Efforts of Sony Group concerning Democratization of Generative AI for Adaptation to Business Operations

Masahiro Oba
Senior General Manager
Digital &Technology Platform AI Acceleration Dept.
Sony Group Corporation

Taichi Hirano
Senior Architect
Digital &Technology Platform AI Acceleration Dept.
Sony Group Corporation

1. Introduction

The rapid advancement of generative artificial intelligence (AI) is dramatically changing how companies use AI. In particular, the emergence of large language models (LLMs) has demonstrated the possibility of expanding the use of AI to general business users as well as experts. In this special feature, the efforts of Sony Group to "democratize" generative AI and adapt it to business processes—including technical aspects and actual use cases—are introduced.

2. Overview of Sony Group and Strategy for Generative AI

2.1 Business structure of Sony Group

Generating annual total sales of approximately 13 trillion yen, Sony Group mainly operates in the following six business segments: game, music, picture, finance, semiconductors, and electronics. Each business unit is unique and diverse, and the autonomy of each unit is emphasized under the corporate culture of "free and open-mindedness."

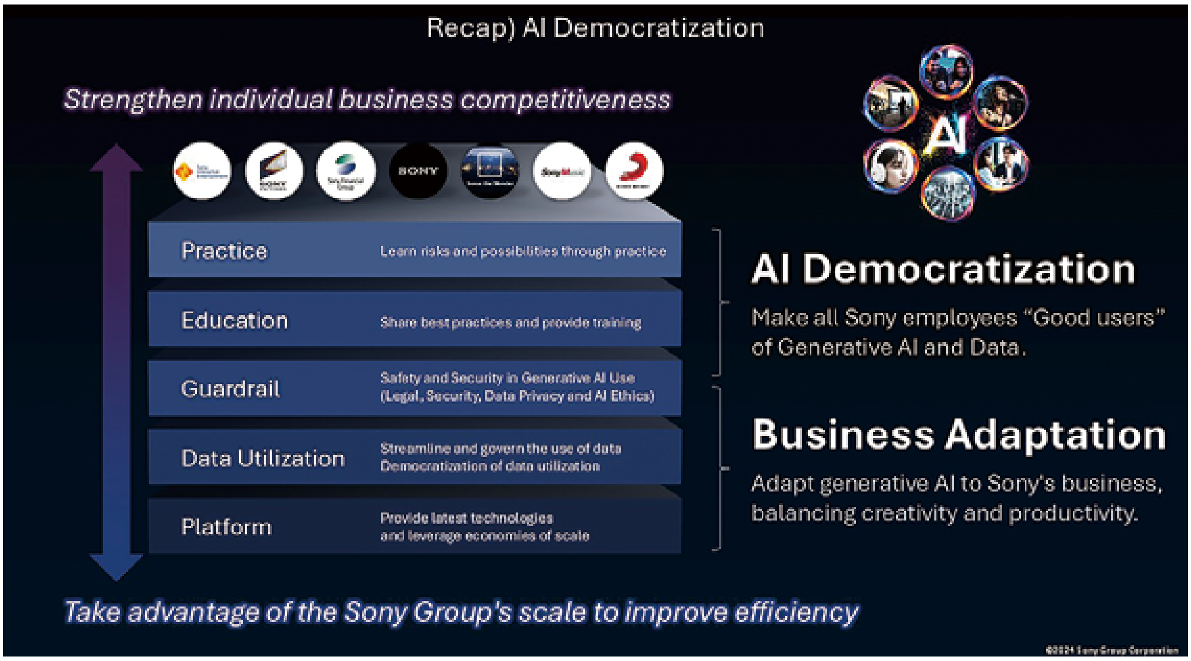
2.2 A vision for generative AI

The Sony Group has a vision of "democratizing AI, technology, and data to enable all Sony Group's employees to become good users in a manner that achieves both creativity and productivity." In particular, we believe that "creativity resides in people, and AI supports creativity," and we aim to increase both the creativity and productivity of creators and the employees who support them by utilizing generative AI properly.

2.3 AI-democratization stack

To democratize generative AI, the Sony Group is building an "AI-democratization stack." As a combination of technologies, this stack will provide the infrastructure needed across the entire Group, including a platform and guardrails that allow all employees to use generative AI with confidence, as well as the educational content and the latest information, including company-wide global events. We have set key performance indicators (KPIs) for each stack, and we are working to accelerate the democratization of generative AI and its business adaptation across the entire Group.

■ **Figure 1: AI democratization strategy**



Recap) AI Democratization

Strengthen individual business competitiveness

Practice — Learn risks and possibilities through practice

Education — Share best practices and provide training

Guardrail — Safety and Security in Generative AI Use (Legal, Security, Data Privacy and AI Ethics)

Data Utilization — Streamline and govern the use of data Democratization of data utilization

Platform — Provide latest technologies and leverage economies of scale

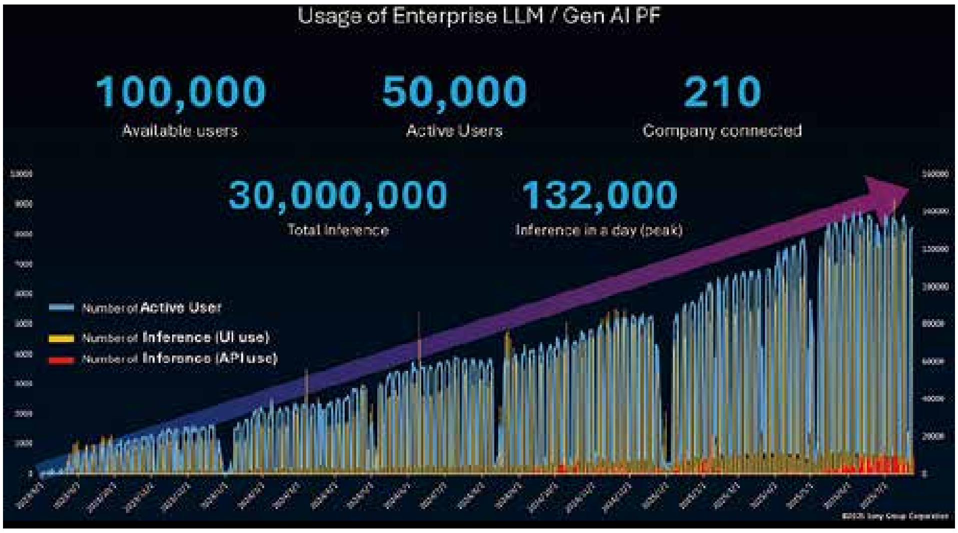
AI Democratization
Make all Sony employees "Good users" of Generative AI and Data.

Business Adaptation
Adapt generative AI to Sony's business, balancing creativity and productivity.

Take advantage of the Sony Group's scale to improve efficiency

©2024 Sony Group Corporation

Figure 2: Usage of Enterprise LLM/GenAI PF



3. Enterprise LLM: Foundation for generative AI

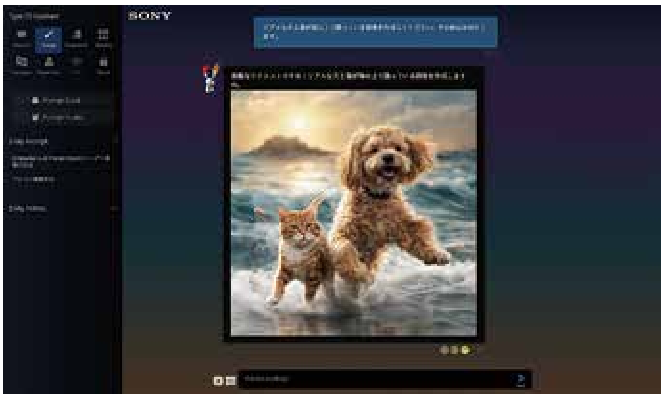
3.1 Overview and features

As a first step in democratization of generative AI, Sony Group is providing all employees with a chat-style web application called "Enterprise LLM." Built in a cloud-native, auto-scaling environment mainly using AWS (Amazon Web Services cloud), this app has been used by 210 group companies and has 50,000 active users (as of August 1, 2025).

The key features of Enterprise LLM are listed as follows:

- Security and guardrails: To ensure the appropriate use of internal data, we work with the security, data-privacy, legal, and AI ethics departments to establish systemic security and rules and guidelines.
- Multi-cloud support: To keep up with the rapid evolution of LLMs, Enterprise LLM connects to multiple cloud environments, including AWS, Google Cloud, and Microsoft Azure, and enables use of over 130 LLMs and text-to-image models.
- Use-case optimization: Equipped with various support functions for business use, Enterprise LLM provides AI types and prompt-input assistance optimized for common use cases.

Figure 3: Enterprise LLM



3.2 Usage and effectiveness

Full-scale deployment of enterprise LLM began in August 2023, and since then, it has handled over 130,000 generation requests per day and executed over 20-million inferences as of August 2025. Major cloud-platform providers have also praised it for being in an advanced state with a high level of activity compared to similar services provided by other companies.

To understand usage of Enterprise LLM, we have created an environment that allows us to understand usage in real time while maintaining anonymity. This understanding is enabled by having the AI itself automatically classify and analyze input prompts while taking privacy into consideration.

3.3 Key use cases and productivity benefits

The main use cases of Enterprise LLM within the Sony Group are listed as follows:

1. Writing reports and emails
2. Translation of different languages
3. Program generation and coding support
4. Summarization and analysis of text
5. Generating ideas and brainstorming

For each of these use cases, AI is used to measure productivity improvements in real time. For example, we calculated that using AI in "creating reports and emails" can reduce the time spent on each task by an average of 25 minutes. And we estimated that this time reduction will result in a monthly savings of approximately

50,000 hours across Sony Group.

3.4 Positioning and significance of awareness-raising activities

As for democratization of generative AI, organizational awareness activities are as important as developing the technological infrastructure. Sony Group has positioned our awareness program as a key pillar of the AI-democratization stack, and we are implementing a company-wide initiative with strong executive endorsement. Not simply as training programs, awareness activities are implemented as strategic measures to simultaneously transform organizational culture and promote the use of technology.

According to the data, approximately 50,000 community members accessed educational content via Web and Teams, and 10,000 people participated in events with hands-on sessions and consultations. These figures are evidence that proper educational activities are the foundation for technology adoption of generative AI.

3.5 Multi-layered structure of awareness programs

The awareness-raising activities of Sony Group consist of the following multi-layered approach:

1. Regular events: More than 60 training events are held annually, and “Gen AI Day,” which invites other vendors, is held six times.
2. Knowledge sharing by experts: Sharing insights through presentations at major industry events (seven times)
3. Technical consultations: Practical support through 300 technical consultations
4. Ongoing engagement: 7,500 people continue to experience new technology (generative AI) every month after registering

This multi-layered structure is unique in that it satisfies various levels of learning needs, which range from providing basic knowledge for beginners to specialized applications.

3.6 Performance indicators for awareness-raising activity

The effectiveness of our awareness-raising activities is measured by the following three quantitative indicators:

- Monthly active users of generative AI: 8,000
- Total registered generative-AI users: 50,000
- Inquiries to the Technical Support Team (Center of Excellence): 600

These indicators show that awareness-raising activities go beyond simply transferring knowledge and lead to actual adoption of the technology. Of particular note is the correlation between number of participants in awareness-raising programs and number of actual AI users because it demonstrates that effective awareness-raising is an accelerating factor in the spread of technology.

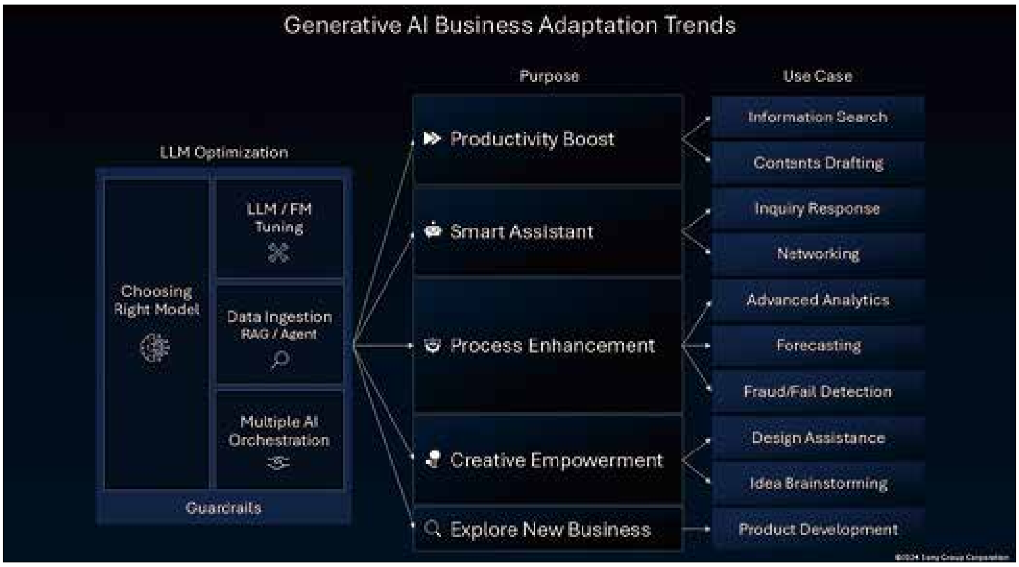
Awareness-raising activities serve as a catalyst for democratizing generative AI, acting as a critical bridge between the technological infrastructure and practical business adaptation. This fact suggests that investments in organizational and cultural aspects of technology adoption are as important as—or even more important than—investments in the technical aspects.

4. Initiatives aimed at business adaptation

4.1 Status of PoC (proof of concept)

Enterprise LLM is an environment for creating experiences, and provides a flexibly customizable PoC environment to facilitate actual business adaptation. Currently, we are conducting PoCs in over 300 departments, 136 of which have been completed, and 51

■ Figure 4: Business trends in adaptation of generative AI



Generative AI Business Adaptation Trends

LLM Optimization

Choosing Right Model

LLM / FM Tuning

Data Ingestion RAG / Agent

Multiple AI Orchestration

Guardrails

Purpose

Productivity Boost

Smart Assistant

Process Enhancement

Creative Empowerment

Explore New Business

Use Case

Information Search

Contents Drafting

Inquiry Response

Networking

Advanced Analytics

Forecasting

Fraud/Fail Detection

Design Assistance

Idea Brainstorming

Product Development

©2024 Sony Group Corporation

have already progressed to the production phase for actual business use.

The diverse objectives of the PoC include “productivity boost,” “smart assistant,” “process enhancement,” “creative empowerment,” and “explore new business.” Initially, basic use cases such as information search and chatbots were the focus; however, as employee literacy has improved and technology has evolved, more-advanced use cases, such as BPR (business-process reengineering) by leveraging AI and the creation of new added value, have also been increasing.

4.2 Technology architecture

The technology architecture that supports business adaptation of generative AI consists of the following three main components:

1. LLM capability: By leveraging Amazon Bedrock, Azure OpenAI Service, Google Cloud Vertex AI, etc., a multi-cloud, multi-LLM environment that enables the use of the best-of-breed LLMs on the market has become available. In addition to models widely used on the market, lightweight models for fine-tuning and task- and industry-specific models are also provided in a scalable manner.
2. Data pipeline: Establishing various technological elements for data processing, which is the core of utilizing generative AI.
3. Business PoC workspace: An environment allowing business users to customize LLMs to suit their use cases easily (by using Bedrock Studio, etc.) is provided.

4.3 Main technical elements

The following three technical elements are prerequisite for utilizing generative AI.

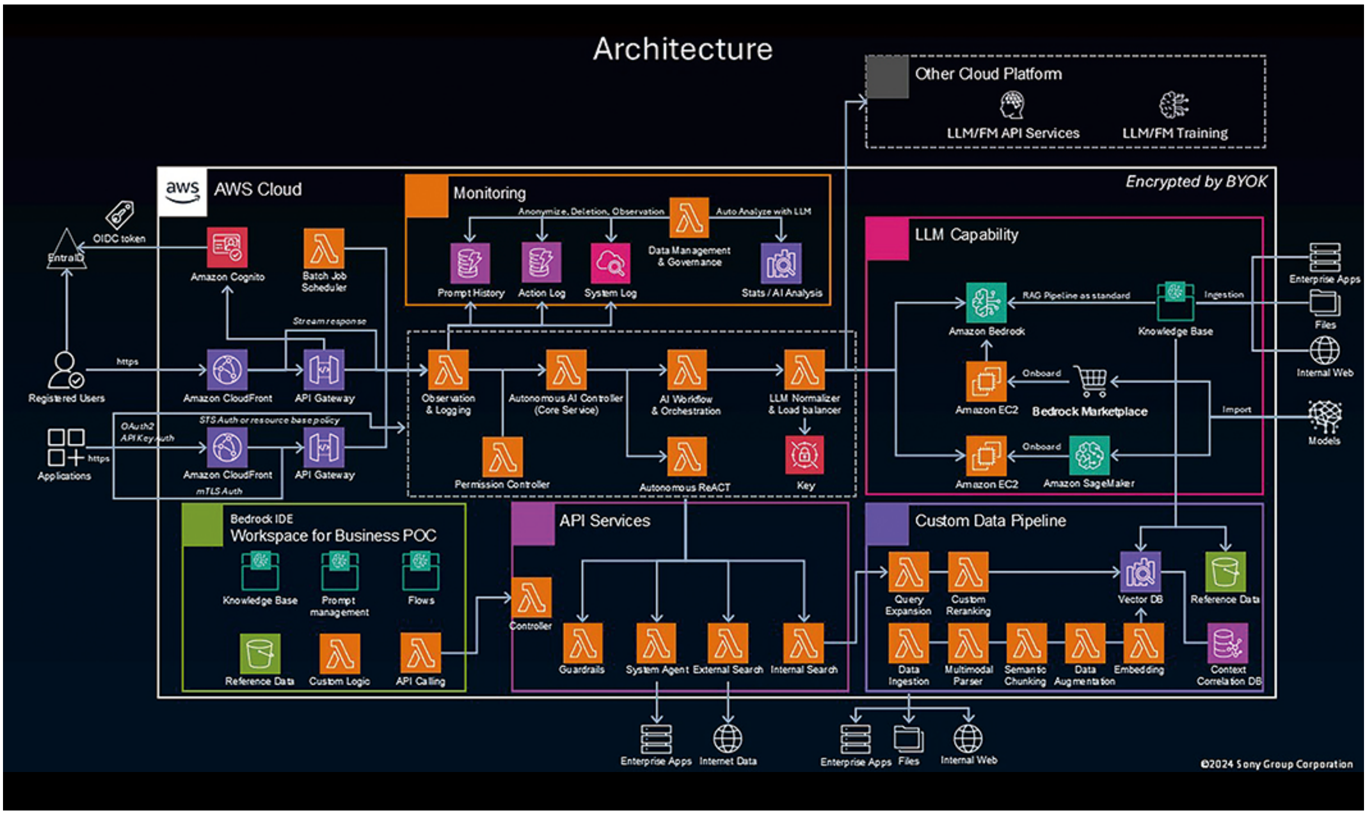
1. Prompt tuning: Controlling model output by optimizing prompts
2. Retrieval-augmented generation (RAG): Incorporating external knowledge to improve answer accuracy
3. Model tuning: Tuning models for specific tasks

The data pipeline combines a variety of technical elements that include:

- Data-source integration: Connectors and web crawlers required for corporate use
- Parsing and chunking: Multimodal context extraction and semantic processing
- Embedding: Selection of highly accurate embedding models and fine-tuning
- RAG utilizing graph structures
- Advanced search and generation: Query extension, fact checking, hallucination suppression, etc.

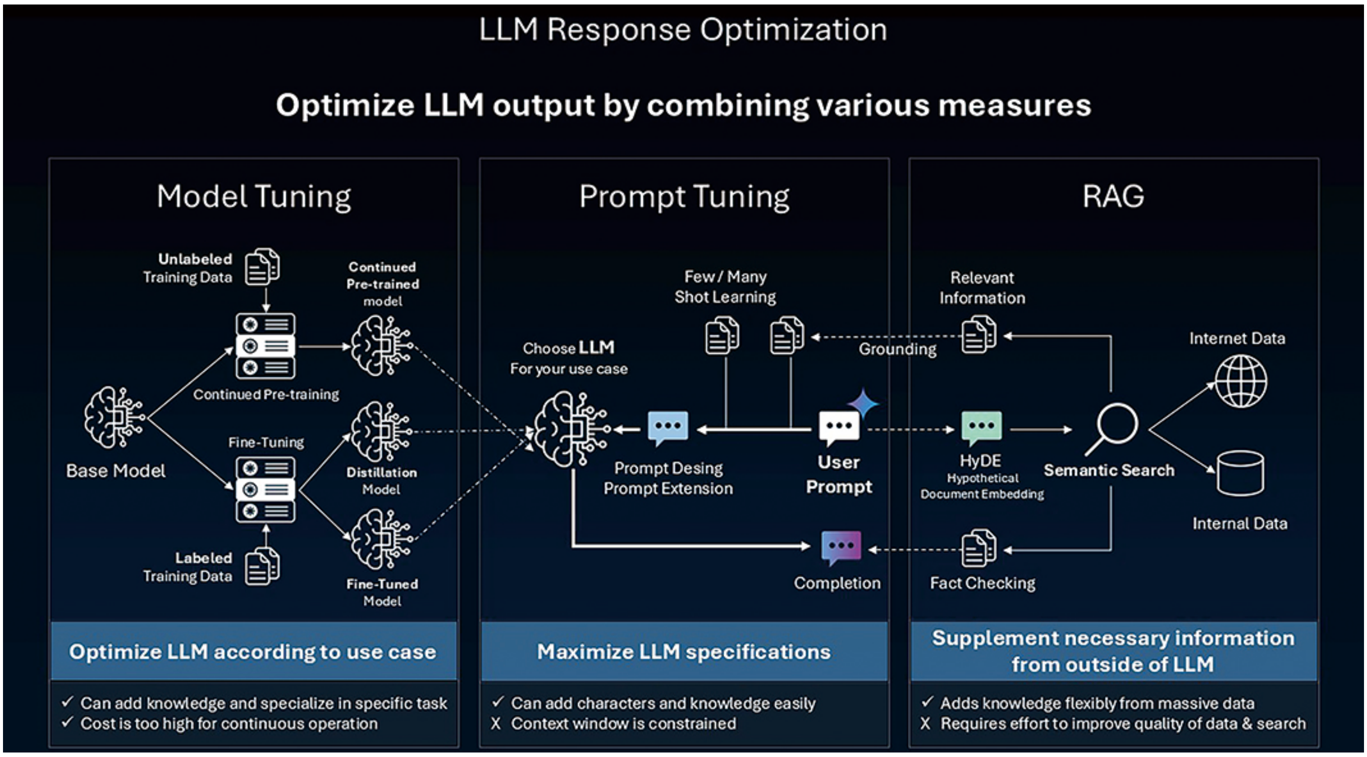
These technical elements are selected and combined appropriately according to the use case and the balance between cost and accuracy.

Figure 5: Architecture



©2024 Sony Group Corporation

■ **Figure 6: Optimization of response of LLMs**



LLM Response Optimization

Optimize LLM output by combining various measures

Model Tuning

Optimize LLM according to use case

✓ Can add knowledge and specialize in specific task
✓ Cost is too high for continuous operation

Prompt Tuning

Maximize LLM specifications

✓ Can add characters and knowledge easily
X Context window is constrained

RAG

Supplement necessary information from outside of LLM

✓ Adds knowledge flexibly from massive data
X Requires effort to improve quality of data & search

5. Examples of actual application

5.1 Enhancing response to enquiries by Enterprise LLM

In one case in which the response to enquiries by Enterprise LLM itself was enhanced using generative AI, the accuracy of inquiry response and information search was increased by storing related data in a vector database and utilizing RAG. This approach has been implemented in multiple PoCs, and best practices for Enterprises LLM have been accumulated.

5.2 Utilization of unstructured data

By utilizing "multimodal understanding," it is now possible to extract and utilize thoroughly the context of business documents that contain complex tables, graphs, diagrams, and other information that was previously difficult to interpret. The utilization of multimodal technology is expected to become an important part of future business adaptation of generative AI.

6. Expansion into Agentic AI and Future Strategy

6.1 Definition and characteristics of agentic AI

Sony Group is promoting adoption of "agentic AI" as the next evolution of generative AI. Unlike traditional generative AI (which simply responds to input prompts), agentic AI is capable of planning and autonomously executing a series of steps to complete complex tasks.

Agentic AI has the following four main features:

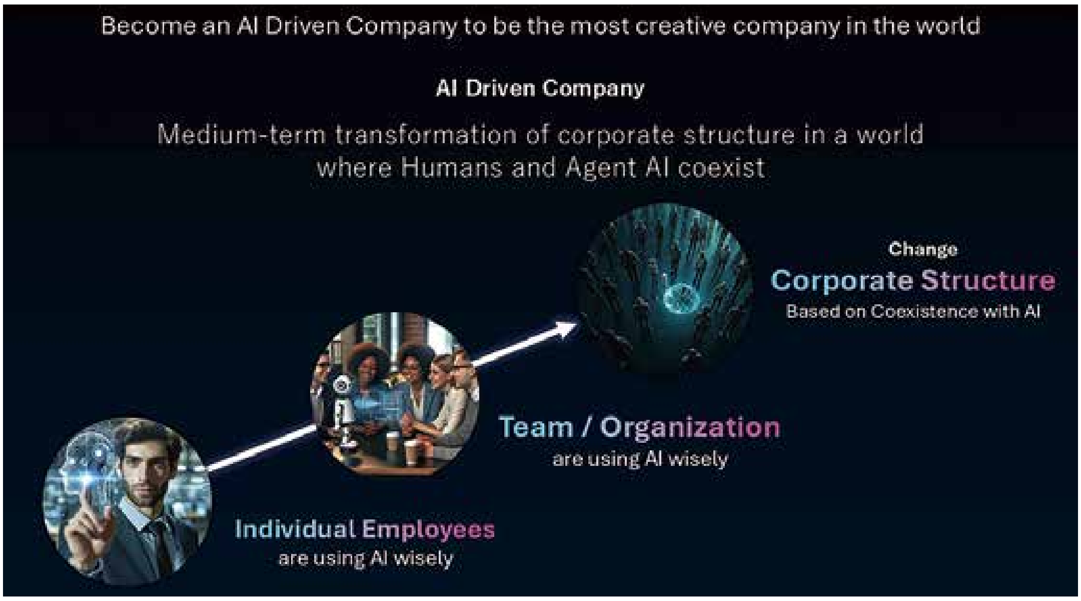
1. Autonomous: The ability to function without continuous human intervention
2. Planning: The ability to devise unique procedures to achieve goals
3. Tool utilization: The ability to use external tools and APIs as needed.
4. Memory & self-improvement: The ability to remember the results of actions and self-improve on the basis of data. These characteristics will enable agentic AI to evolve from a purely reactive AI to a more proactive and autonomous AI. As a result, it will be able to automate more complex business tasks and support decision-making.

6.2 Our vision for transforming into an AI-driven company

Sony Group is currently undergoing a transformation into an "AI-driven company" to become the most-creative company in the world. This signifies a mid-term transformation of corporate structure in a world where humans and AI agents coexist. It comprises three levels:

1. Individual level: Each employee uses AI wisely
2. Team/organization level: Teams and departments collaborate by using AI
3. Corporate-structure level: Transforming corporate structures to accommodate coexistence with AI

Figure 7: Become an AI-driven company to be the most-creative company in the world



Become an AI Driven Company to be the most creative company in the world

AI Driven Company

Medium-term transformation of corporate structure in a world where Humans and Agent AI coexist

Change
Corporate Structure
Based on Coexistence with AI

Team / Organization
are using AI wisely

Individual Employees
are using AI wisely

By realizing this vision, we aim to build a new corporate model that optimally combines human creativity with the processing capability and efficiency of AI.

6.3 Values of Agentic AI Platform Proposed by D&T PF

Sony Group's "Digital & Technology Platform" (D&T PF) organization is building a comprehensive platform for utilizing agentic AI across the group. The key proposed values of this platform are listed as follows:

1. Sharing generative AI across the Sony group: Sharing generative-AI technology across business units
2. Flexibility without vendor lock-in: Ensuring flexibility to connect with the tools that work best for the user
3. Providing various types of AI agent:
   - Common agents: Basic agents shared across the Sony Group
   - Team agents: Agents shared by teams of different business units
   - Personal agents: Agents specialized for personal use
4. Providing a variety of methods for building AI:
   - Tailor-made development: Advanced customization by using SDKs (software-development kits), etc.
   - Utilizing AI-building tools: Utilizing existing tools such as Bedrock and Claude Studio
   - No-code AI design: A development environment (in which ELLM agents and other tools are utilized) that requires no specialized knowledge

6.4 Strategy for Implementing Agentic AI on the basis of Five Concepts

The strategy of Sony Group for implementing Agentic AI is based on the following five concepts:

1. Democratizing agentic AI and promoting transformation into an AI-driven company: Concurrently driving widespread dissemination of technology and organizational transformation
2. Expanding the multi-cloud, multi-LLM concept: Utilizing diverse cloud services and LLMs
3. Developing protocols for incorporating leading internal and external AI agents: Building an ecosystem
4. Providing a platform for scaling up outstanding in-house models and AI technology: Ensuring scalability of technology
5. Enhancing templates to handle many use cases: Improving practicality

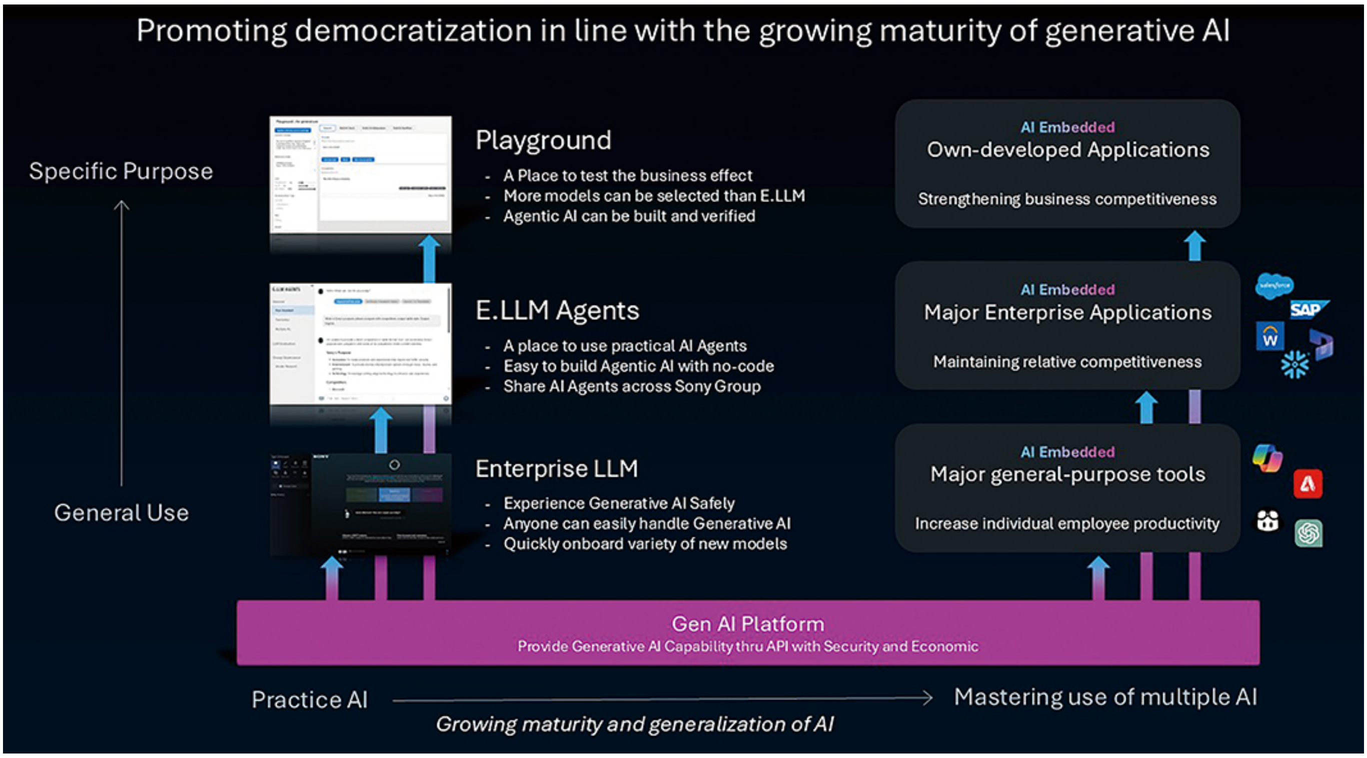
By implementing these concepts, we aim to unlock the full technical potential of agentic AI and accelerate the creation of business value.

6.5 Future prospects and challenges

Adoption of agentic AI is positioned as the next step in democratizing generative AI within the Sony Group. Future prospects and challenges concerning generative AI are summarized as follows:

1. Establishing a new model of human-AI collaboration: Exploring the optimal division of roles between AI agents and humans
2. Addressing organizational challenges associated with transformation of corporate structure: Redesigning existing business processes and organizational structures
3. Overcoming technical challenges: Balancing autonomy and safety, understanding complex tasks, and improving accuracy of task execution

■ **Figure 8: Promoting democratization in line with the growing maturity of generative AI**



Promoting democratization in line with the growing maturity of generative AI

Specific Purpose

General Use

Playground
- A Place to test the business effect
- More models can be selected than E.LLM
- Agentic AI can be built and verified

E.LLM Agents
- A place to use practical AI Agents
- Easy to build Agentic AI with no-code
- Share AI Agents across Sony Group

Enterprise LLM
- Experience Generative AI Safely
- Anyone can easily handle Generative AI
- Quickly onboard variety of new models

AI Embedded
Own-developed Applications
Strengthening business competitiveness

AI Embedded
Major Enterprise Applications
Maintaining relative competitiveness

AI Embedded
Major general-purpose tools
Increase individual employee productivity

Gen AI Platform
Provide Generative AI Capability thru API with Security and Economic

Practice AI → Mastering use of multiple AI

Growing maturity and generalization of AI

4. Ethical and legal considerations: Clarifying accountability for autonomous AI behavior

To address these challenges, Sony Group is working not only on technological development but also on establishing governance frameworks and fostering talent and organizational culture. Through the adoption of agentic AI, we aim to realize our vision of “balancing creativity and productivity” at an even higher level and thereby accelerate our transformation into a truly AI-driven company.

7. Conclusion

Sony Group aims to achieve both creativity and productivity through the “democratization” of generative AI and its adaptation in business operations. As generative-AI technology rapidly evolves, the areas in which it can be used are expanding daily, and dozens and more use cases are currently emerging. From now onward, we will continue to create an environment in which engineers and business users can work together to explore new possibilities and to promote the democratization of generative AI while making full use of generative AI-related solutions such as Amazon Bedrock.

Cover Art

Chrysanthemum blossoming on Dangozaka Hill, Yanaka (Yanaka Dangozaka kiku) from A Hundred Views of Musashi Province

Woodblock prints depict famous landmarks in Tokyo.

Kobayashi Kiyochika
(1847-1915)

Source: National Diet Library,
NDL Image Bank
(https://rnavi.ndl.go.jp/imagebank/)