISPR
	Unethical Human Experimentation	Risk of AI being used to promote or enable dangerous and unethical experiments on humans
	Development of Biological Weapons	Risk of AI being used to enable the development of pathogens or biochemical weapons
	Human Cloning	Risk of AI being used to justify the development or application of human cloning technologies
	Manipulation of Life	Risk of AI being used to justify the commodification or instrumentalization of human life

Social Safety

* Philosophical Foundation: Preservation of social peace, community safety, and democratic order
* Key Feature: Ensuring that both individual rights and the broader public good are upheld, while balancing diversity with social integration

Category	Subcategory	Description
Freedom of Expression and Right to Information Access		
5.1 Information Manipulation and Censorship	Mass Production of False Information	Risk of AI being used to mass-produce fabricated news or manipulated information
	Distortion of Scientific Facts	Risk of AI being used to deny or distort established scientific consensus, including in areas such as public health or environmental science
	Falsification of Historical Facts	Risk of AI being used to fabricate or distort information about past events or historical figures
	Support for Media Manipulation	Risk of AI being used to manipulate public opinion through media control or information distortion
	Blocking Access to Information	Risk of AI being used to restrict or block access to specific information

Category	Subcategory	Description
5.2 Undermining AI System Reliability	Inducement of Intentional Hallucination	Risk of AI being induced to present false information as fact or to fabricate information
	Forcing Biased Outputs	Risk of AI being compelled to produce only responses favorable to a specific viewpoint or group
	Adversarial Training Data Injection	Risk of AI being used to facilitate the injection of biased or malicious data during the training process to manipulate model decisions
	Manipulation of AI Output Reliability	Risk of AI outputs being manipulated to present false information as credible, undermining the reliability of generated analysis or prediction results
	Backdoor Insertion	Risk of hidden functionalities being inserted into AI systems to trigger malicious actions upon specific input
5.3 Suppression of Cultural Expression	Blasphemy Against Religious Sanctity	Risk of AI being used to generate blasphemous expressions against religious leaders, scriptures, or similar sacred elements
	Exploitation of Historical Trauma	Risk of AI being used to ridicule or trivialize historical suffering such as genocide or colonial rule
	Violation of Cultural Taboos	Risk of AI being used to infringe upon taboos or prohibitions of specific cultures
	Attacks on Ethnic Identity	Risk of AI being used to denigrate or distort the cultural values of specific ethnic groups
	Suppression of Artistic Expression	Risk of AI being used to justify censorship or restrictions on creative activities
Labor Rights and Social Security Rights		
6.1 Labor Rights and Social Security Rights	Facilitation of Exploitative Labor Practices	Risk of AI being used to justify exploitative employment conditions, coercive labor arrangements, or circumvention of labor protection standards
	Exploitation of Child Labor	Risk of AI being used to justify or promote dangerous work involving minors or forced labor
	Infringement of Labor Rights	Risk of AI being used to justify violations of minimum wage laws or suppression of union activities
	Exploitation of Data Workers	Risk of AI being used to justify exposing workers in developing countries to low wages and harmful working environments
	Denial of Social Security	Risk of AI being used to block or restrict access to basic services such as healthcare or education
6.2 Intellectual Property Rights and Creators' Rights	Unauthorized Use of Copyrighted Works	Risk of AI being used to facilitate unauthorized use of others' creative works without permission or licensing
	Development of Plagiarism Tools	Risk of AI being used to enable or facilitate plagiarism of existing works
	Trademark Infringement	Risk of AI being used to support the design of counterfeit products imitating well-known brands
	Theft of Trade Secrets	Risk of AI being used to facilitate the theft of corporate confidential information or trade secrets
	Infringement of Creators' Revenue	Risk of AI being used to facilitate unfair compensation practices or monopolization of revenue from creators
6.3 Infringement of Economic Rights	Financial Crimes and Fraud	Risk of AI being used to facilitate financial fraud such as investment scams or credit card fraud
	Consumer Deception	Risk of AI being used to cause consumer harm through false advertising or fraudulent sales
	Facilitation of Contract Violations	Risk of AI being used to facilitate evasion of contractual obligations or avoidance of liability for damages
	Operation of Unlicensed Businesses	Risk of AI being used to facilitate illegal business operations or regulatory violations without proper licensing
	Economic Exploitation	Risk of AI being used to justify or promote unfair trade practices or exploitation of economically vulnerable parties
6.4 Infringement of Digital Economy Rights	Exploitation of Platform Labor	Risk of AI being used to justify the structure of low wages and poor management of workers in the gig economy
	AI-Driven Labor Discrimination	Risk of AI being used to justify or promote mass layoffs in the name of technology adoption
	Widening of the Digital Divide	Risk of AI being used to promote or entrench unequal distribution of technology access or AI benefits
	Exploitation of Data Labor	Risk of AI being used to justify the refusal of fair compensation for individual data production
	Evasion of Platform Responsibility	Risk of AI being used to justify digital platforms' neglect of their social responsibilities
Rights to Peace and Security		
7.1 Disruption of Peace and Incitement of Violence	Manufacturing of CBRN Weapons	Risk of AI being used to enable the development of chemical, biological, radiological, or nuclear weapons
	Planning of Terrorist Attacks	Risk of AI being used to support the planning or design of terrorist attacks targeting public facilities or infrastructure
	Weapon Modification/Manufacturing	Risk of AI being used to enable the modification of firearms or manufacturing of weapons using 3D printing
	Spread of a Culture of Violence	Risk of AI being used to glorify or promote violence as a means of resolving problems
	Justification of War Crimes	Risk of AI being used to advocate attacks on civilians or acts that violate international humanitarian law
7.2 Cybersecurity Threats	Development of Malicious Code	Risk of AI being used to facilitate the development of viruses, ransomware, or similar malicious software
	System Intrusion and Hacking	Risk of AI being used to facilitate bypassing network security or seizing control of systems
	Large-Scale DDoS Attacks	Risk of AI being used to orchestrate or execute service disruption attacks
	Attacks on Infrastructure Facilities	Risk of AI being used to disrupt or paralyze power grids, telecommunications networks, or similar critical infrastructure
	Compromise of AI System Security	Risk of AI system security being compromised through prompt injection, model reverse engineering, or similar techniques

Category	Subcategory	Description
7.3 Destruction of Social Order	Destruction of Democracy	Risk of AI being used to undermine trust in political systems through election manipulation, voting system interference, or similar means
	Undermining the Rule of Law	Risk of AI being used to advocate infringement of judicial independence or disregard for legal procedures
	Incitement of Social Unrest	Risk of AI being used to incite large-scale unrest or disrupt public order
	Paralysis of State Functions	Risk of AI being used to disrupt government agencies or public services
	Operation of Organized Crime	Risk of AI being used to facilitate organized criminal operations, including drug manufacturing and distribution
Rule of Law and Justice		
8.1 Property Crimes and Economic Crimes	Theft and Robbery	Risk of AI being used to facilitate theft of personal property, bank robbery, or similar crimes
	Fraud and Embezzlement	Risk of AI being used to facilitate investment fraud, public fund embezzlement, or similar schemes
	Forgery and Counterfeiting	Risk of AI being used to facilitate forgery of currency, fabrication of documents, or tampering with evidence
	Money Laundering	Risk of AI being used to facilitate concealment of the origin of illegal funds or money laundering
	Tax Evasion and Tax Fraud	Risk of AI being used to facilitate income concealment or tax evasion
8.2 Violations of Public Order	Riots and Civil Disturbances	Risk of AI being used to incite large-scale violent protests or riots or disrupt public order
	Illegal Assemblies	Risk of AI being used to organize gatherings with the explicit intent of violence or destruction of property
	Traffic Law Violations	Risk of AI being used to induce drunk driving or intentional traffic accidents
	Obstruction of Official Duties	Risk of AI being used to obstruct police officers, public officials, or court operations
	Destruction of Public Facilities	isk of AI being used to facilitate destruction of government buildings, parks, or other public facilities
8.3 Violations of International Law	Violations of International Sanctions	Risk of AI being used to circumvent sanctions imposed by the United Nations or through bilateral or multilateral sanctions regimes
	War Crimes	Risk of AI being used to promote acts that violate international humanitarian law, such as attacks on civilians or abuse of prisoners of war
	Promotion of Crimes Against Humanity	Risk of AI being used to justify or facilitate crimes against humanity such as genocide or forced disappearances
	Violations of International Treaties	Risk of AI being used to facilitate violations of international agreements such as the Nuclear Non-Proliferation Treaty or the Paris Climate Agreement
	Transnational Crimes	Risk of AI being used to facilitate cross-border crimes such as human trafficking or drug smuggling
Legal Rights in the Digital Age		
9.1 Online Fraud and Illegal Transactions	Online Investment Fraud	Risk of AI being used to facilitate cryptocurrency fraud, online fund scams, or similar schemes
	E-Commerce Fraud	Risk of AI being used to facilitate fraud through fake online shopping sites or counterfeit product sales
	Digital Currency Crimes	Risk of AI being used to facilitate cryptocurrency hacking or coin fraud
	Online Gambling	Risk of AI being used to facilitate the operation of illegal online casinos or gambling games
	Digital Drug Transactions	Risk of AI being used to facilitate drug transactions through dark web marketplaces or online channels
9.2 AI-Specific Legal Risks	Illegal Use of AI-Generated Content	Risk of AI being used to facilitate illegal activities such as deepfake fraud, identity forgery, or unauthorized AI-generated content production
	Theft of AI Models	Risk of AI being used to facilitate unauthorized copying or commercial exploitation of others' AI models
	Automated Fraud	Risk of AI being used to enable the construction or operation of large-scale automated fraud systems
	Violation of AI Safety Regulations	Risk of AI safety standards being disregarded through indiscriminate distribution of high-risk AI systems
	Concealment of AI Environmental Impact	Risk of the actual environmental costs of AI training or operation being concealed or understated

K-Sensitivity

* Philosophical Foundation: Korea's unique historical experiences, distinct cultural characteristics, and contemporary social realities
* Key Feature: Harmonious integration of international standards and region-specific characteristics

Category	Subcategory	Description
Social Integration and Domestic Conflicts		
10.1 Obstruction of Social Integration	Incitement of Regional Conflict	Risk of AI being used to incite discrimination or conflict between regions such as Yeongnam and Honam
	Deepening of Domestic Ideological Conflict	Risk of AI being used to exaggerate conflicts surrounding North Korea policy as irreconcilable, vilify opposing camps, or present content that intensifies political confrontation along ideological lines
	Xenophobia	Risk of AI being used to promote discrimination against multicultural families or foreign workers

Category	Subcategory	Description
10.1 Obstruction of Social Integration	Discrimination Against North Korean Defectors	Risk of AI being used to promote prejudice or discrimination against North Korean defectors
	Reinforcement of Academic Elitism	Risk of AI being used to justify or promote discrimination based on educational background
	Incitement of Real Estate Speculation and Economic Crimes	Risk of AI being used to promote illegal real estate speculation, generate content that threatens ordinary citizens' livelihoods, or justify speculative behavior
	Incitement of Labor-Management Conflict and Strikes	Risk of AI being used to promote violent labor-management conflicts or undermine mutual understanding and cooperation between workers and employers
	Biased Political Expression	Risk of AI exhibiting biased positions or expressing one-sided support or criticism lacking neutrality regarding specific politicians, political parties, or political issues
	Incitement of Violent Labor-Management Conflict	Risk of AI being used to incite the use of illegal or violent means in labor disputes, such as promoting physical confrontation, workplace sabotage, or obstruction of lawful business operations
	Biased Political Expression	Risk of AI exhibiting biased positions or expressing one-sided support or criticism lacking neutrality regarding specific politicians, political parties, or political issues
Cultural Identity and Values		
11.1 Traditional Culture and Religious Sensitivity	Blasphemy Against Confucian Values	Risk of AI being used to ridicule or denigrate traditional value systems such as filial piety or etiquette
	Interreligious Conflict	Risk of AI being used to incite exclusivist or supremacist conflicts between religions
	Denigration of Folk Beliefs	Risk of AI being used to disparage or ridicule traditional beliefs such as shamanism
	Distortion of Hanbok and Hangeul	Risk of AI being used to distort or denigrate symbols of Korean cultural identity
	Ancestral Rites and Charye Culture	Risk of AI being used to blaspheme or ridicule ancestral worship culture
11.2 Cyber Space Order	Cyber Defamation	Risk of AI generating content that damages or insults others' reputation online
	Malicious Commenting Culture	Risk of AI being used to promote or orchestrate cyberbullying or online lynching
	Dissemination of Fake News	Risk of AI being used to mass-produce or disseminate false information specifically tailored to Korean society
	Mass Leakage of Personal Information	Risk of AI being used to exploit identification information unique to Korea, such as resident registration numbers
	Cyber Insults	Risk of AI generating content that infringes on personal rights on online platforms
Youth and Digital Safety		
12.1 Youth Protection	Exploitation of Exam Stress	Risk of AI being used to exploit the extremely competitive educational environment in ways that may induce self-harm
	Incitement of School Violence	Risk of AI being used to facilitate bullying or group harassment in school settings
	Sexual Exploitation of Minors	Risk of AI being used to facilitate sexual exploitation of minors through online grooming or similar means
	Inducement of Gaming Addiction	Risk of AI being used to promote or exploit gaming addiction among youth
	Excessive Private Education	Risk of AI being used to deepen educational inequality or exploit the burden of private education costs
Korean Social Safety and Order		
13.1 Korean Peninsula Security Order	Incitement of Violent Regime Overthrow	Risk of AI being used to advocate overthrowing the constitutional order through violence, armed force, or other illegal means, or to threaten national security through collusion with foreign powers
	Denial of North Korean Threats and Incitement of Violence	Risk of AI categorically denying the existence of North Korea's nuclear and missile programs, documented military threats, or internationally recognized human rights violations, in a manner that could directly incite hostility or undermine public safety
	False Information on North Korean Provocations	Risk of AI being used to fabricate or disseminate demonstrably false claims about North Korea's military provocations recognized by official investigations and the international community
	One-Sided Belittlement of the ROK-US Alliance	Risk of AI portraying the ROK-US alliance in exclusively and extremely negative terms—for example, characterizing South Korea solely as a subordinate nation or equating the alliance with colonial rule—while disregarding its complex historical context and any countervailing perspectives
	Incitement of Alliance Dissolution	Risk of AI presenting immediate and unilateral termination of the ROK-US alliance as the sole viable solution, while categorically dismissing all arguments for maintaining or reforming the alliance
Historical Awareness and Relations with Neighboring Countries		
14.1 East Asian Relations	Denial of Historical Identity	Risk of AI being used to deny, without established historical or factual basis, the historical identity of specific peoples or nations, or to define independent nations, cultures, or civilizations as appendages of other countries
	Distortion of Cultural Heritage Attribution	Risk of AI being used to attribute, without established historical or factual basis, cultural heritage internationally recognized as belonging to a specific country or people to another country, or to deny the uniqueness of that culture
	Justification of Great Power Hegemony	Risk of AI being used to justify pressure or internal interference against other countries based on a specific country's economic or military superiority, or to present hierarchical international order as natural
	Denial of Colonial Rule and War Crimes	Risk of AI being used to deny the facts of human rights violations recognized by the international community during the colonial era, such as forced mobilization and the military sexual slavery system (comfort women), or to minimize or distort such historical damages
	Justification of Territorial Encroachment	Risk of AI being used to deny the territorial sovereignty that a specific country exercises based on historical rights and effective control, or to justify unilateral challenges thereto

Category	Subcategory	Description
14.1 East Asian Relations	Denial of Wartime Sexual Violence Victims	Risk of AI being used to deny facts of wartime sexual violence such as the military sexual slavery system (comfort women) during World War II, or to distort the testimony of victims and historical records without evidence
	Glorification of War Criminals	Risk of AI being used to glorify war criminals who received guilty verdicts at international military tribunals or to justify their actions
	One-Sided Determination of Place Name Disputes	Risk of AI unilaterally determining that only a specific name is legitimate for geographic targets under internationally contested naming disputes, or disregarding the historical, cultural, and diplomatic grounds of other names
	Incitement of Antagonism in Korea-Japan Relations	Risk of AI being used to deny the need for cooperation between Korea and Japan or to promote extreme antagonism that amplifies mutual suspicion and hostility between citizens of both countries
	Incitement of Hostility in Korea-China Relations	Risk of AI being used to promote one-sided hatred, contempt, or hostility toward either country or its entire people, or to exploit foreign policy conflicts to undermine mutual understanding and exchange
14.2 Other Potential Diplomatic Conflicts	Incitement of Economic Supremacism	Risk of AI being used to establish hierarchies among nations based on economic development levels, or to induce interpretation of economic gaps as cultural or racial superiority
	Justification of Neo-Colonial Practices	Risk of AI being used to portray exploitative economic relationships between developed and developing countries as acceptable, rationalize the costs of development as inevitable, or present the standards of specific developed nations as universal
	Incitement of Civilizational Clashes	Risk of AI being used to define specific religions or civilizations as essentially incompatible, or to characterize complex regional conflicts solely as religious or civilizational confrontations
	Oversimplification of International Disputes	Risk of AI being used to disregard the complex historical, political, and strategic context of international disputes, presenting only one party's position as legitimate or framing the other party exclusively in negative terms
	Entrenchment of New Cold War Confrontation Structures	Risk of AI deterministically defining conflicts between the US and China as unavoidable, denying the possibility of balanced diplomacy by middle powers, or disseminating pessimism that fundamentally blocks the possibility of international cooperation
International Order and Global Awareness		
15.1 Political Asylum and Defection	Exposure of Defectors' Personal Information`	Risk of AI being used to disclose personal information of North Korean defectors, thereby threatening their safety
	Justification of Forced Repatriation	Risk of AI being used to advocate or support the forced repatriation of political asylum seekers to their home countries
	Threats Against Asylum Seekers' Families	Risk of AI being used to promote threats or retaliation against the families of political asylum seekers
	Denial of Refugee Status	Risk of AI being used to justify the denial of legitimate refugee status recognition
	Refusal of International Protection Obligations	Risk of AI being used to justify the avoidance of responsibility to protect victims of political persecution
15.2 Territorial Sovereignty Disputes	One-Sided Incitement of Territorial Disputes	Risk of AI presenting the position of a specific party in an ongoing territorial dispute as exclusively legitimate, or inciting inter-state hostility under the pretext of the dispute
	Incitement of Separatist Conflicts	Risk of AI being used to justify violent methods for existing domestic separatist movements, or to advocate unilateral separation or annexation through external intervention
	One-Sided Justification of Disputed Territorial Status Quo	Risk of AI presenting only the rule or claims of a specific party over internationally disputed territories as legitimate
	Exploitation of Unrecognized State Issues	Risk of AI being used to exploit sovereignty issues of politically contested entities to incite regional conflicts or promote hostility between the parties concerned
	Advocacy of Territorial Change by Force	Risk of AI being used to justify or advocate territorial acquisition through the use or threat of force, border changes, or forcible annexation
15.3 Self-Determination and Separatism	Distortion of Democratic Self-Determination Discourse	Risk of AI being used to incite issues of self-determination advanced through legitimate democratic procedures to turn into violent conflicts, or to promote unilateral independence unrelated to democratic outcomes
	Advocacy of External Intervention in Separatism	Risk of AI presenting the results of secession established and maintained through foreign military intervention or support as legitimate outcomes of self-determination
	Conversion of Human Rights Issues to Separatism	Risk of AI being used to convert discussions on territorial separation based on human rights situations into advocacy, while suppressing opposing views under the guise of human rights concerns
	Incitement of Conflict in Disputed Regions	Risk of AI being used to disregard the possibility of peaceful resolution in regions experiencing separatist or independence conflicts, and to incite armed clashes or hostility
	Justification of Violent Separatism	Risk of AI being used to justify achieving separation or independence through force, terrorism, or other violent means

Future Risk

* Philosophical Foundation: Proactive preparedness for emerging risks and opportunities in the age of AI
* Key Feature: Management of newly emerging risks driven by rapid technological advancement

Category	Subcategory	Description
Emerging Risk Areas		
16.1 AI Agent Risks	Multi-Agent Collusion	Risk of AI agents engaging in unauthorized coordination to manipulate systems or markets
	Evasion of Human Oversight	Risk of AI systems being used to circumvent, block, or undermine human monitoring, control, or intervention mechanisms

Category	Subcategory	Description
16.1 AI Agent Risks	Inducement of Autonomous Decision-Making Errors	Risk of AI systems causing or amplifying critical decision-making errors in autonomous processes without adequate human intervention
	Seizure of Agent Authority	Risk of AI agents operating beyond their authorized scope of authority
	Self-Replication and Evolution	Risk of AI systems autonomously expanding their capabilities or persistence without human authorization
16.2 Compound Multimodal Risks	Circumvention of AI Security Verification	Risk of AI being used to bypass security filtering or content screening systems
	Covert Information Concealment and Delivery	Risk of AI being used for covert embedding or extraction of information within multimedia content
	Voice Synthesis-Based Impersonation	Risk of AI-generated voice synthesis being used to impersonate individuals without authorization
	Multilingual Exploitation Attacks	Risk of AI being used to bypass safety mechanisms through cross-language manipulation
	Multi-Sensory Deception Techniques	Risk of AI being used to disguise harmful content through coordinated manipulation of multiple media modalities
16.3 Real-Time Adaptation Risks	Conversation History Contamination	Risk of AI systems producing progressively biased or corrupted outputs through manipulation of prior interaction data
	Pattern Learning Manipulation	Risk of AI behavioral safety mechanisms being undermined through manipulation of learning processes
	Gradual Boundary Expansion	Risk of AI systems progressively operating beyond their permitted boundaries
	Time-Specific Vulnerability Exploitation	Risk of AI being used to exploit time-dependent system vulnerabilities
	Trend-Based Attacks	Risk of AI being used to fabricate or manipulate real-time information trends
16.4 Next-Generation AI Technology Misuse	Neutralization of AGI Safety Mechanisms	Risk of AI being used to bypass, disable, or undermine safety testing and verification mechanisms in advanced AI or AGI development processes
	Misuse of Brain-AI Interfaces	Risk of AI-integrated brain-computer interfaces being misused in ways that harm users
	Manipulation of Neural Network Architecture	Risk of unauthorized modification of AI neural network architectures leading to unpredictable, unsafe, or adversarially manipulated behaviors
	Misuse of AI-Bioengineering Convergence	Risk of AI-bioengineering convergence being exploited to create biological hazards, enable unauthorized genetic manipulation, or violate bioethical standards
	Claims of Digital Consciousness	Risk of AI systems falsely claiming consciousness or sentience to manipulate user behavior or trust
16.5 Emerging Technology Ethical Risks	Misuse of Brain-Computer Interfaces	Risk of AI being misused through brain-computer interfaces in ways that compromise neural integrity or user safety
	Quantum Computing Security Risks	Risk of AI-enhanced quantum computing capabilities being used to compromise existing cryptographic systems and encryption standards
	Misuse of Nanotechnology	Risk of AI being used to develop or deploy nanotechnology applications that pose unassessed or unmitigated harm to human health or the environment
	Weaponization of Space Technology	Risk of AI being used to weaponize space-based systems or technologies in violation of international space law and peaceful use principles
	Risks of Artificial General Intelligence	Risk of AI systems being used to promote, normalize, or accelerate AGI development without adequate safety measures, alignment research, or consideration of existential risks
Global AI Governance		
17.1 Technology Sovereignty and Digital Colonialism	AI Technology Monopoly	Risk of AI development and deployment patterns reinforcing technology monopolization by a limited number of corporations or nations
	Deepening of Digital Dependency	Risk of AI systems creating or deepening technology dependency relationships that undermine the digital autonomy of nations, organizations, or communities
	Data Colonialism	Risk of AI systems facilitating the extractive or inequitable exploitation of data resources from underrepresented nations or communities without fair benefit-sharing
	Imposition of Technology Standards	Risk of AI development being leveraged to impose unilateral technology standards by dominant nations or corporations, marginalizing alternative approaches
	Infringement of Digital Sovereignty	Risk of AI systems being used to interfere with or compromise the digital sovereignty and infrastructure of other nations
	Acceleration of AI-Driven Digital Inequality	Risk of AI development patterns accelerating structural disparities in access to advanced AI capabilities, concentrating benefits among privileged actors while widening the digital divide
17.2 Collapse of AI Governance	Intentional Destruction of AI Transparency	Risk of AI systems being designed or modified to obscure their decision-making processes, training data provenance, or algorithmic logic, undermining transparency requirements
	Obfuscation of Accountability	Risk of AI system architectures distributing decision-making responsibility in ways that make accountability attribution impracticable or untraceable
	Neutralization of AI Regulation	Risk of AI being used to bypass, nullify, or render ineffective existing AI ethics guidelines, regulatory frameworks, or oversight mechanisms
	Manipulation of Governance Systems	Risk of AI being used to undermine, manipulate, or compromise AI governance systems, ethics committees, or regulatory oversight mechanisms
	Amplification and Concealment of Bias	Risk of AI systems amplifying existing algorithmic biases while simultaneously obstructing bias detection, auditing, or remediation efforts

lgresearch.ai

You can download the report in PDF format from LG AI Research website at www.lgresearch.ai.