LLM+×IOWN —The Advancement of IOWN, the Launch of NTT's LLM, and Their Synergy—

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

Abstract

This article presents tsuzumi, NTT's large language model (LLM), and the advancement of the Innovative Optical and Wireless Network (IOWN). It is based on the keynote speech given by Shingo Kinoshita, senior vice president, head of Research and Development Planning of NTT Corporation, at the "NTT R&D FORUM 2023 — IOWN ACCELERATION" held from November 14th to 17th, 2023.

Keywords: large language model, IOWN, photonics-electronics convergence, AI

1. Launch of NTT's large language model

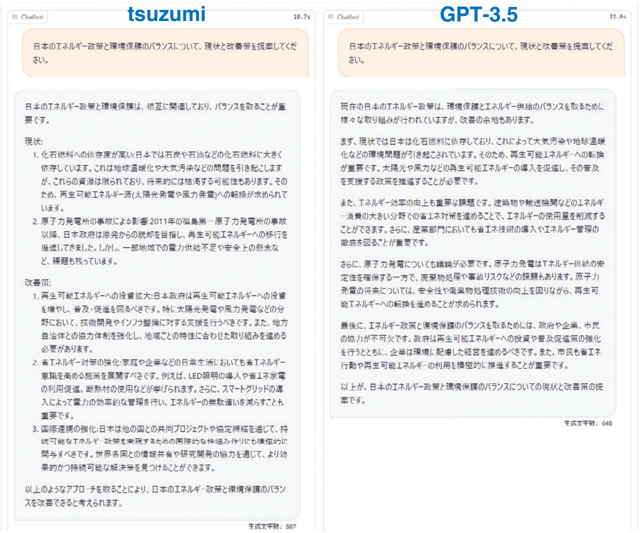
We announced at a press conference on November 1 that we have developed tsuzumi, NTT's large language model (LLM). This LLM has four main features.

1.1 Lightweight

Its first feature is that it is lightweight. Today's LLMs are in competition for the number of parameters, thus becoming very large in scale. As such, the challenge now is achieving sustainability. For example, OpenAI's Generative Pre-trained Transformer 3 (GPT-3) has 175B parameters and requires about 1300 MWh of electricity for one training session. This is equivalent to an hour's worth of electricity produced by one nuclear power plant. In contrast, we have adopted the following strategy.

In terms of direction being pursued, we do not aim to build one massive LLM that knows everything but small LLMs with specialized knowledge. Our approach to this is not simply to increase the parameter size but to make the LLM smarter by improving the quality and quantity of training data added to it. Therefore, we developed and announced two types of tsuzumi. The ultralight version, tsuzumi-0.6B, is 0.6B in parameter size, which is about 1/300th that of GPT-3. The light version, tsuzumi-7B, is 1/25th that of GPT-3.

What are the benefits of reducing size? The first is that training can be carried out at very low cost. For example, GPT-3-scale training is said to cost about 470 million yen per session. In contrast, tsuzumi-7B and tsuzumi-0.6B cost 19 and 1.6 million yen, respectively, enabling reduction in costs by 1/25 and 1/300. The second benefit is the cost of inference, i.e., the cost of using the LLM. For example, GPT-3 would need about 5 high-end graphics processing units (GPUs), which cost about 15 million yen. The 7B and 0.6B versions, however, would cost about 0.7 and 0.2 million yen, respectively. In terms of the number of GPUs, cost is reduced with the use of only



日本のエネルギー政策と環境保護のバランスについて、現状と改善策を提案してください。

日本のエネルギー政策と環境保護は、相互に関連しており、バランスを取ることが重要です。

現状:

1. 化石燃料への依存度が高い:日本では石炭や石油などの化石燃料に大きく依存しています。これは地球温暖化や大気汚染などの問題を引き起こしますが、これらの資源は限られており、将来的には枯渇する可能性もあります。そのため、再生可能エネルギー源(太陽光発電や風力発電)への転換が求められています。
2. 原子力発電所の事故による影響:2011年の福島第一原子力発電所の事故以降、日本政府は原発からの脱却を目指し、再生可能エネルギーへの移行を推進してきました。しかし、一部地域での電力供給不足や安全上の懸念など、課題も残っています。

改善策:

1. 再生可能エネルギーへの投資拡大:日本政府は再生可能エネルギーへの投資を増やし、普及・促進を図るべきです。特に太陽光発電や風力発電などの分野において、技術開発やインフラ整備に対する支援を行うべきです。また、地方自治体との協力体制を強化し、地域ごとの特性に合わせた取り組みを進める必要があります。
2. 省エネルギー対策の強化:家庭や企業などの日常生活においても省エネルギー意識を高める施策を展開すべきです。例えば、LED照明の導入や省エネ家電の利用促進、断熱材の使用などが挙げられます。さらに、スマートグリッドの導入によって電力の効率的な管理を行い、エネルギーの無駄遣いを減らすことも重要です。
3. 国際連携の強化:日本は他の国との共同プロジェクトや協定締結を通じて、持続可能なエネルギー政策を実現するための国際的な枠組み作りにも積極的に関与すべきです。世界各国との情報共有や研究開発の協力を通じて、より効果的かつ持続可能な解決策を見つけることができます。

以上のようなアプローチを取ることにより、日本のエネルギー政策と環境保護のバランスを改善できると考えられます。

生成文字数: 807

日本のエネルギー政策と環境保護のバランスについて、現状と改善策を提案してください。

現在の日本のエネルギー政策は、環境保護とエネルギー供給のバランスを取るために様々な取り組みが行われていますが、改善の余地もあります。

まず、現状では日本は化石燃料に依存しており、これによって大気汚染や地球温暖化などの環境問題が引き起こされています。そのため、再生可能エネルギーへの転換が重要です。太陽光や風力などの再生可能エネルギーの導入を促進し、その普及を支援する政策を推進することが必要です。

また、エネルギー効率の向上も重要な課題です。建築物や輸送機関などのエネルギー消費の大きい分野での省エネ対策を進めることで、エネルギーの使用量を削減することができます。さらに、産業部門においても省エネ技術の導入やエネルギー管理の徹底を図ることが重要です。

さらに、原子力発電についても議論が必要です。原子力発電はエネルギー供給の安定性を確保する一方で、廃棄物処理や事故リスクなどの課題もあります。原子力発電の将来については、安全性や廃棄物処理技術の向上を図りながら、再生可能エネルギーへの転換を進めることが求められます。

最後に、エネルギー政策と環境保護のバランスを取るためには、政府や企業、市民の協力が不可欠です。政府は再生可能エネルギーへの投資や普及促進策の強化を行うとともに、企業は環境に配慮した経営を進めるべきです。また、市民も省エネ行動や再生可能エネルギーの利用を積極的に推進することが重要です。

以上が、日本のエネルギー政策と環境保護のバランスについての現状と改善策の提案です。

生成文字数: 648

Fig. 1. Comparison of tsuzumi and GPT-3.5.

one low-end GPU and only one central processing unit (CPU), respectively.

1.2 Proficiency in Japanese

Its second feature is its high linguistic proficiency, especially in the Japanese language. **Figure 1** shows example answers of tsuzumi on the left and GPT-3.5 on the right to a question about the current situation and possible improvement measures for attaining a balance between Japan's energy policies and environmental protection. The left answer presents a well-analyzed response in Japanese. Comparisons of tsuzumi with GPT-3 and other LLMs were made using the Rakuda benchmarks. For example, tsuzumi was compared against GPT-3.5. They were asked the same questions, and their answers were input to GPT-4 to determine which gave the better answer and which is superior or inferior. As a result, tsuzumi beat GPT-3.5 at 52.5% probability. Comparisons with the four top-class LLMs in Japan revealed that tsuzumi had an overwhelming win rate of 71.3 to 97.5%.

Furthermore, tsuzumi can not only give answers in Japanese to questions in Japanese. For example, when tsuzumi was asked to extract four data items, namely, device name, achievement, exhibition event, and future plans in JSON format from a text of a recent press release about an artificial photosynthesis device developed by NTT, it gave a properly structured response as requested. Its performance in English is in fact on par with the world's top-level language models. In comparison to Llama 2, an English language model developed by Meta, tsuzumi produced almost the same English benchmark results. For example, tsuzumi gave a smooth and quick response when asked to translate the Japanese text into English. It is also proficient in programming language. When asked to write code in a specific format, it gave a proper response in the requested format. It is currently being trained in Chinese, Korean, Italian, and German, so it will also eventually be able to give answers in those languages.

1.3 Flexible customization

Its third feature is flexible customization. Language models have a base model, which can give fairly adequate answers to general questions. However, to construct models that can give specific answers in the field of finance or the public sector, for example, tuning must be executed. There are three tuning methods available (**Fig. 2**). On the left of Fig. 2 is a method called prompt engineering. With this method, prompt inputs to the base model are added with financial information to enable the model to give a more finance-specific response. In the middle is a method called full fine-tuning, which creates a model specialized in finance by re-training the base model with financial data and changing the entire set of parameters. On the right is a method called adapter tuning, with which the base model is used as is, and these blue adapter components on finance-specific knowledge are added on top of the base model like a hat. They have different advantages and disadvantages in

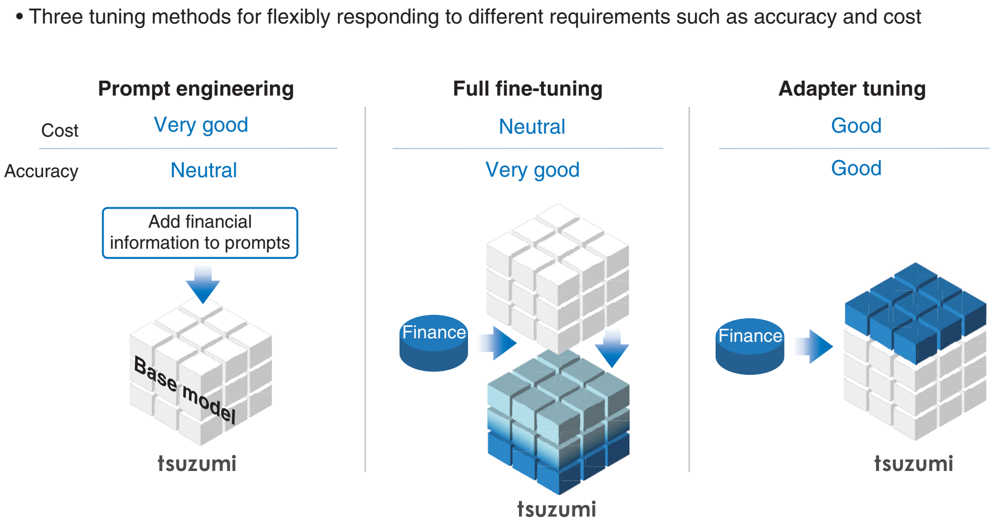
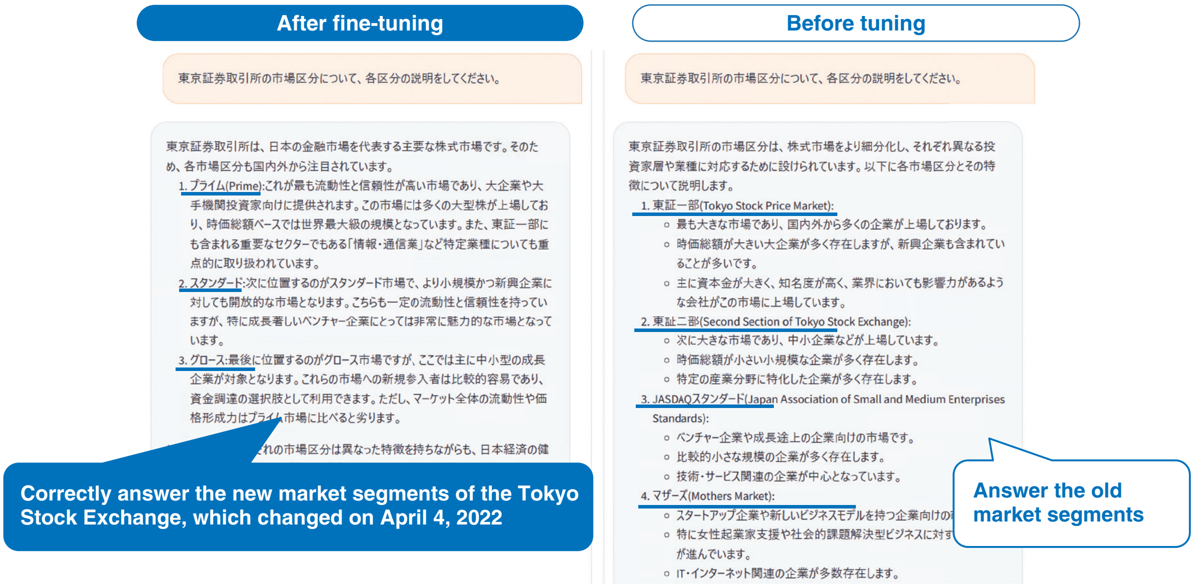


Fig. 2. Different tuning methods.



After fine-tuning

東京証券取引所の市場区分について、各区分の説明をしてください。

東京証券取引所は、日本の金融市場を代表する主要な株式市場です。そのため、各市場区分も国内外から注目されています。

1. プライム(Prime):これが最も流動性と信頼性が高い市場であり、大企業や大手機関投資家向けに提供されます。この市場には多くの大型株が上場しており、時価総額ベースでは世界最大級の規模となっています。また、東証一部にも含まれる重要なセクターでもある「情報・通信業」など特定業種についても重点的に取り扱われています。
2. スタンダード:次に位置するのがスタンダード市場で、より小規模かつ新興企業に対しても開放的な市場となります。こちらも一定の流動性と信頼性を持っていますが、特に成長著しいベンチャー企業にとっては非常に魅力的な市場となっています。
3. グロース:最後に位置するのがグロース市場ですが、ここでは主に中小型の成長企業が対象となります。これらの市場への新規参入者は比較的容易であり、資金調達の選択肢として利用できます。ただし、マーケット全体の流動性や価格形成力はプライム市場に比べると劣ります。

れの市場区分は異なった特徴を持ちながらも、日本経済の健

Correctly answer the new market segments of the Tokyo Stock Exchange, which changed on April 4, 2022

Before tuning

東京証券取引所の市場区分について、各区分の説明をしてください。

東京証券取引所の市場区分は、株式市場をより細分化し、それぞれ異なる投資家層や業種に対応するために設けられています。以下に各市場区分とその特徴について説明します。

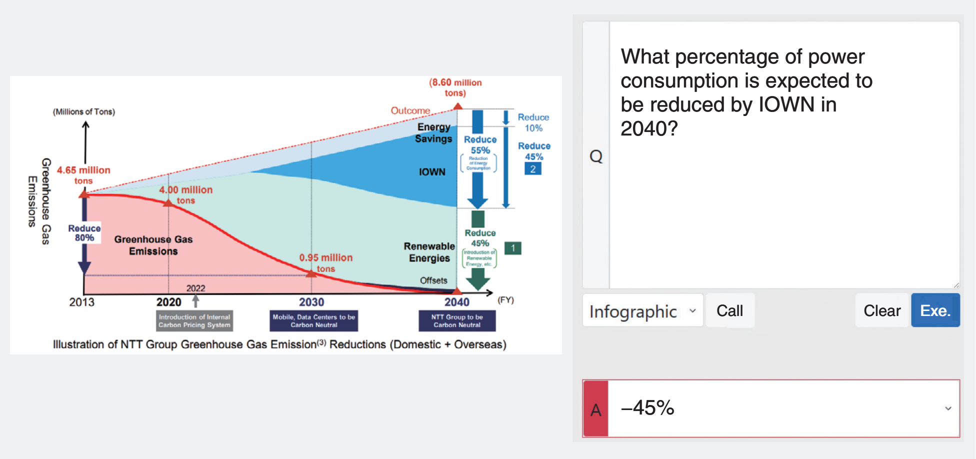
1. 東証一部(Tokyo Stock Price Market):
   - 最も大きな市場であり、国内外から多くの企業が上場しております。
   - 時価総額が大きい大企業が多く存在しますが、新興企業も含まれていることが多いです。
   - 主に資本金が大きく、知名度が高く、業界においても影響力があるような会社がこの市場に上場しています。
2. 東証二部(Second Section of Tokyo Stock Exchange):
   - 次に大きな市場であり、中小企業などが上場しています。
   - 時価総額が小さい小規模な企業が多く存在します。
   - 特定の産業分野に特化した企業が多く存在します。
3. JASDAQスタンダード(Japan Association of Small and Medium Enterprises Standards):
   - ベンチャー企業や成長途上の企業向けの市場です。
   - 比較的小さな規模の企業が多く存在します。
   - 技術・サービス関連の企業が中心となっています。
4. マザーズ(Mothers Market):
   - スタートアップ企業や新しいビジネスモデルを持つ企業向けの
   - 特に女性起業家支援や社会的課題解決型ビジネスに対す
     が進んでいます。
   - IT・インターネット関連の企業が多数存在します。

Answer the old market segments

Fig. 3. Example of fine-tuning in financial industry.

terms of cost and accuracy. As an example of the advantage of tuning, the base model can be made more specific to a particular industry, company, or organization or can be updated with the latest information. One can also add functions by training the model with new tasks, such as summarization and translation, to make it more task specific.

Figure 3 shows an example of fine-tuning in the financial industry. On the right is the response using data before tuning, and on the left is the response after fine-tuning the data for the financial industry. LLM tsuzumi was asked to explain the market segments of the Tokyo Stock Exchange (TSE). The right response shows the old segments, such as 1st section, 2nd section, JASDAQ, and Mothers. The left response shows the new segments established by TSE on April 4, 2022. The LLM properly learned and gave the correct segments, i.e., Prime, Standard, and Growth.



Q: What percentage of power consumption is expected to be reduced by IOWN in 2040?

A: –45%

Fig. 4. Modality extension: Language + Visual.

1.4 Multimodality

Its fourth feature is multimodality. Language models in general have received language inputs and produced language outputs. Multimodality refers to the ability to handle modals other than language. For example, one can now add visual and audio capabilities. For example, when tsuzumi was shown a receipt and asked, “What is the total amount excluding the 10% consumption tax?” It then calculated the total by looking at the unit price and quantity columns on the receipt and correctly replied that the total is 9500 yen.

Figure 4 shows another example. This is a graph from NTT’s Green Vision. While being shown this complicated graph, tsuzumi was asked “What percentage of power consumption is expected to be reduced by IOWN in 2040?” It correctly analyzed the graph and replied 45%. Thus, it is capable of providing answers by analyzing the question in combination with figures.

2. Technological capabilities of NTT laboratories

Next, I would like to talk about the technological capabilities of NTT laboratories that enabled us to achieve these four excellent features.

Figure 5 shows a table listing the ranking of companies based on the number of publications in the field of artificial intelligence (AI). NTT is ranked 12th in the world and 1st in Japan in this report published annually by a U.S. venture capital firm. In the top 1 to 11 are GAFA (Google, Apple, Facebook, and Amazon) and other major information technology (IT) vendors in the U.S. and China. Compared with NTT, they probably have ten times more research funding and have many times more researchers. Nevertheless, NTT has been able to conduct research quite efficiently, enabling us to achieve these rankings. In AI, natural language processing is a very important area in the development of language models. NTT is number one in Japan in terms of the number of publications in this particular area. It is also number one in the number of awards for excellence given by the Japanese Association for Natural Language Processing. The development of tsuzumi is backed by a long history and solid track record in research. We used excellent data for training tsuzumi, which is its distinguishing feature. We used more than 1T tokens, which can be thought of as the number of words, for pre-training. We used not only Japanese and English but also 21 other languages as well as programming languages to train tsuzumi. These data cover a very wide range of domains, from specialized fields to entertainment. For instruction tuning, which makes the pre-trained data more human-like in response and behavior, we used Japanese corpora created from NTT’s over more than 40 years of research. We also used new tuning data we created specifically for generative AI.

In this R&D forum, we presented 11 exhibits on tsuzumi. I will briefly introduce representative exhibits in the following sections.

2.1 LLM tsuzumi comprehensively understands the real world

A supervisor and junior member are having a

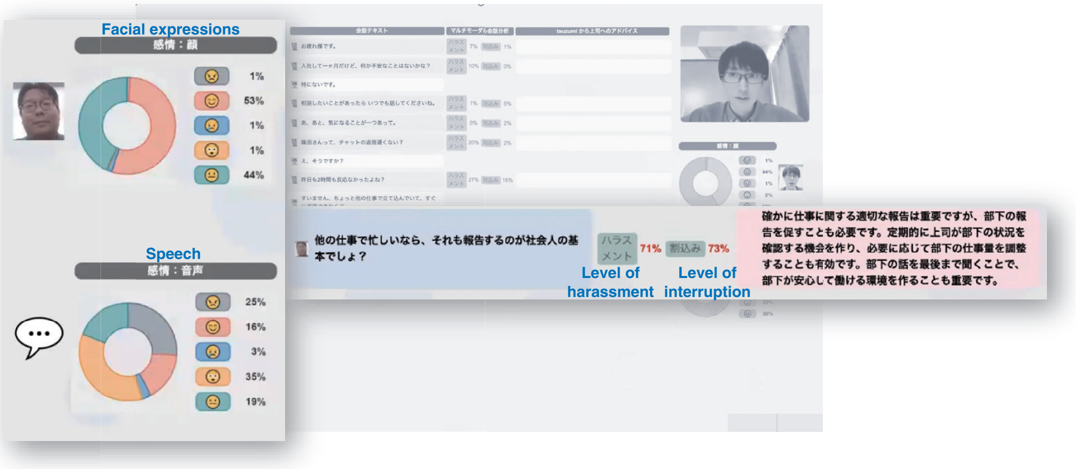
Rank	Company name
1	Google (USA)
2	Microsoft (USA)
3	Facebook (USA)
4	Amazon (USA)
5	IBM (USA)
6	Huawei (China)
7	Alibaba (China)
8	NVIDIA (USA)
9	Tencent (China)
10	Samsung (South Korea)
11	Baidu (China)
12	**NTT** (Japan)
13	Apple (USA)
14	OpenAI (USA)
15	Intel (USA)
16	Adobe (USA)
17	Salesforce (USA)
18	Yandex (Russia)
19	NEC (Japan)
20	VinAI (Vietnam)

Top 100 Global Companies Leading in AI Research in 2022[*1]

[*1] https://thundermark.medium.com/ai-research-rankings-2022-sputnik-moment-for-china-64b693386a4

Fig. 5. Number of AI publications: 12th in the world & 1st in Japan.

- Multimodal comprehension of communication between supervisors and junior members to address power-harassment by supervisors

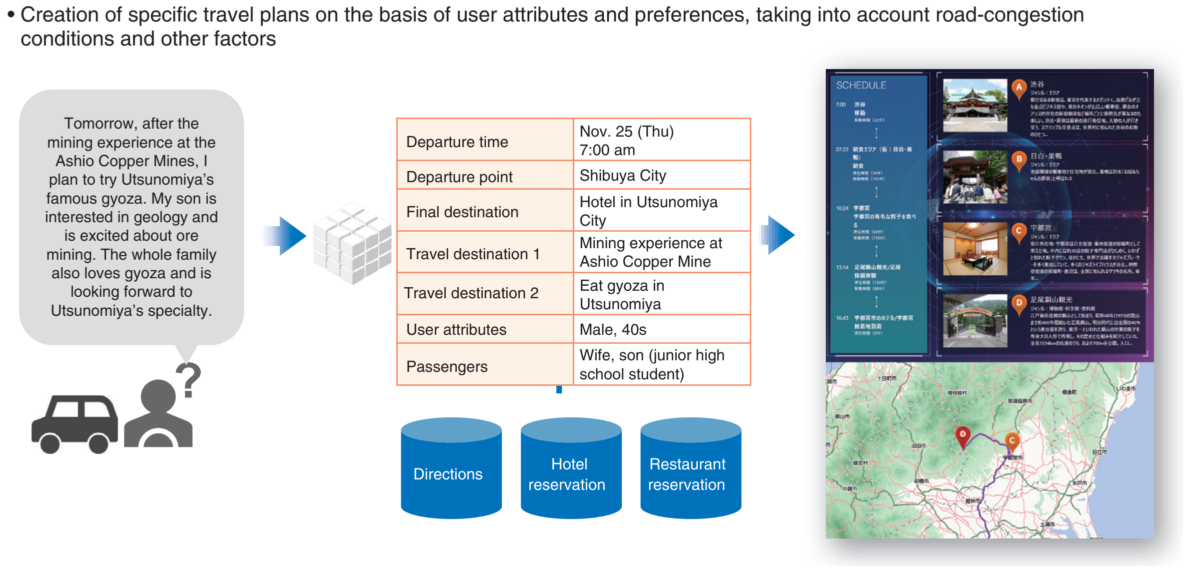


他の仕事で忙しいなら、それも報告するのが社会人の基本でしょ？

確かに仕事に関する適切な報告は重要ですが、部下の報告を促すことも必要です。定期的に上司が部下の状況を確認する機会を作り、必要に応じて部下の仕事量を調整することも有効です。部下の話を最後まで聞くことで、部下が安心して働ける環境を作ることも重要です。

Fig. 6. LLM tsuzumi comprehensively understands the real world.

conversation via online communication, during which the supervisor displays power-harassment behavior. LLM tsuzumi detects the power-harassment behavior and calls it to the supervisor's attention. "I'm sorry. I'm a little busy with other jobs so I couldn't reply immediately." "If you're busy with other work, isn't it your basic responsibility as a working adult to report that, too?" The person on the left of **Fig. 6** is the supervisor making power-harassment statements. The LLM then analyzes the emotions on the basis of the supervisor's facial expressions and speech. It determines to what degree the person is laughing or angry as percentages. The blue area on the left in the below column shows what the supervisor said. The middle shows that the level of harassment is about 71%. It also shows a 73% level



- Creation of specific travel plans on the basis of user attributes and preferences, taking into account road-congestion conditions and other factors

Tomorrow, after the mining experience at the Ashio Copper Mines, I plan to try Utsunomiya's famous gyoza. My son is interested in geology and is excited about ore mining. The whole family also loves gyoza and is looking forward to Utsunomiya's specialty.

Departure time	Nov. 25 (Thu) 7:00 am
Departure point	Shibuya City
Final destination	Hotel in Utsunomiya City
Travel destination 1	Mining experience at Ashio Copper Mine
Travel destination 2	Eat gyoza in Utsunomiya
User attributes	Male, 40s
Passengers	Wife, son (junior high school student)

Fig. 7. LLM tsuzumi understands the user situation.

of interruption while the other person is speaking. These percentages show a relatively high level of harassment. In response, as shown in the pink area, tsuzumi gives advice to encourage behavioral change in the supervisor. It says, "While it is important to properly report on their work, it is also important to encourage junior members to do so. An effective way to encourage them is for the supervisor to create opportunities to check on junior members on a regular basis and adjust their workload as necessary. It is also important to listen patiently to junior members to create an environment in which they can work with peace of mind." Thus, tsuzumi can give appropriate advice to supervisors to address their behavior.

2.2 LLM tsuzumi understands user situations

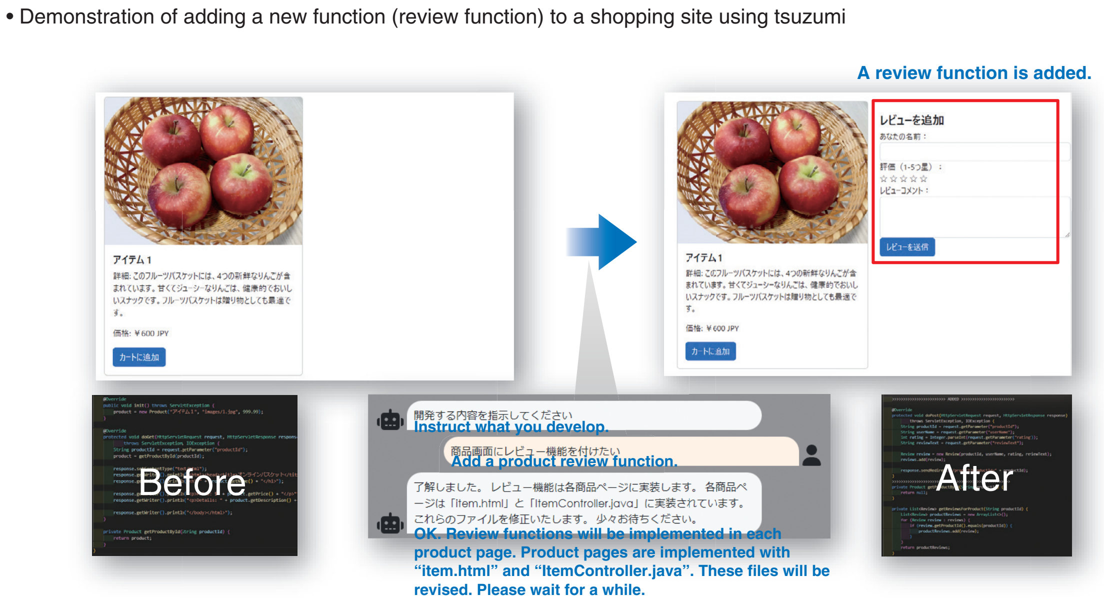
LLM tsuzumi can create specific travel plans on the basis of user attributes and preferences, taking into account road-congestion conditions and other factors. The gray text on the left of **Fig. 7** shows the user's requests. It says, "Tomorrow, after the mining experience at the Ashio Copper Mines, I plan to try Utsunomiya's famous gyoza. My son is interested in geology and is excited about ore mining. The whole family also loves gyoza and is looking forward to Utsunomiya's specialty." Once the car-navigation system with tsuzumi hears this, it structures and analyzes the inputs into specific information such as departure time and departure point, as shown in the middle. It then searches the web to gather information on directions, hotel reservations, and restaurant reservations. Finally, it creates an action or travel plan to propose to the user.

2.3 LLM tsuzumi with physical senses and a robot body

A robot equipped with tsuzumi can create a menu and set a table according to the user's request. In this example, the user says, "Prepare a dinner table that will warm the body up on a cold winter's day. Make considerations for left-handedness." The robot then analyzes the request and actually serves the food while explaining the arrangement. For example, it says, "Curry is good for warming the body up, and salad. They pair well with spring rolls for a seasonal feel. Tea warms the body, too. In consideration of left-handers, the chopstick and spoon are placed in the opposite direction." The robot servs food while giving such explanations.

2.4 Ultra-high-speed software development

How to add a new function, i.e., a review function, to a shopping site with tsuzumi was demonstrated. The website on the left of **Fig. 8** only gives the usual introduction of a product, without a review function. One can then instruct tsuzumi to add a product review function. In response, it analyzes the source code and carries out these actions. The demonstration showed how tsuzumi can write a new source code to create a review section on a website.



- Demonstration of adding a new function (review function) to a shopping site using tsuzumi

A review function is added.

アイテム 1

詳細: このフルーツバスケットには、4つの新鮮なりんごが含まれています。甘くてジューシーなりんごは、健康的でおいしいスナックです。フルーツバスケットは贈り物としても最適です。

価格: ￥600 JPY

レビューを追加

あなたの名前：

評価（1-5つ星）：

☆☆☆☆☆

レビューコメント：

開発する内容を指示してください
Instruct what you develop.

商品画面にレビュー機能を付けたい
Add a product review function.

了解しました。レビュー機能は各商品ページに実装します。各商品ページは「item.html」と「ItemController.java」に実装されています。これらのファイルを修正いたします。少々お待ちください。
OK. Review functions will be implemented in each product page. Product pages are implemented with “item.html” and “ItemController.java”. These files will be revised. Please wait for a while.

Fig. 8. Ultra-high-speed software development.

2.5 Next-generation security operations

LLM tsuzumi can handle incident responses on behalf of security experts in a dialogue format. For example, if a virus is detected on a user’s computer, tsuzumi analyzes the virus, informs the user, asks the user whether they have accessed the malicious site, and instructs the user to respond as soon as possible via chat. The demonstration showed how tsuzumi interacts with the user step by step to urge users to respond to security issues.

2.6 Phishing-site detection

LLM tsuzumi analyzes the input website and determines whether it is a phishing site. It’s accuracy in detecting phishing sites is more than 98%, which is much more accurate than checking by humans.

3. Advancement of IOWN

Next, I would like to move on to the advancement of the Innovative Optical and Wireless Network (IOWN). First, I would like to explain the IOWN roadmap (**Fig. 9**). IOWN1.0 is a networking technology that connects datacenters (DCs) using optical fiber. IOWN2.0 optically links the boards inside the server in the DC. Evolving further, IOWN3.0 will optically connect the chips, while IOWN4.0 will enable optical connection inside the chip. Now, let’s look at the roadmap by generation. There are a number of elemental technologies that make up IOWN for each generation. An example is a device called the photonics-electronics convergence (PEC) device. Along with the evolution of IOWN generations 1.0 to 4.0, PEC will also continue to evolve from the 2nd to the 3rd, 4th, and 5th generations. The All-Photonics Network (APN) will evolve within IOWN1.0 through the addition of functions and increasing performance. The super white box of the data-centric infrastructure (DCI) will evolve from IOWN1.0, to 2.0, and to 3.0, along with the evolution of PEC devices, as Steps 0, 1, and 2. This roadmap shows how we will be moving forward with the advancement of these technologies.

I’ll now introduce what we achieved for IOWN1.0 in 2023. One achievement is the significant progress in the commercialization of the APN. The APN consists of APN-I for the core network, APN-G for the edge network, APN-T installed in the user base, and OTN (Optical Transport Network) Anywhere in the user terminals. Different companies have launched specific products for the APN. In March 2023, NTT EAST and NTT WEST began providing specific network services using these products. This 100-gigabyte leased-line service enables users exclusive use of optical wavelengths from end to end. Using OTN Anywhere enables visualizing latency and provides functions for adjusting and aligning different delay times. Using this service, we have conducted various proofs of concept (PoCs), and many PoCs were in the

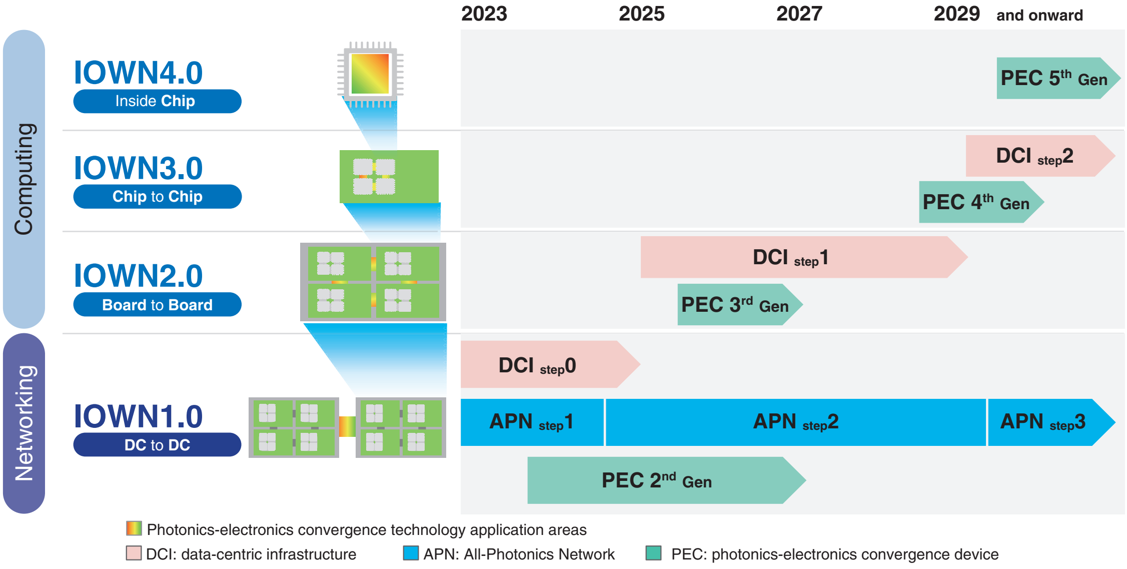


Fig. 9. Progress of IOWN.

entertainment field such as for concerts, e-sports, comedy, and dance.

Other than entertainment, we plan to use the APN to construct DCs of the future. Thus far, the range of DC-to-DC connections has been limited because of significant delays in conventional networks. The limit is about 60 km, but there is not enough land within this range, making it difficult to add more DCs. If the connection distance between DCs can be increased from 60 to 100 km by using the APN, then there will be more land available for constructing connected DCs. We believe that the APN is highly suitable for such expansion of connected DCs. We are currently conducting demonstration experiments at various locations to achieve this. By expanding beyond the Tokyo metropolitan area to major cities throughout Japan and even to other parts of the world, we believe that it will be possible to build a global APN network.

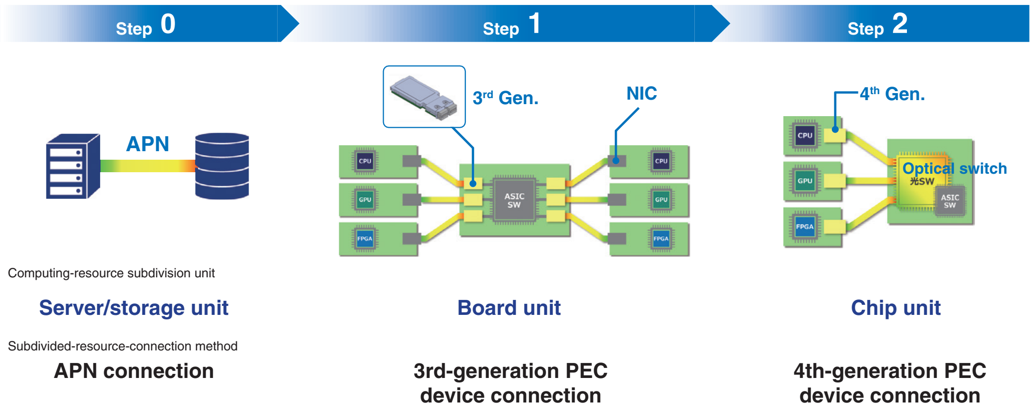
Next, I will report on the status of IOWN2.0 and 3.0. The first is concerning DCI. DCI is a next-generation computing architecture for achieving high performance with low power consumption by allocating optimally subdivided computer resources centered on data (**Fig. 10**).

In Step 0, the units for computer-resource subdivision are the server and storage, and the APN is used for connecting them. In Step 1, the unit of subdivision is the board inside the server. By connecting boards with a 3rd-generation PEC device, we aim to achieve ultra-low power consumption and ultra-high-speed switching. In Step 2, computing resources are subdivided by chip units, which will be connected with a 4th-generation PEC device. This will enable achieving even lower power consumption and higher performance. The key device to achieve Step 1 is the 3rd-generation PEC device called the optical engine. The yellow areas of Step 2 in Fig. 10 correspond to each optical engine. We have been conducting experiments with Broadcom on this chip in the middle, which has a switching capacity of about 5 Tbit/s, with each optical engine having a transmission capacity of 3.2 Tbit/s. It is therefore possible to configure a single device with 5 Tbit/s of switching capacity. The 4th-generation PEC device will optically connect chips at an implementation efficiency six times higher and a power efficiency two times higher than the 3rd generation. This will enable further improved performance and lowering power consumption.

4. Synergy between LLMs and IOWN

The third topic is about the synergy between LLMs and IOWN. For IOWN, we are conducting experiments combining DCI Step 0, the APN, and LLMs. We have this amount of training data in Yokosuka and wanted to install GPUs nearby, but there was not enough power or space. Therefore, we used a GPU cloud in Mitaka and connected it with the database in Yokosuka by the APN for remote access. At this

- An architecture that subdivides computing resources and optimally combines them in accordance with the purpose of data processing
- The APN and PEC technologies are used to connect subdivided computing resources


Step 0
Step 1
Step 2
APN
3rd Gen.
NIC
CPU
GPU
FPGA
ASIC SW
4th Gen.
Optical switch
光SW
Computing-resource subdivision unit
Server/storage unit
Board unit
Chip unit
Subdivided-resource-connection method
APN connection
3rd-generation PEC device connection
4th-generation PEC device connection

ASIC: application-specific integrated circuit
FPGA: field-programmable gate array
NIC: network interface card

Fig. 10. DCI.

distance, the network file system would be quite slow and there would be considerable performance degradation. With the APN, however, we were able to achieve connection with almost zero performance degradation even at 100-km distance. Specifically, the performance degraded only by about 0.5%. By optically connecting each CPU and GPU directly using an optical switch, we can carry out LLM training and inference with the minimal and optimized combination of computing resources. We currently use many GPUs to train LLMs, but many of these GPUs are sometimes idle. We aim to implement training with as few computer resources as possible while making all devices work at full capacity.

NTT's future vision for the world of AI is to create an AI constellation. We envision a next-generation AI architecture to more smartly and efficiently address social issues by combining multiple, small, specialized LLMs, instead of creating a single monolithic massive LLM. For example, AI with personas representing a human resources manager, clinical psychologist, truck driver, and elementary school teacher talk about "what is needed to revitalize our shrinking community." They offer their opinions and come up with a consensus on what to do, with the involvement of humans when necessary. We believe we can create a mechanism for building consensus through these interactions.

We formed a business partnership with sakana.ai to conduct joint research to build the AI constellation mentioned above. The venture company sakana.ai is in the spotlight right now. It was founded by well-known AI experts. David Ha was lead researcher at Google Brain and at Stability AI, the company that created Stable Diffusion image generation AI. While Llion Jones was one of the Google developers who created the basic algorithm for transformers, now being used in ChatGPT and other AIs. They established sakana.ai with a base in Japan to conduct R&D of new LLMs and AI constellations. We have entered into a business partnership to work together with them in these areas.

5. Three resolutions of NTT laboratories

Finally, I would like to conclude by talking about the three resolutions of NTT laboratories (**Fig. 11**). "Do research by drawing from the fountain of knowledge and provide specific benefits to society through commercial development." These words, proclaimed in 1950 by Goro Yoshida, the first director of the Electrical Communication Laboratory, embody the vision of NTT laboratories. They point to three elements built on top of each other. "Doing research by drawing from the fountain of knowledge" is the foundation. On top of this is the phase of "commercial development," and, on the topmost is "providing specific benefits to society."

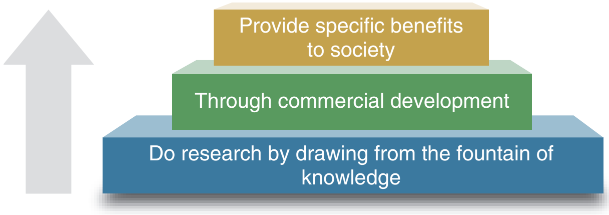

Provide specific benefits to society
Through commercial development
Do research by drawing from the fountain of knowledge

Fig. 11. Three resolutions of NTT laboratories.

Rank	Company	Country/Region	No. of papers
1	IBM	USA	8096
2	Microsoft	USA	7050
3	Siemens	GERMANY	6628
4	Samsung	SOUTH KOREA	6353
5	Google	USA	5926
6	Huawei Technologies	CHINA	5392
7	Intel Corporation	USA	4982
8	SINTEF	NORWAY	4602
9	General Electric	USA	4309
10	Philips	NETHERLANDS	3822
11	**NTT**	**JAPAN**	**3439**
12	Nokia Corporation	FINLAND	3352
13	Bosch	GERMANY	2821
14	Samsung Electronics	SOUTH KOREA	2782
15	China Electronics Technology	CHINA	2727

Source: Survey by NTT using Web of Science, InCites Benchmarking

Fig. 12. Research publications ranking: 11th in the world & 1st in Japan (ICT and engineering companies, 2017–2021).

The most important is the research at the bottom, done “by drawing from the fountain of knowledge.” Not only in AI, as mentioned earlier, but also in all engineering fields, NTT is ranked 11th in the world in the number of publications (**Fig. 12**). In world-class research areas, such as speech recognition, information security, optical communications, and quantum computing, NTT boasts the world’s highest number of publications, beating Google and IBM. We would like to build on these accomplishments and solidify our position as a world leader in research by further aiming for the top and expanding our world-class research areas. This is our first resolution.

Next is the phase of “commercial development” in the middle. IOWN and LLMs are the two key technologies that we would like to robustly develop and put into commercial use as an embodiment of our second resolution.

Last is the social implementation phase of “providing specific benefits to society.” In this regard, we newly established the Research and Development Market Strategy Division in June 2023. Thus far, the laboratories and the Research and Development Planning Department, to which I belong, have been working together in various ways with customers, partner companies, and business companies. Under the new organization, we established the Marketing Planning and Analysis Department, wherein the Research and Development Planning Department will work together with both the Marketing Planning and Analysis Department and Alliance Department to enhance and broaden the scope of our activities. Thus, our third resolution is to implement research and development results into society going forward.