



別府湾会議2023

NTT版LLM tsuzumi

2024年2月27日



執行役員 研究企画部門長

木下 真吾

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

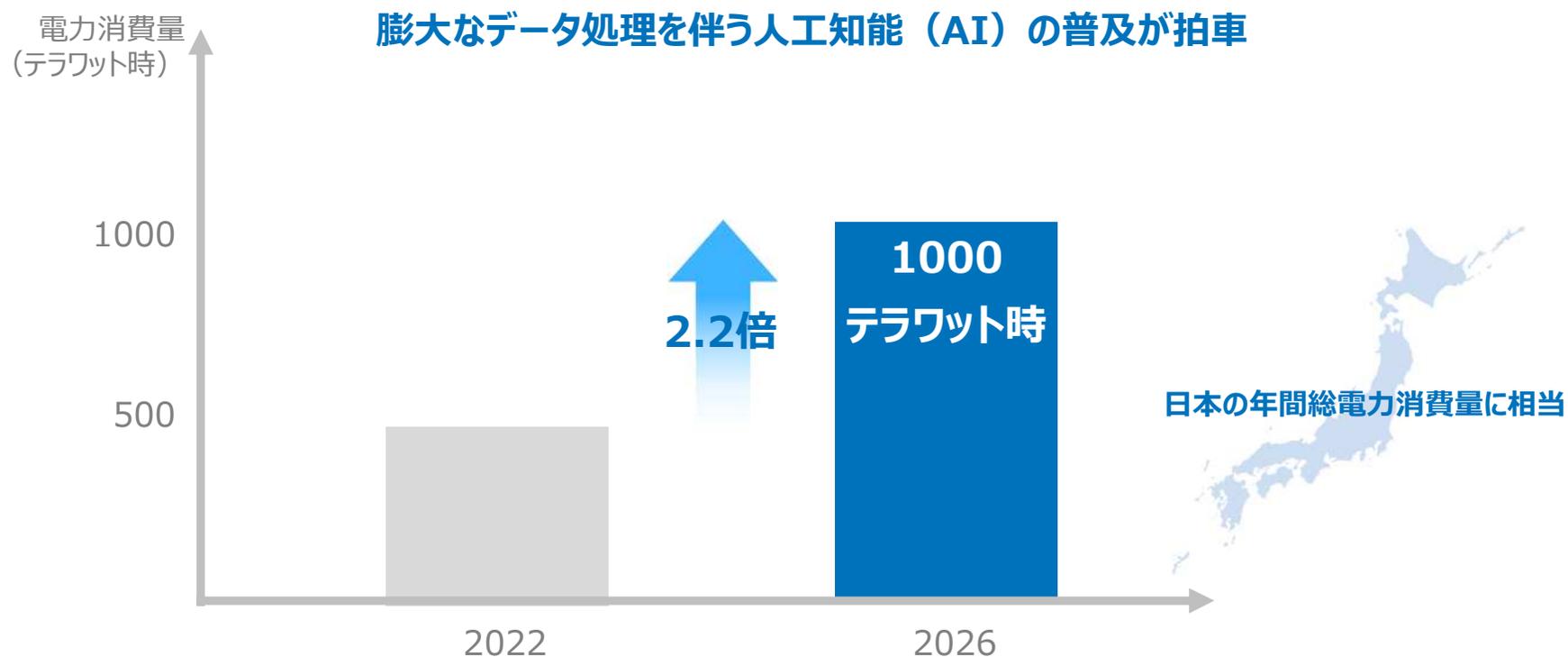
約**1300MWh**^[1] / 1回の学習



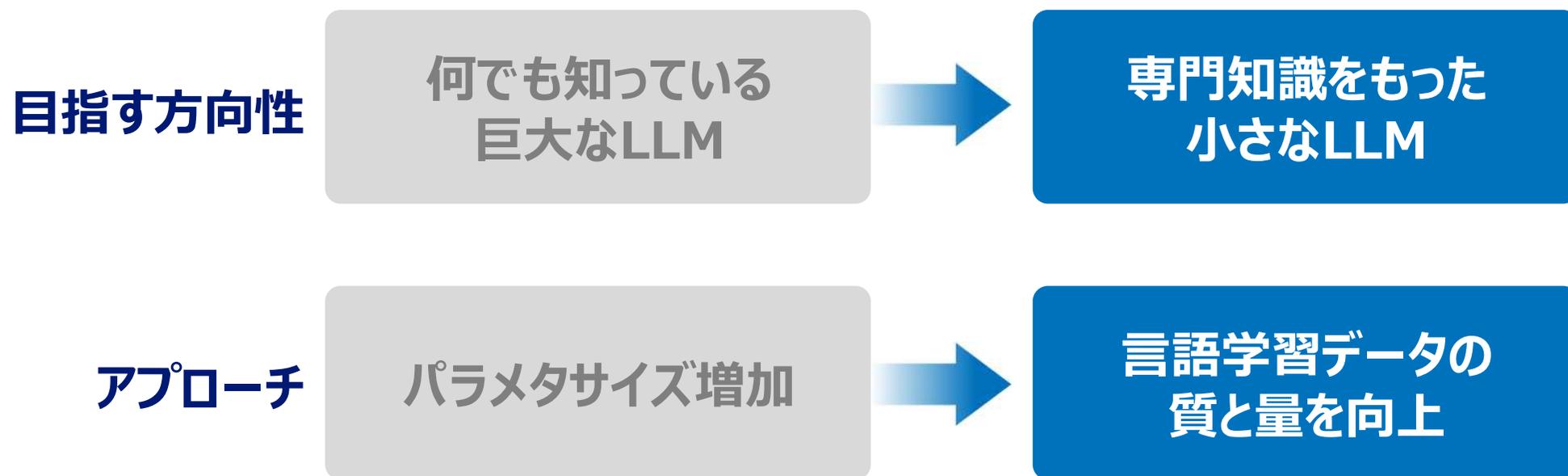
原発1基分 約**1000MWh**

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

データセンターの電力消費量



国際エネルギー機関（IEA）試算



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



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

学習データを300Bトークンとした場合の学習コスト比較 (NTT試算値※)



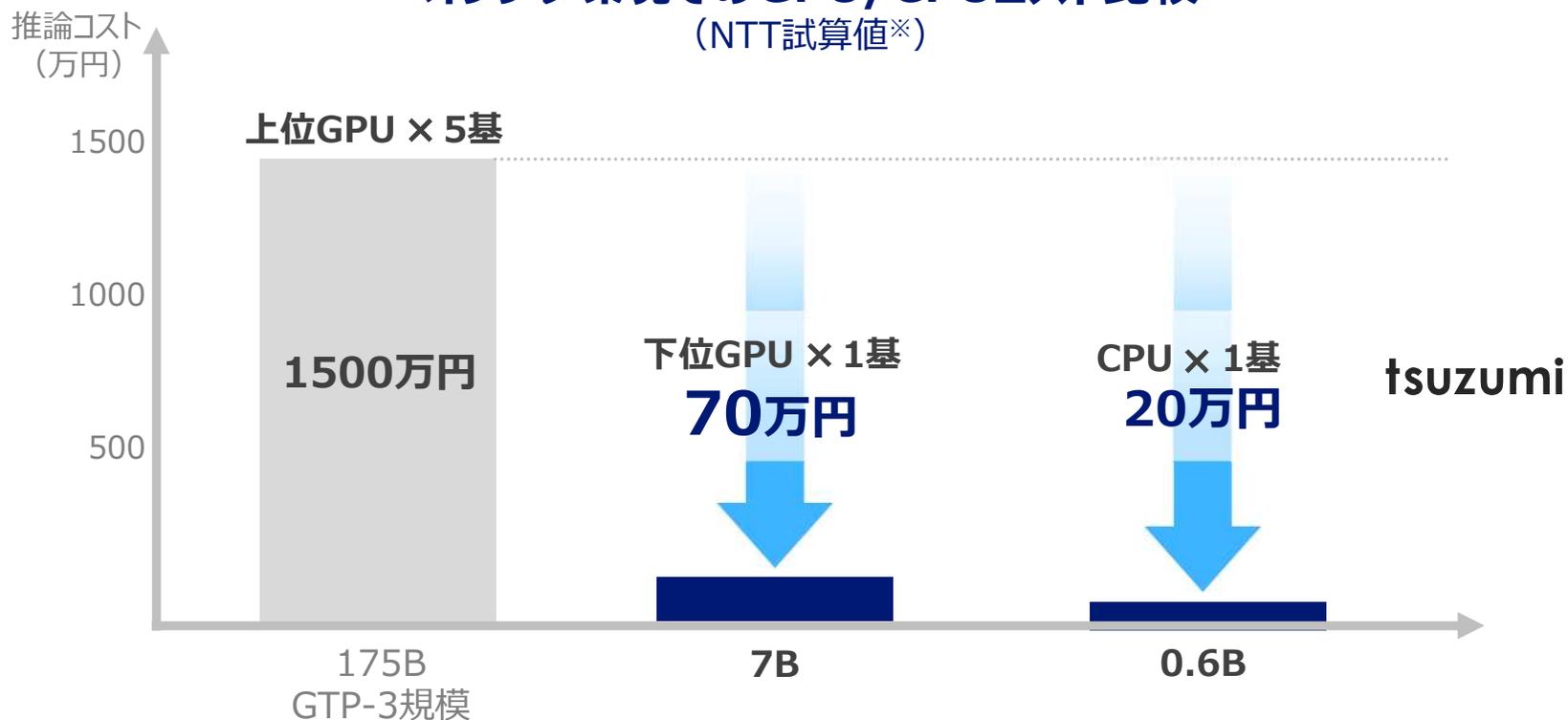
※試算条件
• LLaMA-1 7Bの学習時82432GPU-hoursをベースに、パラメタ数比とトークン数比から各LLMの必要GPU-hoursを算出
• 算出したGPU-hoursとAWSのGPUクラウド料金から学習コストを算出
• AWS GPUクラウド料金: A100-80GB 1ノード (8GPUs) 約14万円/日と想定
• 通常、パラメタサイズが小さい場合、精度を向上させるためには、2-3倍程度の学習データが必要。それに比例しコストも向上

軽量化のメリット② 推論コスト



GTP-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

小さい

特長2 日本語に強い



model

tsuzumi-7b

tsuzumi

model

ChatGPT(3.5-Turbo)

GPT 3.5

Params

Params

Chatbot

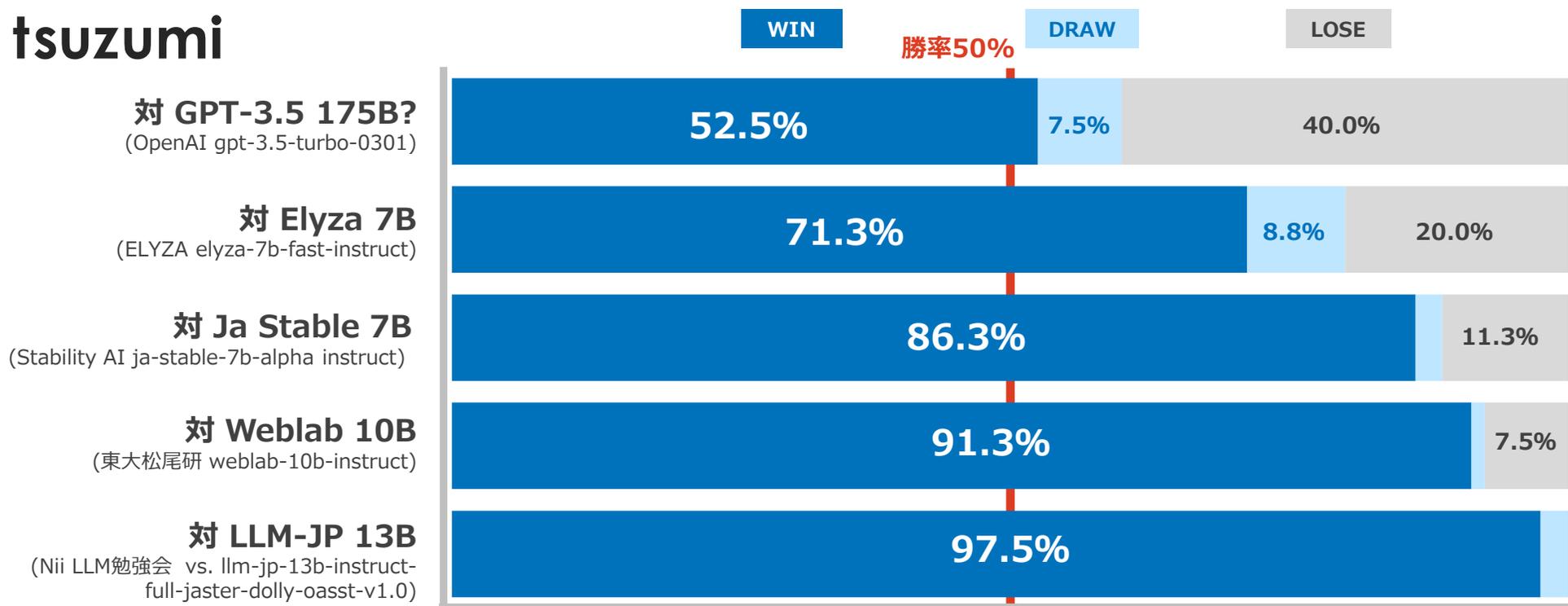
Chatbot

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



※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

Chatbot

Chatbot

高速な推論



- tsuzumi-7B: A100 40GBで 100 tokens / sec を超える

```
oi-at@frep:/bw-files1/llm-at/share/LLAMA-CPP$
```

I

小さい

邦楽器

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応える、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

コスト	◎
精度	△

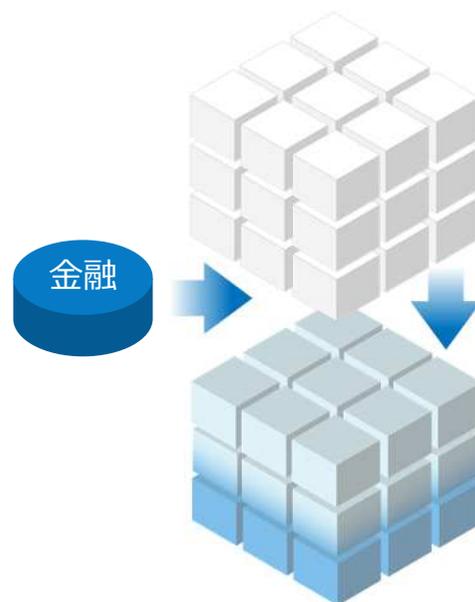
プロンプトに
金融関連情報を付加



tsuzumi

フルファインチューニング

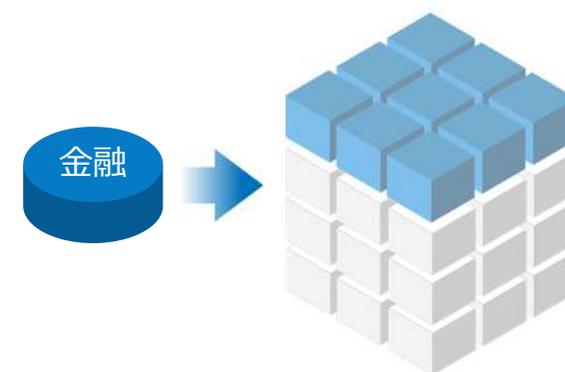
コスト	△
精度	◎



tsuzumi

アダプタチューニング

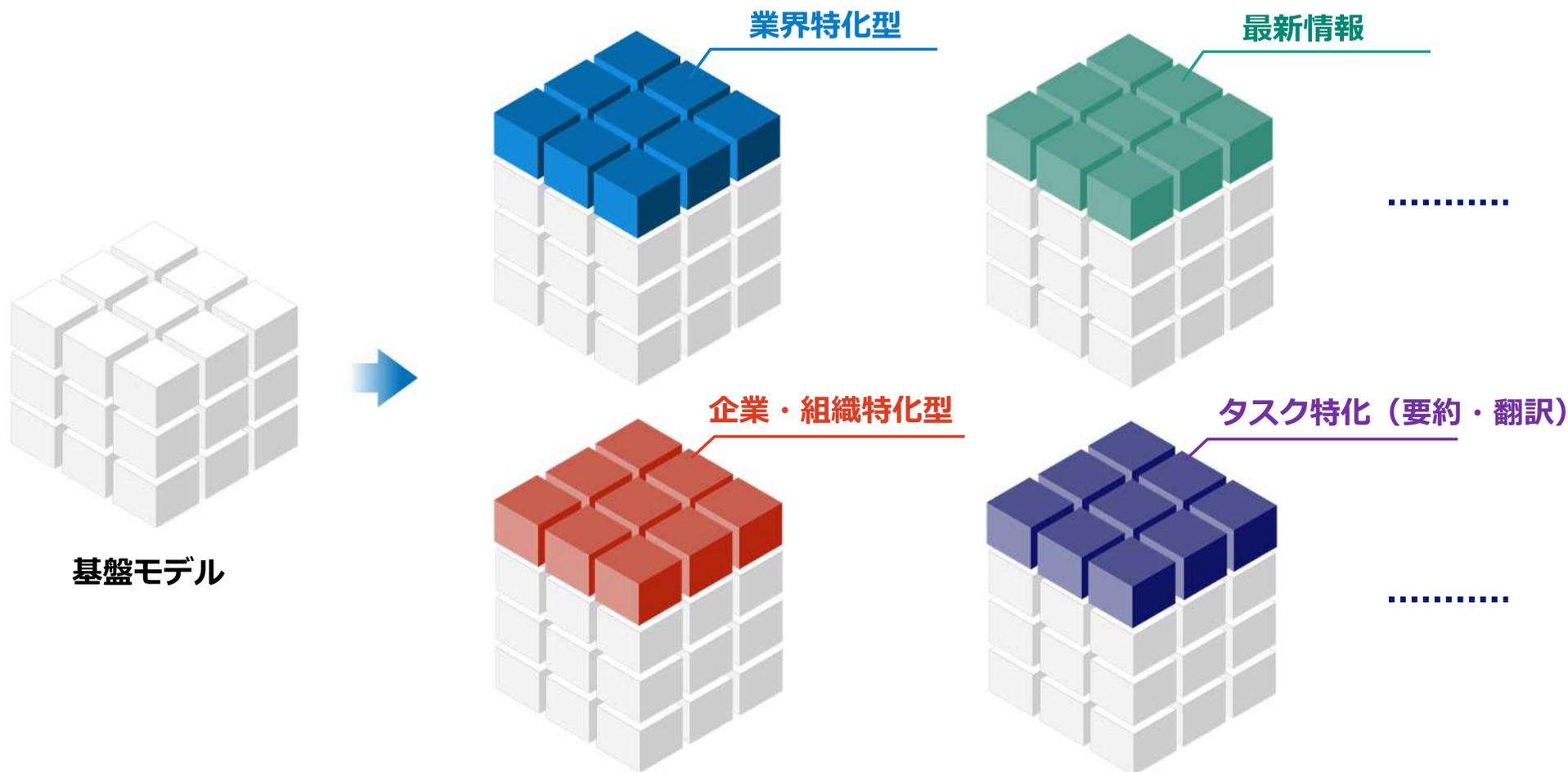
コスト	○
精度	○



tsuzumi

様々なチューニング

業界ごと、組織ごと、個人などカスタマイズを低コストで実現

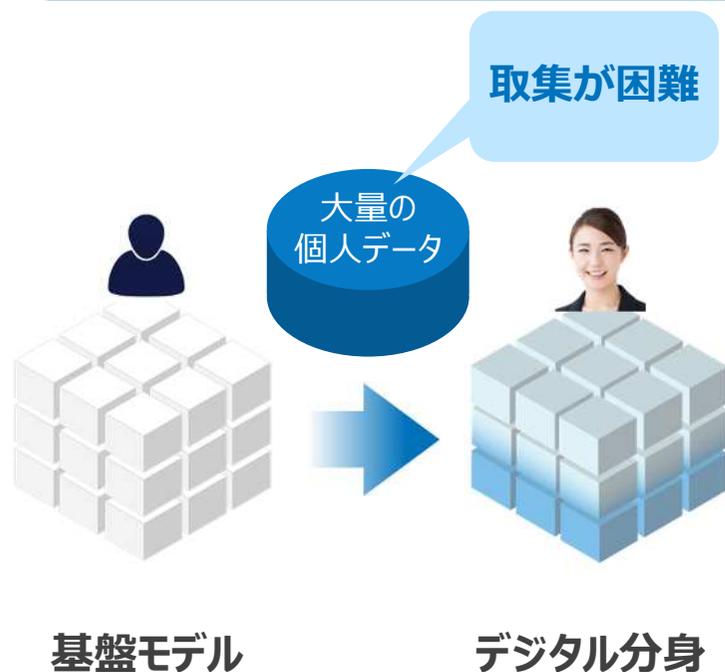


デジタル分身

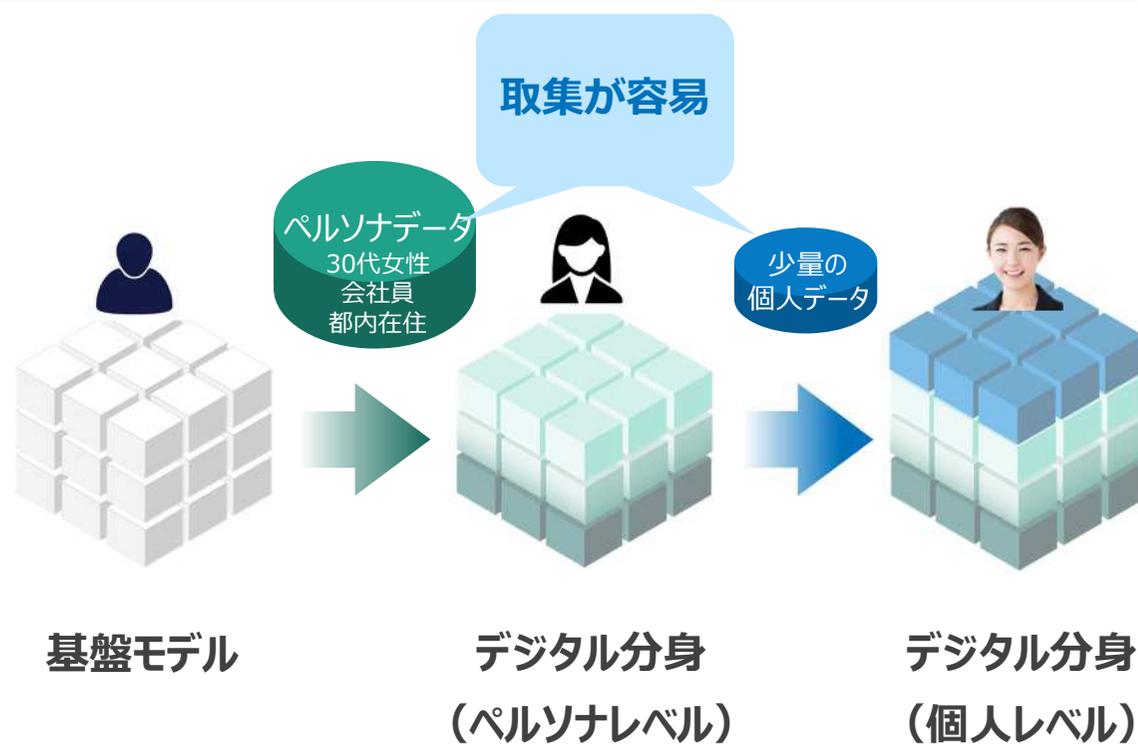


本人のデジタルツインを少量の個人データから生成

従来手法



個人性再現対話技術



小さい

邦楽器

調べ緒により調律が容易

特徴4 マルチモーダル



モーダル拡張 言語 + 視覚



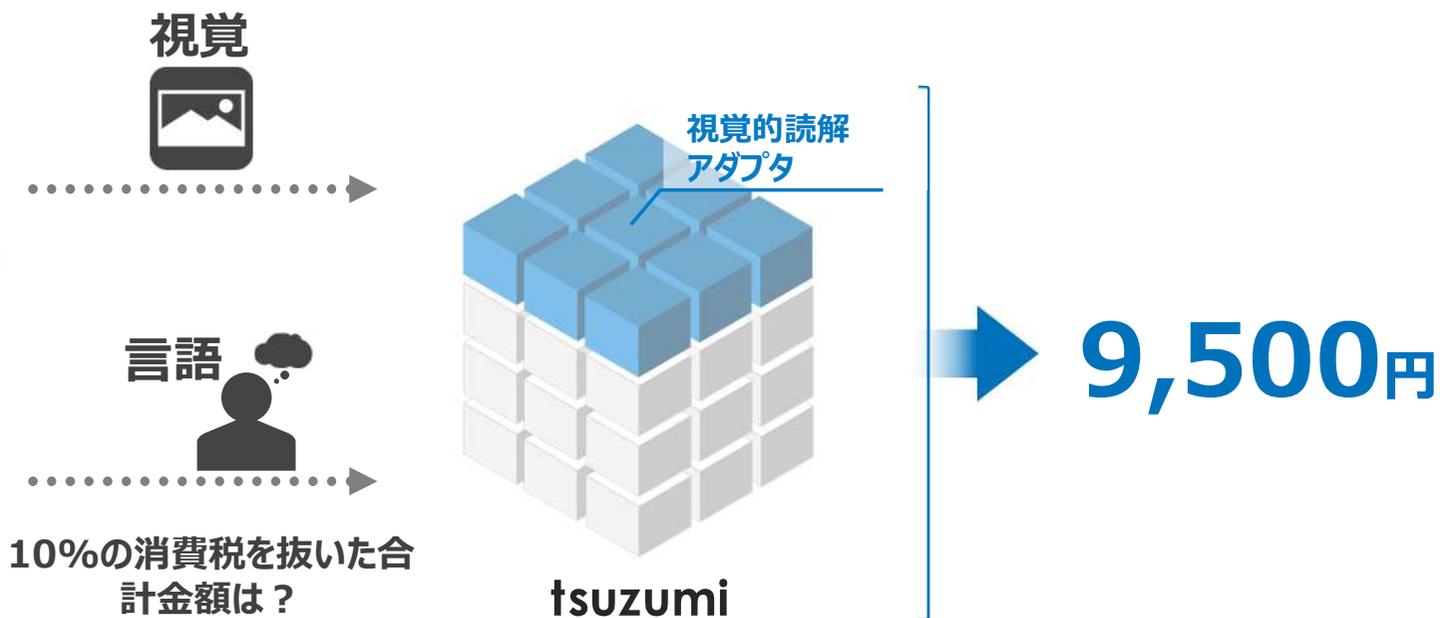
- ・ 言語による質問だけでなく、文書画像を提示しながらの質問が可能
- ・ 請求書や仕様書など画像付き文書を用いる業務やRPA業務に適用可能

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



モーダル拡張 言語 + 視覚

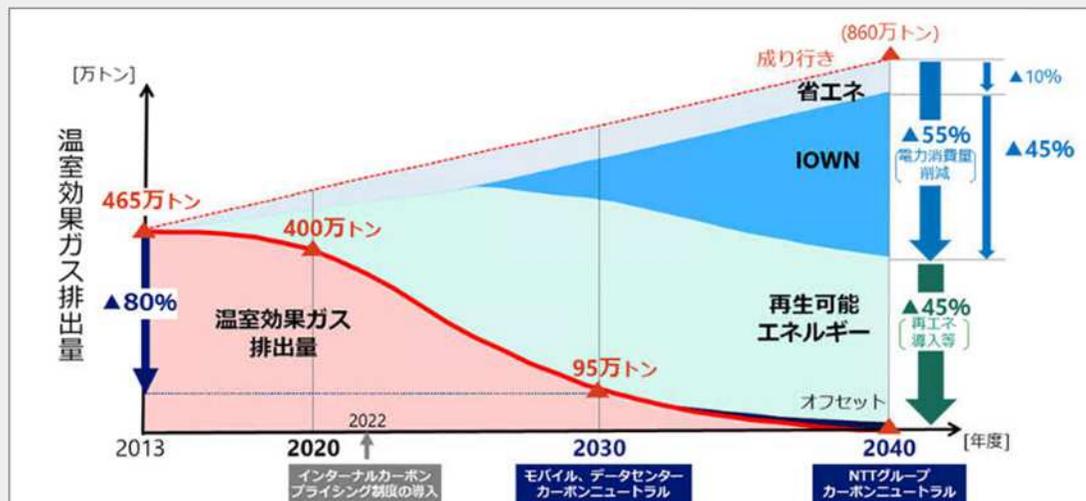


図1 NTTグループ温室効果ガス排出量の削減イメージ(国内+海外)

また、NTTグループでは、通信分野から様々な産業分野へのIOWNの普及・拡大や新たなサービスの提供、再生可能エネルギーの開発強化・導入拡大などにより、自らのカーボンニュートラル実現に向けた取組みを社会へ拡大し、日本政府の「2050年カーボンニュートラル宣言」の実現に貢献します。

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%

GPT-4を上回るNTTの視覚読解技術 (AAAI2024発表)

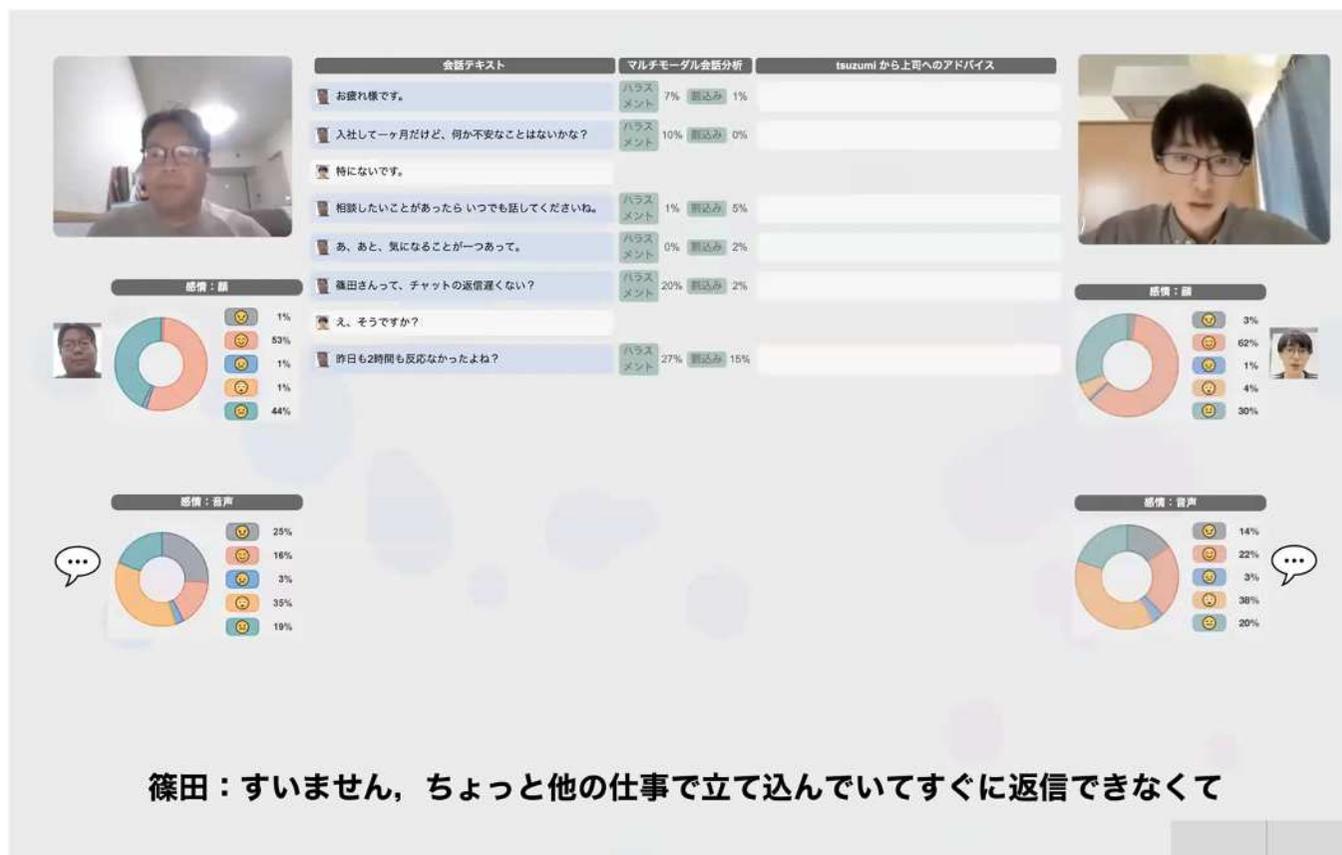


視覚読解ベンチマーク	CORD	InfoVQA	SlideVQA
	レシート質問 	インフォグラフィクス質問 	複数スライドにまたがる質問 
GPT-3.5	30.4	37.8	11.7
GPT-4V	69.4	49.9	21.0
tsuzumi	46.0	50.9	31.9

モーダル拡張 言語 + 視覚 + 聴覚



上司と部下のコミュニケーションをマルチモーダルに理解して、上司のパワハラ行動の改善を促します



モーダル拡張 言語 + 視覚 + 聴覚



上司と部下のコミュニケーションをマルチモーダルに理解して、上司のパワハラ行動の改善を促します

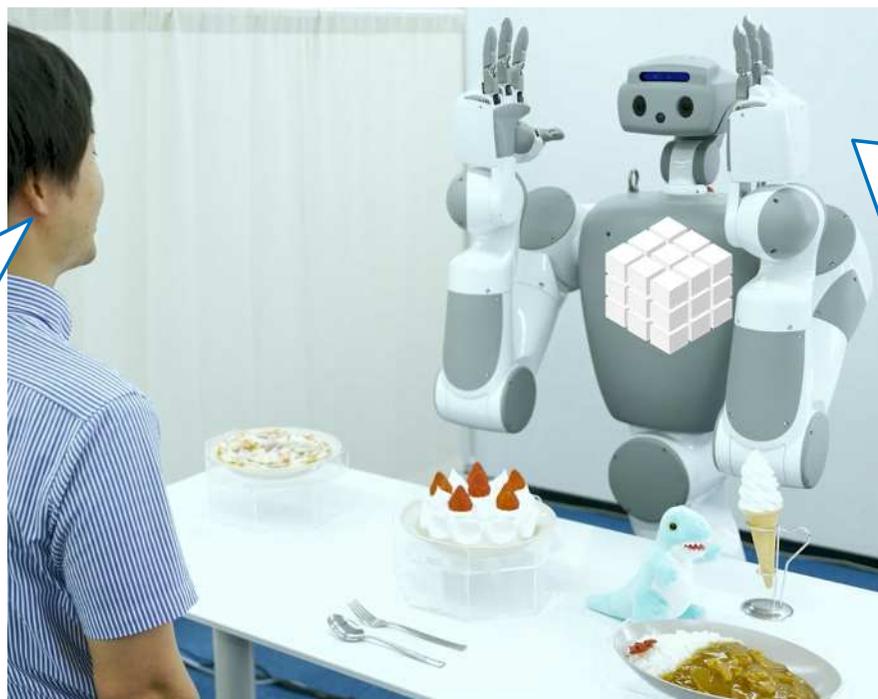


モーダル拡張 言語 + 身体性



- ・ tsuzumi搭載ロボットが、ユーザの要求に応じてメニューやテーブルの配置を考え、理由を説明しながら配膳を行う

寒い冬の日
に温まる
夕食の
テーブルを
作って
ください。
左利きを
意識して



1. 寒い冬の日**に温まるため**、カレーを選びます
2. **サラダ**は軽めの食事としてバランスが良いと思います。また、**季節感もある**ので春巻きと一緒に提供します
3. **お茶**は**体を温める飲み物**なので、この夕食には必要不可欠です。また、食器として箸も用意していますが、また、**左利きであることから**、箸とスプーンは右利きとは**反対の向きに置く**事で、おもてなしの心を伝えることができます。
4. 最後に**デザート**として**どら焼き**を提供し、満足度を高めます。全体的に和風な雰囲気を出したかったため、主食・主菜・副菜ともに日本料理中心となっております

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（TAACL, 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

tsuzumi

IOWNとの連携

IOWN

Innovative Optical and Wireless Network

NTTが2030年頃の実用化に向けて推進している次世代コミュニケーション基盤の構想、
IOWN (Innovative Optical and Wireless Network)

“光”デジタルテクノロジーによる限界打破

Networking

Computing

IOWN 1.0

DC to DC

IOWN 2.0

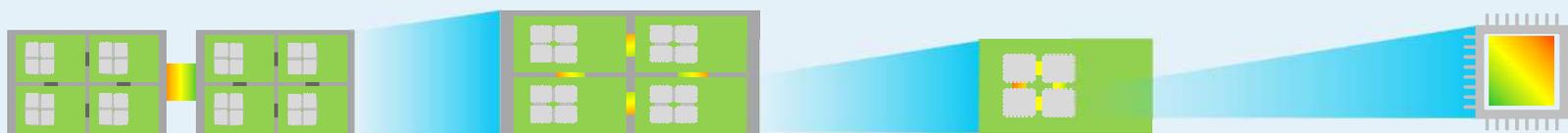
Board to Board

IOWN 3.0

Chip to Chip

IOWN 4.0

Inside Chip



性能目標

低消費電力

電力効率
100倍

大容量・高品質

伝送容量
125倍

低遅延

エンドエンド遅延
1/200倍

IOWN x LLM

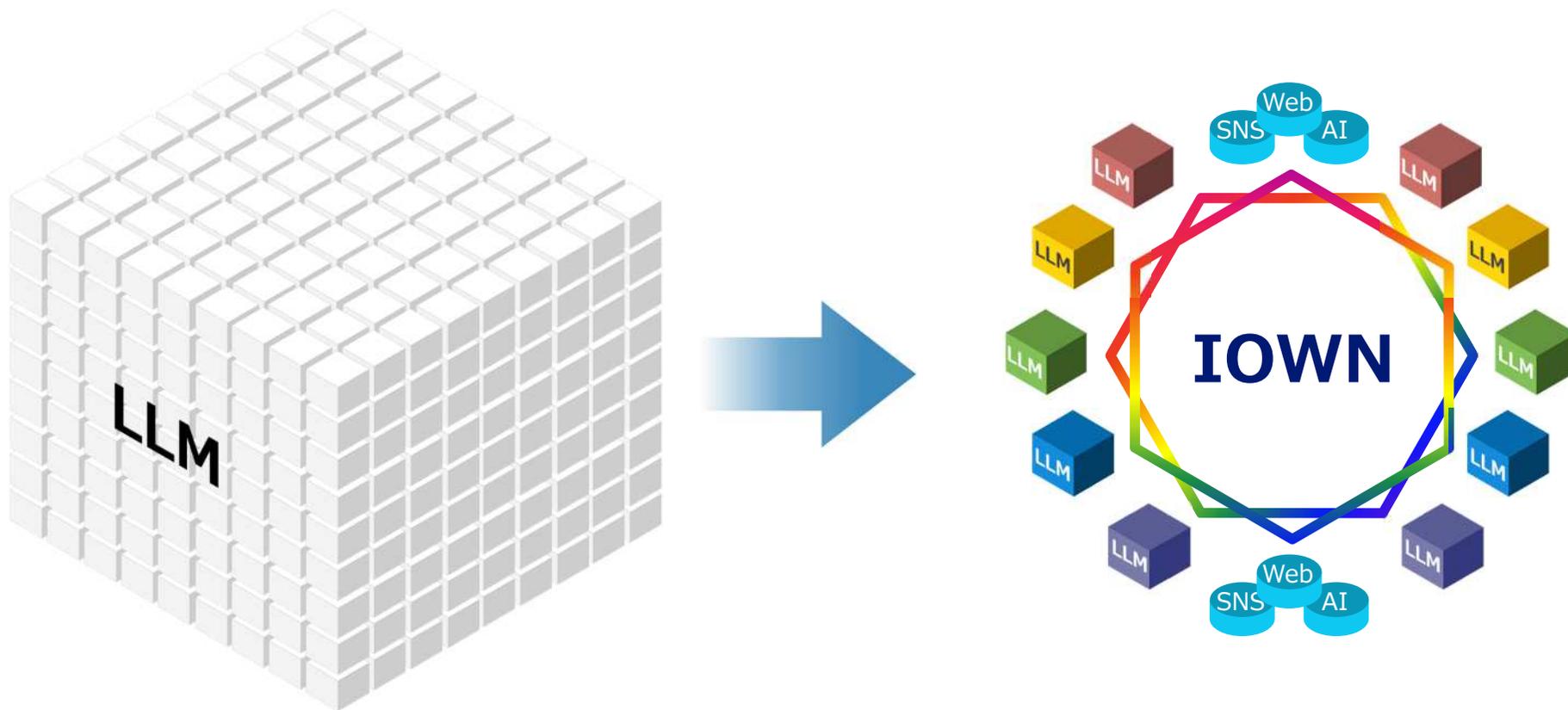


- ・学習データを手元に置いたまま、約100km離れたデータセンターのGPU利用
- ・ローカル環境と遜色のない安全かつ低遅延のLLM学習環境を実現



AIコンステレーション

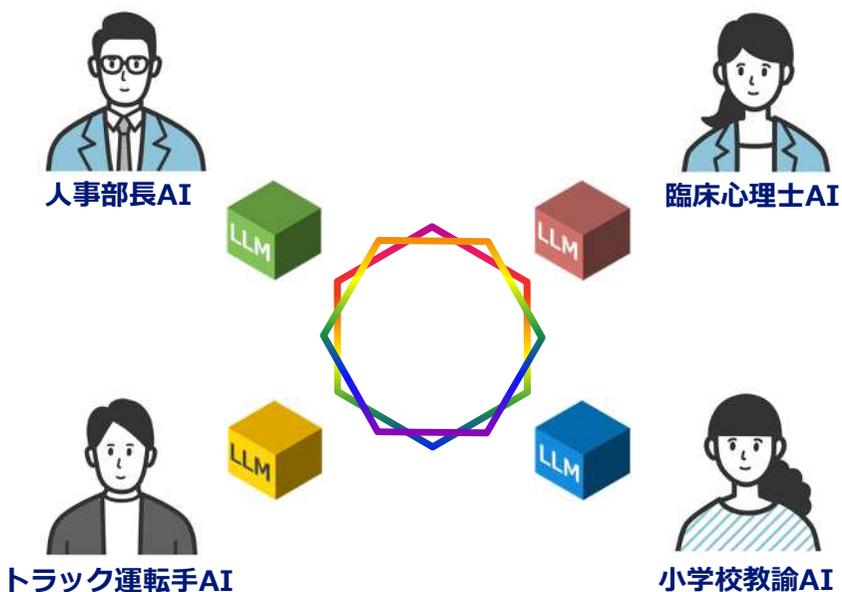
- ・何でも知っている1つの巨大なLLMではなく、専門性や個性をもった小さなLLMの集合知による社会課題解決
- ・大量のLLMの連携基盤としてIOWNが重要となる



大規模&高速な「まちづくりディスカッション」



人口が減っている我が地域の活性化には何が必要ですか？



臨床心理士
心の健康を考慮した地域活性化が必要だと思います。たとえば、心理的サポートを提供する施設やプログラムがあれば。

小学校教諭
それはいい考えですね。教育の面からも、心の健康は非常に重要です。子供たちが安心して学べる環境が必要です。

人事部長
人材育成にも関連しますね。心の健康と教育がしっかりしていれば、地域に貢献する人材が育つと思います。

トラック運転手
交通も大事ですよ。心の健康や教育施設に簡単にアクセスできるような公共交通が必要です。

臨床心理士
確かに、交通の便が良いと、多くのサービスが受けられますね。

小学校教諭
教育と交通が連携すれば、例えば学校バスが運行されると、子供たちが安全に学校に通えます。

Sakana AI社との資本・業務提携



地球にやさしく賢い次世代のLLM群のあるべき姿をめざす

sakana.ai ×  NTT



David Ha, CEO

- Google Brainの日本部門統括として研究を指揮
- Stability AI 研究リード

創業メンバ

2023年8月東京にて創業



Llion Jones, CTO

- ChatGPT等でも利用されている深層学習モデル
"Transformer"の開発者の1人

tsuzumi

ビジネス展開

NTT版LLM **tsuzumi**

既に開始したお客さまとの活用の取り組み

KU:P

京都大学医学部附属病院
KYOTO UNIVERSITY HOSPITAL



東京海上日動

Canon

モビリティ

生命保険

IT

メディア

飲料メーカー

官公庁

自治体

小売

問い合わせ160社超

…等

NTTデータ LITRON連携



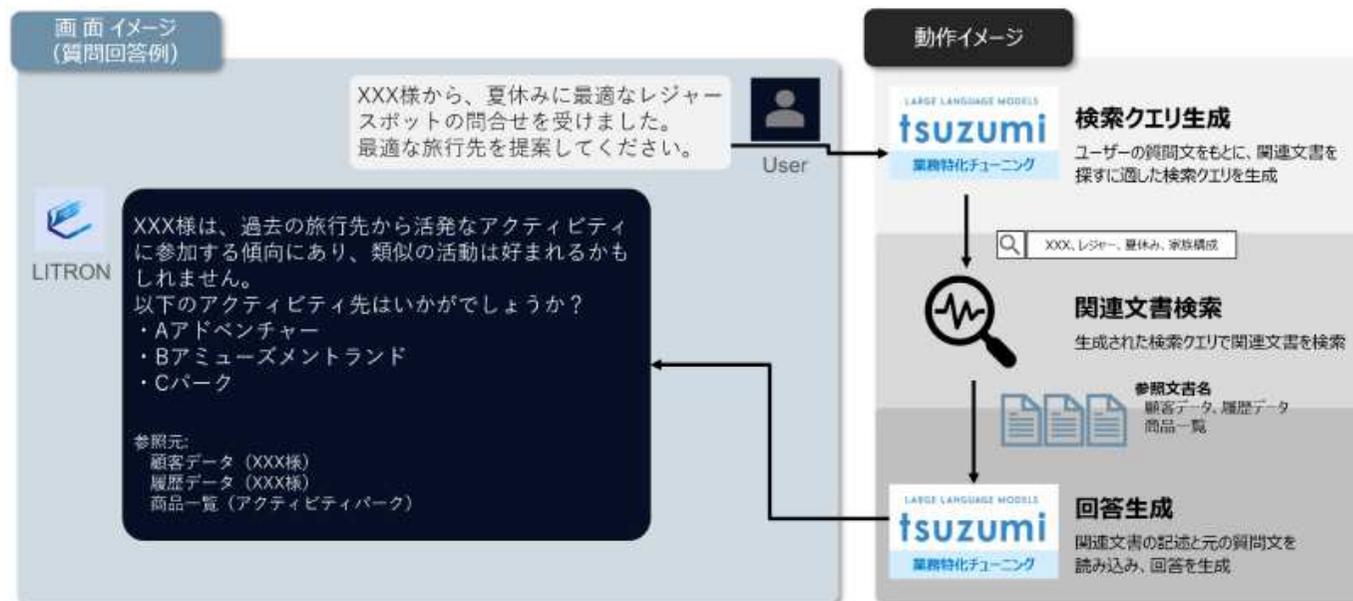
NTT DATA

ホーム > ニュース

NTT版LLM tsuzumiと連携したLITRON®の新サービスを提供開始

～閉域環境でセキュアに低コストで生成AI利用範囲を拡大～

報道発表



NTTテクノクロス ForeSight Voice Mining連携



新卒採用 : Webストア : サイトマップ : English 検索

HOME ソリューション・製品 企業情報 導入事例 ニュース イベント・セミナー コラム

HOME > ニュース >
コールセンターAIソリューション「ForeSight Voice Mining®」にNTT独自の大規模言語モデル「tsuzumi」を活用し、コールセンター業務プロセスの高度化を推進

プレスリリース

コールセンターAIソリューション「ForeSight Voice Mining®」にNTT独自の大規模言語モデル「tsuzumi」を活用し、コールセンター業務プロセスの高度化を推進

テキスト化された対応内容

【オペレーター】 お電話ありがとうございます。コールセンターの林でございます。
【お客様】 すいません、保険料の件で、ちょっと確認したいことがあるんですけども。
【オペレーター】 はい、かしこまりました。ではご契約状況の確認をさせていただきます。お客様はご契約者様ご本人様でいらっしゃいますか。ご契約者番号はお分かりになりますでしょうか。
【お客様】 はい本人です。契約書がちょっと今手元になくて、ないとだめですかね。
【オペレーター】 さようでございますか。ではお客様のお名前をフルネームでお聞かせいただけますか。
【お客様】 はい、ヤマダタロウと申します。
【オペレーター】 はい、かしこまりました。続きまして、生年月日とお住いの住所をお教えいただけますか。
【お客様】 はい。1969年の7月7日です。住所は横浜市西区みなとみらい4の4の5です。
【オペレーター】 はい。ヤマダ様、ご確認できました。保険料の、こういった内容のご確認でございますでしょうか。
【お客様】 はい、9月3日だったかな、接触事故で対応してもらったんですけども、車の修理で保険を使ったので保険料あがるんですね。3等級ダウンと聞いたのですが。
【オペレーター】 はい、保険料の増額があるかどうかですね。はい、保険料が変わりまして、月額で1,260円の増額になります。ただ今回の事故はヤマダさまの過失が小さいので、3等級ダウンではなく1等級のダウンです。
【お客様】 あ、そうなんですね。いつからですか。
【オペレーター】 はい、来月から増額になります。
【お客様】 わかりました。
【オペレーター】 他にご不明な点など、ございますでしょうか。
【お客様】 大丈夫です。ありがとうございました。
【オペレーター】 こちらこそありがとうございました。

項目ごとに整理された要約文

【主題】
保険料の増額についての問い合わせ

【本人確認】
お客様はご契約者本人であることを確認
お客様のフルネーム、生年月日、住所の確認

【お客様の要望】
9月3日の接触事故による保険料の増額についての説明と確認

【オペレーターの説明】
保険料は月額で1,260円増額される
保険料の増額は、お客様の過失が小さいため、1等級のダウンである
増額は来月から適用

【今後の対応】
保険料の増額は来月から適用される

