

Face Matters Simply and Honestly While Communicating. Nurture Intuition Underpinned by Proven Results



Shingo Kinoshita

Senior Vice President, Head of Research and Development Planning, NTT Corporation

Abstract

In 2023, NTT Group announced its new medium-term management strategy: “New Value Creation & Sustainability 2027 Powered by IOWN.” To execute this strategy, NTT laboratories are committed to pursuing the world’s best research and development (R&D) under the following guiding principles: keep researchers motivated and excited, research and develop powerful technology to benefit society in a scalable and sustainable manner, create the future rather than predict it, and nurture intuition and be creative. We interviewed Shingo Kinoshita, NTT senior vice president, head of Research and Development Planning, about the strengths of NTT’s R&D and his mindset as a top executive.

Keywords: R&D, large language model, IOWN

NTT’s research and development (R&D) has over 70 years of history and a wealth of world-class human resources

—Over a year has passed since you became head of the Research and Development Planning Department. Could you tell us about the current state of R&D at NTT?

Since I was appointed head of the Research and Development Planning Department in June 2023, it has been busy but fulfilling days as the research we have been focusing on is now moving into the phase of practical application and business development. For example, after we launched the first commercial service of the Innovative Optical and Wireless

Network (IOWN), IOWN1.0, in March 2023, in November that year, we announced NTT’s large language model (LLM) called “tsuzumi,” and in March 2024, we launched a commercial service of tsuzumi.

Since we are in the critical phase in which the results of our research are developed into a business, we have been collaborating with many people, including our customers and NTT operating companies. Generative artificial intelligence (AI) is advancing rapidly, and to meet the needs of society, R&D must keep up with that pace. We have therefore accelerated R&D on tsuzumi and announced it six months ahead of schedule.

NTT’s R&D expenditures and the number of researchers are actually not that large compared with other global information and communication



technology (ICT) companies. Currently, NTT has approximately 2200 researchers, and R&D expenditures are approximately 120 billion yen. Although these numbers are large in Japan, compared with GAFAM (Google, Apple, Facebook (currently Meta), Amazon, Microsoft), we have a fraction of the number of researchers, and our R&D expenditures are several tenths of theirs.

—How NTT can compete with global companies that have unparalleled resources in terms of R&D expenditures and the number of researchers?

One of NTT R&D's strengths is our abundant human resources, namely, a large number of talented researchers. To make up the above-mentioned difference in R&D expenditures, we are making creative efforts to achieve results, and we in management are supporting our researchers in terms of funding and human resources to the maximum extent possible.

Among ICT-related companies, NTT is ranked ninth in the world in terms of the number of papers published, which is a measure of research capability. In the fields of optical communications, information security, neural engineering, speech recognition, and quantum computing, the number of papers published by NTT ranks first or second in the world, which puts us ahead of Google and IBM. Achieving this ranking in spite of the fact that the number of our researchers and scale of R&D expenditures are significantly

smaller is truly down to the high capability of our researchers. To solidify our position as a world leader in research, we hope to raise our ranking in terms of the number of papers published to fifth place in the next few years.

In addition to having outstanding human resources, NTT's R&D has a history of more than 70 years. GAFAM, for example, have not existed for very long, and their personnel are sometimes in flux. In contrast, we have a long history of research, the fundamentals of which have been handed down from senior to junior researchers over a period of nearly 70 years. At NTT laboratories, researchers from a very wide range of fields, from networking to AI and quantum computing, can interact with each other as colleagues, which facilitates synergy. Interacting with external researchers is called "open innovation," but we can generate a wide range of innovations within NTT laboratories.

High motivation among researchers is also the strength of NTT's R&D. For researchers, financial rewards are important; however, the most-important factors are the freedom of choosing a research theme, a comfortable research environment, and connection with co-researchers. NTT laboratories are home to many world-class researchers in various fields, such as NTT Fellow Tatsuaki Okamoto, who is a world authority on cryptography and blockchain, NTT Fellow Yutaka Miyamoto, who is an expert in high-capacity and scalable optical transport network

technology, and NTT Fellow Takehiro Moriya, who is an expert in speech and audio signal processing and coding. These “top runners” act as role models to pass on their spirit and style of research, and this practice has become the unifying force for our researchers and has formed the foundation of our history. In this sense, NTT’s R&D is a treasure trove of human resources.

Toward the social implementation of IOWN and tsuzumi

—Could you tell us about current major R&D initiatives?

Let me first introduce NTT’s LLM “tsuzumi,” which we announced at a press conference in November 2023. The key features of tsuzumi are lightweight, high linguistic proficiency (especially in the Japanese language), flexible customization, and multimodality.

Regarding the first feature, lightweight, two versions of tsuzumi are available: “tsuzumi-0.6B,” an ultra-lightweight version with 600 million (0.6B) parameters, and “tsuzumi-7B,” a lightweight version

with 7 billion (7B) parameters. OpenAI’s GPT-3, a representative LLM, requires a large-scale computer. For example, training a GPT-3-class model requires an hour’s worth of electricity produced by a single nuclear plant, and using it requires multiple high-end graphics processing units (GPUs). With tsuzumi, both power consumption and the number of GPUs can be reduced to a few tenths of those amounts.

Regarding the second feature, high linguistic proficiency, when comparisons of tsuzumi with GPT-3.5 and other LLMs were made using the Rakuda benchmark, a performance-evaluation method for generative AI, tsuzumi beat GPT-3.5 at over an 80% win rate and beat the four top-ranked Japanese LLMs at an overwhelmingly high win rate.

Regarding the third feature, flexible customization, LLMs are good at answering general questions; however, questions specific to an industry or company can be difficult to answer. To address this difficulty, tsuzumi supports a variety of tuning methods.

Regarding the fourth feature, multimodality, while a typical LLM is asked and answers questions in language, tsuzumi understands not only language but also visual and auditory inputs, such as charts, pictures, and audio.

I should explain why an LLM with these excellent features could be developed in such a short period—it is largely due to the technical capabilities of NTT laboratories. For example, NTT is top ranked in the world and first in Japan in terms of the number of papers presented at the most prestigious international conference in the field of natural language processing, which is one of the most important fields concerning AI. NTT has also won first place in global competitions in machine translation and other fields as well as numerous awards from domestic research associations.

To develop an LLM, it is necessary to prepare a large amount of high-quality training data, and we have prepared training data containing more than one-trillion tokens (the smallest unit used in text analysis), including not only Japanese and English but also 21 other languages and programming languages, for pre-training. This dataset covers a very wide range of fields, from various specialties to entertainment. For instructional tuning that follows pre-training, we created a wide range of new data and used existing training data accumulated over 40 years of research on natural language processing, including translation, summarization, dialogue, and reading comprehension.





—We can't wait to see tsuzumi put into practical use. How will IOWN play a role in that? Could you tell us about the progress in the IOWN initiative?

Since the announcement of IOWN in 2019, we have advanced research, development, and practical application of it and launched “IOWN1.0,” the first commercial service, in March 2023. We are currently conducting R&D with the aim of deploying IOWN in society. Specifically, we are working on the practical application of photonics-electronics convergence (PEC) devices and their application to communications and computing as well as Digital Twin Computing and next-generation general-purpose AI to maximize the use of those devices and applications.

Let me introduce the next milestones, namely, IOWN2.0, 3.0, and 4.0. In IOWN1.0, optical connections are made between datacenters without optical-to-electrical conversion. In IOWN2.0, optical connections will be made between computer boards in datacenters by using second-generation PEC devices. IOWN3.0 will enable optical connections between semiconductor packages inside a computer board by using third-generation PEC devices, and IOWN4.0 will enable optical connections between dies (chips) inside a semiconductor package by using fourth-generation PEC devices. The application area of

optics will be expanded to the inside of semiconductor packages in a manner that achieves overwhelmingly high speed, wide bandwidth, and low power consumption.

IOWN also plays an important role in the R&D of LLMs. For example, during the development of tsuzumi, the training data was located in Yokosuka City (Kanagawa Prefecture) and the GPU cluster was located in Mitaka City (Tokyo), and by connecting the locations with the IOWN1.0 All-Photonics Network, we were able to create an efficient training environment for an LLM that makes the 100-km distance seem irrelevant.

After introduction of IOWN2.0 and beyond, IOWN will become even more important in the R&D of LLMs. Current computing systems have a fixed architecture consisting of central processing units (CPUs) and GPUs, and in some cases, the performance of GPUs cannot be fully exerted due to CPU intervention. On the contrary, IOWN allows the necessary number of CPUs and GPUs to be directly connected by optics on a per-component basis in a flexible architecture. Therefore, dynamic control of combinations of CPUs and GPUs optimized for LLM training and inference will become possible.

Looking further into the future, we pursue an “AI constellation” as NTT’s vision for the world of AI. An AI constellation is a next-generation AI architecture

in which multiple, small, and specialized LLMs are combined, rather than creating one huge, monolithic LLM, that are smarter and more efficient than a gigantic LLM.

For example, we are thinking of creating a system by which AI with various personas, such as a student, senior citizen, elementary-school teacher, parent raising children, and doctor, could discuss the question, “What is needed to revitalize our community, which is experiencing a decline in population?” Each persona could give their own opinion, and those opinions could be combined or a consensus could be reached while a real person could occasionally step in to interact and help reach a consensus. IOWN will play an important role in efficiently coordinating a large number of distributed AIs.

**Create an environment in which everyone can
evaluate each other fairly under four
guiding principles**

—Given your journey thus far, could you tell us what you value as a top executive?

I started my career at NTT as a researcher in 1991. At that time, NTT had just been privatized, and the Internet was in its infancy. Although I had majored in physics at university, I felt the difficulty in research on physics, and the newness of privatized NTT and computers appealed to me, so I jumped at the chance to join an NTT laboratory.

To be honest, I had no knowledge of telecommunications, so I had quite a hard time after I joined that NTT laboratory. Even so, my outstanding senior researchers taught me a lot of things, and I immediately fell in love with research. Because I like new things, when I encountered a new research theme, I got as excited as a child who had received a new toy, and I was very happy when my papers were accepted by academic conferences or cited by other researchers.

I also directed a large-scale project of practical application of research—involving several hundred people—for the Tokyo Olympics. The pleasure of research differs in accordance with the research phase, but each phase, whether research, development, or practical application, has its own unique charms. I’m very happy to be able to contribute to society through the results of our research.

Now as the head of the Research and Development Planning Department, I’m striving to create an environment in which each researcher can engaged in

their research activities by assigning them to research groups in a well-balanced manner while taking into account their individual joys, purposes, and characteristics.

I value simplicity and honesty when dealing with matters. When working in an organization, as a result of consideration and coordination in many directions, matters may become more complicated than necessary or deviate from what they should be. Therefore, researchers should value a well-honed simplicity that is easy to understand and satisfactory to everyone.

When matters become complicated, it can be difficult to correct course, so you may have no choice but to accept that situation. From a medium- to long-term perspective, however, going back to the starting point and returning things to their ideal state is a shortcut to success. I want to face things honestly without giving up on anything.

—How do you think NTT’s R&D will develop in the future?

My guiding principles for NTT laboratories are fourfold: keep researchers motivated and excited, research and develop powerful technology to benefit society in a scalable and sustainable manner, create the future rather than predict it, and nurture intuition and be creative.

At the time when research on telecommunication networks and computers was in its infancy, if companies could make a high-speed product inexpensively, people were sure to use it. In other words, if researchers researched, developed, and commercialized what they considered to be “good products,” they were able to benefit society. However, society is changing rapidly and becoming more complex, and people’s values are diversifying, so the “good products” that researchers think of and the “good products” that society demands do not necessarily match. This state of affairs is especially true in the global market.

Under this circumstance, I want to lead our R&D under the guiding principles that we set forth while keeping in mind the words of Goro Yoshida, the first director of the Electrical Communication Laboratory, “Do research by drawing from the fountain of knowledge and provide specific benefits to society through practical application.” In other words, we will draw from the fountain of knowledge to solidify our position as the world’s best research institute, implement IOWN in society by pursuing practical application, and provide specific benefits to society by striking the right balance between a market-in approach, in

which we plan and promote research on the basis of market needs, and product-out approach, in which we promote research to create a future that the market has not even noticed.

I also value intuition. In the world of research, there is intuition that only researchers who have mastered a research field can have. Intuition underpinned by proven results has the power to move people. That is why I expect our researchers to accumulate results. At the same time, we in management intend to continue developing our ability to accept the intuitions of our young researchers who are accumulating results. We will produce numerous research results in an environment where we can fairly evaluate each other as researchers. I ask our partners and members of society to make the most of our researchers, who are

conducting world-class R&D, and make full use of the results of their research.

Interviewee profile

■ Career highlights

Shingo Kinoshita joined NTT in 1991. In 2008, he became a senior manager in charge of human resources at NTT Information Sharing Platform Laboratories; in 2012, he became a senior manager of the R&D Planning Group at NTT Research and Development Planning Department; and in 2021, he became vice president, head of NTT Human Informatics Laboratories. He assumed his current position in June 2023.