How to use GenAI without compromising data sovereignty

Masakazu Murao

15 November 2024

Iakyara vol.392

Nomura Research Institute, Ltd.

Executive Summary

Generative AI is now being applied to higher-value-added business processes, no longer confined to routine tasks. In this new phase, financial institutions must optimize generative AI to specific tasks, utilizing their own proprietary data. In doing so, they will have to safeguard the sovereignty of their training data.

Masakazu Murao
Expert System Consultant
Financial Process Innovation Planning Department

Optimized GenAI needs

Generative AI (GenAI) is becoming ubiquitous amid growing availability of services that make it readily accessible to the masses. In the financial sector, however, AI use cases are currently still limited to generalist tasks like collecting information, responding to live chat inquiries and preparing minutes of meetings.

To effectively perform tasks tailored to a financial institution, a GenAI model must first be optimized using the institution's specific data. Such optimization is essential for tasks like sales support, which relies on records of past customer interactions; compliance checks, which require access to transaction histories and proprietary checklists; and clerical support, which depends on the institution's unique workflows and procedures.

In response to the rising demand for GenAI applications that leverage closed data—including confidential business information and customer personal data—technologies are becoming increasingly sophisticated in optimizing GenAI for specific tasks.

Data sovereignty

Use of a financial institution's closed data in GenAI applications raises concerns about data sovereignty. Data sovereignty is the concept of retaining data in its country of origin and safeguarding data ownership, access rights and traceability pursuant to that country's laws and regulations. When a financial institution's data are used to train a GenAI model, data sovereignty must be respected for the sake of data protection.

Data sovereignty is heavily dependent on the laws of the locale where data are stored and processed. Regulations governing cross-border data transfers differ by country/region. Certain countries even prohibit data originated within their borders from being stored outside their borders. Some countries have recently been tightening their control over data from the standpoint of their economic security and/or digital trade balance.

As the GenAI era advances, data sovereignty will likely gain even greater importance, as effective data risk management is essential to fully realize GenAI's potential. Protecting sensitive information, such as trade secrets and personal data, is crucial for mitigating risks and ensuring that organizations can harness GenAI's benefits.

GenAI use that does not compromise data sovereignty

For Japanese companies, there are three key prerequisites to using GenAI without compromising data sovereignty.

First, data hosting and GenAI training/inference must be restricted to data centers located in Japan to ensure that they are governed by Japanese law and shielded from foreign legal or regulatory influences. Given the massive computing resources required for GenAI training/inference, more data center capacity (e.g., GPUs) will have to be added in Japan.

Second, intensive data governance and security are needed, including for trade secrets and personal information. Financial institutions are already subject to many governance- and security-related requirements in addition to strict regulatory supervision. Several other GenAI-specific issues also must be addressed, including hallucinations and copyright infringement. Additionally, the requisite governance and security are so broad in scope that integrated risk management is essential and must encompass everything from data centers, other facilities and their power sources to networks, hardware, operating systems, middleware, applications and data. A sophisticated governance regime is needed to operate all of these components to use GenAI.

Lastly, the GenAI build environment must be optimized to specific business processes. At Japanese financial institutions, GenAI models must understand Japan's financial system and be able to learn and/or reference company-specific workflows, business rules and customer information.

GenAI models have recently become more specialized, with distinct strengths suited to specific use cases. For instance, some of the latest models are compatible with retrieval-augmented generation[1] (RAG), a technique that enhances GenAI's performance on task-specific applications. Choosing the right GenAI model from the available options will be crucial for optimizing business processes that deliver high value.

NOTE

1) RAG is a technology that improves the accuracy of large language models' answers by enabling the model to retrieve information from external sources in addition to generating text internally. It is expected to reduce hallucinations (generation of factually false information).

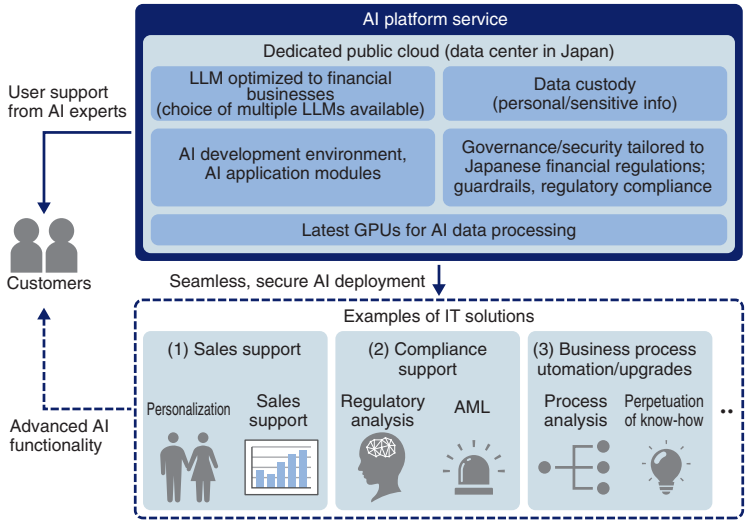
AI platform services

Japanese financial institutions' GenAI use to date has relied heavily on overseas public cloud providers with vast computing resources and extensive service offerings. However, to meet the aforementioned three prerequisites, financial institutions will have to hire personnel with expertise in GenAI models, data centers, GPUs and other hardware, governance and security. Recruiting a sufficient number of qualified personnel is a challenge for any company.

Given these challenges, we anticipate a new GenAI business model emerging in the form of AI platform services that provide shared access to a dedicated public cloud, clearly defined as a secure data-hosting site[2] (see graphic). The services offered would include not only hardware and software but also AI governance/security implementation and user support from AI experts. Such an arrangement would offer the convenience of a public cloud, ensure data sovereignty by virtue of the cloud being used exclusively by the service's subscribers, reduce data center and GPU costs through the shared-usage model, and give users access to high-level operational and technical expertise with respect to AI utilization.

2) The service provider would set up a dedicated public cloud with in its own data center and operate its service under its own controls.

AI platform service concept



Source: NRI

A recent survey[3] conducted by the Ministry of Internal Affairs and Communications reported that Japanese companies are lagging far behind their overseas counterparts in terms of GenAI use. Concerns about challenges such as security risks, copyright infringement and scarcity of expertise loom large in the GenAI space. If AI platform services that ensure data sovereignty were to become available, they may dispel such concerns and expedite GenAI deployment in higher-value-added business processes.

3) *Information and Communications in Japan: White Paper 2024*

about NRI

Founded in 1965, Nomura Research Institute (NRI) is a leading global provider of system solutions and consulting services with annual sales above $4.9 billion. NRI offers clients holistic support of all aspects of operations from back- to front-office, with NRI's research expertise and innovative solutions as well as understanding of operational challenges faced by financial services firms. The clients include broker-dealers, asset managers, banks and insurance providers. NRI has its offices globally including New York, London, Tokyo, Hong Kong and Singapore, and over 17,400 employees.

For more information, visit https://www.nri.com/en

Inquiries to : Financial Market & Digital Business Research Department
Nomura Research Institute, Ltd.
Otemachi Financial City Grand Cube,
1-9-2 Otemachi, Chiyoda-ku, Tokyo 100-0004, Japan
E-mail : kyara@nri.co.jp

https://www.nri.com/en/knowledge/publication/fis/lakyara/