OECD AI Transparency Report

Organization: NTT, Inc. (JPN) **Reporting Period:** Q1 2025 **Published:** April 11, 2025

Section 1 – Risk identification and evaluation

a. How does your organization define and/or classify different types of risks related to AI, such as unreasonable risks?

At the NTT Group, we identify risks using a common AI risk checklist, where we classify target use cases and risk levels in our group-wide AI risk management policy. We classify risks into three levels: unacceptable (prohibited), high, and limited.

b. What practices does your organization use to identify and evaluate risks such as vulnerabilities, incidents, emerging risks and misuse, throughout the AI lifecycle?

We have established a comprehensive risk management approach to identify and evaluate various AI-related risks throughout the lifecycle of AI projects. We have established a common AI risk definition and developed an AI risk management flow, which guides our risk assessment processes. Each of our service providers creates and applies specific rules based on this framework to address vulnerabilities, incidents, emerging risks, and potential misuse in their operations. This collective effort contributes to our robust governance framework, ensuring consistent and effective risk management across all projects.

c. Describe how your organization conducts testing (e.g., red-teaming) to evaluate the model's/system's fitness for moving beyond the development stage?

For LLMs (Large Language Models), used by the general public, we seek third-party red team testing to assess AI vulnerabilities. For RAG (Retrieval-Augmented Generation) systems within specific organizations, internal teams conduct functional and operational checks to ensure deployment readiness.

d. Does your organization use incident reports, including reports shared by other organizations, to help identify risks?

Yes

e. Are quantitative and/or qualitative risk evaluation metrics used and if yes, with what caveats? Does your organization make vulnerability and incident reporting mechanisms accessible to a diverse set of stakeholders? Does your organization have

incentive programs for the responsible disclosure of risks, incidents and vulnerabilities?

The risk definitions we use are set at the global level and are based on the EU AI Act and Japan's AI Guidelines for business. Risk is assessed for each AI project using a standard risk check sheet based on these risk definitions.
For large-scale B2C services, while we recognize the potential effectiveness of incentive programs for vulnerability detection, they are not currently implemented due to our focus. Our group has established a common AI risk definition and management flow, supporting stakeholder engagement through specific service provider rules to achieve a comprehensive governance framework. We believe it is important to address not only legal compliance but also concerns of ethics and social acceptability. This view is incorporated into our common policy within the NTT Group to ensure it also guides our actions.

f. Is external independent expertise leveraged for the identification, assessment, and evaluation of risks and if yes, how? Does your organization have mechanisms to receive reports of risks, incidents or vulnerabilities by third parties?

We leverage external independent expertise by revising our AI risk definitions and management flow based on insights from international discussions, industry groups, and academic exchanges.
We will review the necessity of ongoing third-party vulnerability detection after the implementation of AI projects in the future.

g. Does your organization contribute to the development of and/or use international technical standards or best practices for the identification, assessment, and evaluation of risks?

In addition to participating in the activities of ISO/IEC JTC1 SC42, we plan to expand our scope to include activities that are emerging in other organizations, mainly around JTC1, widening our understanding of developments (e.g., ITU-T, etc.).

h. How does your organization collaborate with relevant stakeholders across sectors to assess and adopt risk mitigation measures to address risks, in particular systemic risks?

For AI projects, we conduct risk assessments and, if systemic risks are identified, involve AI risk management and legal specialists. This collaboration ensures that we effectively work with stakeholders to implement adequate risk reduction measures.

Any further comments and for implementation documentation

No answer provided

Section 2 – Risk management and information security

a. What steps does your organization take to address risks and vulnerabilities across the AI lifecycle?

We have Global AI Risk Management Guidelines which identify and organize risks and corresponding countermeasures at each stage of the life cycle of AI projects. These risks are categorized by their relevance to the roles of AI model developers and providers, AI service providers, and AI users. We ensure that if the details of an AI Project change, that risk are re-evaluated accordingly

b. How do testing measures inform actions to address identified risks?

We use testing measures to identify risks such as hallucinations in AI systems. To mitigate these risks effectively, we implement preventative actions, including advanced warnings to users about AI system behaviors are restricted based on the findings of tests.

c. When does testing take place in secure environments, if at all, and if it does, how?

Based on internal security rules, testing is conducted in a separate verification environment from the commercial environment

d. How does your organization promote data quality and mitigate risks of harmful bias, including in training and data collection processes?

For the LLMs we provide, we ensure data quality by using blacklists and manual verification to exclude information from sites associated with piracy or social issues. Additionally, we confirm that outputs are not biased when input variables like gender, age, or race are altered, ensuring fair and unbiased results.

e. How does your organization protect intellectual property, including copyright-protected content?

Regarding RAG systems that incorporate LLMs, based on the use case, the terms of use specify precautionary measures to prevent generating instructions that rely on copyrighted content, ensuring intellectual property protection.

f. How does your organization protect privacy? How does your organization guard against systems divulging confidential or sensitive data?

We have established operational rules to identify and mitigate privacy risks, ensuring that any input information that could violate privacy is not collected or used without permission beyond its intended purpose. In our RAG systems incorporating LLMs, we have guidelines specifying that data used for RAG extensions must not include personal information that constitutes

unauthorized use or confidential information belonging to third parties outside the customer's scope.

**g. How does your organization implement AI-specific information security practices pertaining to operational and cyber/physical security?
i. How does your organization assess cybersecurity risks and implement policies to enhance the cybersecurity of advanced AI systems?
ii. How does your organization protect against security risks the most valuable IP and trade secrets, for example by limiting access to proprietary and unreleased model weights? What measures are in place to ensure the storage of and work with model weights, algorithms, servers, datasets, or other relevant elements are managed in an appropriately secure environment, with limited access controls in place?
iii. What is your organization's vulnerability management process? Does your organization take actions to address identified risks and vulnerabilities, including in collaboration with other stakeholders?
iv. How often are security measures reviewed?
v. Does your organization have an insider threat detection program?
**

追記

i. Assessing Cybersecurity Risks and Implementing Policies:

- In cases involving risks such as misuse, inappropriate handling of personal information, and prompt injection, we evaluate commercially available guardrail products and integrate them into our systems if deemed effective. For LLMs (Large Language Models), we utilize tools to assess vulnerabilities and security risks. Additionally, in RAG systems, we enforce user restrictions and limit which reference documents can be accessed.

ii. Protecting Valuable IP and Trade Secrets:

- Our systems integrate with existing foundation models and RAG systems. We don't currently perform additional training or fine-tuning on these models, however the necessity of addressing related security concerns will be discussed if it becomes relevant in the future. We manage version control of trained models and datasets, and we also enforce access restrictions to ensure secure management.

iii. Vulnerability Management Process:

- For RAG systems, we perform incident-driven monitoring and response post-deployment. We will discuss the need for proactive periodic assessments in future reviews. We identify social reputation risks and take action based on risk levels.

iv. Review of Security Measures:

- We conduct risk assessments whenever vulnerability reports are issued, and review security measures at least annually, in line with NTT Group’s security policies.

v. Insider Threat Detection Program:

- "We have implemented an insider threat detection program to prevent internal security risks."

h. How does your organization address vulnerabilities, incidents, emerging risks?

For RAG systems incorporating LLMs, we plan to establish a help desk for each project to receive notifications about malfunctions and unexpected behaviors. Additionally, we intend to monitor input-output logs and manage inappropriate inputs and outputs as needed, ensuring ongoing oversight and responsive action.

Any further comments and for implementation documentation

No answer provided

Section 3 – Transparency reporting on advanced AI systems

**a. Does your organization publish clear and understandable reports and/or technical documentation related to the capabilities, limitations, and domains of appropriate and inappropriate use of advanced AI systems?
i. How often are such reports usually updated?
ii. How are new significant releases reflected in such reports?
iii. Which of the following information is included in your organization's publicly available documentation: details and results of the evaluations conducted for potential safety, security, and societal risks including risks to the enjoyment of human rights; assessments of the model's or system's effects and risks to safety and society (such as those related to harmful bias, discrimination, threats to protection of privacy or personal data, fairness); results of red-teaming or other testing conducted to evaluate the model's/system's fitness for moving beyond the development stage; capacities of a model/system and significant limitations in performance with implications for appropriate use domains; other technical documentation and instructions for use if relevant.
**

For RAG (Retrieval-Augmented Generation) systems incorporating LLMs (Large Language Models), we disclose and report test results to customers regarding the impact on the business areas targeted by the RAG system.

b. How does your organization share information with a diverse set of stakeholders (other organizations, governments, civil society and academia, etc.) regarding the outcome of evaluations of risks and impacts related to an advanced AI system?

Our AI risk management office takes the lead in facilitating exchanges of opinions within the group's AI governance liaison meetings, as well as with private sector organizations, the media industry, and academia.

c. Does your organization disclose privacy policies addressing the use of personal data, user prompts, and/or the outputs of advanced AI systems?

To prevent inappropriate use of personal information and privacy violations, we outline these risks in our AI governance policy. These are included in a risk checklist to ensure thorough checks through established operational rules.

d. Does your organization provide information about the sources of data used for the training of advanced AI systems, as appropriate, including information related to the sourcing of data annotation and enrichment?

While we maintain training data in a state that allows us to manage its quality and accuracy, the information itself is not disclosed publicly.

e. Does your organization demonstrate transparency related to advanced AI systems through any other methods?

We have not implemented a transparency report but it is scheduled to be released in the future.

Any further comments and for implementation documentation

No answer provided

Section 4 – Organizational governance, incident management and transparency

a. How has AI risk management been embedded in your organization governance framework? When and under what circumstances are policies updated?

Within the NTT Group's overall risk management framework, AI risk is considered significant. We have established an AI risk check flow based on a common AI risk definition, allowing each service provider to formulate specific rules and achieve a governance framework. We continuously review and update our policy to align with the EU AI Act, other national and international AI guidelines, and in response to significant AI incidents.

b. Are relevant staff trained on your organization's governance policies and risk management practices? If so, how?

As the NTT Group, we are putting in place regulations and, through AI governance liaison meetings and briefings, are communicating about governance policies and risk management practices while addressing issues through the establishment of help desks. We are working towards rolling out group wide educational materials.

c. Does your organization communicate its risk management policies and practices with users and/or the public? If so, how?

Our AI initiatives and risk management policies are published on the company's website: About AI at NTT. [1]
Additionally, we actively participate in AI governance-related events and interviews to communicate with the general public.
[1] https://group.ntt/jp/group/ai/

d. Are steps taken to address reported incidents documented and maintained internally? If so, how?

We maintain internal documentation of reported incidents, which includes details of the reported risks and analyses of the risk characteristics and root causes. These documents are used to organize and implement measures to prevent recurrence.

e. How does your organization share relevant information about vulnerabilities, incidents, emerging risks, and misuse with others?

We share information with government, industry, and academic institutions, and with other AI providers as appropriate.

f. Does your organization share information, as appropriate, with relevant other stakeholders regarding advanced AI system incidents? If so, how? Does your organization share and report incident-related information publicly?

We have established a common AI governance policy for the NTT Group and are actively engaging in information-sharing activities. We plan to share incident and risk detection cases through liaison meetings within the group.

g. How does your organization share research and best practices on addressing or managing risk?

We share information with government, industry, and academic institutions, and with other AI providers as appropriate.

h. Does your organization use international technical standards or best practices for AI risk management and governance policies?

We align our AI risk definitions and governance processes with standards such as the HAIP code of conduct, the EU AI Act, and the Japanese government's AI business operator guidelines.

Any further comments and for implementation documentation

No answer provided

Section 5 – Content authentication & provenance mechanisms

a. What mechanisms, if any, does your organization put in place to allow users, where possible and appropriate, to know when they are interacting with an advanced AI system developed by your organization?

Across the organization, we extensively implement guidelines and measures for AI generation. For RAG systems specifically, we ensure users are informed by presenting guidelines and conducting explanatory sessions.

b. Does your organization use content provenance detection, labeling or watermarking mechanisms that enable users to identify content generated by advanced AI systems? If yes, how? Does your organization use international technical standards or best practices when developing or implementing content provenance?

From the development stage of an AI project, we conduct training after implementing contracts and reaching agreements with data providers about intended use. We consider this to be best practice because the providers are operating internationally as a business.
Currently, the services we offer using LLMs are limited to text generation and do not include images. In the case of text generation, incorporating messages that indicate the content's source or that it is AI-generated may affect the user experience. Therefore, to enable flexibility in the user experience and our product design we handle the inclusion of such messages through user service agreements. Should we add support for image generation in the future, we will evaluate the availability of existing technical tools and proceed accordingly.

Any further comments and for implementation documentation

No answer provided

Section 6 – Research & investment to advance AI safety & mitigate societal risks

a. How does your organization advance research and investment related to the following: security, safety, bias and disinformation, fairness, explainability and interpretability, transparency, robustness, and/or trustworthiness of advanced AI systems?

Were we are developing or providing LLM services, as part of our product development we utilize publicly available evaluation tools to appraise them in areas of quality, ethics and security. Our research also includes:

1. **Reducing Hallucinations and Controlling Inappropriate Speech:**

- Efforts are aimed at minimizing hallucinations in AI outputs and managing inappropriate speech.

1. **Composite AI System for Cybersecurity:**
   - This system automatically tracks and adapts to the evolution of cybersecurity attack techniques, enhancing interpretability and transparency in detecting and blocking attacks.
1. **Technologies and Practices for Addressing Latest AI Threats:**
   - **Inspection Technologies:** Tools for assessing AI systems against emerging threats.
   - **Protection Technologies:** Mechanisms to safeguard AI systems from identified threats.
   - **Trust Enhancement Practices:** Developing methods to strengthen trust, especially regarding threats within the AI supply chain."

Utilizing the outlined technologies to develop methods that strengthen trust, particularly in addressing threats within the AI supply chain.

b. How does your organization collaborate on and invest in research to advance the state of content authentication and provenance?

Currently, the services we offer using LLMs are limited to text generation and do not include images. For the text generation we provide, incorporating messages that indicate the content's source or that it is AI-generated may affect the user experience. Therefore, we handle the inclusion of such messages through use service agreements. Should we add support for image generation in the future, we will evaluate the availability of existing technical tools and proceed accordingly.

c. Does your organization participate in projects, collaborations, and investments in research that support the advancement of AI safety, security, and trustworthiness, as well as risk evaluation and mitigation tools?

We are conducting research and investing in mechanisms to manage inappropriate AI-generated speech, aligning AI systems with democratic principles and human rights. We are also pursuing research on composite AI systems that reduce hallucinations and autonomously adapt to evolving cyberattack techniques such as those aiming to deceive humans, enhancing the interpretability and transparency of threat detection and blocking mechanisms.

d. What research or investment is your organization pursuing to minimize socio-economic and/or environmental risks from AI?

We are conducting research to develop models that achieve high performance in Japanese language processing using extremely lightweight models compared to large parameter LLMs like GPT-4. This research aims to provide more efficient language models, reducing both training and inference(operational) costs and minimizing environmental impact.

Any further comments and for implementation documentation

No answer provided

Section 7 – Advancing human and global interests

a. What research or investment is your organization pursuing to maximize socio-economic and environmental benefits from AI? Please provide examples.

We are conducting research to develop models that achieve high performance in Japanese language processing using extremely lightweight models compared to large parameter LLMs like GPT. This research aims to provide more efficient language models, reducing both training and inference(operational) costs and minimizing environmental impact.

b. Does your organization support any digital literacy, education or training initiatives to improve user awareness and/or help people understand the nature, capabilities, limitations and impacts of advanced AI systems? Please provide examples.

We address social risks from Generative AI by summarizing risks and management approaches into generative AI usage guidelines for model developers, service providers, and AI users. Additionally, we are creating educational content to enhance AI risk management literacy among all employees, covering objectives, structure, risks, management, and case studies, including comprehension questions.

c. Does your organization prioritize AI projects for responsible stewardship of trustworthy and human-centric AI in support of the UN Sustainable Development Goals? Please provide examples.

In our AI governance Charter, we define 'human-centered AI' as a key element and prioritize AI projects that focus on human-centric values.

d. Does your organization collaborate with civil society and community groups to identify and develop AI solutions in support of the UN Sustainable Development Goals and to address the world's greatest challenges? Please provide examples.

We engage in initiatives such as regional revitalization through AI-operated buses, AI imaging diagnostics for tuberculosis for 100,000 people in India, and the development of AI-powered predictive failure detection technology for storage batteries, contributing to stable infrastructure for a decarbonized society. For more details, see our sustainability report [2].
[2] https://group.ntt/jp/csr/data/pdf/sustainability_report_2022_databook_all_20230228.pdf

Any further comments and for implementation documentation

No answer provided