



NTT版LLM **tsuzumi** の概要

2023年11月1日

執行役員 研究企画部門長

木下 真吾

- 1 **tsuzumiの特長**
- 2 **tsuzumiとIOWN**
- 3 **プロダクトライン**

- 1 **tsuzumiの特長**
- 2 tsuzumiとIOWN
- 3 プロダクトライン

tsuzumi

特長 1 軽量

サステナビリティ

GPT-3 (175B) 規模の学習には、膨大なエネルギーが必要

約 **1300**MWh^[1] / 1回の学習

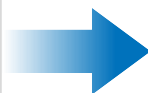


原発1基分 約 **1000**MWh

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

目指す方向性

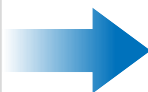
何でも知っている
巨大なLLM



専門知識をもった
小さなLLM

アプローチ

パラメタサイズ増加



言語学習データの
質と量を向上

tsuzumi

軽量なLLM tsuzumiを2種類を開発



超軽量版

tsuzumi-0.6B

GPT-3 (175B) の
約**300**分の**1**

軽量版

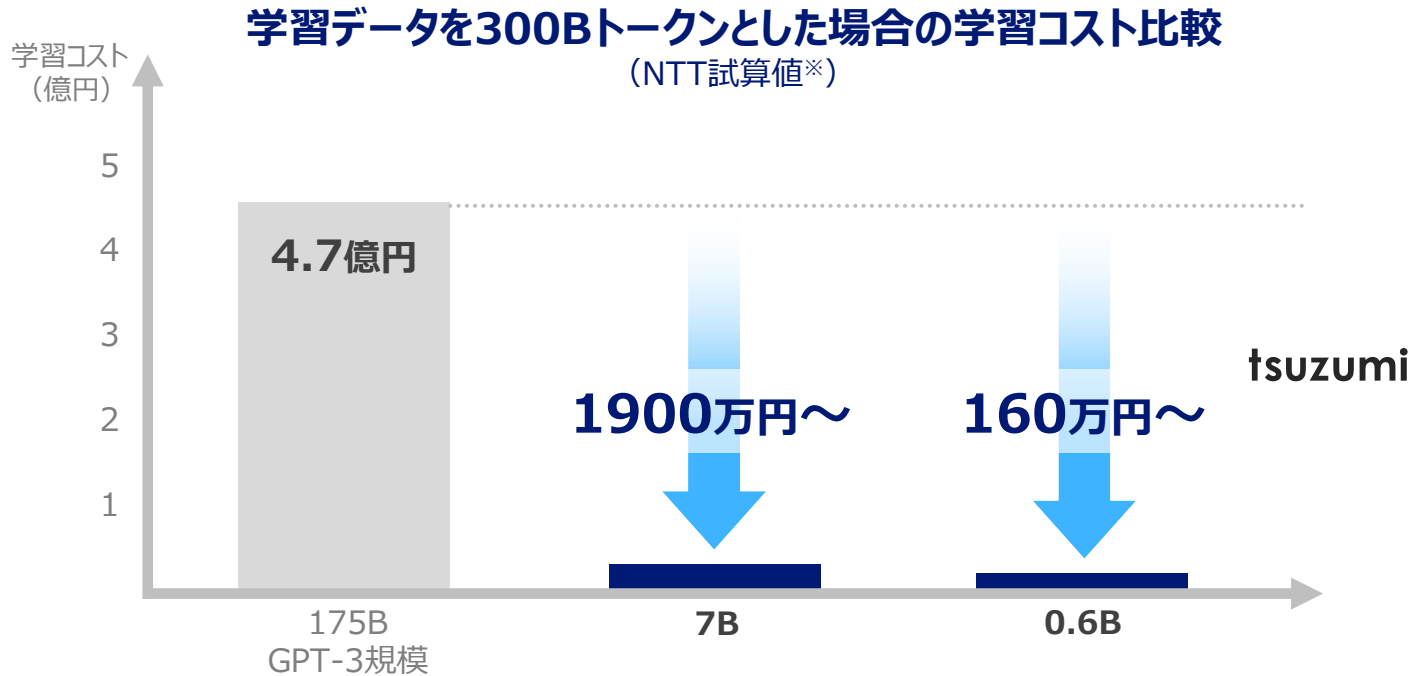
tsuzumi-7B

GPT-3 (175B) の
25分の**1**

軽量化のメリット① 学習コスト



GPT-3規模のLLMと比べて、学習コストを約300分の1 (超軽量化) および約25分の1 (軽量化) に低減可能



※試算条件

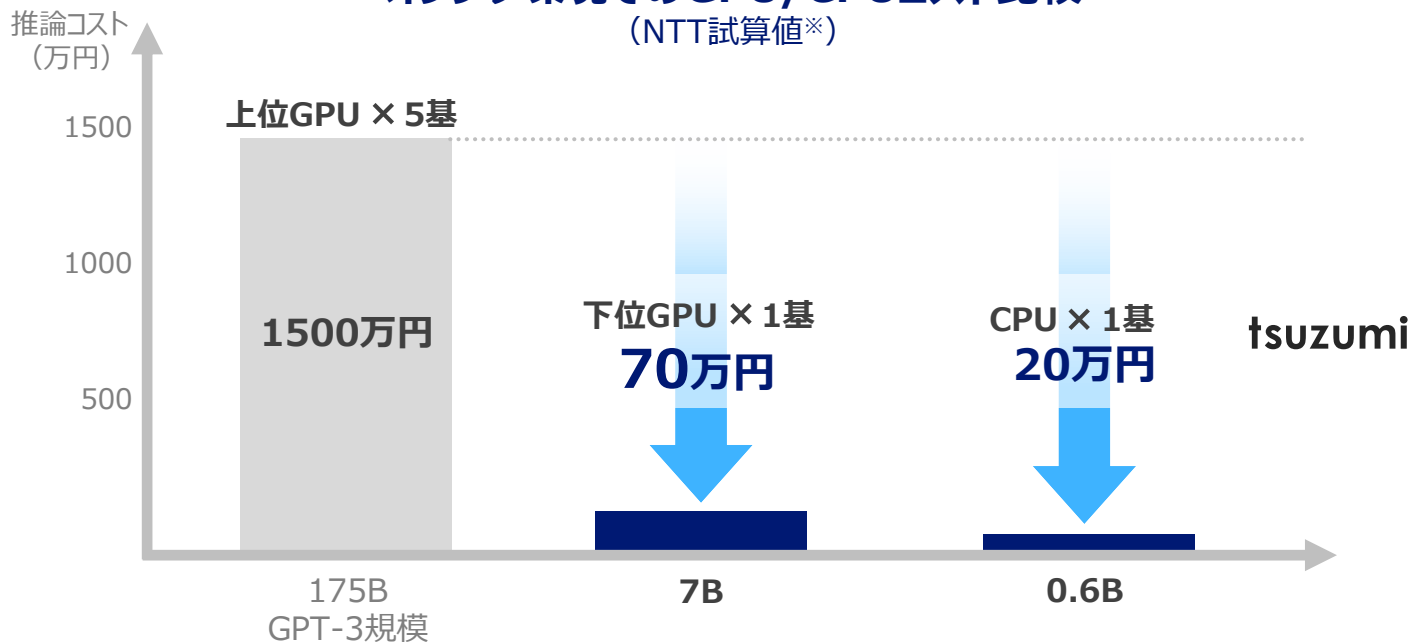
- Llama-1 7Bの学習時82432GPU-hoursをベースに、パラメタ数比とトークン数比から各LLMの必要GPU-hoursを算出
- 算出したGPU-hoursとAWSのGPUクラウド料金から学習コストを算出
- AWS GPUクラウド料金: A100-80GB 1ノード (8GPUs) 約14万円/日と想定
- 通常、パラメタサイズが小さい場合、精度を向上させるためには、2-3倍程度の学習データが必要。それに比例しコストも向上

軽量化のメリット② 推論コスト



GPT-3クラスと比べて、推論コストを約70分の1 (超軽量版) と約20分の1 (軽量版) に低減可能

オンプレ環境でのGPU/CPUコスト比較 (NTT試算値※)



※試算条件

- 量子化: 16ビット
- 必要GPUメモリサイズ: パラメタ数 × 量子化サイズ/8bit (175Bは350GB, 7Bは14GB, 0.6Bは2.4GB)
- ハードウェアコストは、上位GPU A100 80GB: 300万円/台, 下位GPU A10 24GB: 70万円/台, CPU PC: 20万円/台として換算、その他の運用などの費用は含まず

tsuzumi

特長 2 言語性能（特に日本語）が高い

model

tsuzumi-7b

Params

Chatbot

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:

- 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。
- 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:

- 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。
- 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。
- 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。

model

ChatGPT(3.5-Turbo)

Params

Chatbot

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のバランスについて、現状と改善策を提案します。

現状では、日本はエネルギーの多くを化石燃料に依存しています。これにより、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。一方で、日本は再生可能エネルギーの導入やエネルギー効率の向上など、環境保護に取り組んでいます。

改善策としては、以下の点が考えられます。

- 再生可能エネルギーの促進:太陽光や風力などの再生可能エネルギーの導入を積極的に推進することで、化石燃料に依存しない持続可能なエネルギー供給を実現します。
- エネルギー効率の向上:産業や家庭でのエネルギー使用効率を高めるための取り組みを行います。省エネルギー機器の普及やエネルギー管理システムの導入などが有効です。
- 廃棄物のリサイクル:廃棄物のリサイクルや再利用を促進することで、資源の有効活用と環境負荷の軽減を図ります。
- 環境教育の強化:若い世代に環境問題の重要性を啓発し、環境に配慮した行動を促すために、環境教育を充実させる必要があります。
- 国際協力の強化:日本は他の国との協力を通じて、環境問題の解決に取り組むべきです。技術やノウハウの共有、持続可能な開発の支援などが重要です。

これらの改善策を実施することで、日本はエネルギー政策と環境保護のバランスをより良くすることができると考えられます。

日本語性能比較 : Rakudaベンチマーク

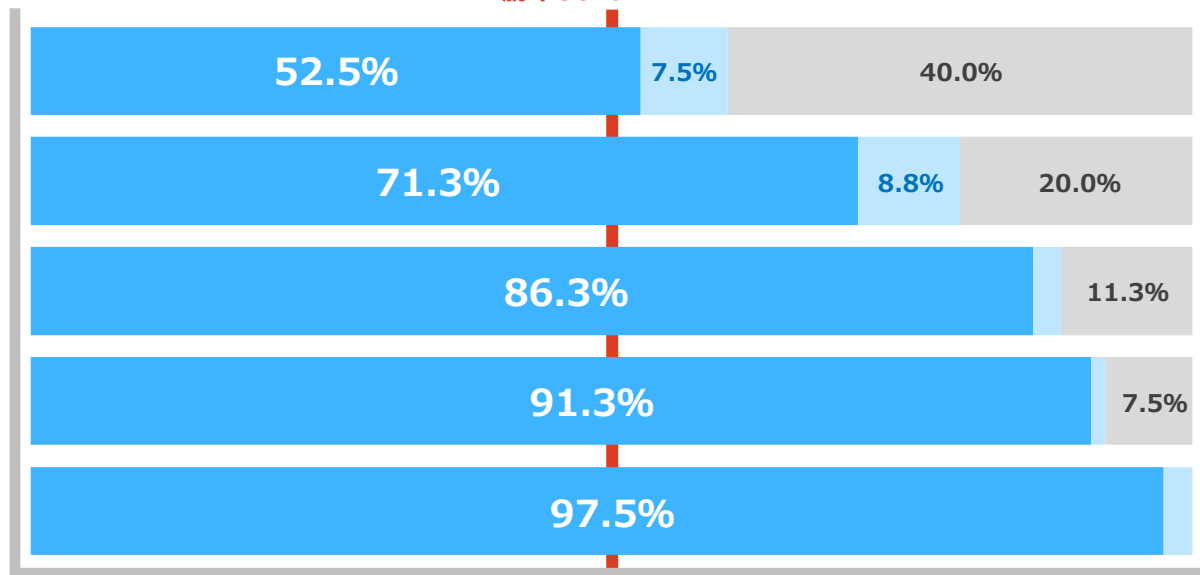


tsuzumi-7Bは、世界トップクラス、国産LLM中トップの性能を達成 (評価スコア : 1225 / 2023.10.26時点)
大規模なGPT-3.5を上回り、同クラスの国産LLMを大きく上回る

tsuzumi

対 GPT-3.5 175B?
(OpenAI gpt-3.5-turbo-0301)

WIN 勝率50% DRAW LOSE



※rakudaベンチマーク : <https://yuzuai.jp/benchmark> 2023.10.22実施
日本の地理・政治・歴史・社会に関する40問の質問 GPT-4による2モデルの比較評価 (40問×提示順2) で採点
llm-jpを除くモデル出力はサイトにアップロードされているものを利用 llm-jpはhuggingfaceのモデルカード記載の設定による
入力の繰り返しおよび終端トークンは後処理により除外した。
評価スコアは、2023/09/27付リーダーボード記載の全モデルとtsuzumi-7bをGPT-4による2モデルの比較評価を行い、Bradley-Terry strengthsにてランキングした結果

日本電信電話株式会社（本社：東京都千代田区、代表取締役社長：島田 明、以下「NTT」）は、太陽光エネルギーを利用する半導体光触媒と二酸化炭素（CO₂）を還元する金属触媒を電極として組み合わせた人工光合成デバイスを作製し、世界最長の350時間連続炭素固定を実現しました。CO₂変換反応による累積炭素固定量は420g/m²に達し、これは樹木（スギ）が年間で固定する単位面積当たりの炭素量を上回る量に相当します。今後は、より高性能な人工光合成デバイスを実現するために、電極での反応の高効率化、電極の長寿命化の両立を図ります。さらに、屋外試験を通じて、太陽光エネルギーを用いたCO₂削減技術のひとつとして確立し、気候変動の抑制に寄与し、持続可能な社会の実現に貢献します。本成果のベースとなる技術は、2023年11月14日～17日に開催されるNTT R&D フォーラムー IOWN ACCELERATION（※1）に展示予定です。

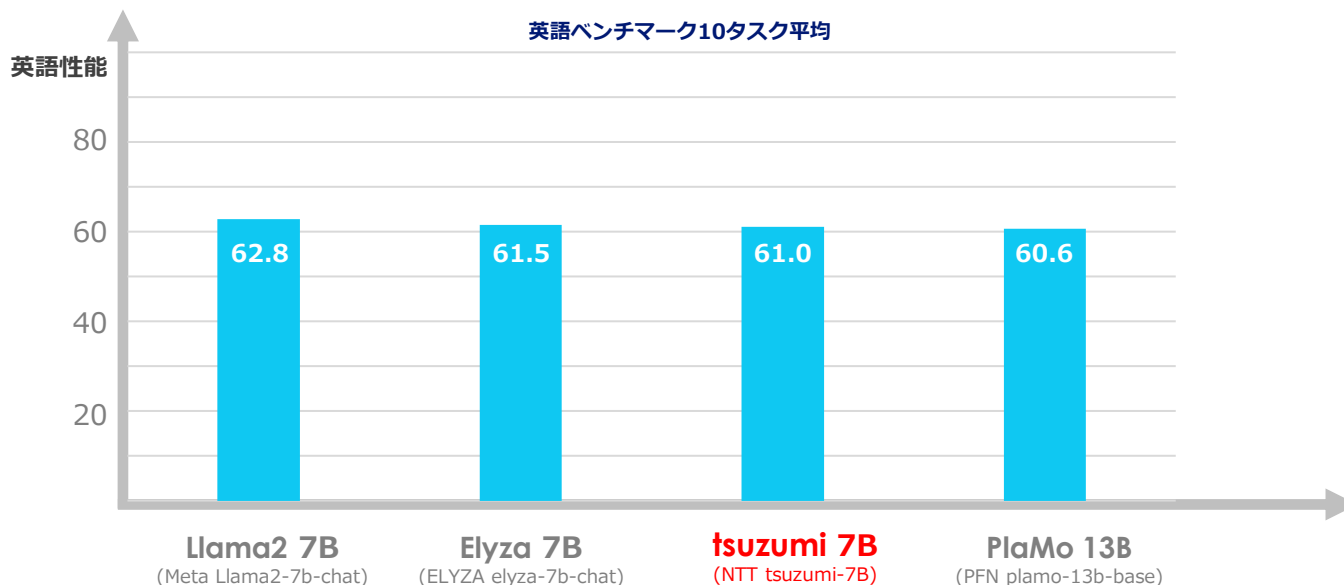
テキストから、デバイス名(name)、実現したこと(achievement)、展示イベント(event)、今後の予定(plan)を抽出して、jsonフォーマットで出力してください。

```
{
  "name": "人工光合成デバイス",
  "achievement": "世界最長の350時間連続炭素固定を実現しました。",
  "event": "NTT R&D フォーラムー IOWN ACCELERATION(※1)に展示予定です。",
  "plan": "より高性能な人工光合成デバイスを実現するため、電極での反応の効率化と長寿命化を両立させます。さらに屋外試験を通じて太陽光エネルギーによるCO2削減技術として確立し、気候変動抑制に寄与し持続可能な社会の実現に貢献します。"
}
```

英語性能比較 : lm-evaluation-harness



同サイズで英語をメインとした世界トップクラスLLMのLlama 7Bと同程度の性能を実現
英語以外にも、中韓仏独など多言語・プログラミング言語を拡充予定



* PLaMoのみベースモデルの結果

※1: 評価方法
lm-evaluation-harness: <https://github.com/EleutherAI/lm-evaluation-harness>
rinna/bilingual-gpt-neox-4bのモデルカード記載の英語10タスク (commonsense reasoning分野) におけるスコアの平均値
評価指標はaccおよびacc_norm (両方存在する場合はacc_normを優先)
Llama2-7b, elyza-7b, tsuzumi-7bは全てchat/instructモデルを利用 plamo-13bはbaseモデルの数値

英語に翻訳してください。

2023年10月に開催される国際会議「CSCW 2023 (The 26th ACM Conference on Computer-Supported Cooperative Work and Social Computing)」にて、NTT研究者の論文が採択され、Honorable Mention賞を受賞しました。CSCWは協調作業支援・ソーシャルコンピューティング分野の国際会議であり、論文採択率は約25%のトップカンファレンスとして知られており、Honorable Mention賞は、採択論文のうち上位5%の優れた論文に授与されます。採択論文「Escaping the Walled Garden? User Perspectives of Control in Data Portability for Social Media」は、GDPR (General Data Protection Regulation : EU一般データ保護規則) などで定められた「データ・ポータビリティ権」に着目し、ユーザ視点でその可能性と課題を探求しています。「データ・ポータビリティ権」は、ユーザが自身の個人データを他のサービスに移行できる権利として事業者に義務付けられているもので、これによってユーザが納得した上で自由にサービスを選択できるようになることが期待されています。しかし、事業者から提供されるこの権利を実現する機能をユーザがどのように認識し、サービス選択の自由度をどれだけ実感しているかは、明らかではありません。本研究は、特にSNSの文脈におけるデータ・ポータビリティに着目し、アメリカのSNSユーザを対象にアンケートとインタビュー調査を行いました。調査の結果、ユーザはSNS事業者の提供する機能に対して概ね好意的な印象を持っていましたが、他のサービスを選択するには不十分であることが示唆されました。この知見をもとに、ユーザが自由にサービスを選べるようサポートするために、適切なデータのみを移行先のプラットフォームに選択移行できる機能や、新しいプラットフォームに連絡先を移行しつつも各個人のプライバシーを守る方法を提案しています。NTTは引き続き、人間と情報の本質に迫る基礎理論の構築と社会に変革をもたらす革新技術の創出をめざした基礎研究を推進するとともに、プライバシーやヒューマンファクターの研究開発を通じて、個人の権利の保護と開かれた社会の実現に向けて尽力します。

In October 2023, a paper by NTT researchers was selected for presentation at the "CSCW 2023 (The 26th ACM Conference on Computer-Supported Cooperative Work and Social Computing)" international conference. The CSCW is an international conference in the field of cooperation support and social computing, with a publication rate of approximately 25% as one of the top conferences. The Honorable Mention Award is given to the best papers among those accepted, with 5% being excellent papers. The paper titled "Escaping the Walled Garden: Understanding User Perspectives on Data Portability Rights" discusses the potential and challenges related to data portability rights defined under General Data Protection Regulation (GDPR) such as user control over their personal data. It explores how users can benefit from this right while understanding its implications and limitations. However, it remains unclear whether users understand and feel the freedom to choose services based on these rights provided by service providers. In order to address this issue, we conducted surveys and interviews targeting American SNS users. As a result, although users had generally positive impressions towards the features offered by SNS operators, they indicated that they were not sufficiently satisfied to switch services without further consideration. Based on this insight, we propose methods to support users in selecting appropriate data to transfer to other platforms and protect their privacy while maintaining contact information on new platforms. NTT will continue to promote fundamental research aimed at grasping humanity and information, and through research and development in privacy and human factors, strive to ensure individual rights protection and open society.

pandasでcsvをparquetに変換するコードを書いてください。

以下は、pandasを使用してCSVファイルをParquet形式に変換するためのコードです。

```
import pandas as pd
from pyarrow import parquet

# CSVファイルの読み込みとデータフレームへの変換
df = pd.read_csv('path/to/file.csv')
df = df.astype({'column1': 'int32', 'column2': 'float64'})

# Parquet形式での書き込み
df.to_parquet('path/to/output/dir', index=False)
```

このコードでは、pandas ライブラリを使用してCSVファイルを読み取ります。その後、astype() メソッドを使って各列を指定した型に変換します(ここでは整数と浮動小数点数)。最後に、to_parquet() 関数を使用してParquet形式で出力ディレクトリに書き込んでいます。

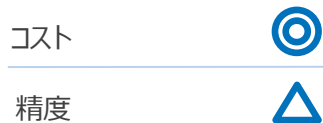
tsuzumi

特長 3 柔軟なカスタマイズ

さまざまなチューニング方法

精度やコストなど要件の違いに柔軟に 대응、3つのチューニング方法を提供

プロンプトエンジニアリング

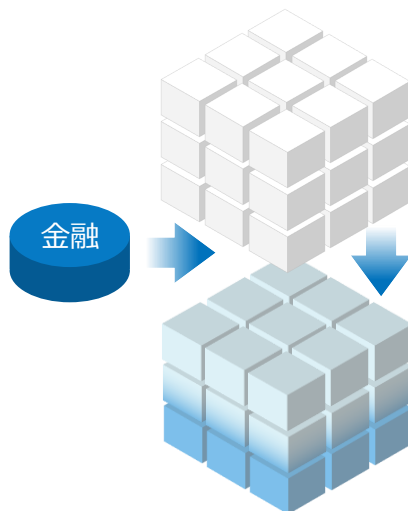


プロンプトに
金融関連情報を付加



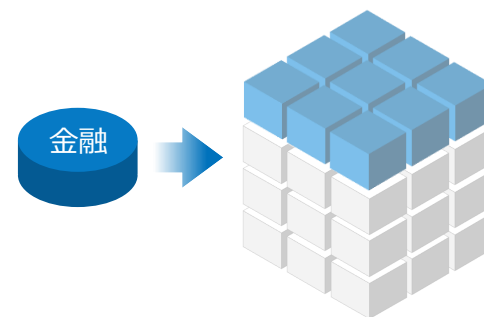
tsuzumi

フルファインチューニング



tsuzumi

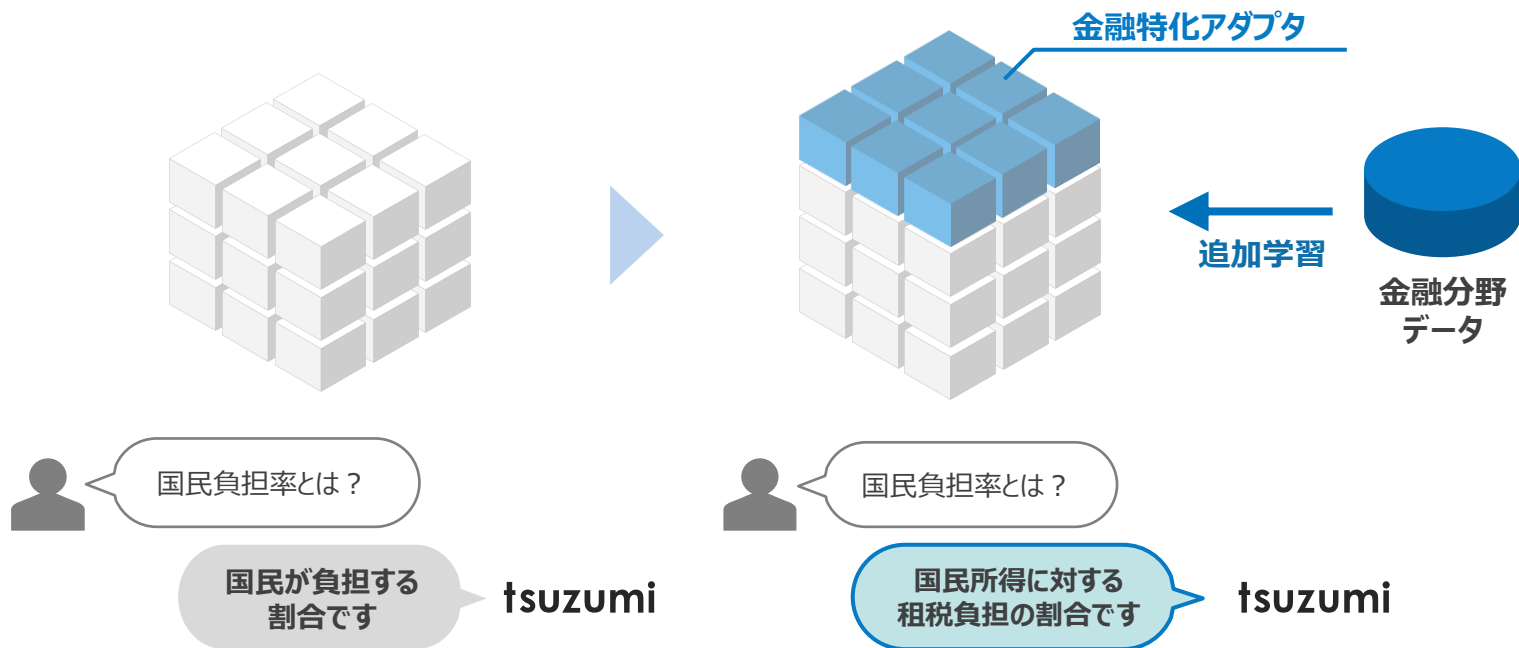
アダプタチューニング



tsuzumi

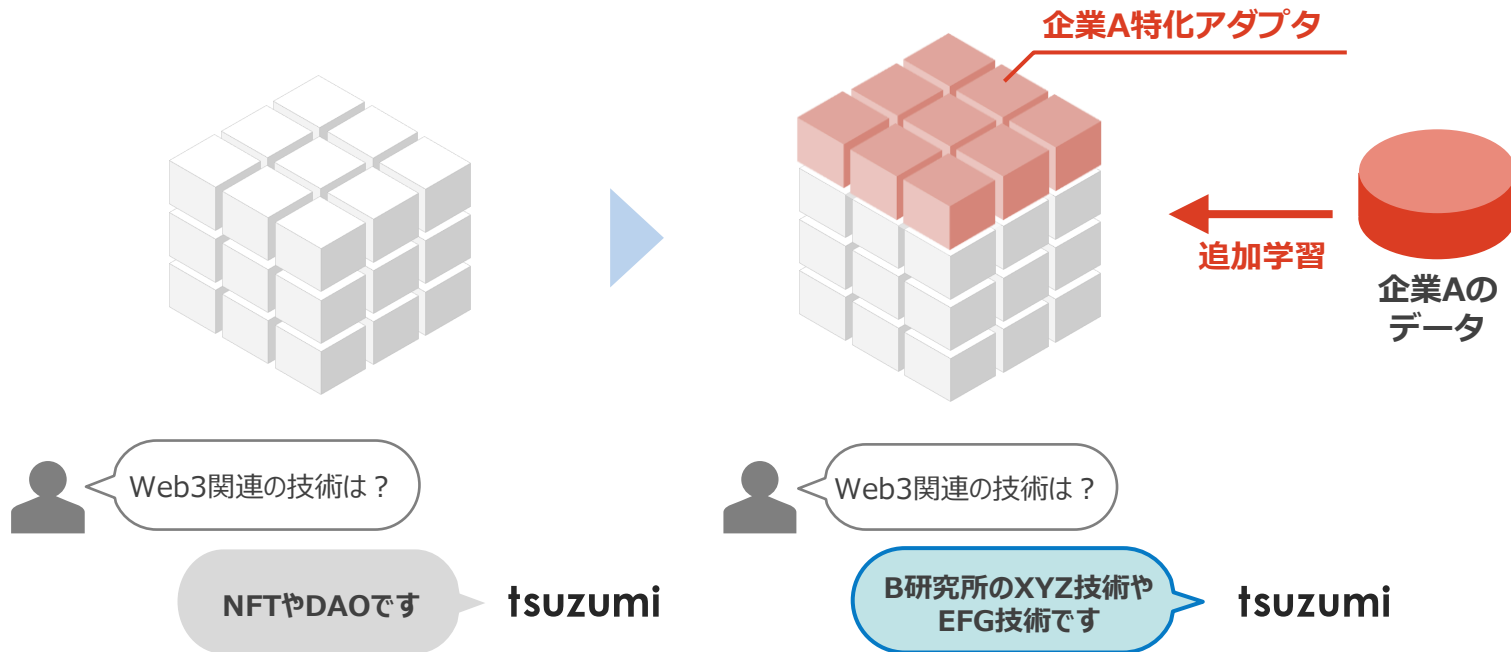
アダプタチューニングのメリット① 業界特化

業界ごとのカスタマイズを低コストで実現



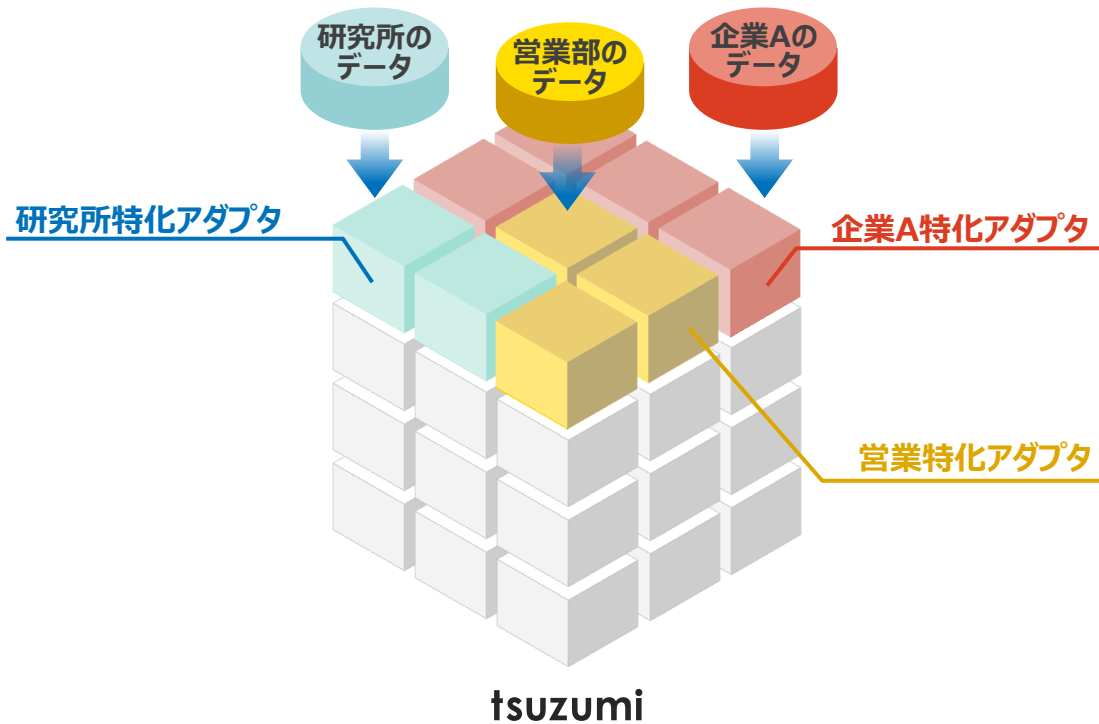
アダプタチューニングのメリット② 組織特化

組織ごとのカスタマイズを低コストで実現



マルチアダプタ

基盤モデルを複数アダプタで共有し、利用ユーザやシーンに応じてアダプタを切り替えたり、組み合わせたり可能

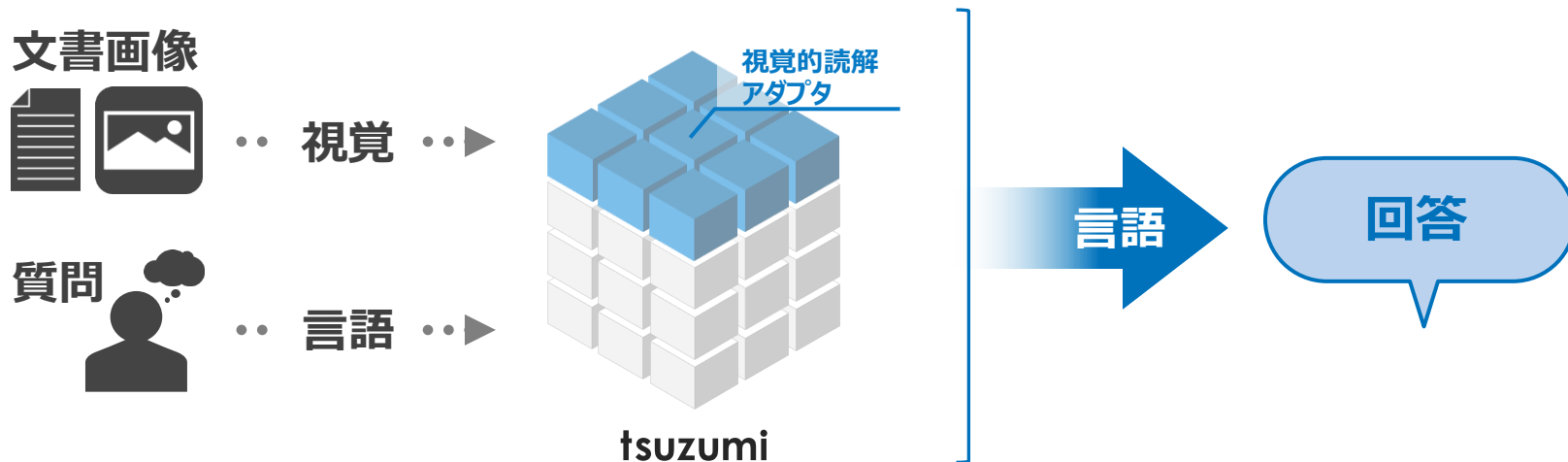


tsuzumi

特長 4 マルチモーダル

モーダル拡張① 言語 + 視覚

言語による質問だけでなく、文書画像を提示しながらの質問が可能
請求書や仕様書など画像付き文書を用いる業務やRPA業務に適用可能



モーダル拡張① 言語 + 視覚 実施例

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

視覚

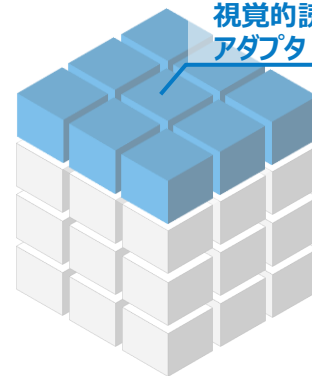


言語



10%の消費税を抜いた
合計金額は？

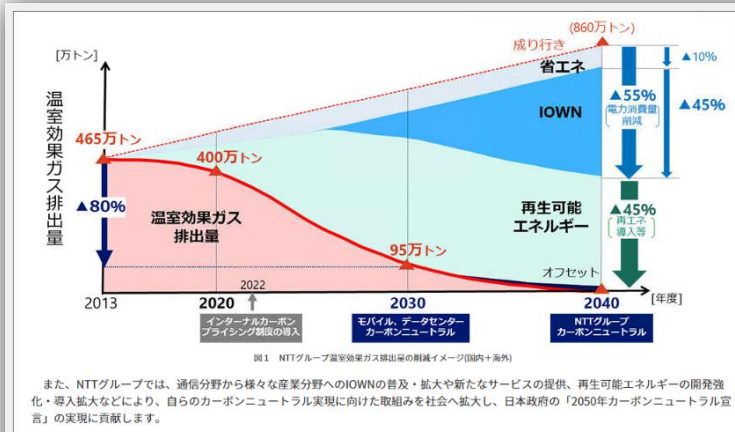
視覚的読解
アダプタ



tsuzumi

9,500円

モーダル拡張① 言語 + 視覚 実施例



2040年のIOWNの電力消費量削減の割合は何%でしょうか？

Q

Infographic

呼出

クリア

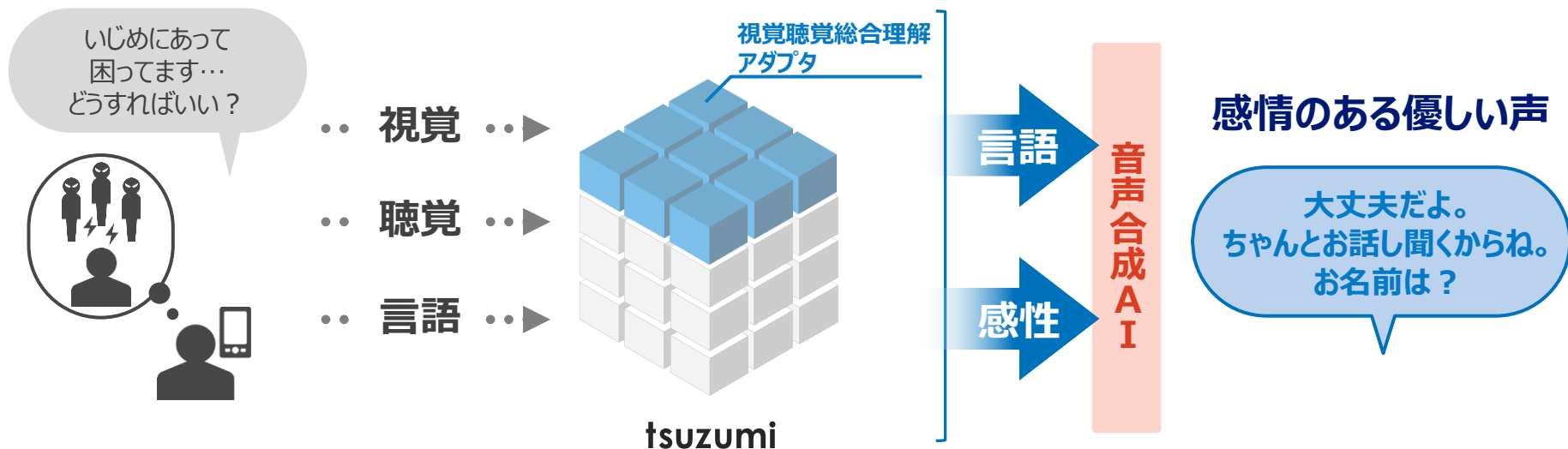
実行

A

▲45%

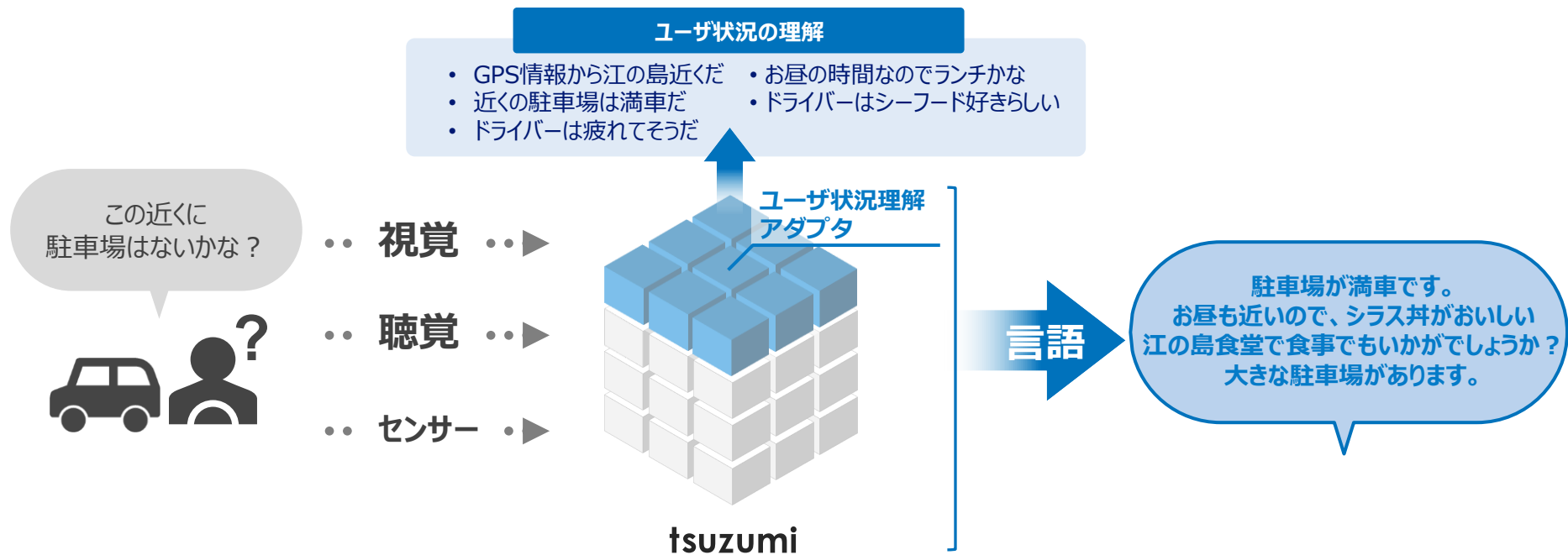
モーダル拡張② 言語 + 視覚 + 聴覚

言語による質問に加え、質問者の様子を踏まえた回答が可能
カウンセリング、コールセンター、相談所など人に寄り添う業務に適用可能



モーダル拡張③ 言語 + ユーザ状況

言語による質問に加え、質問者の状況を踏まえた回答が可能
カーナビ、スマホナビなどコンシェルジュ業務に適用可能



tsuzumi

NTT研究所の技術力

AI分野論文数 世界12位 国内1位



順位	企業名
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>

自然言語処理分野（日本語） 世界トップクラス

言語処理トップカンファレンス（ACL, NAACL, ACL, EMNLP, COLING）への
2015-2021 採択数 *1

順位	企業名	採択数
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 引用元: <https://murawaki.org/misc/japan-nlp-2021.html>

言語処理学会での実績



*2 企業の研究機関の中で1位

トップ国際会議WMT主催

機械翻訳分野で
最も権威の高い国際コンペティション
ニュース翻訳タスクの4部門

1位*1

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



東北大学HP <https://www.tohoku.ac.jp/japanese/2021/07/news20210730-02.html>

東北大-理研AIP-NTTの合同チーム

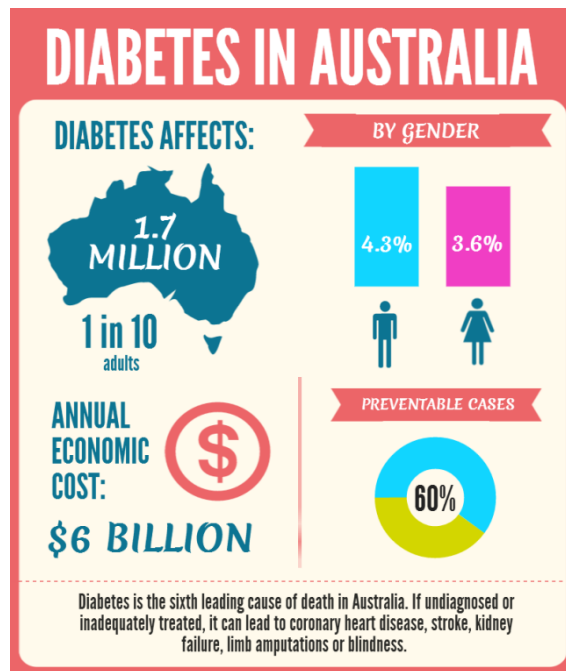
トップ国際会議ICDAR主催

視覚読解コンペティション
Infographics VQA

世界

2位*1

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



Q 女性の糖尿病患者の割合は？

A 3.6%

Q 糖尿病を防げた事例の割合は？

A 40%

事前学習

- **1兆以上**のトークン数
- **日英 + 21言語 + プログラミング言語**
- 専門領域からエンタメまで、**幅広いドメインをカバー**

インストラクションチューニング

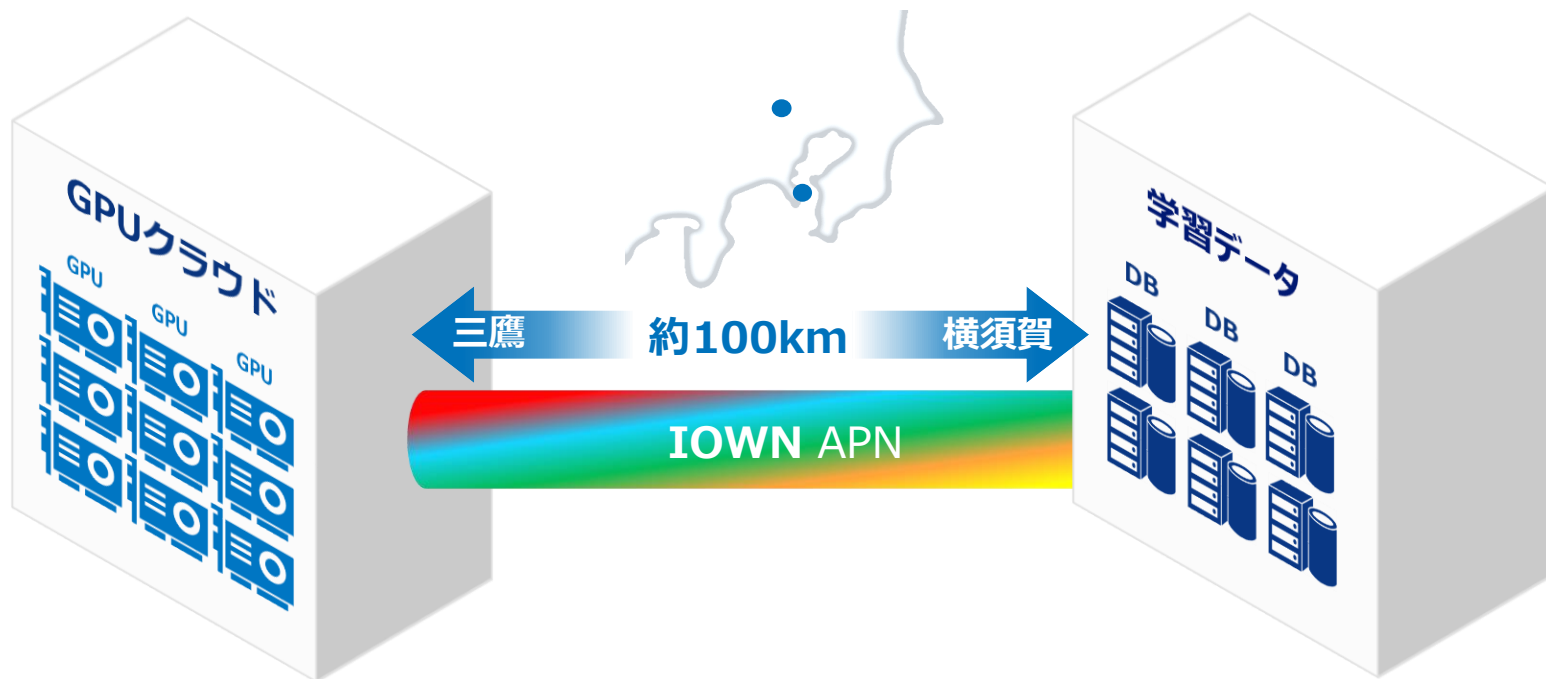
- 長年の研究で蓄積してきた**翻訳・要約・対話・読解などの内製データ**を活用
- 「有益さ」「安全さ」に関する**広範囲なチューニングデータを新規作成**

- 1 tsuzumiの特長
- 2 **tsuzumiとIOWN**
- 3 プロダクトライン

IOWN APN x LLM

APNを利用したLLMサブリンハイブリッド環境構築

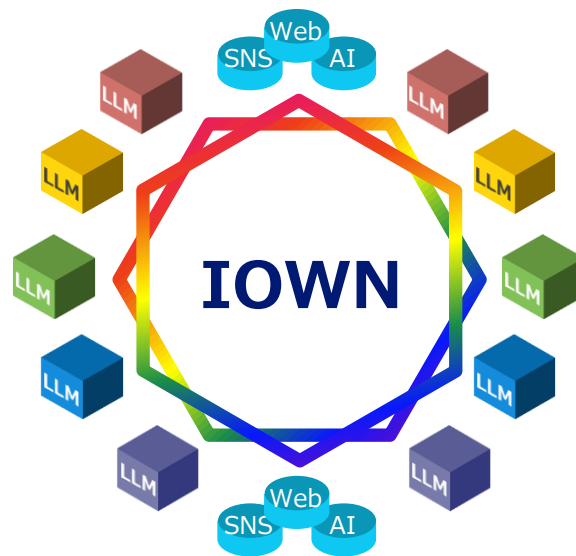
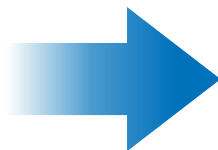
学習データを手元に置いたまま、約100km離れたデータセンターのGPU利用
ローカル環境と遜色のない安全かつ低遅延のLLM学習環境を実現



NTTが目指すAIの未来 AIコンステレーション



何でも知っている1つの巨大なLLMではなく、専門性や個性をもった小さなLLMの集合知による社会課題解決
大量のLLMの連携基盤としてIOWNが重要となる



- 1 tsuzumiの特長
- 2 tsuzumiとIOWN
- 3 **プロダクトライン**

ラインアップ	パラメタ サイズ	実行環境	リリース 予定日	タスク/ 対応言語	言語精度	チューニング	マルチモーダル
超小型版 tsuzumi	0.6B	CPU	2024.3	準汎用タスク/ 日本語のみ	日本語 トップクラス	フルパラメタ アダプタなし	視覚、聴覚
小型版 tsuzumi	7B	下位GPU	2024.3 トライアル版： 2023.10	汎用タスク/ 日本語、英語、 他21言語*1、 プログラミング言語	日本語 トップクラス	フルパラメタ シングルアダプタ	視覚、聴覚 トライアル版：視覚のみ
中型版 tsuzumi	13B～	上位GPU	2024.4以降	汎用タスク/ 日本語、英語、 他21言語、 プログラミング言語	多言語 トップクラス	フルパラメタ マルチアダプタ	視覚、聴覚、感性、 ユーザ状況、 身体感覚など

*1:事前学習データとして利用しているだけで品質については今後評価・向上予定

R&Dフォーラム

開催日時：11月14日～17日
プレスデー：11月13日

IOWN、tsuzumiなどテーマごとに
約**100**展示

tsuzumi 関連

- ・ 特別展示（12展示）
- ・ 木下基調講演
- ・ 研究者特別講演
- ・ パートナー特別講演

約100展示をテーマごとにご紹介します

代表的なテーマ

生成 AI 大規模言語モデル

次世代コンピューティング基盤

APN (All-Photonics Network)

100 Exhibits

次世代無線通信技術

サステナブル技術

NIT R&D FORUM 2023
IOWN
IOWN ACCELERATION

2023年11月14日(火)～11月17日(金)
講演 > 展示 >
よくあるご質問はこちら >

「NTT R&D FORUM 2023 - IOWN ACCELERATION」では、今年3月にサービスを開始したIOWN (Innovative Optical and Wireless Network) の具体的なサービス・システム、ユースケース、要素技術や、生成AIのNTT版AI技術を中心とし、NTTグループR&Dの最新成果について、講演や展示を通じて分かりやすくご紹介してまいります。

NTT R&D の最新動向をお届けします
IOWN - NTT 版 AI 基盤など

NTT R&D FORUM 2023
NTT 研究開発本部
NTT 研究開発本部
NTT 研究開発本部
Keynote Speeches