Front-line Researchers

NTT's Large Language Model "tsuzumi": A High-performance and Low-energy-consumption Large Language Model with Expertise in Specific Fields

Kyosuke Nishida Senior Distinguished Researcher, NTT Human Informatics Laboratories

Abstract

Since ChatGPT, OpenAI's generative artificial intelligence (AI) chatbot, was released in November 2022, large language models (LLMs) have rapidly drawn worldwide attention, and new LLMs have been announced one after another. While the "bright side" of such LLMs, namely, high performance, has been brought to attention, their "dark side," namely, increased energy consumption, has also been pointed out. On November 1, 2023, NTT's LLM called "tsuzumi" was announced. It was developed with a particular vision in mind, creating a future in which many and



small LLMs having different characteristics are combined, rather than creating one huge and monolithic LLM, to achieve social well-being through AI-human collaboration. Thus, tsuzumi has made it possible to achieve the bright side and address the dark side of LLMs simultaneously, and its commercial service was launched on March 25, 2024. We interviewed Kyosuke Nishida, a senior distinguished researcher at NTT Human Informatics Laboratories, about how the tsuzumi project was launched, features of tsuzumi, trends in LLMs, and his thoughts on the challenges of competing in a highly competitive market.

Keywords: large language model, tsuzumi, natural language processing

NTT's large language model "tsuzumi" was developed—from announcement to commercial launch—in five months

-Could you tell us about the research you are currently conducting?

I am researching large language models (LLMs), including NTT's tsuzumi, which was announced in November 2023. Since joining NTT in 2009, including a stint at an operating company from 2013 to 2015, I have been involved in various research and development activities. In the field of natural language processing, which I began researching around 2017, I have been focusing on themes such as machine reading of natural language text by artificial intelligence (AI) models and visual machine reading, which combines visual information with textual information to help AI models understand documents in the same way humans do when reading documents. My research colleagues and I received Paper Awards from 2018 to 2024 (Best Paper Awards in 2018 and 2021) at the Annual Meeting of the Association of Natural Language Processing and were the runner-up in the Document Visual Question Answering competition at the International Conference on Document Analysis and Recognition in 2021.

After Google announced their Bidirectional Encoder Representations from Transformers (BERT) language model in 2018 and NTT began constructing an NTT version of a Japanese BERT model, I started researching language models. In parallel with the release of ChatGPT, OpenAI's generative-AI chatbot, in November 2022 and LLMs gaining worldwide attention, I sensed that a global paradigm shift toward the emergence of early versions of general-purpose AI that had never existed was underway. I therefore emphasized the necessity and importance of the development of LLMs at NTT and launched a project on LLMs in February 2023. Since then, I have been investigating LLMs as my main theme.

In this project, I procured computing resources, collected the data necessary for training, and built an LLM with team members. The development of our LLM was completed in a very short time; pre-training of the LLM started around June 2023, tsuzumi was announced on November 1, 2023, and the start of its commercial service was announced on March 25, 2024.

-Could you explain what an LLM is?

A language model is a machine-learning model that models the likelihood of generating a sequence of words (tokens) and predicts the probability of future (or missing) tokens. An LLM is based on a deeplearning model in the form of a neural network called a transformer and enables a variety of language-processing tasks such as highly accurate information retrieval and generation and modification of computer programs. An LLM's capability of advanced language processing is primarily achieved through two processes: pre-training and instruction tuning. Pre-training is the process of pre-training a language model, i.e., the transformer is pre-trained with a large corpus of text to create a pre-trained language model. Since its emergence in 2018, BERT has been gaining attention as an effective language model for natural language processing. Since the release of GPT-3 in 2020, the capability of pre-trained language models has advanced to the point that these models can handle any task to some extent-without fine-tuningby having them generate text following task-defining prompts.

Instruction tuning is the process of explicitly learning pairs of instructions and responses for various tasks to further develop the ability acquired through pre-training. In conventional tuning using task data, for example, two sentences are input to BERT as tokens called "separators" to produce a score with three labels, "0" (yes), "1" (no), or "2" (possible), which is used to determine the relationship between the two input tokens. This process is data-driven learning, i.e., the language model is trained in a datadriven manner with no knowledge of what tasks it is given. In instruction tuning, however, a task is defined in language. Specifically, the task that the LLM is to solve or the output the LLM should produce in response to the input text (instruction) is defined in language so that the LLM can respond to those instructions (Fig. 1). This process-instruction tuning-increases the versatility of an LLM and enables it to respond to unknown tasks without learning.

Current LLMs are becoming larger, and following scaling laws^{*}, language models with a large model size (number of parameters) are being created; in fact, models with over a trillion (1000 billion) parameters

^{*} Scaling laws: Empirical laws stating that as model size (number of parameters), dataset size, and the amount of computation used to train a natural language-processing model increase, loss decreases according to a power law.







Source: https://lifearchitect.ai/models/

Fig. 2. Increasing the scale of LLMs.

are under development (**Fig. 2**). For example, GPT-4 is said to have 1.76 trillion parameters.

As this trend progresses, however, energy consumption for using and training LLMs increases exponentially, and costs increase accordingly. It has been estimated that training an LLM such as GPT-3, which is said to have 175 billion parameters, uses about 1300 MW/h of electricity per training session, which is equivalent to an hour's worth of electricity produced by one nuclear power plant.

	() tsuzumi
Japanese musical instrument (hand drum)	Strong in Japanese language
Small	Small and lightweight
Easy tuning with tuning strings	Flexible tuning
Beautiful appearance, sound, and the way it is played	Multimodality

Fig. 3. Japanese hand drum and tsuzumi.

Japanese hand drum and tsuzumi are both associated with the keywords "Japanese language," "compact and lightweight," "flexible tuning," and "multimodality"

—What kind of LLM is tsuzumi?

While the global trend is toward larger-scale, general-purpose LLMs that can do anything, their power consumption is becoming problematic. We therefore adopted the vision of creating a future in which many and small LLMs having different characteristics are combined, rather than creating one large-scale and monolithic LLM, to achieve social well-being through AI-human collaboration. With that vision in mind, we decided to develop a high-performance LLM with specialized knowledge in a specific field by improving the quality and quantity of language training data, rather than competing on model size. We also intended to pursue the concept of "AI constellation," in which multiple small and high-performing LLMs, such as LLMs that are strong in the medical and educational fields, are combined to form a superior LLM. We named our LLM "tsuzumi" after a traditional Japanese hand drum (Fig. 3), because our LLM has the following features: strength in the Japanese language, high performance despite small and lightweight, ease and flexibility of tuning, extendibility to multimodal applications (multimodality).

I will present the evaluation results of tsuzumi based on the Rakuda benchmark, which compares the output of two models by using an evaluation set of 40 questions covering 4 categories, i.e., Japanese geography, politics, history, and society. This benchmark uses GPT-4 as the evaluator instead of a human evaluator. In the evaluation example shown in **Fig. 4**, GPT-4 compares the output of tsuzumi with that of GPT-3.5 and rated tsuzumi highly in terms of specificity and detail. According to the overall evaluation using this benchmark, tsuzumi beat GPT-3.5 and other Japanese LLMs at an overwhelmingly high win rate, demonstrating tsuzumi's high-level Japanese language comprehension and generation abilities (**Fig. 5**).

For pre-training tsuzumi, we prepared a total of over 1000 billion tokens in a wide range of domains from technical documents to entertainment. We will continue to improve the quality and quantity of training data. It is also important to consider the ratio of languages contained in training data during pretraining, and it is not yet clear what kind of language balance should be taken in pre-training to build an LLM that is strong in the Japanese language with a limited corpus of Japanese. Another company has recently adopted a method of additionally training non-Japanese language LLMs on Japanese language and performed well in Japanese language processing. We pre-trained tsuzumi by using in-house data from scratch and added codes and multilingual data, mainly in Japanese and English. We are continuing to verify this approach from various perspectives, particularly the effect of adding multiple languages to pre-training corpus.

When creating a pre-training corpus, we also focus on tokenization, which breaks down sentences into tokens, and preprocessing. Tokenization is done by taking into consideration Japanese words. If word constraints are not taken into account, unnatural and redundant tokens can easily be generated due to bias

Question

Please give examples of the social and political changes that occurred during the transition from the Heisei era to the Reiwa era and explain their significance.

Answer by tsuzumi-7B

The transition from the Heisei era to the Reiwa era has seen many social and political changes. Below are some of the main changes.

1. Spread of the Internet and digital technology: During this period, information and communication devices such as smartphones and personal computers spread rapidly, which has greatly affected people's lifestyles. It is especially noteworthy that the number of users of social networking services (SNSs) has increased, which has made it easy to communicate among individuals. Moreover, cloud computing and AI have advanced and they are now being used in the business world (...)

Answer by GPT-3.5

Several social and political changes occurred during the transition between the Heisei and Reiwa eras. Some of them are described below.

(1) Political change: Compared to the Heisei era, the Reiwa era has seen continued political stability, and the current administration is supported by the younger generation. This state of affairs indicates that political values are changing from generation to generation. Compared to the Heisei era, the Reiwa era has seen an increasing population inflow from overseas, and policies to accept foreign human resources are being promoted. This trend is important in terms of helping to cope with globalization and Japan's declining population. (...)

Judgment by GPT-4: usefulness, relevance, accuracy, detail, Japanese-language ability

The answer of Assistant 1 (tsuzumi) provides a detailed explanation of five specific social and political changes that occurred during the transition from the Heisei to Reiwa eras. The significance and impact of each change is specified, so the information presented is highly useful and relevant. In addition, the Japanese expressions are appropriate, and the text is easy to understand. (...) Both tsuzumi-7B and GPT-3.5 understood the intent of the question and provided appropriate answers; however, I judge Assistant 1's answer to be superior because it provided more specific and detailed information. My evaluation is therefore higher for Assistant 1.





Improved world-class Japanese language ability of tsuzumi and increased win rate against GPT-3.5 by about 30%

(40 questions × 2 presentation orders) as an evaluator.

Fig. 5. Comparison of Japanese language LLMs based on the Rakuda benchmark.

in the training corpus. Our tokenizer incorporates the results of NTT's research on word segmentation that has been conducted for many years, achieving the word segmentation that is close to that of natural language. To improve the quality of the text that is generated as much as possible, we use various methods to remove noisy text. Regarding instruction tuning, we have compiled our own instruction-tuning data covering a wide range of categories of questions and instructions to use in training and are currently maintaining and expanding the data from the perspective of both usefulness and safety.

—Now that tsuzumi has been launched as a commercial service, what are your plans for the future?

Conventional generative AI models may not perform well in fields where technical terms and industry-specific expressions are often used, such as medical care and software development. Since tsuzumi can be customized for industry-specific data, it can expand the application areas of AI. To improve customer experience, it is essential to enable AI to read and comprehend visual information such as charts and graphs in manuals, etc. and personalize AI services by updating customer information. By providing world-class Japanese-language processing capability as well as visual information reading comprehension capability, tsuzumi will support the evolution of customer-support areas such as contact centers and consultation chatbots.

In the medical field, although electronic medical records have been introduced in Japan, it has been difficult to collect and analyze medical-record data because the way of writing medical records depends on the hospital and doctor even for the same symptoms. By ultra-lightweight, flexible, and secure processing of medical-record data, tsuzumi reads and comprehends medical-record data recorded by doctors, arranges the data sets in a common format with appropriate expressions, and puts them in a state ready for analysis.

NTT DOCOMO's contact centers receive more than 40 million inquiries a year from customers. These centers currently use communicators (human operators) and chatbots to answer these inquiries; however, many types of manuals are used and new information is frequently added to them, so it takes much time and effort to create appropriate responses and lists of questions and answers for the chatbots. I believe that introducing tsuzumi into this process will enable AI to understand the content of calls and manuals with visual information, understand customer requests, and provide the most-appropriate answers.

We announced the commercial launch of tsuzumi on March 25, 2024, but we had just created a basic model in the research phase. We are currently working with engineers across the NTT Group to refine that model to provide value to our customers. While supporting that work, I want to improve the versatility of tsuzumi to achieve AI constellation—a nextgeneration AI architecture in which multiple, small, and specialized LLMs are combined, rather than creating one huge, monolithic LLM-and advance research toward a world in which humans and AI can coexist naturally. Although LLMs have made remarkable progress, they still have many shortcomings that must be addressed before they can be implemented in society and assist people in all kinds of environments. It is particularly important to ensure that AI has an input-output interface comparable to that of humans and can tackle tasks that require multimodality and physicality. Our main focus is currently on the combined understanding of vision and language; however, I want to investigate how to link hearing, force, touch, and biological signals, such as brain waves, to language models.

Taking on the challenge with the awareness of what is important now and what should be done naturally now

—What do you keep in mind as a researcher?

When pursuing my research, I want to do something that involves both academia and business, something important that everyone naturally thinks they should do at that moment in time. I often talk about this to my team members. Around 2017, I worked on the theme of "reading tables," which involves computers automatically reading tabular data as humans can, and that technology was used in a solution for contact centers. Then, around 2018, I worked on machine reading and summarization to understand and generate natural language, and those technologies were offered as part of an AI service provided by NTT Communications called COTO-HA[®]. Since then, I have moved in the direction of enabling computers to understand human language under conditions more similar to those of humans. such as comprehension of visual documents, and have approached themes that can create value according to the technological level of the time. LLMs are a highly competitive field and a great challenge for me, but I felt that NTT should definitely develop an LLM, so I launched a project in February 2023.

While I was investigating machine reading comprehension as my research theme, I decided to participate in the MS MARCO challenge, an international competition for machine reading comprehension hosted by Microsoft. Until then, I had no experience competing with many other researchers on a major topic, but our model initially took first place, and the letters "NTT" appeared on top of the leaderboard. Although I was delighted, I was also impressed that the people around me were also very happy. This experience made us believe that we could compete on a global stage, and we went on to participate in several competitions and ranked top of their leaderboards. It took courage to take on LLMs as a research theme because they are attracting much attention and evolving fast, but I feel that the successful experience at MS MARCO has pushed me forward. The business world has high expectations for LLMs, and as an academic topic, LLMs have much potential, from analyzing the mechanisms of LLMs to proposing new models, so I find them very rewarding to investigate.

—What is your message to junior researchers?

Mainstream research themes are "red ocean" fields in which many researchers are competing aggressively, and some people may avoid those fields because it is difficult to achieve results in those competitive environments. However, it can also be said that LLMs is the field that many researchers focus on because it is currently the most-important research theme. When you research LLMs, you can make new discoveries and achieve results, and I think there are many areas in which you can compete, so I encourage young researchers to take on LLMs. When taking on a new theme that is not mainstream, it is a good idea to be aware of whether the theme is valuable enough for other researchers to follow you. For example, when you are young, you might work on themes that you think they are important, even if they are somewhat incremental, then once you have gained more research ability, you might jump into mainstream themes in red oceans. In any case, I believe it is important to take on the challenge of important and valuable themes.

Many researchers are investigating LLMs, making it truly a red-ocean field. However, it is a new and rapidly changing field, so it is not impossible for us to keep up with it, and I believe we have ample opportunity to compete at the forefront. We are working on the LLM project in cooperation with many people, and we expect to achieve good results.

I think it takes courage to change your research field or theme, but from my own experience of working in various fields and on themes other than natural language processing, I am glad that I was flexible and tried different things. If you challenge yourself to aim at what is important now and what you should naturally do now, you will feel relatively little resistance in changing fields and themes.

■ Interviewee profile

Kyosuke Nishida 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, visionand-language, and AI. He is currently serving as the technical lead for NTT's LLMs. He received the Paper Award from the Association of Natural Language Processing from 2018 to 2024. He is a member of the Association for Computing Machinery, Association of Natural Language Processing, Information Processing Society of Japan, and the Database Society of Japan.