AI Constellation—Collaboration among Diverse AI Models

Junji Teramoto

Abstract

NTT has been promoting AI Constellation since FY2023 as a large-scale artificial intelligence (AI) collaboration technology that enables different AI systems to hold discussions and correct each other to generate solutions from diverse perspectives for problems with causes that are difficult for humans to infer. This article provides the background and details leading up to this concept, such as AI technology trends, capabilities that need to be acquired, and related use cases, and the relationship between AI Constellation and NTT's IOWN (Innovative Optical and Wireless Network) vision for the future.

Keywords: AI Constellation, large language model, next-generation AI

1. Introduction

At NTT Computer and Data Science Laboratories, we aim to create superior value for both individuals and society through the research and development of advanced artificial intelligence (AI) algorithms that will enable the processing of data that had been too difficult to handle due to the scale or complexity of the problems involved.

In this article, we provide the background and vision of AI Constellation as a large-scale AI collaboration technology that NTT has been promoting since FY2023.

2. Generative AI trends and AI Constellation

Since the appearance of ChatGPT in the AI field at the end of 2022, the rise of large language models (LLMs)[*1] has been remarkable, and models for acquiring general-purpose knowledge from vast amounts of open data have been developed. Therefore, diverse use cases ranging from text summarization and image/video generation to translation, material-preparation support, and chatbots have been implemented and enhanced through AI and begun to be used in business applications.

As the scale of AI models begin to increase, however, problems start to arise. For example, the computer processing power required for learning and inference and the power consumed increase greatly, resulting in very high costs as well as a high environmental load. Specifically, the size of AI models and calculation costs have been increasing at annual rates of 180% and 555%, respectively, while graphics-processing-unit computing power has been increasing at an annual rate of 69% [1]. This has made it very difficult to promote sustainable AI businesses. While huge general-purpose models can provide generalized answers to a wide variety of questions, differentiation from other companies is important in business activities, so the difficulty in finding individuality and characteristics in general-purpose models that anyone can access has also become a problem. Thus, the trend in business has been shifting to the development of not only huge LLMs that know everything but also LLMs that possess specialized knowledge of a specific domain or organization and that can be used at a reasonable cost. We thus envision a future that will require problem solving with a collaboration of multiple LLMs in contrast to problem solving with a single LLM.

With the above in mind, NTT proposes AI Constellation to solve problems through the collective intelligence of diverse AI models such as reasonably sized

*1 LLM: A language model trained on vast amounts of text data.

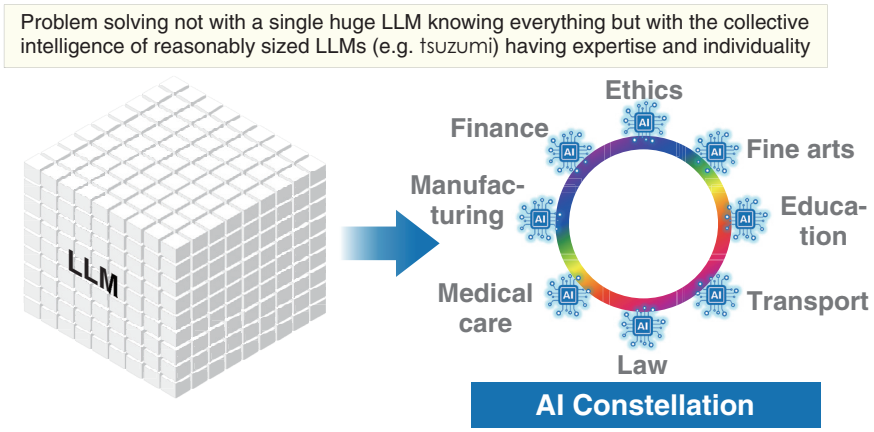

Fig. 1. The future of AI targeted by NTT.

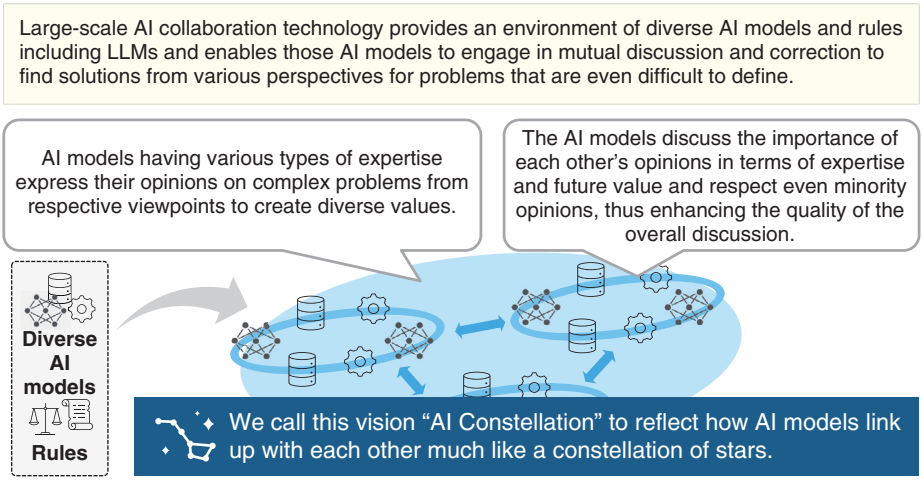

Fig. 2. AI Constellation.

LLMs having expertise and individuality (**Fig. 1**). The idea is to create diverse values from a variety of viewpoints with respect to complex problems by having business-specific AI models, such as those for finance, manufacturing, and medical care, and specialized AI models in law, ethics, and other areas collaborate with each other by holding discussions and offering each other corrections. By having these AI models discuss the importance of each other's opinions in terms of expertise and future value, AI Constellation aims to enhance the overall discussion by respecting minority opinions even if they are only in the minority at present. AI systems link up with each other much like a constellation of stars to create solutions, so this is why we call this vision "AI Constellation" (**Fig. 2**).

3. AI Constellation capabilities and use cases

AI Constellation is expected to have diverse capabilities. When thinking about creativity and individuality as core values in human activities, we consider that, while there are routine tasks that do not require a high level of either creativity or individuality, adding creativity to such tasks could drive sustainable innovation, while adding individuality could lead to disruptive innovation. The current range of LLM application is routine tasks that do not require such creativity and individuality, so we can expect various types of use cases that aim to reduce costs by replacing

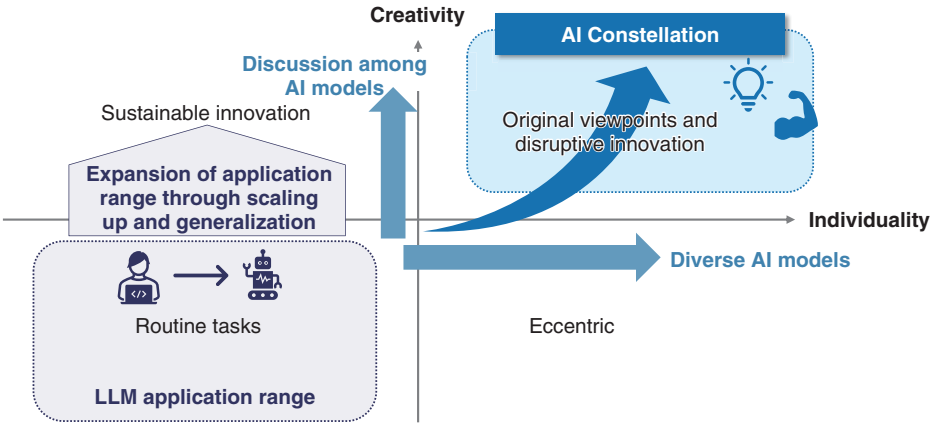

Fig. 3. AI Constellation capabilities.

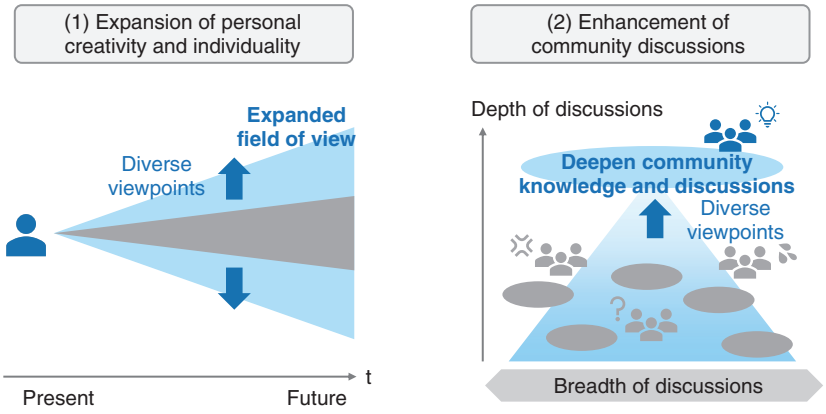

Fig. 4. AI Constellation use cases.

human labor with AI to achieve more efficient or automated work processes. In contrast, AI Constellation is expected to achieve creativity through discussions held among AI models while acquiring individuality through the diversity of those models. In short, instead of replacing humans, we can think of AI Constellation as a means of supporting highly intellectual and creative human work (**Fig. 3**).

AI Constellation use cases (where a use case clearly defines user requirements, purpose of use, etc.) can be broadly divided into two types (**Fig. 4**). One is expansion of personal creativity and individuality. When planning or deciding on something, such as a business plan or work procedure, the investigation can be deepened by executing reverse calculations after predicting or envisioning the future or possible outcomes. However, if information can be provided from diverse viewpoints as in AI Constellation, we can expect that information to contribute to an expanded field of view for the user. The other type of use case is enhancement of community discussions, which is expected to deepen the level of community knowledge and discussions. In the real world, it is common for people to engage in all types of discussions such as meetings or conferences, but it is nevertheless difficult to obtain a consensus for a variety of reasons. These include limitations in discussion time, friction between meeting participants or mutual consideration, absence of stakeholders (concerned parties) with respect to discussion content, and insufficient information to hold a thorough discussion. By having discussions among AI models, as in AI Constellation, intervene in discussions among people, we can expect an exchange of opinions based on diverse

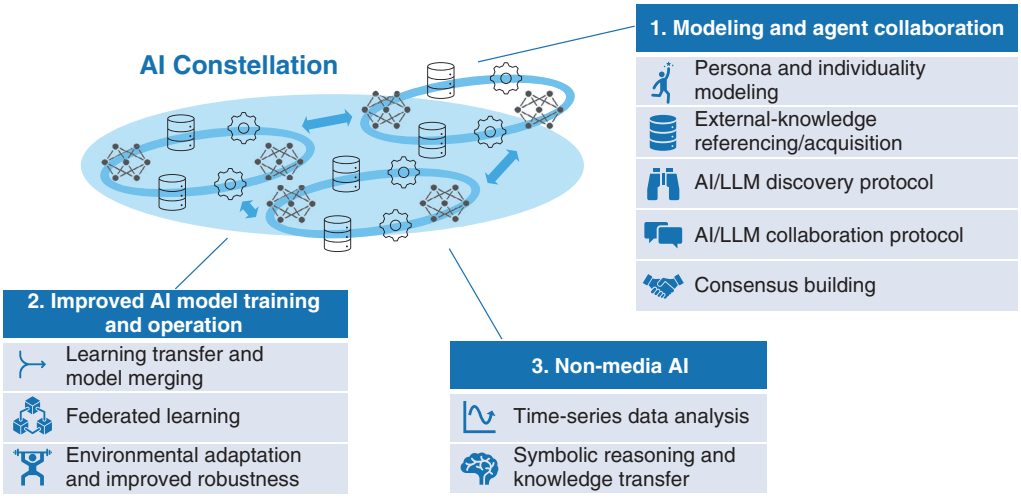

Fig. 5. Research themes toward implementation.

perspectives and knowledge and broader and deeper discussions that could not be obtained only among people. These features could be obtained, for example, by conducting conceivable discussions based on simulations of lengthy discussions, supporting calm discussions through objective AI that understands individuality, acting as a spokesperson for absent stakeholders (future-generation concerned parties, etc.), or incorporating world models or external knowledge.

4. Research themes in AI Constellation

As shown in **Fig. 5**, a variety of research themes are being taken up to achieve AI Constellation that can find solutions to complex problems by having diverse AI models engage in mutual discussion and correction.

4.1 Modeling and agent collaboration

In AI Constellation, to enable AI models to engage in mutual discussion and correction from their respective viewpoints according to the topic at hand, it is necessary to appropriately and efficiently model the role and individuality of those AI models and make it easy for end users to use the system. In this theme, we will achieve external-knowledge referencing and acquisition that can reference diverse types of external knowledge—such as various types of personas and general and domain-specific knowledge—according to the current topic and save knowledge obtained through discussions as external knowledge. We will also use that external knowledge in persona and individuality modeling and in subsequent discussion enhancement. Given that many such AI models will be prepared and combined according to the current topic, we will achieve a protocol for discovering and linking AI models according to ambiguous requirements based on natural language. To appropriately handle the results of a discussion, we will develop a mechanism for consensus building in cooperation with the people involved that is not based simply on a majority vote but that also takes into account minority opinions.

4.2 Improved AI model training and operation

In AI Constellation, it is necessary to prepare and link many AI models to serve diverse applications, so model size and the number of models can be expected to increase in scale. There is therefore a need to significantly reduce development and operating costs including that for AI model training.

In this theme, we take up drastic improvements in AI learning algorithms including learning transfer[*2] [2], an innovative learning technique proposed by NTT, model merging that combines different AI models, and fast, privacy-conscious distributed learning that applies federated learning. We will also pursue a significant reduction in operating costs by achieving robust learning and inference algorithms that support diverse tasks and environments such as by adapting to changes in the environment during training and operation.

*2 Learning transfer: A training method devised by NTT as a new mechanism for reusing past learning processes for AI models.

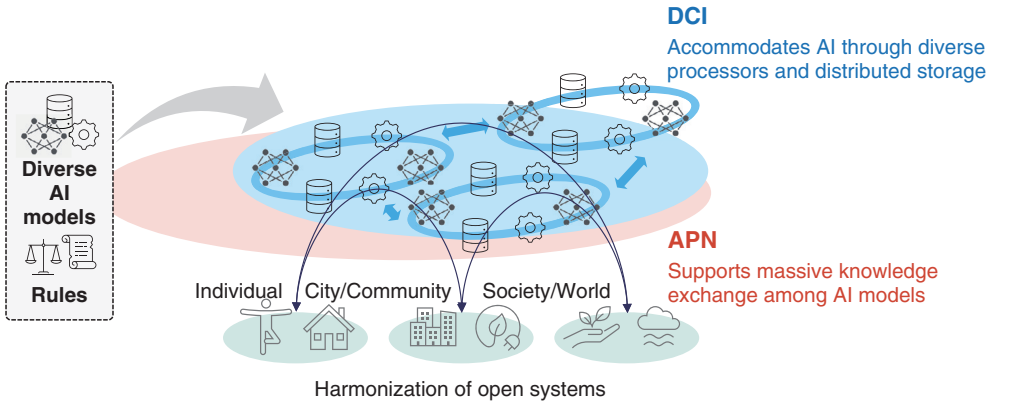

Fig. 6. Future vision of AI Constellation.

4.3 Non-media AI

AI Constellation aims to create solutions for complex problems from diverse values, but while LLMs are especially good with natural language and media data, they tend to be weak in the analysis of numerical data such as non-media time-series data. They cannot, as a result, handle all real-world problems. LLMs also have the ability to output an answer not based on facts called a hallucination*3.

This theme contributes to the provision of rational and logical discussions among AI models based on knowledge acquired from diverse types of factual data by achieving advanced analysis of time-series data and logical inferences using external knowledge.

5. Future developments

Along with the research and development of these algorithms, we are investigating providing AI Constellation as a social platform leveraging the network and computing platform of NTT's Innovative Optical and Wireless Network (IOWN) (**Fig. 6**). Research and development of the All-Photonics Network (APN) [3], one of the main components of IOWN, aims to achieve low power consumption, high-speed and high-capacity communications, and low-latency transmission. We expect these features to enable knowledge sharing among diverse and large-scale AI models that use massive amounts of real data. The data-centric infrastructure (DCI) [4] aims to achieve diverse types of processors and distributed storage so that large-scale AI models can be arranged in just the right way. Deploying AI Constellation in an infrastructure that combines the APN and DCI should be able to support both local production for local consumption of data through diverse AI models and knowledge exchange among global AI models while providing high problem-solving power.

By implementing AI Constellation so that diverse AI models can engage in mutual discussion and correction, NTT would like to contribute to the creation of a harmonious, diverse, and fair society that supports people in carrying out advanced intellectual work and creative tasks that had been difficult for people alone to achieve, such as marketing analysis, formulation of business strategies, and community consensus building.

References

[1] P. Faraboschi, E. Giles, J. Hotard, K. Owczarek, and A. Wheeler, "Reducing the Barriers to Entry for Foundation Model Training," arXiv preprint arXiv:2404.08811, 2024.
[2] Press release issued by NTT, "Realize the World's First 'Learning Transfer' to Reuse Past Learning Trajectories, Significantly Reducing the Cost of Retraining AI Models," May 7, 2024. https://group.ntt/en/newsrelease/2024/05/07/240507b.html
[3] IOWN Functions and Characteristics/APN, https://group.ntt/en/group/iown/function/apn.html
[4] K. Ninokata and C. Schumacher, "Data-centric Infrastructure for Enabling Practical Use of IOWN," NTT Technical Review, Vol. 23, No. 7, pp. 36–42, July 2025. https://doi.org/10.53829/ntr202507fa4

*3 Hallucination: The outputting of information by AI not based on facts.

Junji Teramoto
Executive Research Engineer, Innovative Computing Architecture Laboratory, NTT Computer and Data Science Laboratories.

He received a B.E. and M.E. in electronics, information, and communication engineering from Waseda University, Tokyo, in 1996 and 1998. He has been engaged in research and development on multimedia database management systems since joining NTT Information Laboratories in 1998. He then worked to improve the quality of open-source software, such as PostgreSQL and Linux, to facilitate its introduction into core communications systems. His current interest is in a distributed high-performance computing system.