Nomura Research Institute Group

NEWS RELEASE

Jan. 11, 2024

Nomura Research Institute, Ltd.

NRI Digital, Ltd.

NRI Group to Launch “Private LLM” Generative AI Solution Designed to Minimize Data Leak Risks

Tokyo, January 11, 2024 - Nomura Research Institute, Ltd. (Headquarters: Chiyoda Ward, Tokyo, Chairman and President & CEO, Representative Director: Shingo Konomoto, “NRI”) and NRI Digital, Ltd. (Headquarters: Yokohama City, Kanagawa Prefecture, President & COO: Masakazu Amamiya, “NRI Digital”) intent to launch providing a generative AI solution called “Private LLM (Large Language Model)” to minimize the risk of data leaks in spring or later in 2024. This solution caters to institutions that require an especially high level of information security control.

■ Runs on private clouds and in on-premises environments, enabling confidential and sensitive information to be handled safely

When it comes to the use of generative AI, the extent to which confidential and sensitive information should be transmitted to public LLMs (as represented by OpenAI’s[*1] GPT-4[*2], for instance) has become a major concern. With this solution, open-source LLMs such as Meta’s[*3] Llama 2[*4] can be run on private cloud services being operated at NRI’s data centers or in on-premises environments involving information systems owned and managed by companies themselves. As a result, this solution stands as a bastion of security for confidential and sensitive information. It can even be adapted to high-level information security control requirements.

■ Tailored LLMs to meet business needs of individual companies

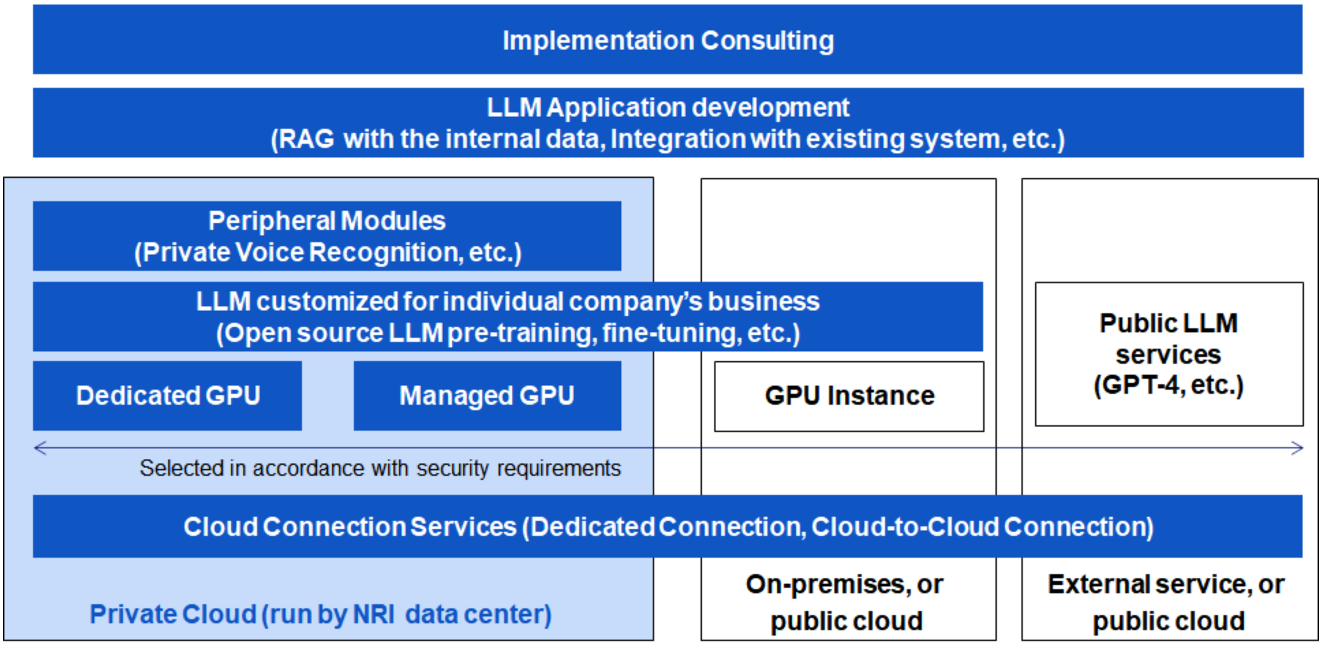
While Open-source LLMs such as Llama 2 may not yet match their public LLMs in performance, by being customized using a company’s proprietary data (pre-training[*5], fine-tuning[*6] etc.), such LLMs potentially can exhibit a business-ready level of performance depending on the nature of the task involved (see the in-house demonstration results and reference materials below). With this solution, LLMs can be customized in a way that both minimizes the risk of data leaks during training (which is even more confidential) and is most suited to a given company's operations.

■ Expanding horizons with"Private voice recognition" and other peripheral modules

Combining voice recognition features with LLMs greatly expands the range of applications to include things like responding to inquiries at call centers or in-person, and thus peripheral modules will also be provided, including a "private voice recognition" module for minimizing the risk of leaks of audio data that can identify specific individuals.

The solution's menu system is depicted in the graphic below.

Chart: "Private LLM" Solution Menu System


Implementation Consulting
LLM Application development
(RAG with the internal data, Integration with existing system, etc.)
Peripheral Modules
(Private Voice Recognition, etc.)
LLM customized for individual company's business
(Open source LLM pre-training, fine-tuning, etc.)
Dedicated GPU
Managed GPU
GPU Instance
Public LLM
services
(GPT-4, etc.)
Selected in accordance with security requirements
Cloud Connection Services (Dedicated Connection, Cloud-to-Cloud Connection)
Private Cloud (run by NRI data center)
On-premises, or
public cloud
External service, or
public cloud

Text outlined in white indicates content provided by NRI.

GPU: Graphics Processing Unit. Required for running LLMs.

RAG: Retrieval Augment Generation. A method for getting LLMs to generate accurate responses with reference to external factual data.

■ Proven Efficiency in NRI's in-house operations

To verify this solution's performance capabilities, NRI applied it to its own in-house accounting administrative procedure support. Using a training set of 60,000 examples for customization, fine-tuning was performed on Llama 2 which was being run on a GPU installed at an NRI data center, and the experiment yielded a business-ready level of higher performance. As a result, some of the relevant operations were replaced by the solution, leading to a 60% reduction in the amount of time required for Q&A creation.

NRI and NRI Digital will continue to be proactively engaged in developing various services that enable the use of generative AI that is optimized to meet the needs of companies and individuals.

[*1] OpenAI: a company that conducts AI research and deployment. For details, see the following website.
https://openai.com/

[*2] GPT-4: one of the large language models developed and provided by OpenAI.

[*3] Meta: for details, see the following website.
https://about.meta.com/

[*4] Llama 2: a large language model developed by Meta. Publicly available for commercial use (including the base model).

[*5] Pre-training: training an LLM with generic language patterns or knowledge.

[*6] Fine-tuning: making fine adjustments to an LLM to tailor it for specific tasks.

Inquiries about this news release:
Kayano Umezawa, Miku Funayama
Corporate Communications Department
Nomura Research Institute, Ltd.
Tel: +81-3-5877-7100
E-mail: kouhou@nri.co.jp

Inquiries about the Solution:
Tomoyasu Okada
AI Solution Promotion Department
Nomura Research Institute, Ltd.

Hiroyuki Nakamura
DX Planning Unit
NRI Digital, Ltd.
E-mail: private-llm@nri.co.jp

For Reference

Table 1: Public LLM and NRI’s Solution “Private LLM”: A Feature Comparison

	Public LLM	NRI’s Solution “Private LLM”
Base model	Closed-source (e.g., OpenAI’s GPT-4)	Open-source (e.g., Meta’s Llama 2)
Provided on	External service	Private cloud On-premises Public cloud
Performance	State-of-the-art at the time	Less than external services (gradually getting closer)
Security	Equivalent to public clouds	Equivalent to on-premises, private clouds (adaptable for high-level security control)
Customizability	Scope of customization is limited	Freely customizable

Table 2: Prospective Llama 2 Performance Improvements Through Customization

Task Category	Use Case	Llama2-7B	Llama2-70B
Q&A (without RAG)	Help desk, contact center	＊＊	＊＊
Q&A (with RAG)	Help desk, contact center	＊＊	＊＊＊
Summarizing	Creating response logs or meeting minutes	＊＊	＊＊
Creating Q&A from text	Assisting creation of FAQs	＊＊	＊＊＊
Natural sentence generation	Creating manuals or guide sentences	＊＊	＊＊
Dialogue	Advice, counseling	＊＊	＊＊
Style conversion	Customer service support (e.g., polite expressions)	＊＊	＊＊
Sentiment analysis	Customer service or SNS etc. sentiment analysis	＊＊＊	＊＊＊
Document classification/tagging	Document attribute classification or tagging	＊＊＊	＊＊＊
Abnormality detection	Call log or email compliance monitoring	＊＊	＊＊＊
Code generation	App development, data analysis	＊＊	＊＊＊

7B=7 billion, 70B=70 billion (no. of parameters)

＊＊＊＝business-ready

＊＊　＝problems found