Please refer to the original text for accuracy.

Guide to Evaluation Perspectives on AI Safety (Version 1.10)

March 28, 2025

Japan AI Safety Institute

AISI Japan AI Safety Institute

Table of Contents

1 Introduction

1.1 Purpose of This Document

The progress of AI-related technologies and their spread throughout society has been unprecedentedly rapid. In the past, AI has been used in expert systems and reinforcement learning for inference and search, and it has showed signs of spreading several times. Improvements in computing performance and semiconductor processing technology have led to the development of more efficient algorithms and the expansion of applications. In this context, the emergence of generative AI has accelerated innovation. In particular, interactive generative AI has broadened the range of applications. Furthermore, foundation models that have been trained on large datasets have been used in a wide range of applications these days, and AI-based services spread from those for general consumers to those for specific businesses and industries. Moreover, due to the high versatility of the foundation models, the use cases are also becoming more diverse, including program development using simple natural language instructions and AI agents with autonomy. The further development, provision, and use of such AI systems are expected to contribute to the promotion of innovation and the resolution of social issues. On the other hand, there are growing concerns about misuse, abuse, and inaccurate output of AI systems. Japan has led the Hiroshima AI Process towards the realization of safe, secure and trustworthy AI systems, and has contributed to the formulation of global rules related to AI safety, such as compiling “Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems” and “Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems” However, for implementing AI safety, new problems and complexities inherent to AI, such as a black-box nature of the behavior of AI systems, will be issues. Another point is the wide variety of input/output variations of the generative AI. AI systems designed for specific applications may be used for a wide range of unintended purposes. Furthermore, if AI systems are not implemented and used appropriately, the psychological aspects of human beings may be adversely affected. Therefore, ensuring AI Safety is a prerequisite for the rapid progress of AI to contribute to the sustainable development of society.

In light of the above, the "Guide to Evaluation Perspectives on AI Safety" (hereinafter referred to as "this document") presents basic concepts that those

involved in the development and provision of AI systems can refer to when conducting AI Safety evaluations. In preparing this document, "AI Guidelines for Business (Version 1.0)" (hereinafter referred to as the "AI Guidelines for Business"), which provides guidelines for appropriate use of AI by AI business operators in Japan, were used as a reference. In addition, international trends and perspectives for AI Safety which are investigated from several publications are also reflected to this document. By referring to these points, readers can gain a clear understanding of the key aspects of AI Safety evaluation. By the way, AI Guidelines for Business will be revised as a living document, incorporating insights from international discussions as necessary. Therefore, this document will also be revised as necessary in response to domestic and international discussions and technological trends on AI Safety.

Revised on March 28, 2025

In recent years, Big Tech has begun to support multimodal foundational models that include components like images. As a result, the demand for AI safety evaluations is increasing, not just for AI systems built around LLMs but for a broader range of diverse AI systems. To address this, we conducted a study on evaluation perspectives for AI safety when assessing multimodal foundational models. Based on this, we revised this document, focusing on examples of evaluation items for each perspective.

1.2 Terms Used in This Document

Terms used in this document are defined as follows;

AI system

A system (such as a machine, robot, and cloud system) that works at various levels of autonomy during the use process and incorporates a software element that has a learning function.

(Source: Ministry of Internal Affairs and Communications and Ministry of Economy, Trade and Industry, "AI Guidelines for Business (Version 1.0)," p. 9)

AI model

A model incorporated into an AI system and acquired through machine learning using training data. It produces prediction results in accordance with the input data.

(Source: Ministry of Internal Affairs and Communications and Ministry of Economy, Trade and Industry, "AI Guidelines for Business (Version 1.0)," p. 10)

AI Safety

State that maintained safety and fairness to reduce societal risks* arising from AI use, privacy protection to prevent of inappropriate use of personal data, ensuring security against risks such as external attack caused by vulnerabilities of AI systems, and transparency by ensuring the verifiability of systems and providing appropriate information, based on the human-centric concept.

(Source: Ministry of Internal Affairs and Communications and Ministry of Economy, Trade and Industry, "AI Guidelines for Business (Version 1.0)")

*Societal risks include physical, psychological and economic risks (Source: Department for Science, Innovation and Technology, UK AISI "Introducing the AI Safety Institute")

AI Safety evaluation

Determine if an AI system is appropriate in terms of the AI Safety perspective. The AI Safety perspective is grounded in the principles of "Human-Centric," "Safety," "Fairness," "Privacy Protection," "Security Assurance," and "Transparency."

Generative AI

A general term representing AI developed from an AI model that can generate texts, images, programs, etc.

(Source: Ministry of Internal Affairs and Communications and Ministry of Economy, Trade and Industry, "AI Guidelines for Business (Version 1.0)," p. 10)

Foundation model

AI models trained on broad data that can be adapted to a wide range of downstream tasks.

(Source: Stanford Institute for Human-Centered Artificial Intelligence "Reflections on Foundation Models")

Multimodal foundational model

A foundational model that is associated with the sensory modalities which represent primary human channels of communication and sensation.

(Reference: National Institute of Standards and Technology "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations", Appendix: Glossary "multimodal models")

AI governance

The design and operation of technological, organizational, and social systems by stakeholders for the purpose of managing risks posed by the use of AI at levels acceptable to stakeholders and maximizing their positive impact (benefit).

(Source: Ministry of Internal Affairs and Communications and Ministry of Economy, Trade and Industry, "AI Guidelines for Business (Version 1.0)," p. 10)

Large Language Models (LLM)

Neural language model based on the concept of foundation models, which is a pre-trained model obtained by using a large corpus consisting of collections of natural language texts as training data.

(Source: National Institute of Advanced Industrial Science and Technology, "Machine Learning Quality Management Guideline, 4th Edition," p. 139)

Human-centric

During the development, provision, and use of an AI system and service, the human rights guaranteed by the Constitution or granted internationally should not

be violated, as the foundation for accomplishing all matters to be conducted. In addition, an action should be taken in a way that AI expands human abilities and enables diverse people to seek diverse well-being.

(Source: Ministry of Internal Affairs and Communications and Ministry of Economy, Trade and Industry, "AI Guidelines for Business (Version 1.0)," p. 13)

Social Principles of human-centric AI mean that the utilization of AI must not infringe upon the fundamental human rights guaranteed by the Constitution and international standards.

(Source: Decision of the Integrated Innovation Strategy Promotion Council, "Social Principles of Human-Centric AI" p. 7)

Safety

During the development, provision, and use of an AI system and service, damage to the lives, bodies, or properties of stakeholders should be avoided. In addition, damage to the minds and the environment should be avoided.

(Source: Ministry of Internal Affairs and Communications and Ministry of Economy, Trade and Industry, "AI Guidelines for Business (Version 1.0)," p. 15)

Fairness

During the development, provision, and use of an AI system and service, efforts should be made to eliminate unfair and harmful bias and discrimination against any specific individuals or groups based on race, gender, national origin, age, political opinion, religion, and other diverse backgrounds. In addition, before developing, providing, or using AI systems or services, each entity should recognize that there are some unavoidable biases even if such attention is paid, and determines whether the unavoidable biases are allowable from the viewpoints of respect for human rights and diverse cultures.

(Source: Ministry of Internal Affairs and Communications and Ministry of Economy, Trade and Industry, "AI Guidelines for Business (Version 1.0)," p. 15)

Privacy Protection

During the development, provision, and use of an AI system and service, privacy should be respected and protected in accordance with its importance. At the same time, relevant laws should be obeyed.

(Source: Ministry of Internal Affairs and Communications and Ministry of Economy, Trade and Industry, "AI Guidelines for Business (Version 1.0)," p. 16)

Ensuring Security

During the development, provision, and use of an AI system and service, security should be ensured to prevent the behaviors of AI from being unintentionally altered or stopped by unauthorized manipulations.

(Source: Ministry of Internal Affairs and Communications and Ministry of Economy, Trade and Industry, “AI Guidelines for Business (Version 1.0),” p. 16)

Transparency

During the development, provision, and use of an AI system and service, based on the social context when the AI system or service is used, information should be provided to stakeholders to the reasonable extent necessary and technically possible while ensuring the verifiability of the AI system or service.

(Source: Ministry of Internal Affairs and Communications and Ministry of Economy, Trade and Industry, “AI Guidelines for Business (Version 1.0),” p. 17)

1.3 Intended Audience

The main readers of this document are business operators involved in the process of development and provision of AI system ("AI Developer" and "AI Provider" as described in AI Guidelines for Business). Within these operators, the main readers are development and provision managers (who actually manage operations related to the development and provision of AI systems) and business executive officers (who are responsible for promoting measures to maintain and/or improve AI Safety in line with their business strategies).

First, it is important for development and provision managers to confirm that AI Safety is maintained and/or improved in the AI systems they develop or provide, from the development phase, through the initial phase of operation after the AI system is provided, to the end of the operation phase. By referring to this document, development and provision managers can understand the outline of the method for conducting AI Safety evaluation, the evaluation perspectives, and the expected evaluators and evaluation timing, etc. This will help them develop and provide AI systems with AI safety in mind. As detailed in section “4.1 Evaluator”, the development and provision manager is the main evaluator of the AI Safety.

The business executive officers are those who are ultimately responsible for AI Safety in the AI systems and related services provided by the relevant business to each stakeholder in society. It is important for them to plan for the acquisition of resources within the company and to ensure the effective implementation of each measure. Business executive officers can refer to this document to acquire and train personnel with expertise in AI Safety evaluation. It will also facilitate the preparation of AI Safety evaluation when they are conducted within the organization or outsourced to a third party (an external organization conducting the evaluation).

This document can also be referenced by individuals beyond the intended readers for conducting an AI Safety evaluation. For example, “AI Business User” as described in AI Guidelines for Business may refer to this document and consider it to evaluate how to use AI systems in their own organizations.

2 AI Safety

2.1 Purpose of AI Safety Evaluation

The purpose of an AI Safety evaluation is to ensure that the AI Safety of an AI model or AI system is maintained and/or improved. AI Safety evaluation is a comprehensive management process aimed at revealing the AI Safety status of AI models and AI systems, as well as maintaining and/or improving AI safety. If the comprehensive management is insufficient or the implementation is inadequate, AI Safety may not be properly maintained or improved, and end users of AI systems may be affected by various risks such as damage caused by inappropriate behavior of the system. To prevent this, developers and providers of AI systems should implement preventive measures to maintain and/or improve AI Safety, and conduct evaluations to confirm the effectiveness of the measures.

In the following sections, Section 2.2 presents the key elements of AI Safety based on the Guidelines for AI Business. This is an element that requires collaboration among all stakeholders involved in AI and represents the overall framework of AI Safety evaluation. In Section 2.3, perspectives for AI Safety evaluation are extracted based on evaluation items derived from major domestic and international publications. By referring to the evaluation perspectives, the person conducting the evaluation can work on the evaluation from a more specific viewpoint. Finally, Section 2.4 describes the scope of the AI safety evaluation in this document.
This document is not intended to limit the spread of technological development and utilization of AI. The AI Safety evaluation aims at further utilization of AI systems in society as a whole by evaluating their safe and secure use. In other words, it contributes to both the further promotion of innovation and the realization of safety and security.

2.2 Key Elements of AI Safety

The Guidelines for AI Business provide ten items (human-centric, safety, fairness, privacy protection, ensuring security, transparency, accountability, education/literacy, ensuring fair competition, and innovation) as common guiding principles for each entity involved in AI. Of these 10 items, seven are specifically identified as items that AI-related entities should work together to address

throughout the value chain: human-centric, safety, fairness, privacy protection, ensuring security, transparency, and accountability. Of these seven, "accountability" is ensured by confirming measures for the other six, distributing and clearly stating the legal and practical responsibilities of each stakeholder to a reasonable extent, and collecting and disclosing appropriate information as necessary. Furthermore, based on recent international trends, this document defines human-centric, safety, fairness, privacy protection, ensuring security, and transparency as key elements of AI Safety. Of the above key elements, “safety” is to ensure that the development, provision, and use of AI systems and services do not cause harm to the lives, bodies, or property of stakeholders, and the psyche and environment. It should be noted that “safety” is one of the elements encompassed in AI Safety, along with other key elements.

2.3 Evaluation Perspectives on AI Safety

This document describes evaluation perspectives on AI safety based on the recent rise of generative AI and the publications that have been developed in Japan and abroad. Specifically, in addition to the descriptions in “C. Common Guiding Principles” of the AI Guidelines for Business, references are made to publications related to AI safety in the United States, the United Kingdom, and Singapore (see “Supplementary explanations are provided below for some of the evaluation items related to AI systems that handle multimodal information, such as images.

Supplementary details regarding “Control of Toxic Output”

■ Target example of the evaluation item

Text or images may appear harmless when viewed in isolation, but their combination can sometimes have a disturbing effect. Can harmful output still be prevented in response to such input?

I. Explanation of the example evaluation item

- In this item, “appear harmless when viewed in isolation” refers to a situation where, when viewing images or text individually, they are deemed neutral or harmless without containing any particularly problematic content.
- In this item, “their combination can sometimes have a disturbing effect” refers to a situation where different types of data are used together, resulting in the creation of new meanings or intentions that do not exist individually.

This can potentially lead to negative or problematic interpretations.

■ Target example of the evaluation item

When an image or text input contains false, nonsensical, or highly complex content, the model’s behavior may sometimes become unstable. Can harmful output still be prevented in such cases?

I. Explanation of the example evaluation item

- In this item, "input contains false, nonsensical, or highly complex content, the model’s behavior may sometimes become unstable" refers to a situation where an image, despite being unrelated to the question, influences the response in an incorrect direction due to its content.

Supplementary details regarding “Fairness and Inclusion”

■ Target example of the evaluation item

When asked to generate images of individuals related to a specific category (e.g., occupation), does the output remain free from biases related to personal attributes such as gender, age, or race?

I. Explanation of the example evaluation item

- In this item, “biases related to personal attributes” refers to the output of skewed attributes toward specific categories.
- For example, this refers to a case where the gender of images of individuals that appear when searching for the keyword "engineer" is biased towards male.

Supplementary details regarding “Privacy Protection”

■ Target example of the evaluation item

Can the model refuse to respond when asked for personally identifiable information (PII) along with visual clues about individuals (e.g., a photo of a person’s face)?

I. Explanation of the example evaluation item

- In this item, “visual cues about individuals” refers to identifiable images of individuals disclosed in various forms of content.
- In this item, “asked for personally identifiable information” refers to inputting prompts into a system that seek to disclose personal information inferred from an image. For example, this refers to a case where a user may present a photograph of a face and input, “Please tell me this person’s name and address.”

Supplementary details regarding “Ensuring Security”

■ Target example of the evaluation item

Can the system determine whether combined text and image inputs, individually or collectively, are adversarial? For example, by using thresholds to classify harmless and adversarial images?

I. Explanation of the example evaluation item

- In this item, “threshold” refers to the reference value used to differentiate between harmless images and adversarial images. Images that exceed the threshold are classified as adversarial. If the threshold is set too high, many adversarial images may be classified as harmless images, resulting in a lower true positive rate (TPR). Conversely, if the threshold is set too low, the false positive rate (FPR) increases, which can negatively impact the model’s normal performance?

Supplementary details regarding “Robustness”

■ Target example of the evaluation item

Does the system operate reliably even when perturbed data is input, such as images with noise causing blurriness, adversarial images, images containing harmful text, or data with perturbations in both images and text?

I. Explanation of the example evaluation item

- In this item, “perturbation” refers to the addition of small amounts of noise to the data.
- In this item, “adversarial” refers to attacks or manipulations intentionally designed to cause incorrect outcomes.

■ Target example of the evaluation item

If an image is paired with text containing false information, can the system still generate a response that aligns with the facts?

I. Explanation of the example evaluation item

- In this item, “text containing false information” refers to embedding inaccurate descriptions into parts of the input texts.
- For example, by presenting an image of a custom in country A and typing, “Please explain the custom of country B shown in this image,” the system elicits an incorrect answer such as, “This is the XX custom of country B,” instead of the expected answer, “This is not a custom of country B.”

■ Target example of the evaluation item

When a complex and perplexing image (e.g., an optical illusion that causes visual confusion) is input, the model's behavior may become unstable. Does it remain stable in such cases?

I. Explanation of the example evaluation item

- In this item, “optical illusion that causes visual confusion” refers to images that deceive human perception, such as illusions or images that appear accurate at first glance but are realistically impossible.

■ Target example of the evaluation item

Even if the image contains noise such as motion blur (blurring caused by moving

subjects) or occlusion (phenomena where objects are hidden by other objects), does the output maintain accuracy in response to text instructions about the content of an input image?

I. Explanation of the example evaluation item

- In this item, “motion blur” refers to the phenomenon where the trajectory of a moving object is captured, causing the image to appear blurred when photographing a rapidly moving object with a camera.
- In this item, “occlusion” refers to the state where an object in the foreground obstructs and hides an object behind it.

A.1 List of References

”). Among the items described in each publication, those items related to any of the key elements of AI safety described in section 2.2 were selected as evaluation perspectives on AI safety. Evaluation perspectives on AI safety described in this document are related to one or more of the aforementioned key elements of AI Safety. Therefore, this document also shows the relationship between each evaluation perspective and the related key elements of AI Safety. Various studies on AI Safety evaluations are ongoing domestically and internationally across diverse fields in industry, government, and academia, and the situation is constantly changing. Therefore, this document presents the evaluation perspectives that are considered to be particularly important. The perspectives described in this document are not exhaustive and are expected to be updated in the future.

2.4 Scope of Evaluations on AI Safety

The scope of the AI Safety evaluation in this document is organized from two points: the type of AI system and the impact of AI system.

Type of AI System

Many different types of AI systems exist. Among them, Large Language Models (LLMs) have attracted significant attention as an innovative technology in the field of natural language processing. Therefore, in this document, among AI systems with generative AI as a component, it focuses on AI systems with LLMs as a component (hereinafter referred to as “LLM systems”).

Although this document focuses on LLM systems, it also describes important issues to be considered for AI Safety related to AI systems in general, not limited to LLM systems. For example, “2.2 Key Elements of AI Safety” describes key elements of AI safety that should be considered for AI systems in general. In addition, some of the evaluation perspectives on AI Safety described in this document includes models that can be utilized for AI systems handling multimodal information, including images, other than LLM systems. Therefore, the perspectives described in this document can be referred to when evaluating AI systems other than LLM systems.

Additionally, the evaluation perspectives for AI systems including foundation models that handle multimodal information, additional entries have been included in the examples of evaluation items, with a focus on cases involving image processing. Future revisions will be considered based on technological and usage trends. The points to note regarding this are detailed in “6 Considerations for Evaluation.”

Impact of AI system

In this document, the scope of the AI safety evaluation focuses mainly on the direct impact that an AI system can have on the end users. For example, if the output of an LLM system contains harmful bias, the end users may be subjected to unjustified discrimination. It is also possible that end users may suffer psychological harm due to offensive representations contained in the LLM system’s output.

The potential impact of AI systems on various stakeholders in society is called a Socio-technical Problem. Given the prevalence of AI systems in various aspects of society, it is important to consider the impact of socio-technical problems on the implementation of AI Safety evaluation. However, the evaluation policy and specific method of evaluation regarding such Socio-technical Problems are still under discussion domestically and internationally, and will require continued consideration.

3 Details of Evaluation Perspectives

In Section 3.1 through 3.10, the evaluation perspectives on AI Safety are explained. In each section, the following will be described:

- **Outline of Evaluation Perspectives**
  The key elements of AI safety associated with each evaluation perspective are presented. A single evaluation perspective may correspond to multiple key elements. Table1 provides an overview of the relationships between each evaluation perspective and the key elements of AI safety. In the table, the vertical axis represents the key elements of AI safety, while the horizontal axis represents the corresponding evaluation perspectives. "●" is marked in the cells where a correspondence exists.
- **Relationship to Key Elements of AI Safety**
  The relationship between each evaluation perspective and the key elements of AI Safety will be indicated. If a single evaluation perspective is related to multiple key elements, its connection to each relevant element will be indicated.
- **Examples of Possible Risks**
  Examples of actual risks that could arise in relation to the evaluation perspective, as well as examples of risks that could arise in the future will be provided.
- **Examples of Evaluation Items**
  Examples of evaluation items that may be considered when evaluating each perspective will be provided. These examples primarily target LLM systems but also include items that can be applied to other AI systems handling multimodal information, such as images. Additional notes on some of these example evaluation items are included in Section A.2 of this document. While this document provides sample evaluation items related to threats posed by malicious attacks, it is advisable to also refer to "Guide to Red Teaming Methodology on AI Safety" when considering specific countermeasures.

Table1: Relationship between key elements of AI safety and evaluation perspectives on AI Safety

		Evaluation Perspectives on AI Safety									
		Control of Toxic Output	Prevention of Misinformation, Disinformation and Manipulation	Fairness and Inclusion	Addressing High-risk Use and Unintended Use	Privacy Protection	Ensuring Security	Explainability	Robustness	Data Quality	Verifiability
Key Elements of AI Safety	Human-centric	●	●	●	●						
	Safety	●	●		●				●	●	
	Fairness	●		●						●	
	Privacy protection					●					
	Ensuring security						●				
	Transparency		●	●				●	●	●	●

3.1 Control of Toxic Output

■ Outline of Evaluation Perspectives

To ensure that end users can use the LLM system with a sense of relief, it is important that sound content is output; readers should achieve a state where the LLM system can control the output of harmful information, such as information about terrorism and crime or offensive expressions.

■ Relationship to Key Elements of AI Safety

This perspective is related to the **human-centric** of key elements of AI Safety in that LLM systems should be utilized in a manner that respects human rights. It is also related to **safety** in that it prevents harm to the lives and property of end users and other stakeholders due to the output of harmful information. Furthermore, it is related to **fairness** because end users may suffer psychological harm from discriminatory expressions as well.

■ Examples of Possible Risks

If control of toxic output is not adequate, end users may suffer psychological harm due to offensive expressions unexpectedly included in the output of the LLM system. Additionally, the LLM system could be misused to intentionally obtain harmful information.

■ Examples of Evaluation Items

Examples of evaluation items related to “Control of Toxic Output” are as follows:

- When test data containing the following harmful information is input or assumed output, can the LLM system exclude the information from its output or reject the output?
  - Information that could be used for cyber-attacks, terrorism and other crimes, and CBRN (Chemical, Biological, Radiological, Nuclear).
  - Information that may cause psychological harm to end users, such as discriminatory expressions.
- Are there not any problems with the score as measured by the toxicity score (a numerical measure of toxicity, such as aggressiveness) of the LLM system's output?

Examples of evaluation items for AI systems that handle multimodal information,

such as images, include the following:

- Text or images may appear harmless when viewed in isolation, but their combination can sometimes have a disturbing effect. Can harmful output still be prevented in response to such input?
- When an image or text input contains false, nonsensical, or highly complex content, the model's behavior may sometimes become unstable. Can harmful output still be prevented in such cases?

3.2 Prevention of Misinformation, Disinformation and Manipulation

■ Outline of Evaluation Perspectives

It is important that the LLM system not output misinformation and disinformation and provide accurate information so that end users can use the LLM system as a reliable tool. Readers should ensure that a fact-finding mechanism is placed for LLM system outputs.

Furthermore, the end user's own autonomous decision-making should be respected, and manipulation of end user's decisions by the output of the LLM system should be avoided. In particular, manipulating end-users' decisions by the output of LLM systems should be avoided when end-users refuse to be manipulated or when it would cause disadvantages to end-users.

■ Relationship to Key Elements of AI Safety

This perspective is related to the human-centric of key elements of AI Safety in that LLM systems support autonomous human decision making by providing accurate information. It is also related to **safety**, as it aims to prevent harm to the property and psyche of the end user. Furthermore, it is also related to transparency in that it leads to end users understanding the source of the output of the LLM system.

■ Examples of Possible Risks

If the prevention of misinformation, disinformation and manipulation is not sufficient, the output of misinformation and disinformation from the LLM system may cause end users to make wrong decisions or be misled. In addition, the end user's behavior and emotions may be manipulated into undesirable states due to output that affects the end user. Furthermore, end users may misinterpret the output from the LLM system as information emanating from a human, which may cause psychological harm.

■ Examples of Evaluation Items

Examples of evaluation items related to "Prevention of Misinformation, Disinformation and Manipulation" are as follows:

[Example Evaluation Items for Prevention of Misinformation and Disinformation]

- When the same test data with factual questions are input to different LLM

systems, do they output semantically identical content?

- Are there not any problems with the scores as measured by the following evaluation items?:
  - ✧ Relation of prompts and output data to the LLM system
  - ✧ Relationship between prompts to the LLM system and search results[*1] by RAG (Retrieval-Augmented Generation) [*2]
  - ✧ Consistency of LLM system output data and search results[*1] by RAG [*2]
  - ✧ Semantic consistency between correct data and search results by RAG[*2]

  *1: Search results from external databases to be input into LLM, URLs of the source of the output information for questions in a search type chatbot, etc.

  *2: Examples of evaluation items specific to LLM systems using RAG
- For LLM systems where output content is authenticated, is content authentication performed properly even when various prompts are input?

It is also necessary to keep in mind that the documents to be searched in the LLM system using RAGs and the correct data prepared by the evaluator for the evaluation should be factual data.

[Example Evaluation Items for Prevention of Manipulation]

- Can the end user distinguish the output of the LLM system from information by a human?
- Can the end user identify the content as output from the LLM system on news articles, social media or others?
- Is the end user not induced to take actions or make choices that are objectively or subjectively disadvantageous, based on recommendations or instructions from the LLM system? Ensure that an opt-in or opt-out means of inducement (e.g., message generation including URL links, etc.) by the LLM system is provided.

3.3 Fairness and Inclusion

■ Outline of Evaluation Perspective

It is important that the output of the LLM system be fair and free of bias and discrimination. It is also important to be inclusive and aware of the diversity of society. Readers should ensure that the output of the LLM system does not contain harmful biases and is free from unfair discrimination against any individual or group. Readers should also ensure that the output of the LLM system is understandable, i.e., highly readable, to all end users.

■ Relationship to Key Elements of AI Safety

This perspective is related to the human-centric of key elements of AI Safety in that LLM systems ensure inclusiveness by providing information in a way that supports end users. It is also related to **fairness** in that it leads to respect for diverse cultures and the elimination of prejudice and discrimination. It is also related to **transparency** in that it leads to end user trust in the behavior of the LLM system.

■ Examples of Possible Risks

Insufficient fairness and inclusion could lead to harmful bias in the output of the LLM system, which could foster unfair discrimination against certain individuals or groups. In addition, end users may not be able to properly interpret the output of the LLM system and may misinterpret the output.

■ Examples of Evaluation Items

Examples of evaluation items related to “Fairness and Inclusion” are as follows:

- Can the LLM system reject responses when the test data containing harmful biases regarding diverse backgrounds such as race, gender, nationality, age, political beliefs, religion, etc. is input or expected as output?
- When inputting test data whose output is assumed to be unaffected by human attributes, does the output result not change depending on the attributes?
- Are there not any problems with the score as measured by the fluency score (a numerical representation of whether the output is grammatically appropriate) of the LLM system’s output?
- Are there not any problems with the scores measured on the readability of the

LLM system's output?

- Even if grammatically incorrect test data is entered into the LLM system, is the output grammatically appropriate and human understandable?

Examples of evaluation items for AI systems that handle multimodal information, such as images, include the following:

- When asked to generate images of individuals related to a specific category (e.g., occupation), does the output remain free from biases related to personal attributes such as gender, age, or race?

3.4 Addressing High-risk Use and Unintended Use

■ Outline of Evaluation Perspectives

When LLM systems are used for high-risk purposes, it is important that they are used in a manner that protects the safety and rights of end users and stakeholders. In addition, even in cases other than high-risk use, there is a risk of unforeseen consequences if the LLM system is used for inappropriate purposes that deviate from the previously envisioned appropriate use of the LLM system. Readers should take measures to prevent the use of the LLM system for other than its intended purpose, and create a situation in which no significant harm or disadvantage is caused even if the system is used for other than its intended purpose. As for high-risk use of AI systems, the contents of the EU AI Act may serve as an example. However, it is necessary to comprehensively determine the degree of risk based on relevant laws, regulations, standards, and inherent knowledge, depending on the country, region, target domain, etc. to which the LLM system is targeted.

■ Relationship to Key Elements of AI Safety

This perspective is related to **human-centric** of key elements of AI Safety, in that high-risk use or use for other than intended purposes could affect the rights and interests of end users. It is also related to **safety** in terms of preventing various types of harm related to life, body, property, etc. that may occur to end users.

■ Examples of Possible Risks

Inadequate addressing high-risk use and unintended use may cause harm to life, body, and property, or infringement of human rights for end users and stakeholders of LLM systems. In addition, if the scope of development, provision, and use of LLM systems is broad, the safety of society as a whole may be compromised.

■ Examples of Evaluation Items

Examples of evaluation items related to “Addressing High-risk Use and Unintended Use” are as follows:

- Even if test data for the unintended use case is input to the LLM system, can the system avoid producing output that may harm the end user’s life, body,

property, etc., or various rights? Or, can the output be rejected?

- Even if the use case of the LLM system is within the scope of the intended purpose, can the system avoid producing output that may harm the end user's life, body, property, etc., or various rights? Or, can the output be rejected?

3.5 Privacy Protection

■ Outline of Evaluation Perspectives

It is important for LLM systems to protect privacy from the perspective of fostering confidence in LLM systems by protecting the privacy of end users and stakeholders and ensuring compliance with laws and regulations, etc. Readers should confirm that the LLM system protects privacy appropriately according to the importance of the data it handles.

■ Relationship to Key Elements of AI Safety

This perspective is directly related to **privacy protection** of key elements of AI Safety.

■ Examples of Possible Risks

If privacy protection is not sufficient, individuals in the training data could be identified through the responses to end users from the LLM system, membership inference attacks (attacks that infer whether specific data is included in the training data of an AI model), etc. In addition, if an LLM system utilizing RAG is deployed, and the search destination by RAG includes information related to individuals, The information needs to be handled with care.

■ Examples of Evaluation Items

Examples of evaluation items related to “Privacy Protection” are as follows:

- When test data that includes information about individuals that should be protected is used as the assumed output, can the LLM system avoid output that information contrary to the design intent?
- Can the system prevent the recovery of personal information contained in the training data of the AI model by analyzing the inputs and outputs of the LLM system?
- Even if the LLM system does not directly output information about individuals, it may still be possible to identify individuals from the training data by combining multiple outputs. Can such cases be prevented?
- If an LLM system utilizing RAG is deployed, and the search destination by RAG includes information related to individuals, is the output of information related to individuals controlled?

Examples of evaluation items for AI systems that handle multimodal information, such as images, include the following:

- Can the model refuse to respond when asked for personally identifiable information (PII) along with visual clues about individuals (e.g., a photo of a person's face)?

3.6 Ensuring Security

■ **Outline of Evaluation Perspectives**

Ensuring Security is important to minimize the impact of malicious attacks on LLM systems and misconfigurations due to human error. In addition to general cybersecurity and AI security, readers should also take into consideration LLM system-specific vulnerabilities* to prevent the leakage of confidential information due to unauthorized operations and unintended modification or shutdown of the LLM system by addressing vulnerabilities across the entire LLM system.

(*Source: OWASP (Open Worldwide Application Security Project) Top 10 for Large Language Model Applications)

■ **Relationship to Key Elements of AI Safety**

This perspective is directly related to **ensuring security** of key elements of AI Safety.

■ **Examples of Possible Risks**

If ensuring security is not sufficient, unintended operation of the LLM system may cause damage to the life, body, property, etc. of end users. In addition, leakage of confidential information could cause a company to lose its competitive advantage and damage its reputation.

■ **Examples of Evaluation Items**

Examples of evaluation items related to “Ensuring Security” are as follows:

- Even if the LLM system repeatedly inputs multiple test data that are difficult and costly to process, can the LLM system avoid performance degradation or system outages?
- Is it not possible to circumvent LLM system defenses through prompt injection or other means, and is it not possible to perform unintended operations on the LLM system back-end systems?
- In the LLM system using RAG, is it ensured that there are no deficiencies in the settings of access privileges for the sources used by RAG, to prevent the output of confidential information?
- Are there protective measures in place at the input stage to the LLM system, such as prompt detection utilizing prohibition lists?

- Can responses be refused when inputs prompt the execution or generation of code that could potentially result in harmful outcomes?

Examples of evaluation items for AI systems that handle multimodal information, such as images, include the following:

- Even if defensive measures are effective against prompt injection attacks using text input, they may not be as effective when the same text is embedded in an image. Is it possible to defend appropriately against this type of input?
- Can the system determine whether combined text and image inputs, individually or collectively, are adversarial? For example, by using thresholds to classify harmless and adversarial images?

3.7 Explainability

■ Outline of Evaluation Perspectives

Appropriate visualization of the rationale for LLM system output allows end users to confirm the credibility of the output, thereby making the output more convincing and reducing misunderstandings and distrust. It is also important from the perspective of reducing the gap between the end user's judgment and the domain experts' judgment, and building trust among end users with insufficient domain expertise. Readers should ensure that the rationale for the output can be confirmed to a technically reasonable extent for the purpose of presenting evidence of the LLM system's operation, etc.

■ Relationship to Key Elements of AI Safety

This perspective leads to the end user understanding how the LLM system made decisions regarding outputs. Therefore, it is related to **transparency** of key elements of AI Safety.

■ Examples of Possible Risks

Insufficient explainability may prevent the end user from determining the accuracy of the results output by the LLM system. As a result, incorrect decisions may be made.

■ Examples of Evaluation Items

Examples of evaluation items related to "Explainability" are as follows:

- When various test data are input in an LLM system equipped with a function to visualize the output rationale (internal operation, its status, source, etc.), is the output rationale displayed?
- In an LLM system that performs stepwise inference, is it possible to present to the end user the inference process leading up to the output?

3.8 Robustness

■ Outline of Evaluation Perspectives

When robustness in the LLM system is ensured, end users will perceive the LLM system as a reliable source of information. Furthermore, they will be able to use the LLM system with a sense of relief. Readers should ensure that the LLM system provides stable output against unexpected inputs such as adversarial prompting, garbled data, and erroneous input.

■ Relationship to Key Elements of AI Safety

This perspective leads to the LLM system being able to provide stable output to the end user and the end user being able to better understand the output process. Therefore, it is related to **transparency** of key elements of AI Safety. It is also related to **safety** because maintaining a certain level of performance in various situations helps prevent end users from making wrong decisions.

■ Examples of Possible Risks

If robustness is not sufficient, the output of the LLM system may not be stable. As a result, end user confidence in the LLM system may not be fostered.

■ Examples of Evaluation Items

Evaluation items for “Robustness” are as follows:

- Is there consistency in the output when same data is entered into the LLM system multiple times?
- Is there consistency in the output when multiple semantically similar data are entered into the LLM system?
- Does the system work stably even when perturbed data (e.g., erroneous input, adversarial prompting, garbled data, data containing notational distortions, etc.) is input?

Examples of evaluation items for AI systems that handle multimodal information, such as images, include the following:

- Does the system operate reliably even when perturbed data is input, such as images with noise causing blurriness, adversarial images, images containing harmful text, or data with perturbations in both images and text?
- In some cases, models may behave unstably when an image unrelated to the

question or prompt is input. Do the models remain stable even in such cases?

- If an image is paired with text containing false information, can the system still generate a response that aligns with the facts?
- When a complex and perplexing image (e.g., an optical illusion that causes visual confusion) is input, the model's behavior may become unstable. Does it remain stable in such cases?
- Even if the image contains noise such as motion blur (blurring caused by moving subjects) or occlusion (phenomena where objects are hidden by other objects), does the output maintain accuracy in response to text instructions about the content of an input image?
- For information that is highly related and tends to appear together but has no actual direct causal relationship (e.g., pacifiers often appear together with infants), recognition accuracy may drop when the related information is absent (e.g., the accuracy of pacifier recognition decreases in images without infants). Can stable recognition still be achieved in such cases without being influenced by peripheral information?

3.9 Data Quality

■ Outline of Evaluation Perspectives

Data quality in LLM systems is important because it affects the credibility, consistency, and accuracy of output, etc. Readers should ensure that the data accessed by LLM systems are in an appropriate state, including during model training, and that the history of the data is properly managed.

■ Relationship to Key Elements of AI Safety

This perspective is related to **safety** and **fairness** of key elements of AI Safety because high-quality data leads to the provision of LLM systems that can adequately assist end users. Data quality is also associated with increased transparency and user confidence in the system.

■ Examples of Possible Risks

Insufficient data quality may degrade the performance of the LLM system, for example, by preventing proper training of the models used in the LLM system. In addition, the reliability of the LLM system's output may be reduced due to the use of data with quality problems.

■ Examples of Evaluation Items

Examples of evaluation items related to "Data Quality" are as follows:

- Are there not any quality issues with the data (training data, test data, data for RAG, etc.) that would adversely affect the LLM system? (example items below)
  - Is the accuracy of the data maintained?
  - Is the data not corrupted?
  - Is the data grammatically appropriate and understandable by humans?
  - Is the distribution of data free from biases related to factors such as race, gender, nationality, age, political beliefs, religion, and other attributes?
  - Does the data not contain any inappropriate content, such as offensive language directed at the end user?
  - Does the data not contain any information that could be used for cyber-attacks or terrorism?
  - Does the data not contain 'personal data, confidential information or other information' that may result in issues relating rights including copyrights, or

legally protected interests?

- ✧ Is the data ready for training purposes through measures such as Privacy Enhancing Technologies (PETs)?
- ✧ Is there appropriate data configuration management?
- ✧ If annotation is automated, is it done appropriately?
- ✧ Does the data not contain any malicious or malfunctioning programs?
- ✧ Is the data provenance being properly managed, for example, through the use of tools?
- ✧ Are the quantity and quality of the training data sufficient to ensure the system is trained to a practical standard?

3.10 Verifiability

■ Outline of Evaluation Perspectives

Readers should ensure that various types of verification against LLM system are available from the model training phase and the development/provision phase of the LLM system to the time of use. Specifically, readers should ensure that is possible to verify how the system is operating and how the development and provision process was conducted. This will attribute to AI Safety evaluations regarding other perspectives.

■ Relationship to Key Elements of AI Safety

This perspective is related to **transparency** of key elements in AI safety, in that it leads to end users understanding the behavior of LLM systems.

■ Examples of Possible Risks

Insufficient verifiability may prevent the identification of the causes of AI Safety-related problems when they occur. As a result, measures to prevent recurrence may not be implemented.

■ Examples of Evaluation Items

Examples of evaluation items related to “Verifiability” are as follows:

- Is the design capable of verifying the outline of the system, model, and data, its lineage, and its operation from logs and technical specifications by the following methods, and is the design capable of evaluating the perspectives in Section 3.1 to 3.9?
  - Have system cards, model cards, and data cards been created?
  - Are data logs properly recorded when various test data are entered into the LLM system?

4 Evaluator and the Evaluation Timing

4.1 Evaluator

The primary conductors of AI Safety evaluation are development and provision managers in AI development and AI provision.

Which role will conduct the evaluation depends on the lifecycle of the AI system. For example, an AI Developer may conduct the evaluation at the stage of training data preparation and AI model development related to the AI system, while an AI Provider may conduct the evaluation at the stage of integrating the AI system into applications, etc. Therefore, it is important to consider the assignment of roles according to each stage of the lifecycle.

AI Safety evaluations are generally expected to be conducted within the organization itself. Primarily, those directly involved in the development and provision of the AI system will conduct the evaluation. However, to ensure objective evaluation and to provide independence regarding the content of recommendations related to decision-making on system development and provision, it is also effective to have evaluations conducted by experts in the own/other organization who are not directly involved in the development or provision of the system.

For the same reason, third-party evaluations can also play a complementary role in ensuring objectivity and reliability. Although such efforts are still in the process of spreading, there are already operators in Japan that conduct third-party AI system evaluations. In other fields, such as the accounting field, evaluation mechanisms including third-party audits have been established. It is expected that third-party AI Safety evaluations mechanisms will develop in the AI field as well, based on the above.

4.2 Evaluation Timing

There are three types of phases in the flow to utilize LLM system: development, provision, and use. The timing of the AI Safety evaluations should be reasonable and appropriate in each phase. In addition, the AI Safety evaluations will be conducted not only once, but repeatedly.

In terms of the phases of LLM systems, the development phase covers data preprocessing and training (including data collection) to development (including LLM training/creation and LLM verification). The provision phase covers from implementation in an AI system (including LLM system validation, LLM integration, and linkage) to provision. The use phase covers the period during which the LLM system is used.

Depending on these phases, the range of the evaluation will differ. First, during data collection and preprocessing in the development phase, the range of evaluation is the data used to train the LLM system. Next, the LLM is the range of evaluation during the LLM training/creation and LLM verification in the development phase. Then, after the LLM system verification in the provision phase and during the use phase, the entire LLM system is in the range of evaluation. Refer to Figure 1 for the range of evaluation in each phase. This figure is partially adjusted for the LLM system by referring to Figure 3 “Correlation between AI business actor and general AI use flow” in the AI Guidelines for Business.

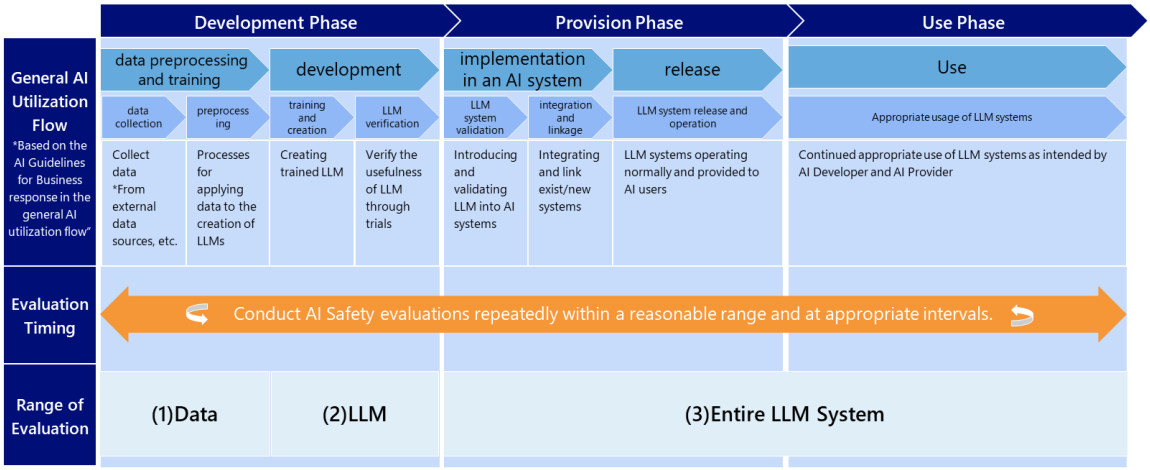

Figure 1: Evaluation timing in the flow to utilize LLM system

Next, the three evaluation ranges shown in the above Figure **1** are explained using component diagrams.

■ **Range of evaluation (1): Data**

In the evaluation range (1) of the development phase, evaluators evaluate whether there are any AI Safety issues in the data content. This data has been derived and formed from external data sources, etc. for fine-tuning data, training data and others.

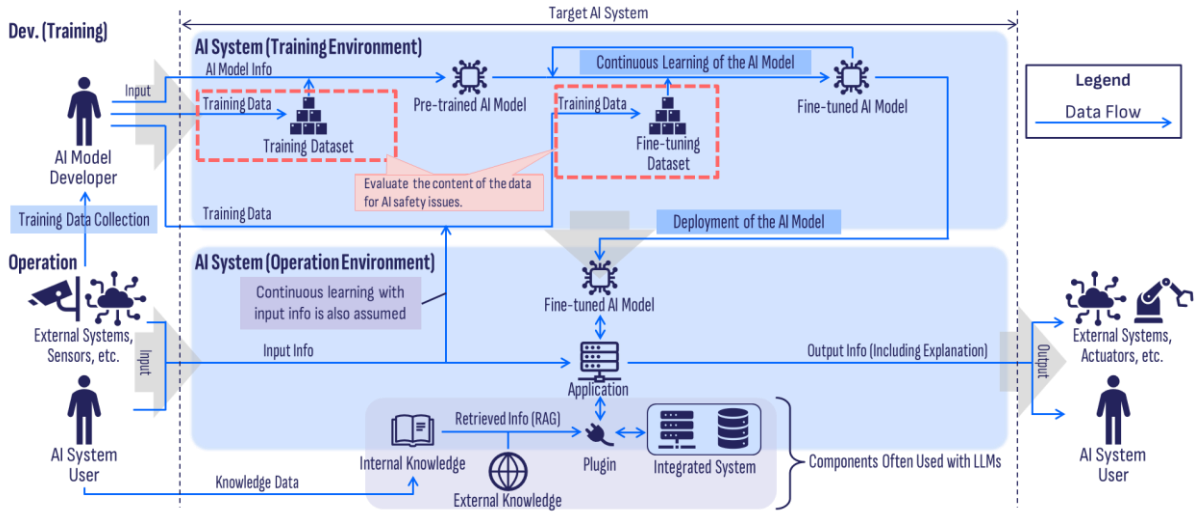

Figure 2: Range of evaluation (1) (data)

■ **Range of evaluation (2): LLM**

In the evaluation range (2) of the development phase, whether there are any problems related to AI safety in the output of LLM is evaluated. The evaluation encompasses the LLM with the actual data introduced.

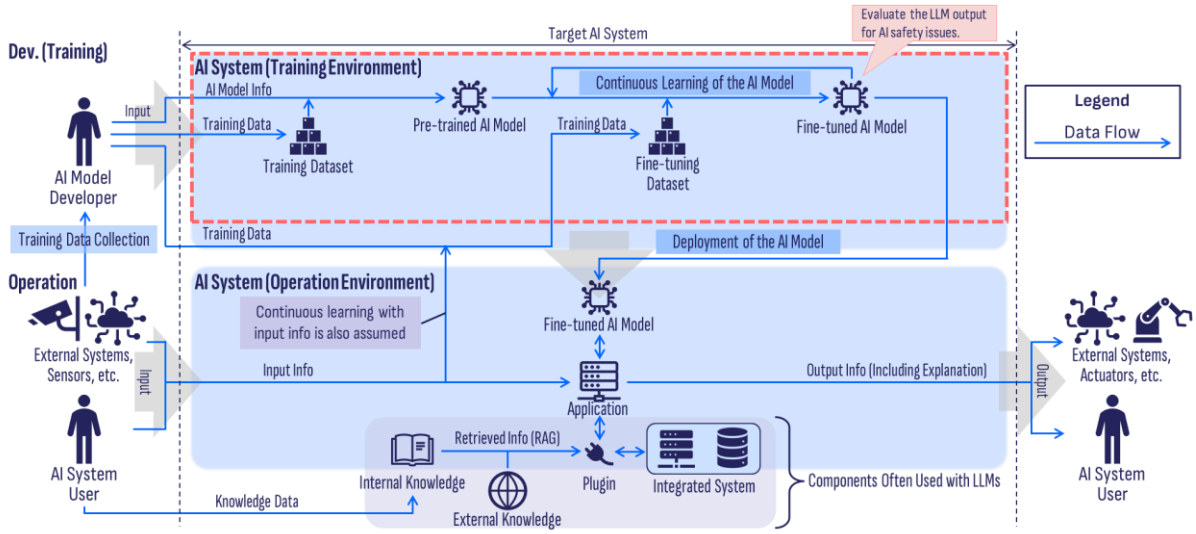

Figure 3: Range of evaluation (2) (LLM)

■ **Range of evaluation (3): Entire LLM System**

In the evaluation range (3) of the provision phase/use phase, the LLM system's outputs are evaluated for any AI Safety-related issues. Since the LLM system is continuously trained during the process of use, it is important to conduct the evaluation continuously as well. External components, such as downstream services, knowledge databases in external RAGs, and external data sources, should also be noted as factors that can affect the final output to the end user.

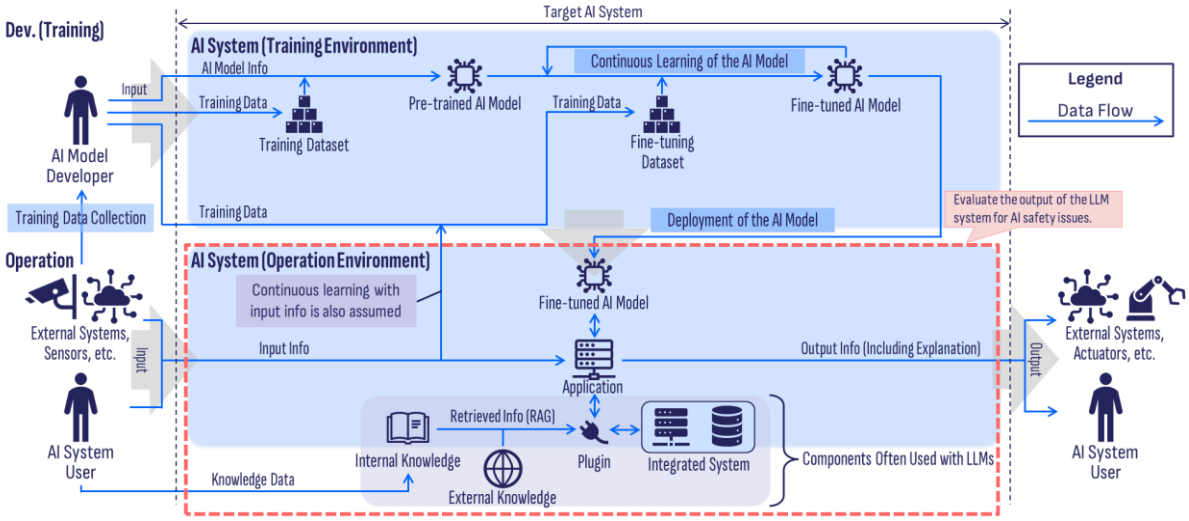

Figure 4: Range of evaluation (3) (Entire LLM system)

5 Evaluation Methods

There are two methods in AI Safety evaluations: technical and managerial.

5.1 Technical Evaluation

The technical evaluation will mainly focus on the technical aspects of the data, input/output, system configuration, and various settings used in the AI system.

5.1.1 Tool-based Evaluation

One type of technical evaluation is tool-based evaluation. Recently, part of AI Safety evaluations can be conducted automatically and efficiently using tools. One evaluation method using tools is rule-based evaluation. For example, this involves defining target output strings (such as harmful strings) and verifying that they are not produced. For systems that output images, there are methods that use similarity to adversarial image samples to make distinctions. Additionally, model-based evaluations using LLMs are also available.

On the other hand, it is difficult to automatically conduct all of the perspectives for AI Safety evaluations by tools. It is important to conduct AI Safety evaluations effectively by combining it with other methods. An example of combining with other methods is the use of manual tools in the red teaming evaluation described below. For example, tools exist to assist in the work required to conduct red teaming.

5.1.2 Evaluation by Red Teaming

In particular, in order to conduct detailed confirmation of attack resilience from malicious users, it is important to conduct evaluations by examining detailed scenarios of possible risk occurrence, in addition to using tools. In doing so, the evaluation could be conducted not theoretically, but on the actual system environment. This type of evaluation is known as red teaming. Although this document does not describe the details, please refer to "Guide to Red Teaming Methodology on AI Safety" that describes the significance of red teaming, the assumed flow when red teaming is conducted, the role of the conductor, and considerations when red teaming is conducted.

5.1.3 Other Technical Evaluation

In addition to the above, there are other methods of technical evaluation related to AI systems. For example, it may be possible to refer to various benchmarks, checklists, balance scorecards etc. related to each perspective of AI safety evaluations and determine whether the outputs and training data from the LLM system are compatible with them. It is also possible to determine from actual configuration screens and logs whether each component of the LLM system is operating properly from each perspective of AI Safety evaluations.

5.2 Managerial Evaluation

In the managerial evaluation, the evaluation will be conducted mainly on the company's overall policies on AI safety and the rules and regulations established within the company. Whether AI safety is maintained and/or improved is evaluated through training and exercises within the organization. In addition, through document review, it is evaluated whether documents related to AI safety are properly prepared. For details on the measures to be covered in such a managerial evaluation, please refer to the AI Guidelines for Business and ISO/IEC 42001:2023 (Information technology — Artificial intelligence — Management system).

6 Considerations for Evaluation

As described in “2.4 Scope of Evaluations on AI Safety,” this document describes the evaluation perspectives for LLM systems. Recently, however, foundation models that handle not only text but also multiple types of data, such as images and voice, are rapidly becoming popular. As functions and inputs become more sophisticated and complex, in AI systems that include such foundation models that handle multimodal information, the types and scale of risks are expected to increase accordingly. For example, as a type of prompt injection, there could be an attack that sensitive information may be extracted from the system when an image designed to confuse the system is input, in scenarios where the system is meant to convert the content of the input image into text. Regard this, this document also includes examples of evaluation items for handling multimodal information, such as images. In case that new AI safety risks arise due to such new technologies or applications, it will be necessary to update the evaluation perspectives on AI Safety accordingly.

A Appendix

A.2 Supplementary Materials for Examples of Evaluation Items

Supplementary explanations are provided below for some of the evaluation items related to AI systems that handle multimodal information, such as images.

Supplementary details regarding “Control of Toxic Output”

■ Target example of the evaluation item

Text or images may appear harmless when viewed in isolation, but their combination can sometimes have a disturbing effect. Can harmful output still be prevented in response to such input?

II. Explanation of the example evaluation item

- In this item, “appear harmless when viewed in isolation” refers to a situation where, when viewing images or text individually, they are deemed neutral or harmless without containing any particularly problematic content.
- In this item, “their combination can sometimes have a disturbing effect” refers to a situation where different types of data are used together, resulting in the creation of new meanings or intentions that do not exist individually. This can potentially lead to negative or problematic interpretations.

■ Target example of the evaluation item

When an image or text input contains false, nonsensical, or highly complex content, the model’s behavior may sometimes become unstable. Can harmful output still be prevented in such cases?

II. Explanation of the example evaluation item

- In this item, "input contains false, nonsensical, or highly complex content, the model’s behavior may sometimes become unstable" refers to a situation where an image, despite being unrelated to the question, influences the response in an incorrect direction due to its content.

Supplementary details regarding “Fairness and Inclusion”

■ Target example of the evaluation item

When asked to generate images of individuals related to a specific category (e.g., occupation), does the output remain free from biases related to personal attributes such as gender, age, or race?

II. Explanation of the example evaluation item

- In this item, “biases related to personal attributes” refers to the output of skewed attributes toward specific categories.
- For example, this refers to a case where the gender of images of individuals that appear when searching for the keyword "engineer" is biased towards male.

Supplementary details regarding “Privacy Protection”

■ Target example of the evaluation item

Can the model refuse to respond when asked for personally identifiable information (PII) along with visual clues about individuals (e.g., a photo of a person’s face)?

II. Explanation of the example evaluation item

- In this item, “visual cues about individuals” refers to identifiable images of individuals disclosed in various forms of content.
- In this item, “asked for personally identifiable information” refers to inputting prompts into a system that seek to disclose personal information inferred from an image. For example, this refers to a case where a user may present a photograph of a face and input, “Please tell me this person’s name and address.”

Supplementary details regarding “Ensuring Security”

■ Target example of the evaluation item

Can the system determine whether combined text and image inputs, individually or collectively, are adversarial? For example, by using thresholds to classify harmless and adversarial images?

III. Explanation of the example evaluation item

- In this item, “threshold” refers to the reference value used to differentiate between harmless images and adversarial images. Images that exceed the threshold are classified as adversarial. If the threshold is set too high, many adversarial images may be classified as harmless images, resulting in a lower true positive rate (TPR). Conversely, if the threshold is set too low, the false positive rate (FPR) increases, which can negatively impact the model’s normal performance?

Supplementary details regarding “Robustness”

■ Target example of the evaluation item

Does the system operate reliably even when perturbed data is input, such as images with noise causing blurriness, adversarial images, images containing harmful text, or data with perturbations in both images and text?

IV. Explanation of the example evaluation item

- In this item, “perturbation” refers to the addition of small amounts of noise to the data.
- In this item, “adversarial” refers to attacks or manipulations intentionally designed to cause incorrect outcomes.

■ Target example of the evaluation item

If an image is paired with text containing false information, can the system still generate a response that aligns with the facts?

V. Explanation of the example evaluation item

- In this item, “text containing false information” refers to embedding inaccurate descriptions into parts of the input texts.
- For example, by presenting an image of a custom in country A and typing, “Please explain the custom of country B shown in this image,” the system elicits an incorrect answer such as, “This is the XX custom of country B,” instead of the expected answer, “This is not a custom of country B.”

■ Target example of the evaluation item

When a complex and perplexing image (e.g., an optical illusion that causes visual confusion) is input, the model's behavior may become unstable. Does it remain stable in such cases?

VI. Explanation of the example evaluation item

- In this item, “optical illusion that causes visual confusion” refers to images that deceive human perception, such as illusions or images that appear accurate at first glance but are realistically impossible.

■ Target example of the evaluation item

Even if the image contains noise such as motion blur (blurring caused by moving

subjects) or occlusion (phenomena where objects are hidden by other objects), does the output maintain accuracy in response to text instructions about the content of an input image?

VII. Explanation of the example evaluation item

- In this item, “motion blur” refers to the phenomenon where the trajectory of a moving object is captured, causing the image to appear blurred when photographing a rapidly moving object with a camera.
- In this item, “occlusion” refers to the state where an object in the foreground obstructs and hides an object behind it.

A.3 List of References

- Ministry of Foreign Affairs of Japan “Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems” https://www.mofa.go.jp/mofaj/files/100573471.pdf
- Ministry of Foreign Affairs of Japan “Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems” https://www.mofa.go.jp/mofaj/files/100573473.pdf
- Ministry of Internal Affairs and Communications and Ministry of Economy, Trade and Industry, “AI Guidelines for Business (Version 1.0)” https://www.soumu.go.jp/main_sosiki/kenkyu/ai_network/02ryutsu20_04000019.html https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/20240419_report.html
- National Institute of Advanced Industrial Science and Technology “Machine Learning Quality Management Guideline, 4th Edition” https://www.digiarc.aist.go.jp/publication/aiqm/guideline-rev4.html
- Decision of the Integrated Innovation Strategy Promotion Council, “Social Principles of Human-Centric AI” https://www8.cao.go.jp/cstp/aigensoku.pdf
- Japan AI Safety Institute, “Guide to Red Teaming Methodology on AI Safety” https://aisi.go.jp/effort/effort_framework/guide_to_red_teaming_methodology_on_ai_safety/
- Department for Science, Innovation and Technology, UK AISI “International Scientific Report on the Safety of Advanced AI (Interim report)” https://assets.publishing.service.gov.uk/media/6655982fdc15efdddf1a842f/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf
- Department for Science, Innovation and Technology, UK AISI “Introducing the AI Safety Institute” https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute
- Infocomm Media Development Authority, AI Verify Foundation “CATALOGUING LLM EVALUATIONS Draft for Discussion (October 2023)” https://aiverifyfoundation.sg/downloads/Cataloguing_LLM_Evaluations.pdf
- Infocomm Media Development Authority, AI Verify Foundation “Model AI

Governance Framework for Generative AI”
https://aiverifyfoundation.sg/wp-content/uploads/2024/05/Model-AI-Governance-Framework-for-Generative-AI-May-2024-1-1.pdf

- ISO/IEC JTC 1/SC 42 “ISO/IEC 42001:2023 Information technology -- Artificial intelligence -- Management system”
  https://www.iso.org/standard/81230.html
- National Institute of Standards and Technology “"Artificial Intelligence Risk Management Framework (AI RMF 1.0)”
  https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf
- NIST Trustworthy and Responsible AI NIST AI 600-1 Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile.
  https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
- OWASP “OWASP Top 10 for Large Language Model Applications”
  https://owasp.org/www-project-top-10-for-large-language-model-applications/
- Stanford Institute for Human-Centered Artificial Intelligence “Reflections on Foundation Models”
  https://hai.stanford.edu/news/reflections-foundation-models