Overview of NTT's LLM tsuzumi

November 1st 2023

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

Overview of NTT's LLM

1 Features of tsuzumi

2 tsuzumi and IOWN

3 Product Lineup

tsuzumi

Feature 1. Lightweight

The challenge of increasing the scale of LLMs

Sustainability

Training in the scale of GPT-3 (175B) requires a massive amount of energy.

Ca. 1300 MWh[1] per training session

Ca. 1000 MWh from one nuclear power plant

[1] https://gizmodo.com/chatgpt-ai-openai-carbon-emissions-stanford-report-1850288635

Strategy for tsuzumi

Pursued direction	A massive LLM that knows everything	**Small LLMs with specialized knowledge**
Approach	Increase parameter size	**Improve the quality and quantity of language training data**

tsuzumi

Two types of lightweight tsuzumi LLMs have been developed:

Ultralight version

tsuzumi-0.6B

Ca. **1/300**
of GPT-3 (175B)

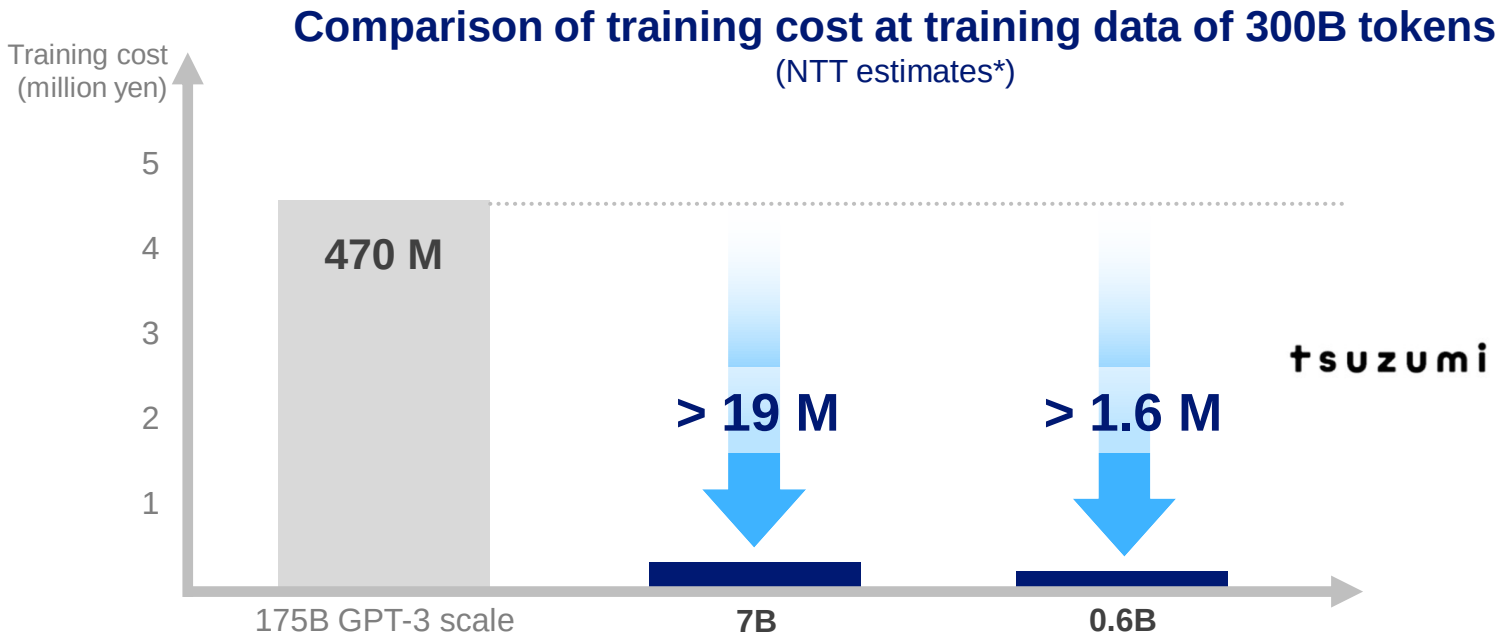
Light version

tsuzumi-7B

1/25
of GPT-3 (175B)

Benefits of reduced weight (1) Training cost

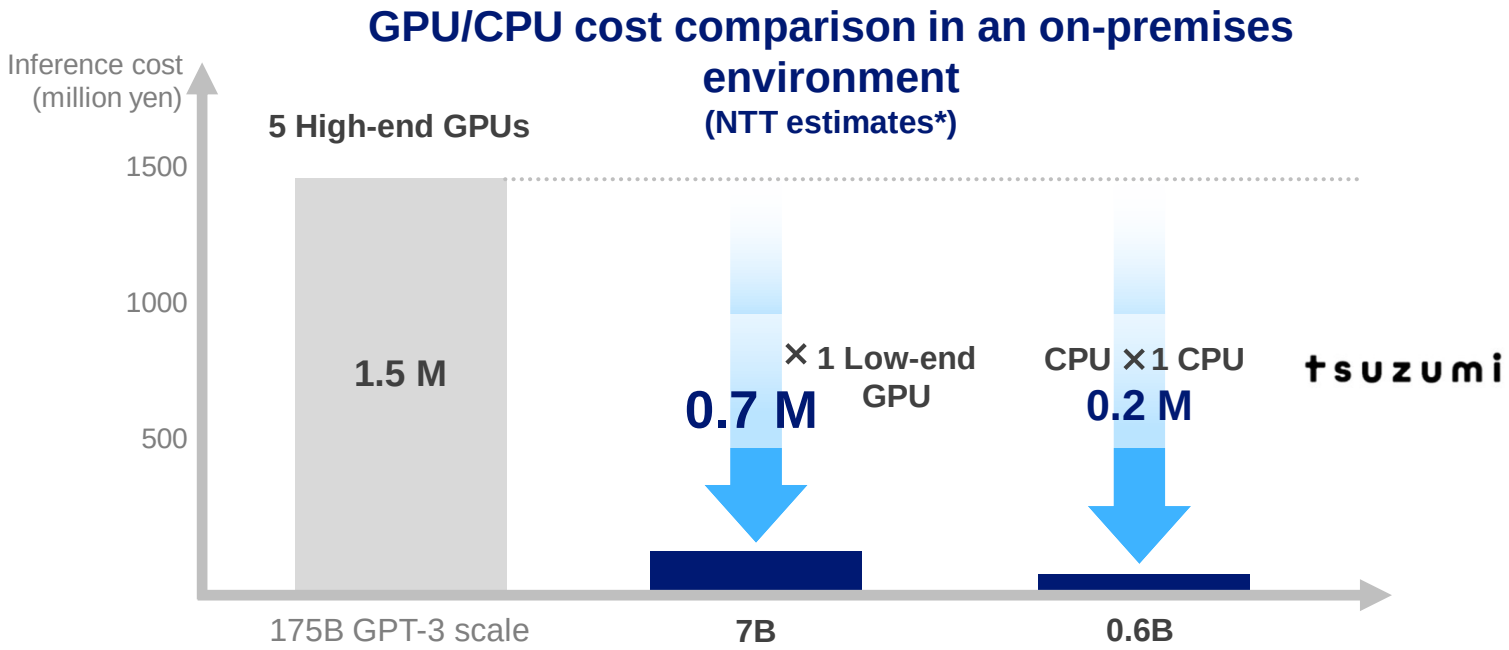
Compared with a GPT-3-scale LLM, training cost can be reduced by about 250 to 300 times.



*Calculation conditions
- The required GPU-hours for each LLM was calculated from the ratio of parameters and the ratio of tokens based on 82,432 GPU-hours for training with Llama-1 7B.
- The training cost was calculated from the calculated GPU-hours and AWS GPU cloud fee.
- AWS GPU cloud fee: A100-80GB 1 node (8 GPUs) assumed to cost approximately 0.14 million yen/day
- Normally, when the parameter size is small, about 2 to 3 times the training data is required to improve the accuracy. The cost also becomes proportionately higher to this.

Benefits of reduced weight (2) Inference cost

Compared with a GPT-3 class LLM, inference cost can be reduced by about 20 to 70 times.



*Calculation conditions
- Quantization: 16 bits
- Required GPU memory size: Number of parameters x quantization size/ 8 bit (350 GB for 175B, 14 GB for 7B, 2.4 GB for 0.6B)
- Hardware cost was converted based on: high-end GPU A100 80 GB: 3 M yen/unit, low-end GPU A10 24 GB: 0.7 M yen/unit, and CPU PC: 0.2 M yen/unit; excluding other operating costs.

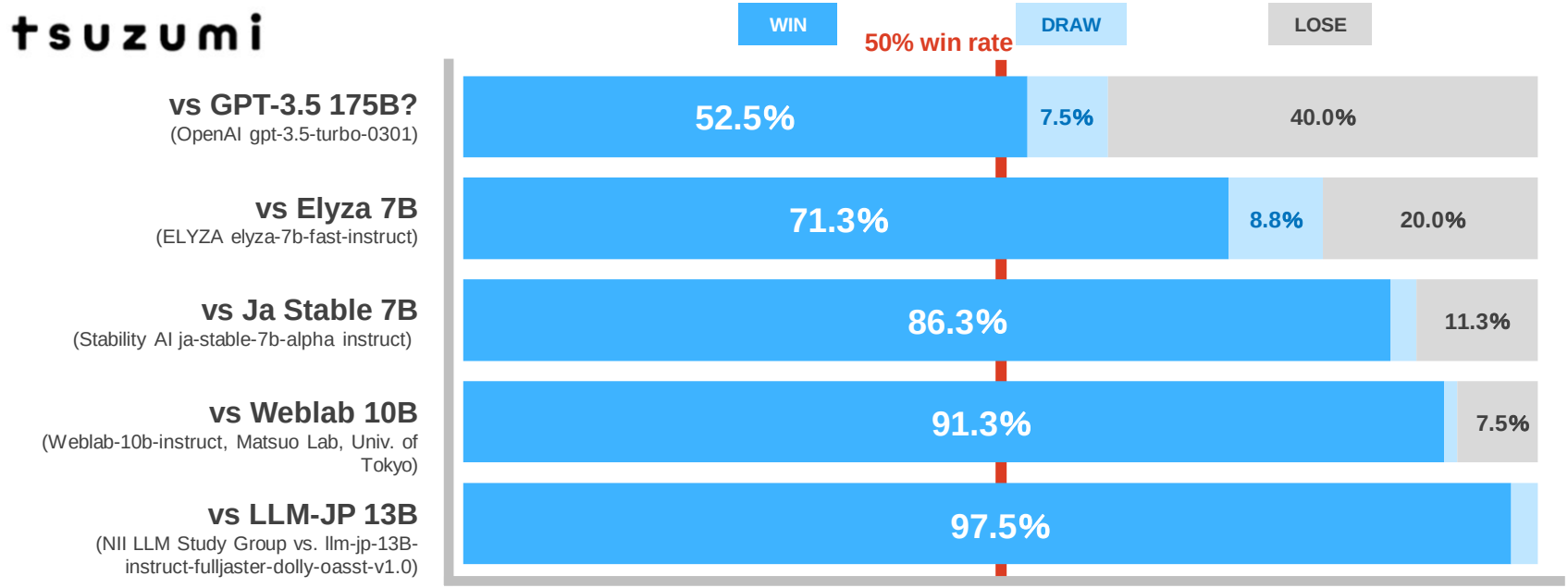
tsuzumi

Feature 2. Proficiency in Japanese

Japanese Proficiency Comparison: Rakuda Benchmark

tsuzumi-7B achieved world-class performance
Surpassed the large-scale GPT-3.5 and significantly outperformed domestic LLMs of the same class

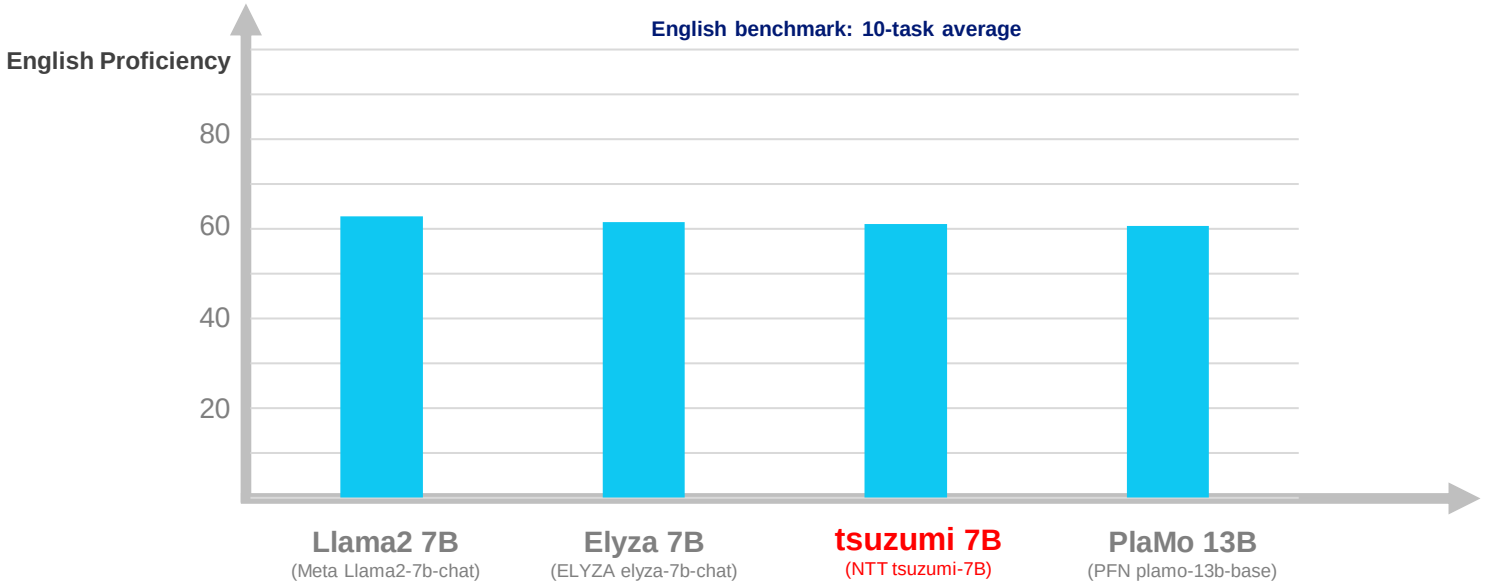
tsuzumi



*Rakuda benchmark: https://yuzuai.jp/benchmark (Oct. 22, 2023)
Forty questions on Japanese geography, politics, history, and society; Scoring based on two-model comparison by GPT-4 (40 questions x 2 orderings);
Except for llm-jp, model output uploaded on the site were used in the evaluation; llm-jp was based on Huggingface model card description settings;
Input repetition and termination tokens were excluded by post-processing

English Proficiency Comparison: lm-evaluation-harness

Achieved the same level of performance as the world's top class LLM, Llama 7B, in same-size comparison mainly in English



*Results for base model for PLaMo only

*1: Evaluation method
lm-evaluation-harness: https://github.com/EleutherAI/lm-evaluation-harness
Average score for 10 English tasks (common sense reasoning field) on the model card of rinna/bilingual-GPT-neox-4b
Metrics were acc and acc_norm (if both exist,acc_norm was preferred)
chat/instruct model was used for Llama2-7b, elyza-7b, and tsuzumi-7b. Result for plamo-13b is that of the base model.

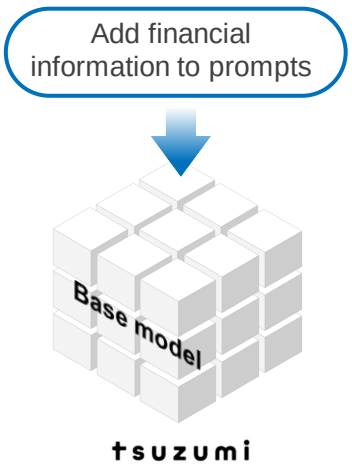
tsuzumi

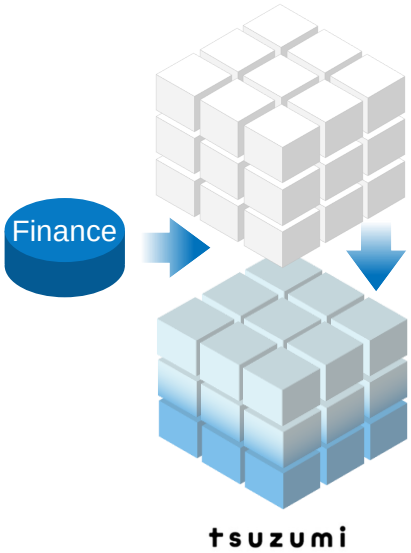
Feature 3. Flexible customization

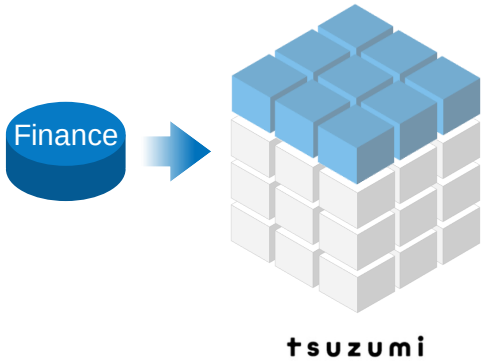
Different Tuning Methods

Three tuning methods are provided to flexibly respond to different requirements such as accuracy and cost.

	Prompt engineering	Full fine-tuning	Adapter tuning
Cost	◎	△	○
Accuracy	△	◎	○

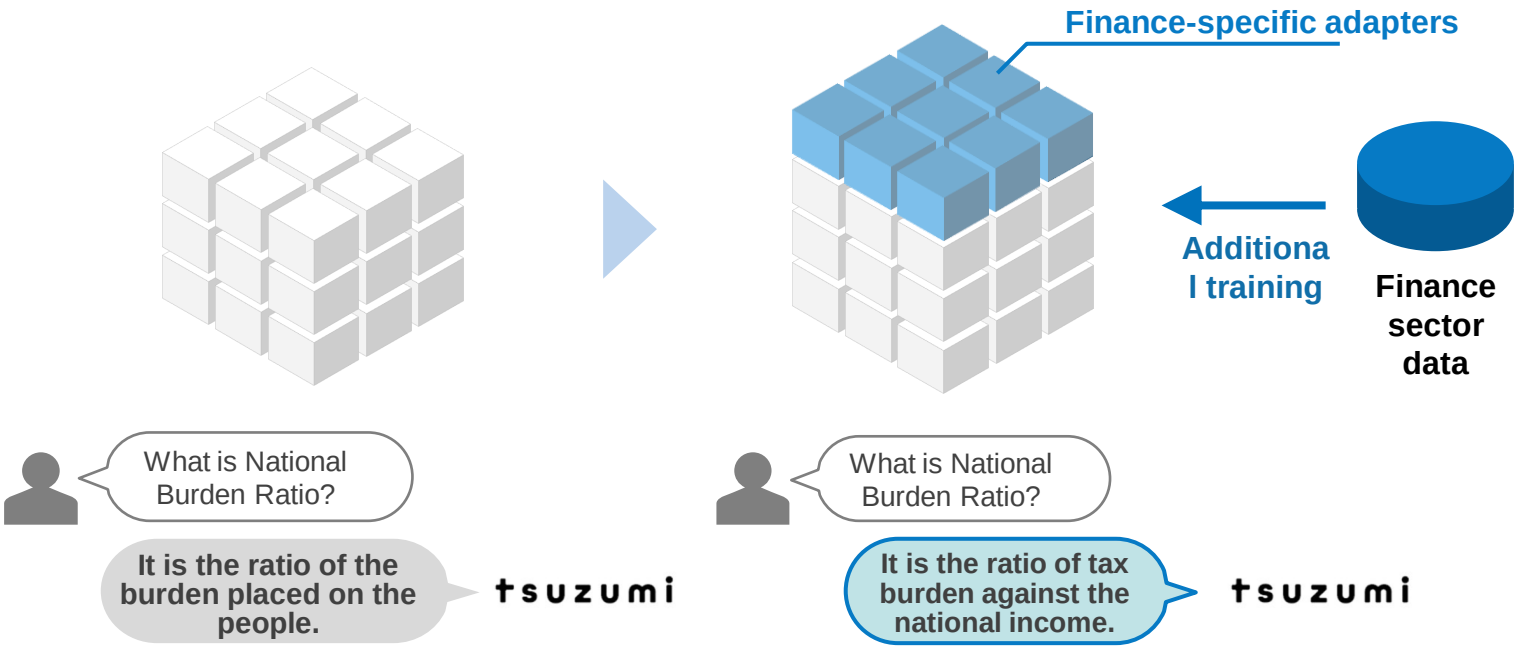






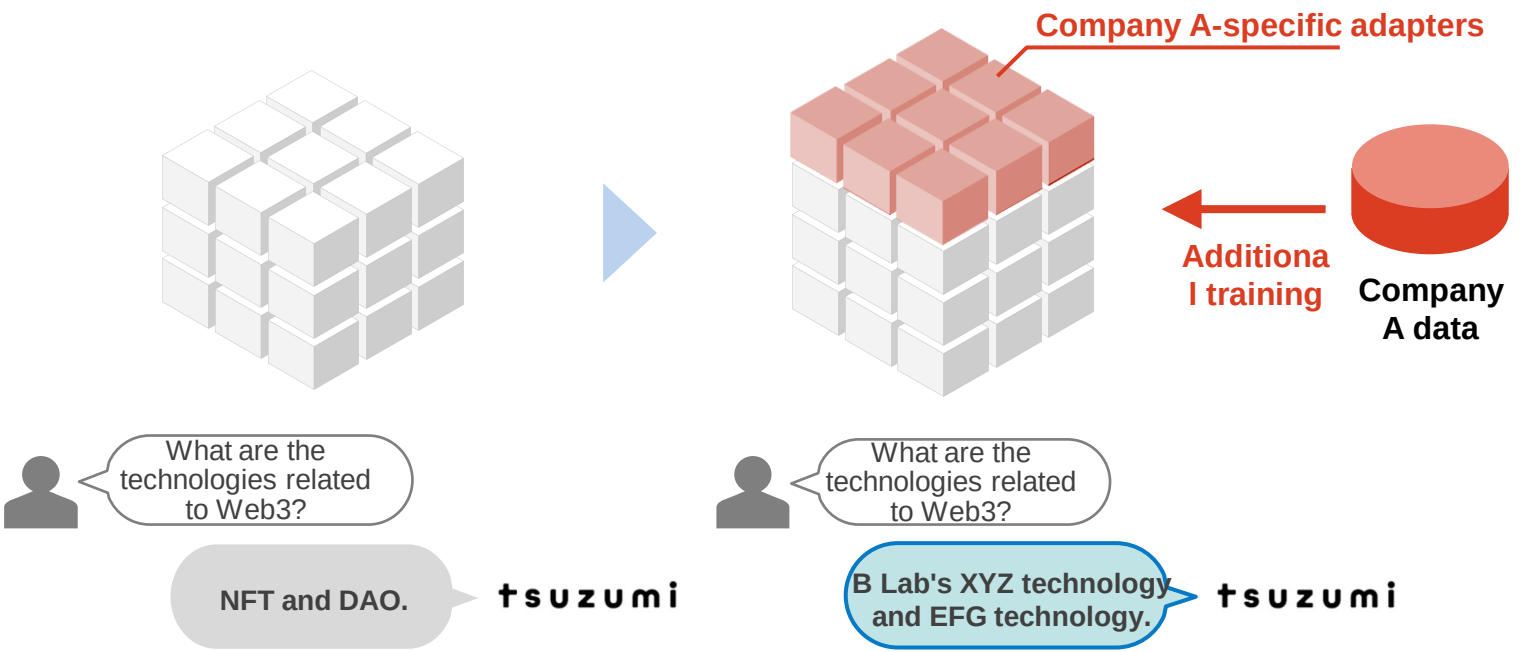
Benefits of Adapter Tuning (1) Industry Specialization

Enables industry-specific customization at low cost



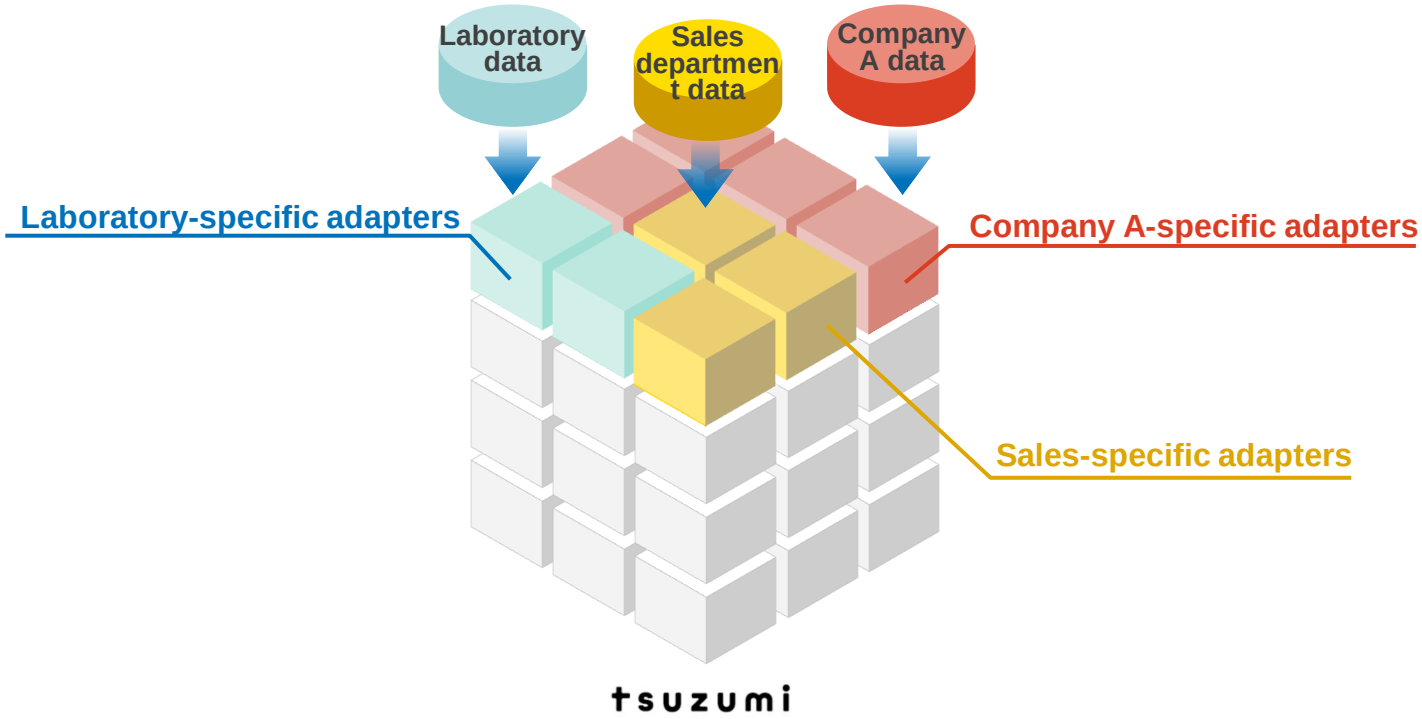
Benefits of Adapter Tuning (2) Organizational Specialization

Enables organization-specific customization at low cost



Multiple Adapters

Enables sharing the base model with multiple adapters, and switching and combining adapters in accordance with the user or scenario.

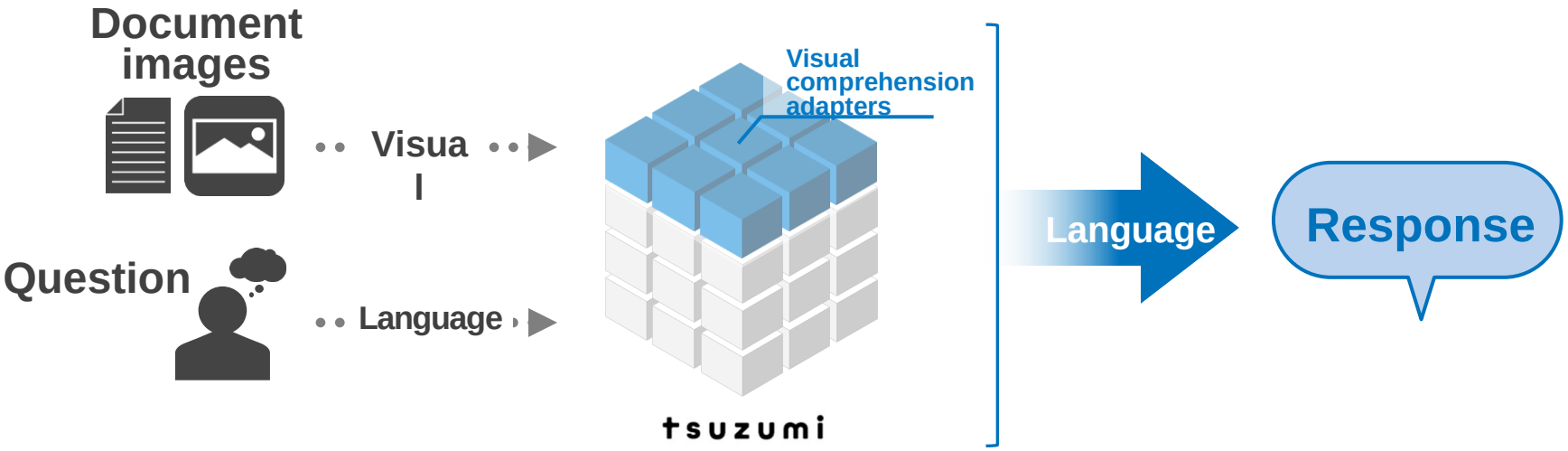


tsuzumi

Feature 4. Multimodality

Modality extension (1) Language + Visual

Enables asking questions not only based on language, but also while presenting document images
Applicable to tasks that involve the use of documents with images such as invoices and specifications and for RPA operations



Modality extension (1) Language + Visual: Implementation Example

ABC株式会社
山田　太郎　　様

発行日：2022年10月12日
株式会社XYZ
〒242-0000
神奈川県〇〇市〇〇
〇〇〇ビル
TEL: 012-345-6789

下記の通り、お見積り申し上げます。

合計金額（税込）　10,450　円

品名	数量	単価	金額
商品A	1	3,000	3,000
商品B	3	1,500	4,500
商品C	1	2,000	2,000

Visual

Language

What is the total amount excluding the 10% consumption tax?

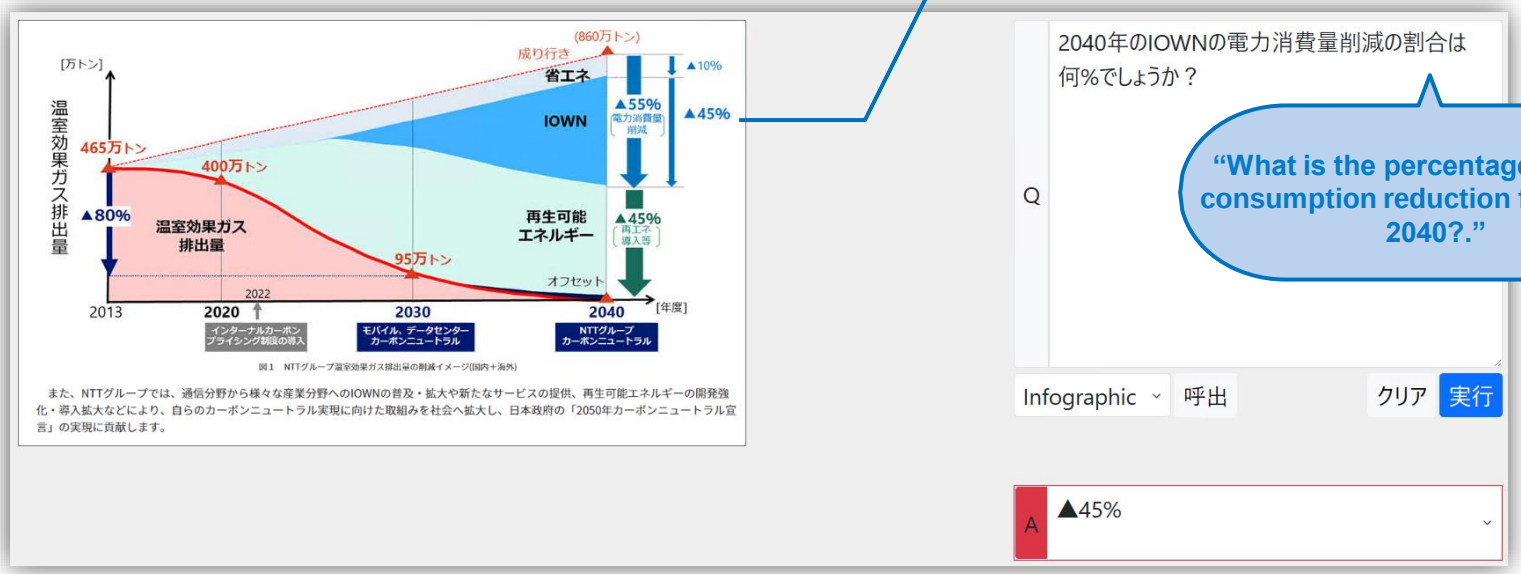
Visual comprehension adapters

tsuzumi

9,500 yen

Modality extension (1) Language + Visual: Implementation Example

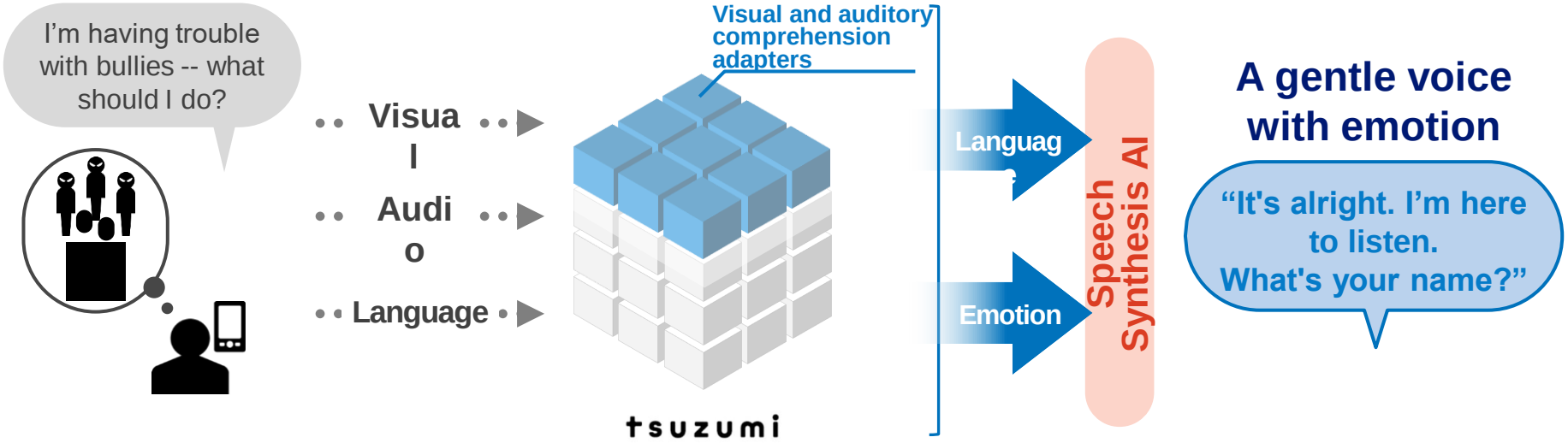
The correct answer is here.



"What is the percentage of power consumption reduction for IOWN in 2040?."

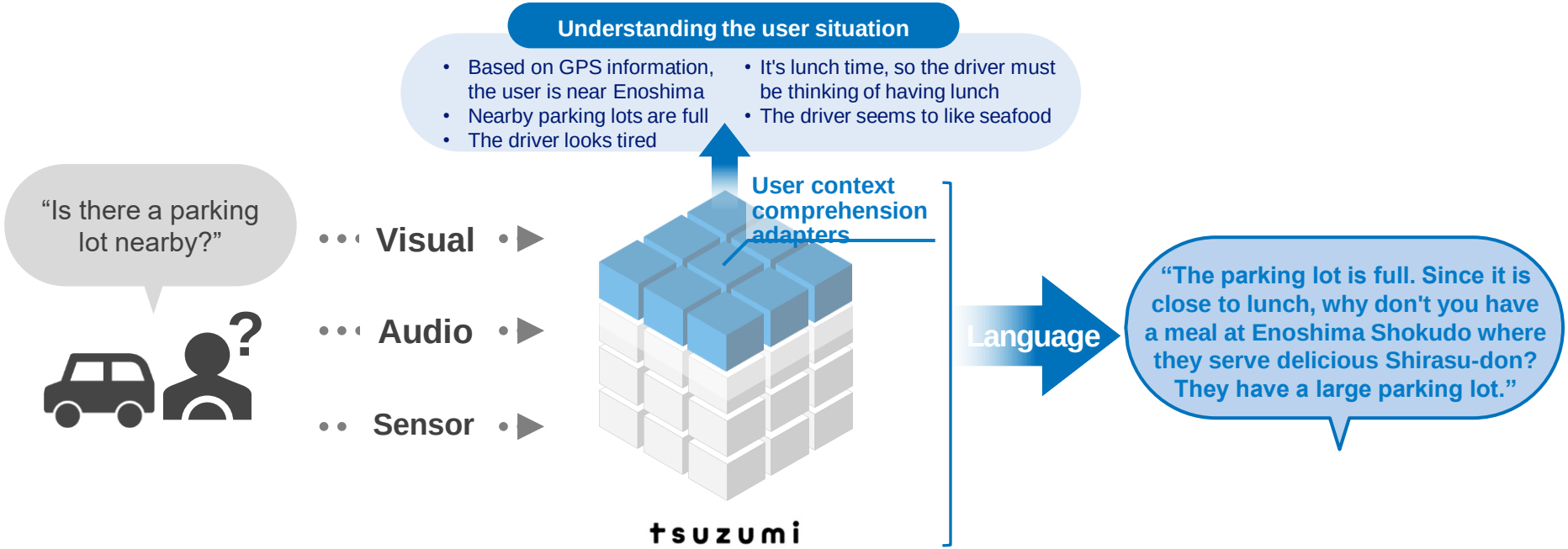
Modality extension (2) Language + Visual + Audio

In addition to language-based questions, enables answering questions based on the condition of the questioner
Applicable to tasks that involve closely working with people, such as counseling, call centers, and consultation centers



Modality extension (3) Language + User Situation

In addition to language-based questions, enables answering questions based on the situation of the questioner
Applicable to concierge services such as car navigation systems and smartphone navigation systems



tsuzumi

NTT Laboratories Technological Capability

Number of AI publications: 12th in the world and 1st in Japan

Rank	Company
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

Natural Language Processing Research: 1st in Japan

Number of papers accepted in top language processing conferences (TACL, NAACL, ACL, EMNLP, COLING) **2015-2021**[*1]

Rank	Company	No. of papers
1	**NTT**	**25.89**
2	Yahoo!	15.35
3	IBM	5.50
4	Fuji Xerox	4.41
5	Google	3.45
6	Fujitsu	2.98
7	PFN	2.51
8	NHK	2.38
9	NEC	1.63
10	Studio Ousia	1.20

*1 Reference: https://murawaki.org/misc/japan-nlp-2021.html

Track record in the Japanese Association for Natural Language Processing

NTT

Number of awards for excellence
in the last 10 years

1st

Machine Translation International Competition: 1st in the World

Sponsored by top international conference, WMT

The most prestigious international competition
in the field of machine translation

1st[*1]

in 4 categories for news translation tasks

*1 https://aclanthology.org/2020.wmt-1.12/

Tohoku University HP https://www.tohoku.ac.jp/japanese/2021/07/news20210730-02.html

Joint team from
Tohoku University, RIKEN, AIP, and NTT

Visual Comprehension International Competition: 2nd Place

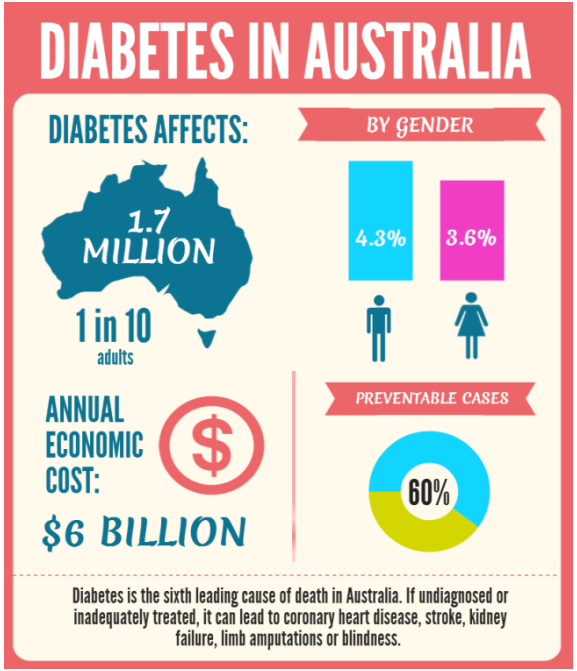
Sponsored by top international conference, ICDAR

Visual Comprehension Competition

Infographics VQA

2位*1 in the world

*1 https://icdar2021.org/program-2/competitions/



Q What is the percentage of women among diabetes patients?

A 3.6%

Q What is percentage of cases where diabetes was prevented?

A 40%

High-Quality, Large-Scale Training Data

Pre-training

- **More than 1 trillion** tokens
- **Japanese-English + 21 languages + programming languages**
- **Covers a wide range of domains** from specialized fields to entertainment

Instruction Tuning

- Leverages **in-house data on translation, summarization, dialog, and comprehension** accumulated over many years of research
- Newly created **wide-ranging tuning data** on “benefit” and “safety”

Overview of NTT's LLM

1 Features of tsuzumi

2 **tsuzumi and IOWN**

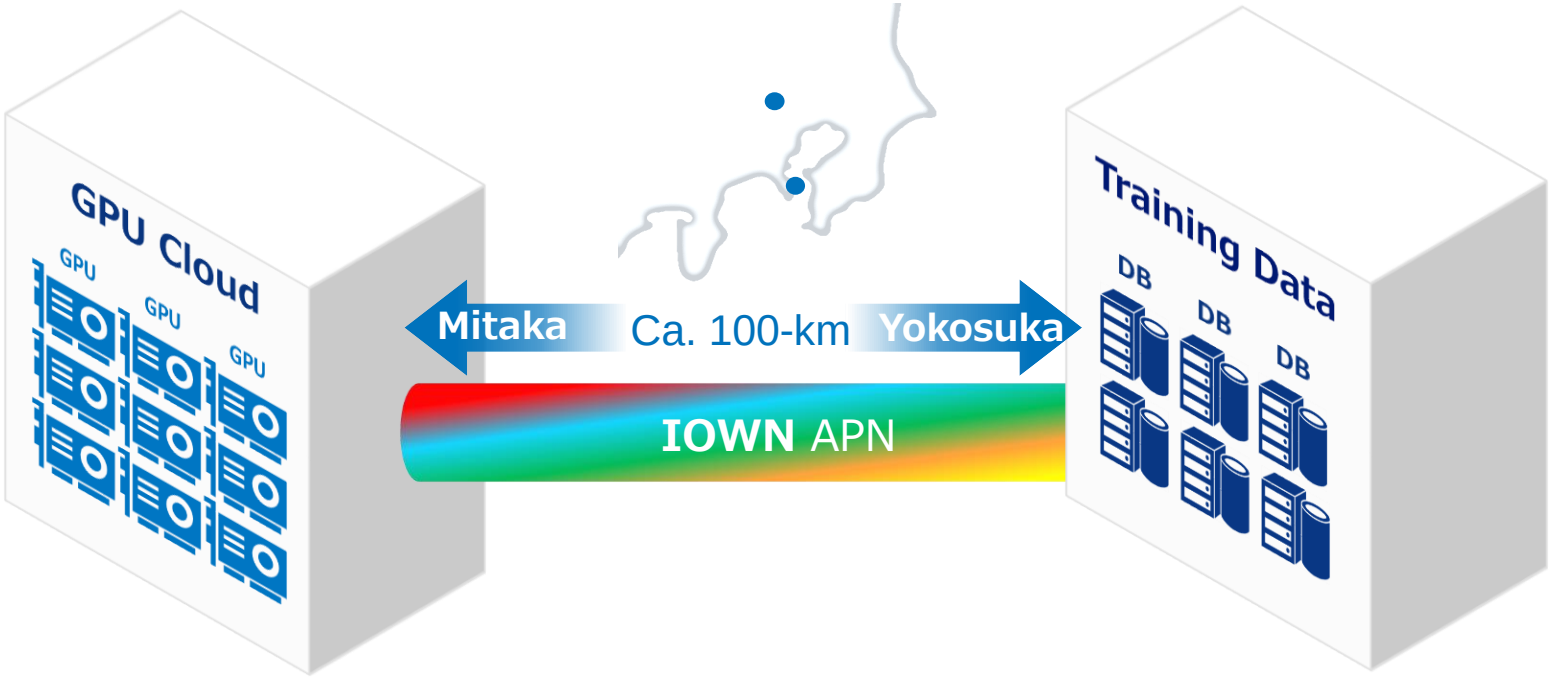
3 Product Lineup

IOWN APN x LLM

Construction of LLM sovereign hybrid environment using APN
Using GPUs in data centers hundreds of kilometers away while keeping training data at hand.
Creation of a safe, low-latency LLM training environment that is comparable with the local environment.

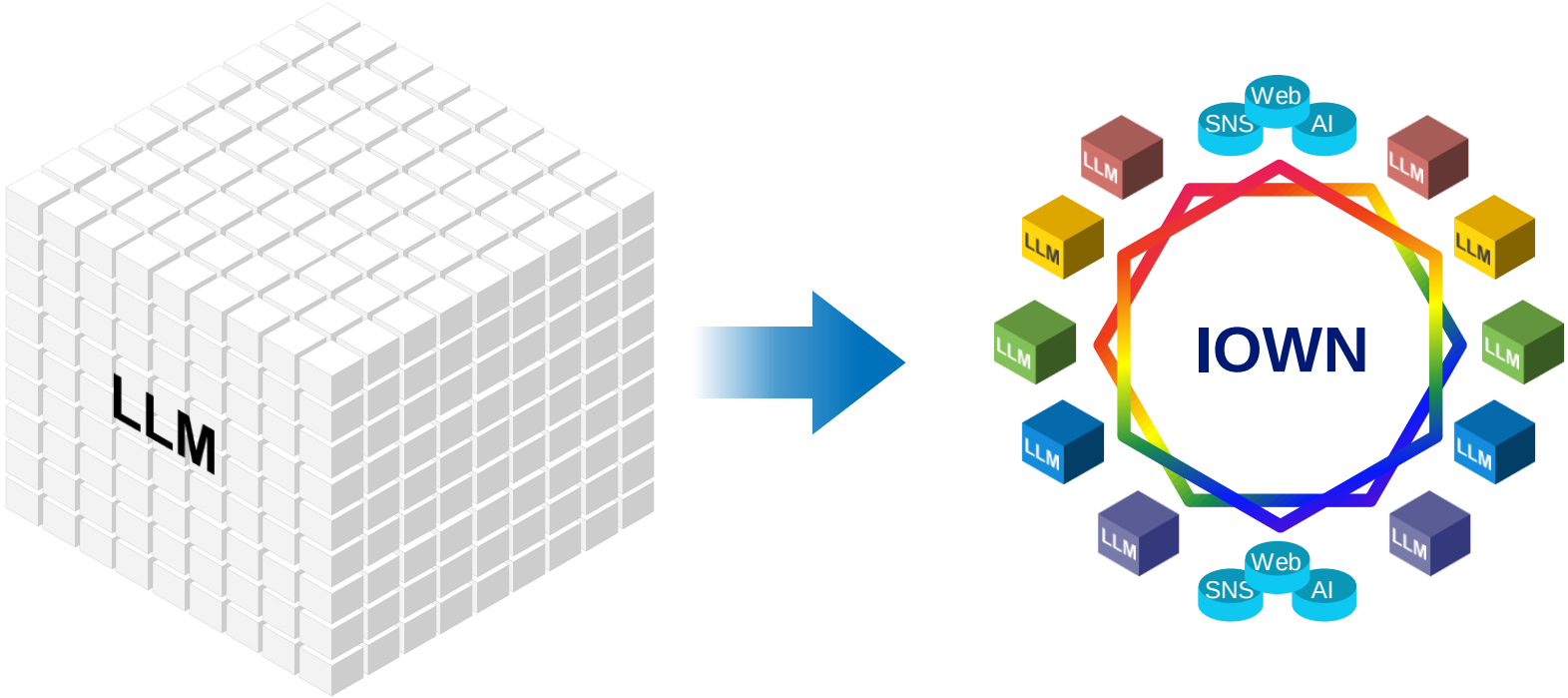
IOWN APN x LLM

Construction of LLM sovereign hybrid environment using APN
Using GPUs in data centers hundreds of kilometers away while keeping training data at hand.
Creation of a safe, low-latency LLM training environment that is comparable with the local environment.



NTT's Vision for the Future of AI: A Constellation of AI

Solving social issues through the collective wisdom of small LLMs with specific expertise and individuality, rather than creating a massive LLM that knows everything



Overview of NTT's LLM

1 Features of tsuzumi

2 tsuzumi and IOWN

3 **Product Lineup**

tsuzumi Product Lineup

Product	Param Size	Works on	Tentative Release Date	Tasks/ Language	Language Accuracy	Tuning	Multimodality
Ultra-lightweight tsuzumi	0.6B	CPU	Mar. 2024	Pre-General Tasks/ Japanese Only	Japanese Top-Level	Full Parameter No Adapters	Visual, Audio
Lightweight tsuzumi	7B	Low Grade GPU	Mar. 2024 For Trial : Oct. 2023	General Tasks/ Japanese, English Other 21 Languages.*1、Programming Languages	Japanese Top-Level	Full Parameter Single Adapter	Visual, Audio For Trial: Visual only
More Larger Models	13B～	High Grade GPU	After Apr. 2024	General Tasks/ Japanese, English Other 21 Languages.*1、Programming Languages	Multi-Lingual Top-Level	Full Parameter Multiple Adapters	Visual, Audio, Emotion, User Situation、Physical Sensations, etc.

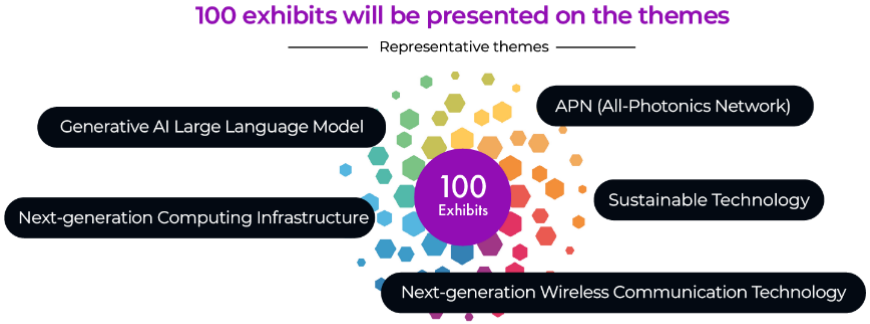
*1: Used only for pre-training data. The accuracy is to be evaluated and improved.

R&D Forum

R&D Forum

November 14 -17,2023
For Press: November 13

100 Exhibits
IOWN, tsuzumi, etc.



Related to tsuzumi

- **12** Real Venue Exhibits
- **Keynote Speeches by**
  - Kinoshita (Senior VP)
  - Our Researchers
  - Our Partners