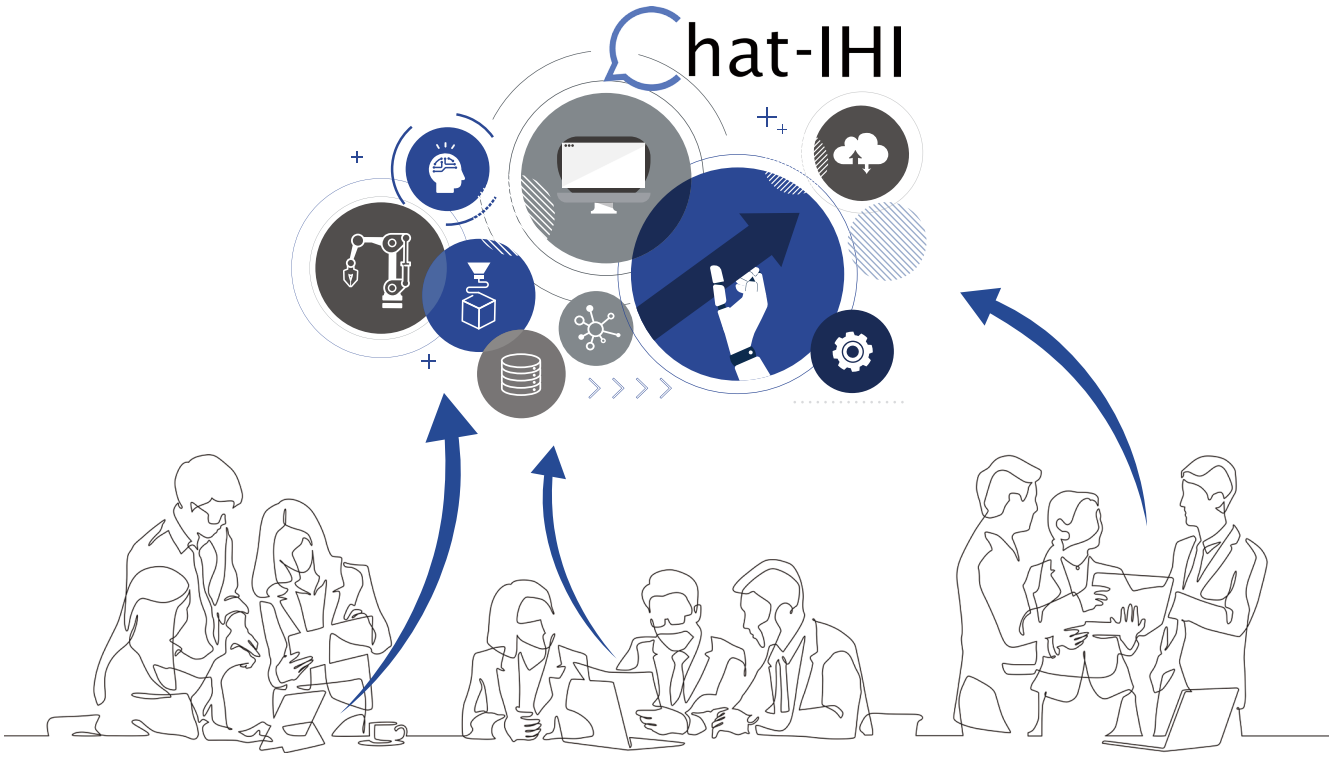
IHI Group's Efforts to Promote DX and Use Generative AI

Using generative AI for business purposes and developing systems based on RAG and SLM technologies

The IHI Group is working on using generative AI, such as large language models (LLMs), to promote digital transformation (DX). This article introduced our initiatives to use LLMs in three phases and the progress of the initiatives.



Conceptual diagram of Chat-IHI utilization

Introduction

The IHI Group is working on digital transformation (DX) to innovate work processes to streamline business operations, reduce lead times, and make other improvements, as well as to innovate business models to solve social issues (refer to "Using Large Language Models to Achieve DX" IHI Engineering Review, Vol. 58, No. 1, 2025). Generative AI, including large language models (LLMs), which has been gathering attention in recent years, is regarded as a key technology for realizing labor savings.

The IHI Group has been introducing the use of LLMs in three phases since FY2023. In phase one, we used an unmodified LLM open to the public that has only learned publicly available information. In phase two, we created and introduced a system to reference business data within the Group. In phase three, we developed a system to deploy LLMs in each business unit. Since the previous report, the

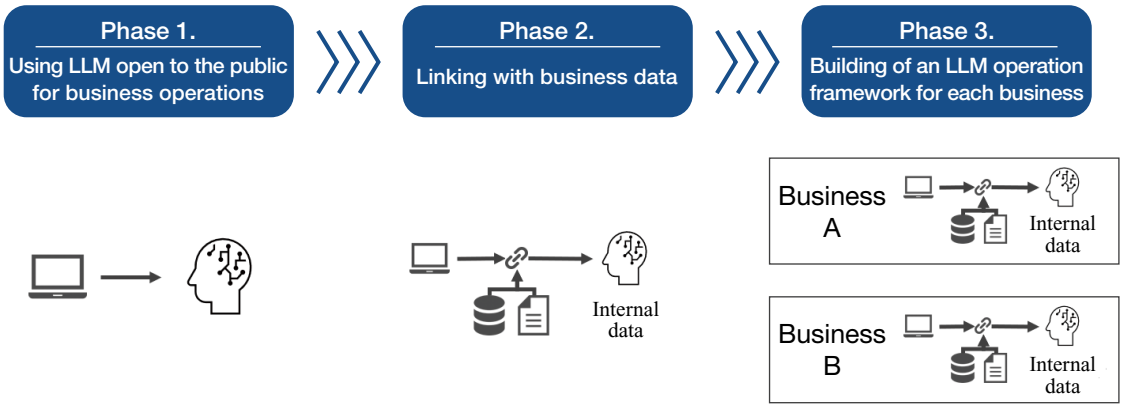
LLM introduced in phase one has become more widely used, and we conducted verification in several departments in phase two. In phase three, we verified a system for departments that handle highly confidential information as a new initiative. This article introduces the progress of initiatives to use LLMs and future challenges in the IHI Group, but these initiatives have been conducted only in Japan.

Adoption of Chat-IHI in the IHI Group

If a user were to simply start typing into an interactive generative AI chatbot service open to the public on the Internet, such as ChatGPT, the LLM underlying the service would learn internal information of the Group from the user's input, which would lead to a risk of information leakage by that information being provided to external users, and other such problems. To address this issue, we developed a cloud-based ChatGPT service in the Group named Chat-IHI that is only accessible to the IHI Group in Japan by using Azure OpenAI Service provided by Microsoft. This environment is configured to neither let Microsoft monitor user-input data nor use it for model learning. This significantly reduced the risks of input information being leaked to the outside or being used for future learning. Since we launched Chat-IHI in June 2023, we have continued to incorporate the latest LLM and enhance its functionality, such as image loading and useful templates. The number of users has continued to increase and now over 4,000 users, which corresponds to approximately 30% of the people who can use Chat-IHI, use it every day (as of March 2025). We plan to continue to periodically add functions in the future to increase the number of users. Chat-IHI can be used only in Japan, but in other countries, LLM tools, such as internal versions of ChatGPT and Microsoft 365 Copilot, are introduced as soon as the necessary environment is prepared in each region or company.

We have enhanced activities to promote its use in parallel with system improvements so that Chat-IHI can be effectively used for business operations. Since we released Chat-IHI, we have put effort into its growth such as by opening the Chat-IHI portal site to consolidate and disseminate information helpful for using Chat-IHI for business operations and posting precautions for use, user guides, and FAQs there. As the use of Chat-IHI expands, we obtained effective use cases in business areas. For example, some users were able to grasp the contents of free-text fields in a questionnaire more rapidly, while others developed programs more quickly. We also formed a task force that consists of the information system department, which provides Chat-IHI, as well as the DX promotion departments in our four business areas, HR department, and technology development department to further accelerate the use of Chat-IHI in these business areas. With these efforts, we have created and deployed use cases where operation time was reduced by using generative AI, including Chat-IHI, as described later.

To collect effective use cases in the Group, we organized a



Three phases to use LLMs in the IHI Group

contest to award internal users who effectively use the generative AI in their operations from February to March 2025. In this contest, applicants posted use cases of LLM tools, such as Chat-IHI and Copilot, with their prompts and the resulting improvements to business operations, and the best uses were selected by the votes of employees. Even though the contest only ran for two weeks, approximately 60 use cases were posted. These use cases show that LLM tools were used for survey work such as patent surveys, training design and other planning work, and developing Microsoft Excel VBA macros and other programs. It was also made clear that LLMs are used in various scenes, actually streamlining business operations. For example, one applicant created a program in a day that normally would take two weeks and reduced the time to process data from 90 hours to five minutes by using the program, and another reduced the time to convert images into text data from an hour to five minutes. Gradually, more and more employees are voluntarily starting to use LLMs for business operations.

We organized the collected use cases and posted them on the portal site so that many users can leverage them. In addition, we will determine pilot workplaces to conduct trials for further business operation improvements using the use cases expected to produce great improvement.

We have also started deploying content suited to users' learning levels and purposes to effectively disseminate collected use cases and important points of use to users. This content includes hands-on videos for beginners and mid-level learners and explanatory videos by use scenario, such as writing minutes and questionnaire analysis. Many employees are using these contents to streamline business operations here and there in the Group. However, LLMs are expected to be used in many more scenarios besides the presented use cases. We will utilize use cases that have been successful on the individual level throughout the organization and incorporate them into work processes to produce effects such as further reductions in operation time.

Status of data linkage in the IHI Group

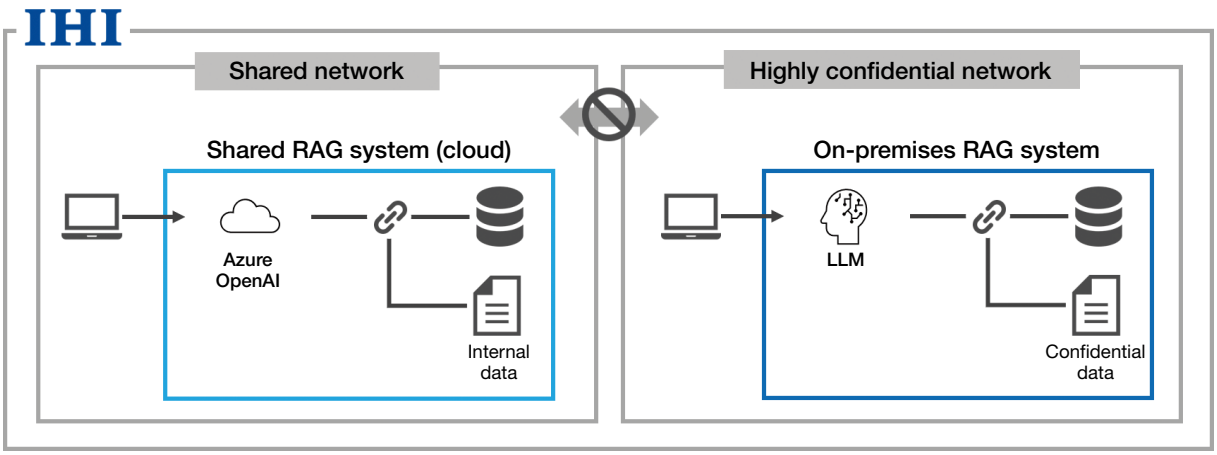
Chat-IHI can be used to innovate the work processes in the Group, for example, by applying internal data such as internal regulations to answers from Chat-IHI. To allow for this usage, we worked on developing and introducing a system that returns knowledge on business operations using a technology to connect LLMs with external databases and information sources, more specifically Chat-IHI with databases accumulating internal data. This technology is called retrieval augmented generation (RAG).

We started providing a RAG system into which the internal regulations had been loaded for test purposes in a secure environment equivalent to that of Chat-IHI. This is expected to result in the effect of reducing the time spent searching through internal regulations. We also received many feedback comments from users, including a request for incorporation of other data and an idea that inputting the business data of their department would expand the range of applicable tasks. In response to these comments, we started providing an environment where users can register their own data. Employees in eight departments tried out this environment and used it for business operations, including regulations search, risk assessment based on accident case reports and other information sources, referencing a maintenance manual by service representatives, and checking of emails exchanged with customers. Since LLMs can hallucinate (generate information not based on facts), users must verify whether answers are true or false. As this system allows users to easily check the original documents that the LLM referenced, internal data can be included in answers from the LLM while reducing the influence of hallucinations, which is a disadvantage of LLMs. This resulted in a certain level of business operations streamlining.

Progress of building the operation framework for each business area

As an enterprise group consisting of four business areas, the IHI Group must manage confidential information as well, for example, by shielding confidential information in a specific business area from the other business areas, while sharing nonconfidential information across the business areas. We decided to prepare a system environment and framework to operate LLMs in each business area as a solution to this situation. Currently, we are conducting a technology survey and designing the framework to build the system. Departments that handle extremely confidential data are conducting verification in an on-premises environment that supports unique security settings according to customer requests. For this environment, we are considering using a small language model (SLM), which can operate even with a small amount of computing resources. With a smaller number of parameters, SLMs can operate without a large server.

We built a small on-premises RAG system using hardware with graphics processing units (GPUs) and verified whether information necessary for design could be extracted efficiently using the design standard document of a product to advance the use of the SLM for business operations. Compressing the SLM through quantization reduced the memory usage and accelerated document creation processing, although the accuracy decreases to some extent. During accuracy verification, we compared the results of searching the design standard document and answers generated by the SLM in response to users' questions, such as "What is XX made of?" with model answers prepared in advance. A user who participated in the verification from the design department appreciated the SLM, saying "The SLM I used is smarter than I thought. It will be useful as a design assistant if the appropriate information is input." The SLM seems to be usable for practical operations, but its applicable range is still limited for the time being. We will start this system for business operations to which it is applicable and attempt to enhance its accuracy and convenience, considering



Shared RAG system and on-premises RAG system

work processes. At the same time, we will aim for full-fledged deployment to business operations by applying the system to more documents.

Summary and future outlook

This article introduced our initiatives to promote DX in three phases, namely, the use of LLMs for business operations, the linkage with internal data, and the building of a framework to operate LLMs in each business area.

In phase one, we created a basis for improving business operations by creating and deploying use cases through the activities of an inter-departmental task force. We will continue to introduce use cases of generative AI to promote the spread of LLM technology in the future through the organization of events such as contests and hands-on training, as well as our portal site. We will develop technologies that contribute to the streamlining of business operations using the challenges certain departments face and deploy the resulting technologies to other departments to streamline the business operations of the entire IHI Group. We will further enhance business efficiency through dissemination activities supported by technology development.

In phase two, we deployed a system to search through internal data using RAG, taking countermeasures to mitigate hallucinations. Including internal data in answers from LLMs streamlined business operations to a certain extent in multiple departments. Toward further use of LLMs for business operations, we will use search technologies for higher accuracy and will also apply technologies such as fine tuning to have LLMs learn new information, to cover technical terms in each area. In addition, we will work on using tools; for example, knowledge graph where knowledge can be systematically handled to consider relations between documents, such as regulations. We will also endeavor to introduce technologies to raise the overall accuracy of answers, including an AI agent that autonomously solves problems.

In phase three, we have begun the verification of an on-premises RAG environment to build the LLM operation framework in each business area. Besides the above-described technology applications, we are also working on selecting the optimal SLM in this fast-evolving field and improving system accuracy with an aim to maximize the effects of improvement of business operations for each business area.

The IHI Group is expanding the innovation of work processes in the Group using generative AI. We will also focus on initiatives to solve social issues.