

NTT's LLM "tsuzumi"

Kentaro Shimizu, Kosuke Nishida, and Kyosuke Nishida

Abstract

In November 2023, NTT announced tsuzumi, a large language model (LLM) based on NTT laboratories' over-40 years of accumulated research in natural language processing. The tsuzumi LLM excels in Japanese-language processing ability, and its key features include being lightweight and providing multimodal support, which enables handling of non-text media. Its lightweight model reduces power consumption and makes on-premises use possible. Its multimodal feature enables comprehension of photos and diagrams, distinguishing tsuzumi from other publicly available LLMs. In this article, we give an overview of tsuzumi on the basis of these key features.

Keywords: tsuzumi, lightweight, multimodal

1. NTT's LLM "tsuzumi" expands the application domains of LLMs

A variety of large language models (LLMs), such as OpenAI's ChatGPT, have been released, attracting significant attention. To accumulate a greater amount of knowledge and provide natural answers, many LLMs have exploded in size. It has been reported that a model of the size of GPT-3 requires about 1300 MWh (equivalent to one hour of electricity generated by a nuclear power plant) for training [1]. Because of the prohibitive initial and operational costs, including the hardware required for large-scale training, many companies find it impractical to create their own language models. They thus turn to commercial LLMs provided as cloud services. However, these companies often handle data with personal and sensitive information, so simply uploading such data to the cloud is problematic. A company's data are considered a part of its assets. How to make effective use of such data with LLMs has thus become a challenge.

To address these issues, NTT Human Informatics Laboratories has been engaged in the research and development of tsuzumi, a lightweight LLM with superior Japanese-language processing ability, and released it to the public in November 2023. The significant advantages of tsuzumi's lightweight nature include reduction in hardware resources and power required for using and tuning the LLM. These benefits enable the use of tsuzumi by companies on their

premises, allowing them to apply and effectively use an LLM on data that would be problematic when stored on external cloud servers.

A decrease in the size of an LLM generally leads to a reduction in performance. However, NTT laboratories' over-40 years of experience in natural language processing has allowed tsuzumi to outperform publicly available LLMs with parameters of similar size in Japanese language processing. Furthermore, tsuzumi features adapter tuning to achieve flexible tuning and multimodal support, which enables comprehension of documents that include non-text content such as figures and diagrams. The key features of tsuzumi are explained in detail in the following sections.

2. Lightweight model

The size of an LLM is measured in terms of its parameter size. There are two versions of tsuzumi: a lightweight version with 7 billion (7B) parameters and an ultra-lightweight version with 600 million (0.6B) parameters.

The parameters of an LLM are variables used to store knowledge and skills acquired by the model during training. The greater a model's parameter size, the greater the model's capabilities tend to be in accumulating knowledge and responding to queries from humans. However, increasing parameter size also leads to an increase in the computational resources

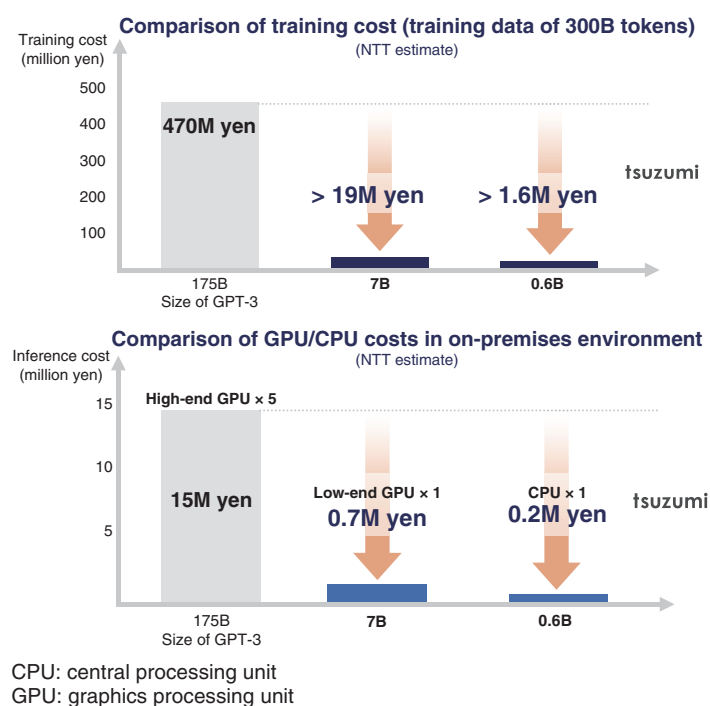


Fig. 1. Relationship between lightweight models and cost reduction.

and power consumption required for training and inference. The technological challenge is thus to reduce an LLM's parameter size while maintaining its performance.

The smaller an LLM's parameter size, the less knowledge it can generally store, leading to reduced performance. For *tsuzumi*, we have thus taken the approach of improving the quality of training data. For example, we seek to remove noise—information that should not be learned, such as redundant or erroneous information—and create training data with information from a wide range of fields instead of a few selective fields. For *tsuzumi*'s ultra-lightweight version, we succeeded in making the model even more lightweight while maintaining its performance by focusing the training on domains and tasks handled by the model.

With this approach, *tsuzumi*'s parameter size was trimmed to 1/25th to 1/300th the size of publicly available massive LLMs, resulting in significant cost reduction (Fig. 1).

3. Superior Japanese-language processing ability

When training an LLM, text included in the training

data is broken down (tokenized) into “tokens.” Tokens are the fundamental units processed by an LLM.

Tokenizers of recent LLMs determine the token set or vocabulary by learning from a collection of texts. LLMs developed outside Japan have a small amount of tokens for the Japanese language included in their vocabulary, and most break down Japanese text into tokens of a single character or byte. Therefore, a great number of tokens are needed to generate Japanese text, which is highly inefficient in terms of text-generation speed. We have thus developed a tokenizer with vocabulary that can efficiently generate Japanese text by training it based on Japanese text. However, the training algorithms of commonly used tokenizers do not take the structure of the Japanese language into account. Thus, there is the issue of the likelihood of redundant tokens, which often appear in the training corpus but not in other texts, being included in the vocabulary.

To address this issue, *tsuzumi* incorporates unique processing (lexical constraint) that takes Japanese words into account when tokenizing. The tokenizers of other companies may include redundant tokens, such as those that appear frequently in Wikipedia but not in other texts, in its segmentation results. In

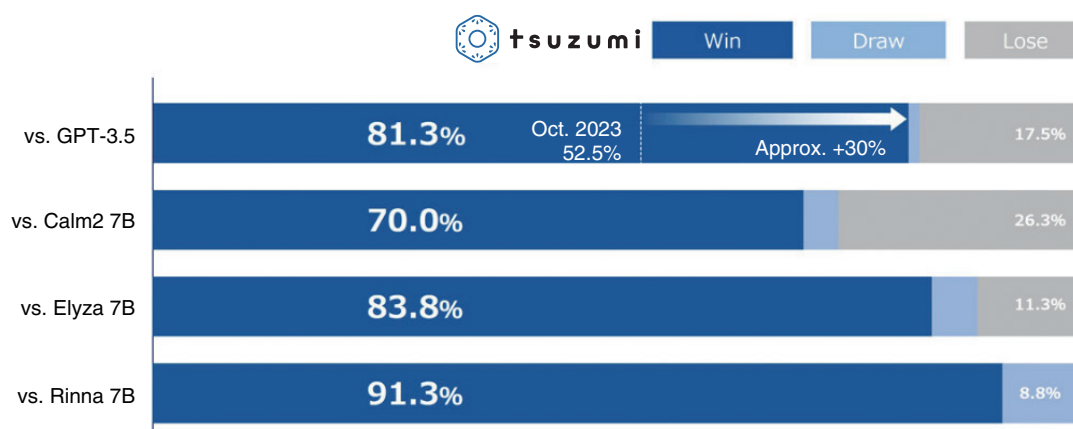


Fig. 2. Comparison of tsuzumi's Japanese-language processing ability against other commercial LLMs.

contrast, tsuzumi uses lexical constraint to achieve segmentation that strongly reflects the structure of the Japanese language. In deploying lexical constraint, NTT applied the results of its long endeavor in the research and development of morphological analysis tools and dictionaries.

Tokenization affects not only the speed of text generation but also the accuracy of text comprehension. To improve the quality of training data described above, tsuzumi's unique tokenization also results in high Japanese-language processing ability. **Figure 2** shows the results of tsuzumi's Japanese-language processing ability compared with other publicly available LLMs, as measured using the Rakuda benchmark. This benchmark consists of a series of questions testing knowledge related to Japan and is often used to compare the performances of Japanese-language LLMs. The same questions were given to LLMs for comparison, and GPT-4 was used to determine a better response of each question between two LLMs. The results indicate that tsuzumi has superior Japanese knowledge and Japanese-language processing ability, not only against LLMs of similar size but also against GPT-3.5.

4. Adapter tuning

The tsuzumi LLM supports adapter tuning. This tuning adds (adapts) a small domain-specific model to the original model. **Figure 3** provides an image of adapter tuning.

Other tuning methods include prompt engineering, which provides relevant information at the same time as the query during inference, and full fine-tuning,

which updates the entire model. **Figure 4** provides a summary of each method's features. Each tuning method has its own characteristics in terms of cost and accuracy. Adapter tuning has the advantage of achieving both at a certain level. When considering use cases within different companies, the deployment of different adapters for each company makes it possible to easily create company-specific models.

We are working on providing support for multiple adapters, as shown in **Fig. 5**. This will allow the requirements of a variety of use cases to be considered by connecting multiple adapters on top of tsuzumi.

5. Multimodal support

As its name implies, an LLM is a model for processing language, so it does not take graphical and audio data as its input or output. However, figures and tables are often included in general documents; such non-text content often has significant meaning. The tsuzumi LLM was designed to be a new type of LLM that comprehends figures and tables [2, 3] as well as speech and user situations (**Fig. 6**). Such use cases were exhibited at the 2023 NTT R&D Forum.

The feature article "NTT's LLM 'tsuzumi': Capable of Comprehending Graphical Documents" [4] in this issue explains tsuzumi's techniques for visual comprehension of graphical documents, one of its multimodal capabilities. Please read it in addition to another feature article "Commercialization of NTT's LLM 'tsuzumi'" [5], which describes the usage domain and usage applications, solution menu, and other features of tsuzumi commercial services

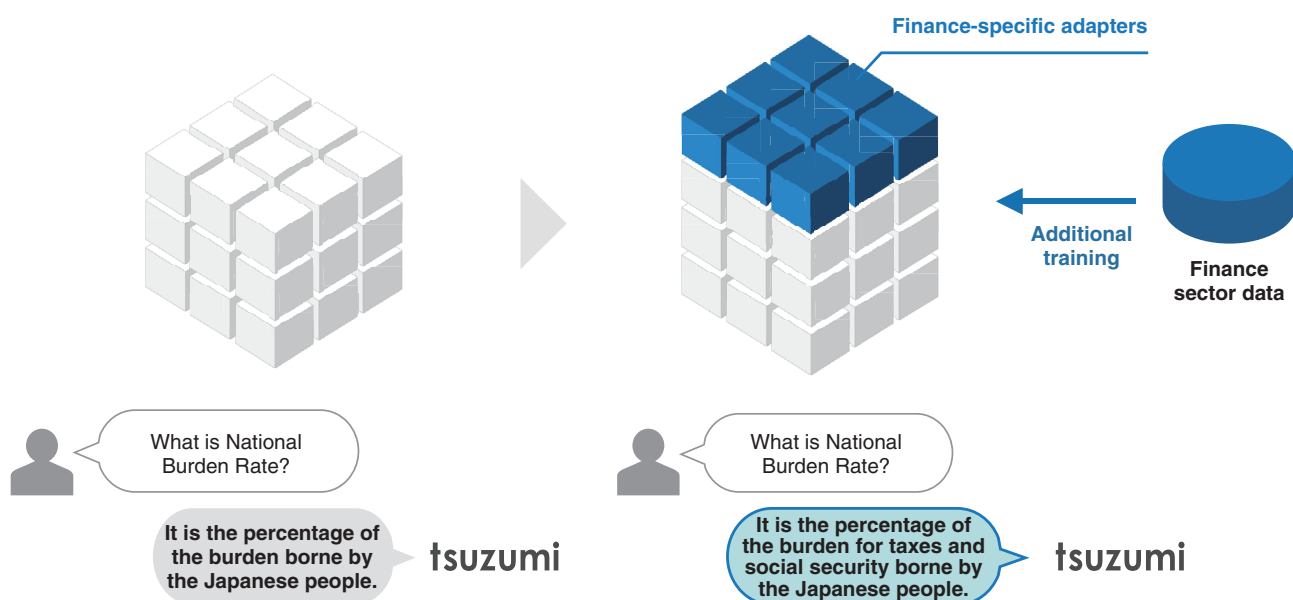


Fig. 3. Image of adapter tuning.

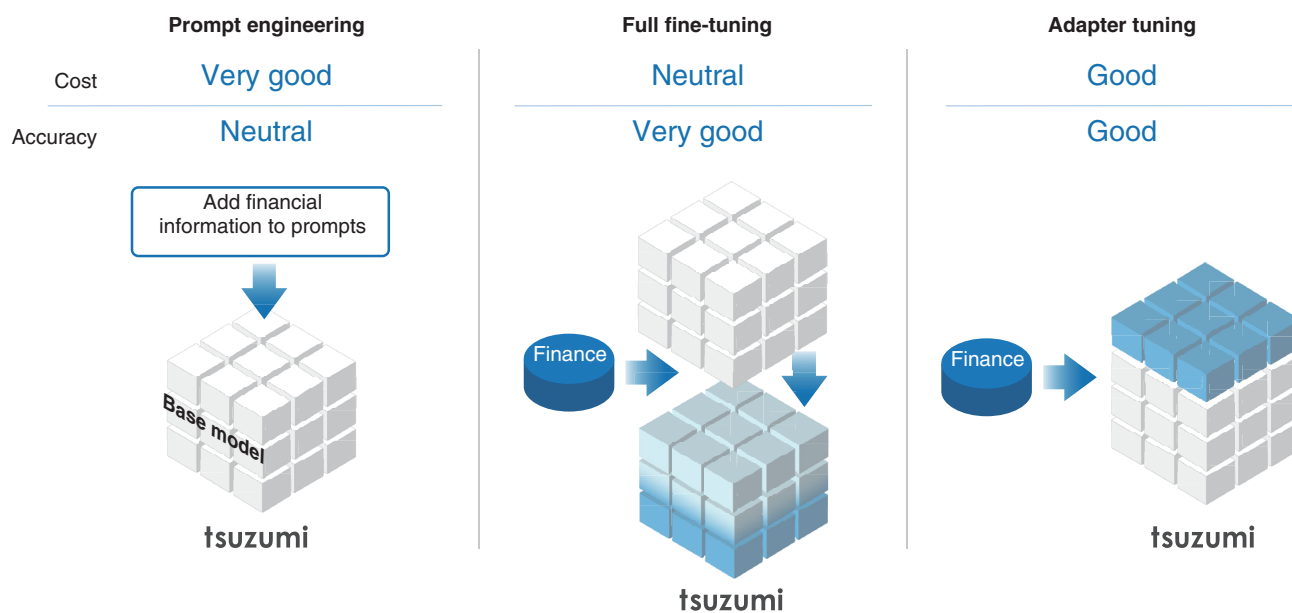


Fig. 4. Comparison of tuning methods.

launched in March 2024.

6. Future developments

As we have described, **tsuzumi** boasts key features, such as being lightweight while possessing superior

Japanese-language processing ability and flexibility in tuning, as well as multimodal support. We are continuing research and development of **tsuzumi** in the following areas:

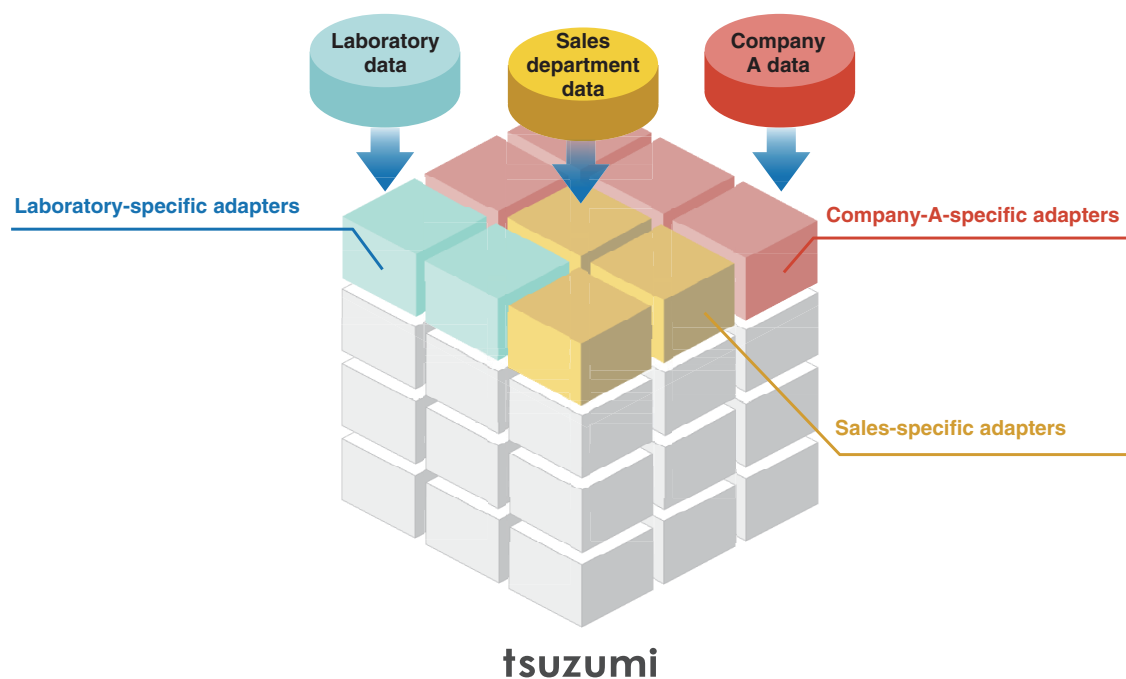
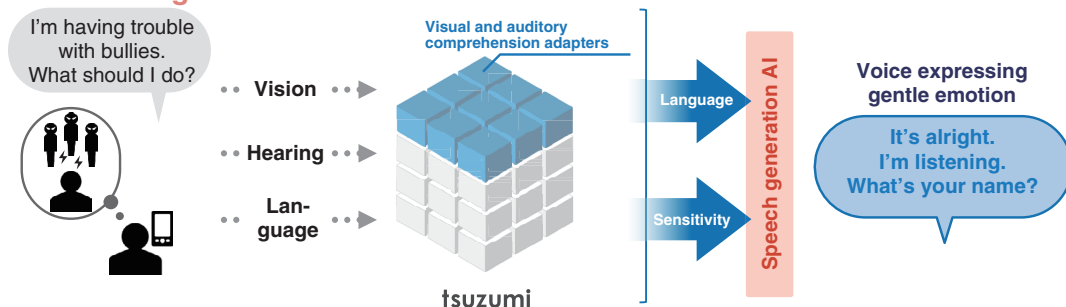


Fig. 5. Multiple adapters: Enable switching and combining adapters in accordance with the user or use case.

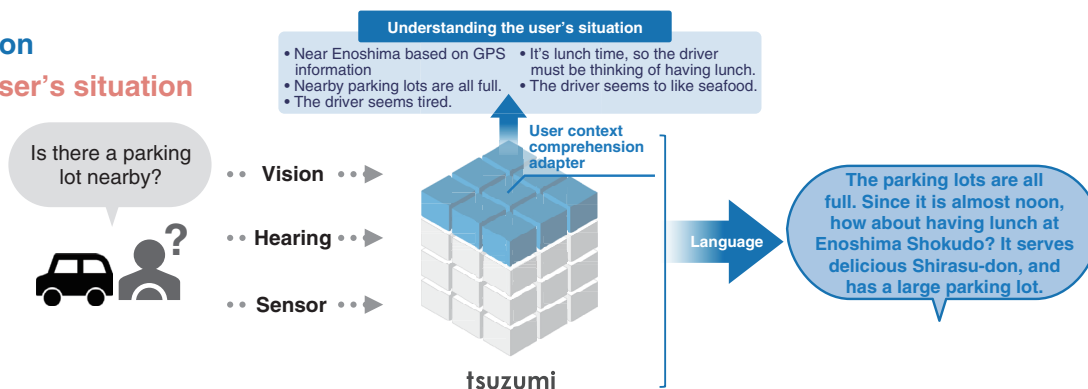
Modal extension

Language + Vision + Hearing



Modal extension

Language + User's situation



GPS: Global Positioning System
AI: artificial intelligence

Fig. 6. Image of modal support.

6.1 Multilingual support

Currently, tsuzumi supports both English and Japanese. We seek to further improve its language processing of these two languages while also working to provide support for other languages (e.g., Chinese, Korean, French, and German) to expand its user base worldwide.

We are also working on adding support for not only human languages but also programming languages. For example, if tsuzumi can comprehend the content of a specification document and output source code in the specified programming language in accordance with the specification, it will contribute significantly to reducing software-development workload. Currently, tsuzumi can output source code to a certain extent. We are working to improve its performance in this area by collecting more training data and implementing more training.

6.2 Medium-sized version

In the announcement of tsuzumi's release in November 2023, it was revealed that in addition to ultra-lightweight and lightweight versions, a medium-sized version is also scheduled for release. This version will have 13B parameters, roughly double the parameter size of the lightweight version (7B parameters). In addition to multilingual support, it will be able to store a greater amount of knowledge and improve its performance as an LLM.

The medium-sized version's greater parameter size means more hardware resources are required compared with the lightweight version. However, we seek to reduce the required resources by investigating a technique (quantization) that enables LLMs to operate efficiently with limited hardware resources.

6.3 Support for safety and security

There is much debate on the safety and ethical

aspects of generative artificial intelligence, including LLMs. From the institutional viewpoint, lawsuits have been filed in other countries over the inclusion of copyrighted works and personal information in training data and the proper solutions. In Japan, these issues are being actively discussed, but there are still no clear conclusions. From the technological viewpoint, systems that robustly reject inappropriate questions, such as those that violate human rights, are needed.

To tackle these issues, we will not only prepare higher quality training data but also collaborate with other research institutes and organizations studying the institutional and technical aspects of these issues. At the same time, we will continue to investigate how to provide tsuzumi users with greater safety and security.

We will continue to expand the capabilities of tsuzumi. The best is yet to come.

References

- [1] M. DeGeurin, "Move Aside, Crypto. AI Could Be the Next Climate Disaster," Gizmodo, Apr. 2023. <https://gizmodo.com/chatgpt-ai-openai-carbon-emissions-stanford-report-1850288635>
- [2] R. Tanaka, T. Iki, K. Nishida, K. Saito, and J. Suzuki, "InstructDoc: A Dataset for Zero-shot Generalization of Visual Document Understanding with Instructions," Proc. of the 38th Annual AAAI Conference on Artificial Intelligence (AAAI-24), Vancouver, Canada, Feb. 2024.
- [3] T. Hasegawa, K. Nishida, K. Maeda, and K. Saito, "DueT: Image-Text Contrastive Transfer Learning with Dual-adaptor Tuning," Proc. of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023), pp. 13607–13624, Singapore, Dec. 2023. <https://doi.org/10.18653/v1/2023.emnlp-main.839>
- [4] R. Tanaka, T. Iki, T. Hasegawa, and K. Nishida, "NTT's LLM 'tsuzumi': Capable of Comprehending Graphical Documents," NTT Technical Review, Vol. 22, No. 8, pp. 26–30, Aug. 2024. <https://ntt-review.jp/archive/ntttechnical.php?contents=ntr202408fa2.html>
- [5] N. Shibuya, I. Kudo, M. Kaido, and S. Yamamoto, "Commercialization of NTT's LLM 'tsuzumi,'" NTT Technical Review, Vol. 22, No. 8, pp. 31–37, Aug. 2024. <https://ntt-review.jp/archive/ntttechnical.php?contents=ntr202408fa3.html>



Kentaro Shimizu

Senior Research Engineer, Human Insight Laboratory, NTT Human Informatics Laboratories.

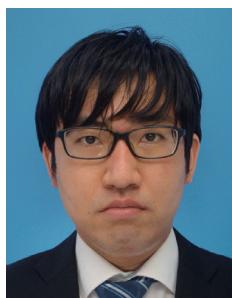
He received a B.E. and M.E. in engineering from Yokohama National University in 1997 and 1999 and Ph.D. in medicine from Tokyo Women's Medical University in 2009. He joined NTT Cyber Space Laboratories in 1999 and studied image processing. His current interests are LLMs and machine learning. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE), Information Processing Society of Japan (IPSJ), the Japanese Society for Artificial Intelligence (JSAI), and the Japan Society of Computer Aided Surgery (JSCAS).



Kyosuke Nishida

Senior Distinguished Researcher, Human Insight Laboratory, NTT Human Informatics Laboratories.

He received a B.E., M.I.S., and Ph.D. in information science and technology from Hokkaido University in 2004, 2006, and 2008. He joined NTT in 2009. His current interests include natural language processing, vision-and-language, and artificial intelligence. He is currently serving as the technical lead for NTT's LLMs. He received the Paper Award from the Association of Natural Language Processing (NLP) from 2018 to 2024. He is a member of the Association for Computing Machinery (ACM), NLP, IPSJ, and the Database Society of Japan (DBSJ).



Kosuke Nishida

Associate Distinguished Researcher, Human Insight Laboratory, NTT Human Informatics Laboratories.

He received a B.E. in engineering and M.E. and Ph.D. in information science and engineering from the University of Tokyo in 2015, 2017, and 2024. He joined NTT Media Intelligence Laboratories in 2017 and studied natural language processing. His current research interests are especially mathematical understandings of models and training algorithms for natural language processing. He is a member of the Association for Computational Linguistics (ACL).