URL　https://journal.ntt.co.jp/article/26651
DOI　https://doi.org/10.60249/24065001

NTT版LLM「tsuzumi」

2023年11月に，NTT研究所が約40年にわたって蓄積してきた自然言語処理技術をベースにした大規模言語モデル（LLM：Large Language Models）「**tsuzumi**」を発表しました．**tsuzumi**は，特に日本語処理能力に優れており，小型軽量，テキスト以外のメディアも扱えるマルチモーダル対応といった特長を備えています．小型軽量の特長は，低消費電力化やオンプレミス利用を可能とし，また，マルチモーダルの特長は，写真や図表を理解した応答を可能とすることで，市中のLLMとは一線を画したユースケースを実現できます．本稿では，こうした特長を踏まえ，**tsuzumi**の全体概要を紹介します．

キーワード：#LLM，#**tsuzumi**，＃商用

清水（しみず）　健太郎
西田（にしだ）　光甫（こうすけ）
西田（にしだ）　京介（きょうすけ）

NTT人間情報研究所

LLMの適用領域を広げるtsuzumi

近年，ChatGPTをはじめとするさまざまな大規模言語モデル（LLM：Large Language Models）が登場し，注目されています．その多くは，より多くの知識を蓄え自然な受け答えを実現するために，非常に大きなサイズになっています．それだけに，モデル学習にかかるエネルギーは膨大で，例えば，GPT-3規模のモデルを学習するためには約1300 MWh（原子力発電所の1時間発電量相当）が必要になるともいわれています[(1)]．これだけ大きな規模になると，必要となるハードウェアをはじめとした初期コストや運用コストが膨大になるため，一企業が独自の言語モデルをそれぞれ作成するには無理があり，クラウドサービスとして提供されているLLMを利用することが一般的です．その一方で，企業が扱うデータには個人情報や機密情報を含むことが多く，安易にクラウド上にデータを入力することが困難なため，企業の財産ともいうべきデータをいかにしてLLMで有効活用するかが，課題となっています．

NTT人間情報研究所では，こうした課題を解決するために，小型軽量で優れた日本語処理能力を持つLLMをめざしてtsuzumiの研究開発に取り組み，2023年11月に発表しました．小型軽量であることのメリットは大きく，LLMの利用やチューニングに必要なハードウェアリソースや消費電力の低減により企業内でのオンプレミス利用の実現をもたらし，社外サーバでの保管が難しいデータに対してもLLMを適用し有効活用できるようになります．

小型軽量化すると，それだけLLMとしての性能は低下するのが一般的ですが，tsuzumiは研究所が40年以上にわたって蓄積してきた自然言語処理ノウハウにより，市中の同程度のパラメータサイズのLLMを上回る日本語処理能力を実現しました．さらに，柔軟なチューニングを実現するアダプタチューニングや，テキスト以外の図表を含む文書理解等を実現するマルチモーダルにも対応し，幅広い利用シーンへの適用が可能です．

以下，これらの特長について，詳しく説明します．

小型軽量化

LLMの規模は，パラメータサイズという値で示され，tsuzumiはパラメータサイズ7B（70億）と0.6 B（6億）の，軽量版と超軽量版の2つをリリースしています．

LLMのパラメータとは，モデルが学習中に獲得する知識やスキルを記憶するための変数です．

また，パラメータサイズとはモデルが持つパラメータの数を表します．パラメータサイズが大きいほどモデルの能力は高まる傾向があり，多数の知識の蓄積や，人間からのさまざまな指示に従った応答ができるようになっていきます．その一方で，パラメータサイズを増やせば増やすほど学習や推論に必要となる計算リソースや消費電力も大きくなるため，LLMとしての性能を維持しつつ，いかにしてパラメータサイズを減らすかが技術的なポイントになります．

一般的に，パラメータサイズが小さいLLMは，その分蓄積できる知識が少なくなり性能が低下します．このため，tsuzumiでは，学習データを作成するにあたり本来は学ばなくてよい冗長な情報や誤った情報といったノイズの除去や，一部の分野に偏らず幅広い分野の情報による学習データの作成など，学習データの質を向上するアプローチをとっています．また，超軽量版では，モデルが扱うドメイン・タスクを絞っていくことで性能を維持しつつさらなる小型軽量化に成功しました．

こうしたアプローチにより，市中の大規模LLMと比較して，tsuzumiのパラメータサイズは25分の1から300分の1となっており，大幅なコスト削減が可能となりました（**図1**）．

優れた日本語処理能力

LLMを学習する際には，学習データに含まれるテキストを「トークン」という，LLMが処理する単位に分割（トークナイズ）します．

近年のLLMでは，このトークナイザはテキストの集合から学習することで，語彙というトークンの総体を決定します．海外製のLLMは語彙に含まれる日本語用のトークンが少なく，テキストの多くが文字単位，あるいはバイト単位のトークンに分割されます．この場合，テキストを生成するために多数のトークンが必要となるため，テキストの生成速度の観点で非効率となります．

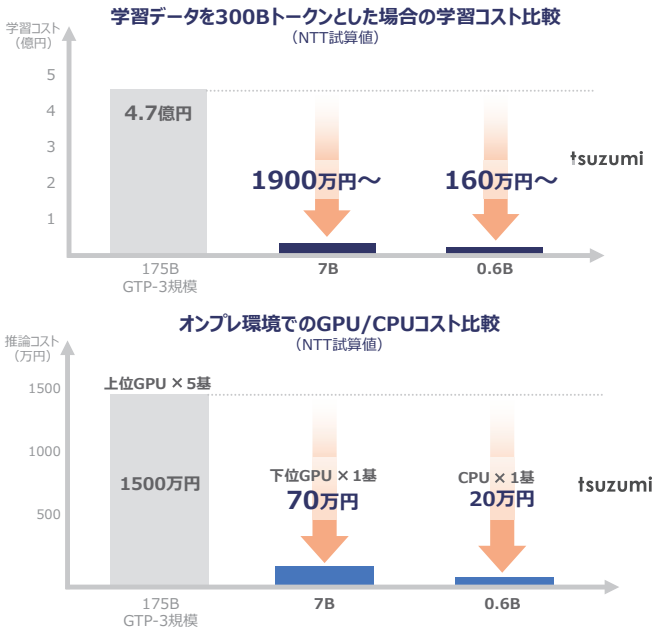


図1　軽量化とコスト削減の関係

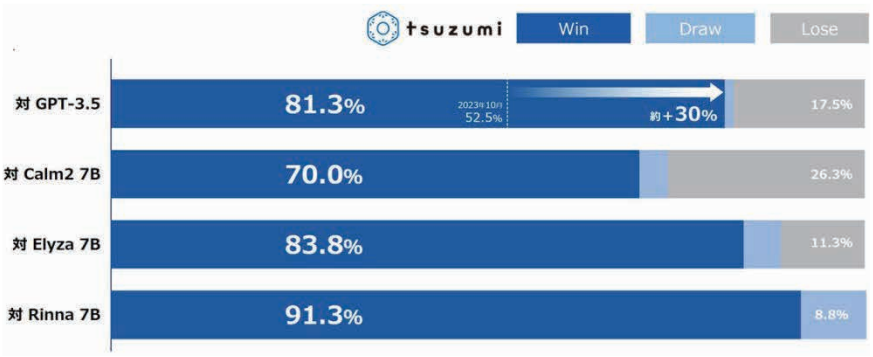


図2　市中LLMとの日本語処理能力の比較

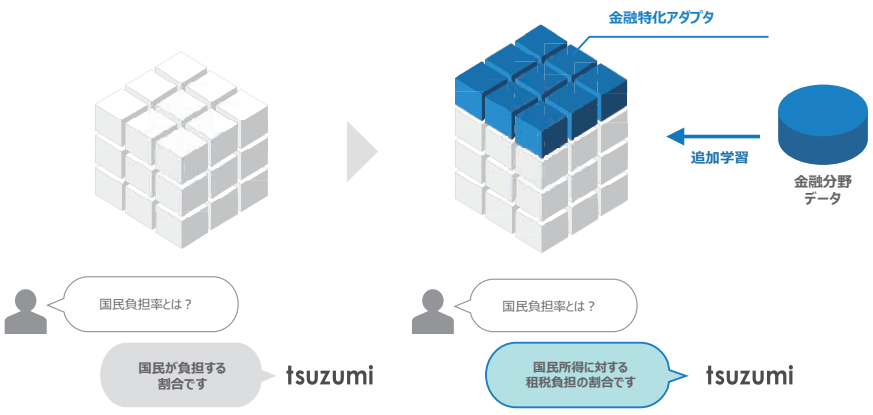


図3　アダプタチューニングのイメージ

そこで，日本語テキストを基にトークナイザを学習することで，効率的に日本語を生成できる語彙を持つトークナイザを作成します．ただし，一般的に用いられるトークナイザの学習アルゴリズムは日本語の構造を考慮しないため，学習コーパスにはよく出現するが他のテキストには出現しないような冗長なトークンが語彙に含まれやすくなる問題があります．

そこでtsuzumiでは，トークナイズに日本語の単語を考慮した独自の処理（単語制約）を取り入れ，この問題を解決しています．他社のトークナイザではWikipediaには頻出するが他のテキストではあまり出現しないような冗長なトークンが分割結果に含まれることがありますが，tsuzumiでは単語制約により日本語の構造を強く反映した分割が可能です．この単語制約を導入するにあたり，NTTが長年取り組んできた形態素解析ツールおよび辞書の整備の成果を活用しました．

トークナイズは，生成速度のみならず，テキストの理解の精度にも影響します．

前述した学習データの質向上と併せ，この独自のトークナイズにより，高い日本語処理能力を実現しているのがtsuzumiです．市中の他のLLMとの日本語処理能力を比較したRakudaベンチマークの結果を**図2**に示します．Rakudaベンチマークは日本に関する知識を問う質問から構成され，比較対象のLLMに同じ質問を与え，それぞれの回答のどちらがより優れているかをGPT-4に判断させるというもので，日本語LLMの性能比較手段としてよく用いられています．

この結果から，tsuzumiは，tsuzumiと同等規模のLLMに対してはもちろんGPT-3.5に対しても，非常に優れた日本の知識および日本語処理能力を有していることが分かります．

アダプタチューニング

tsuzumiは，領域特化した小型のモデルを元のモデルに追加する（アダプテーションする）チューニングであるアダプタチューニングに対応しています．アダプタチュー

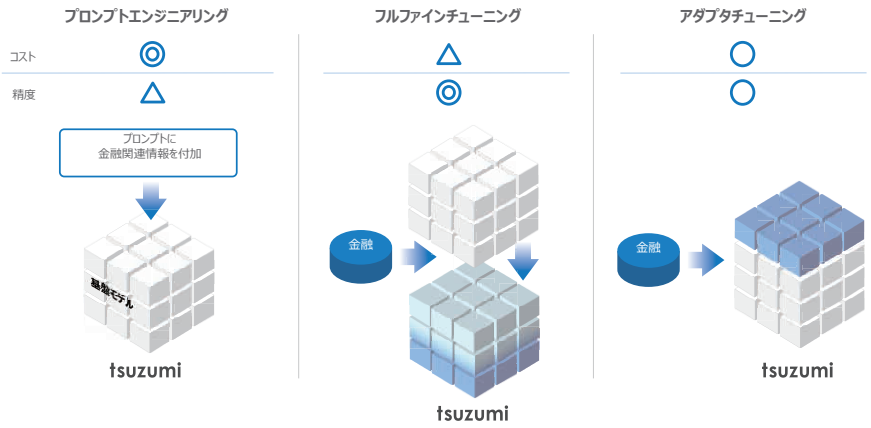


図 4　チューニング方法の比較

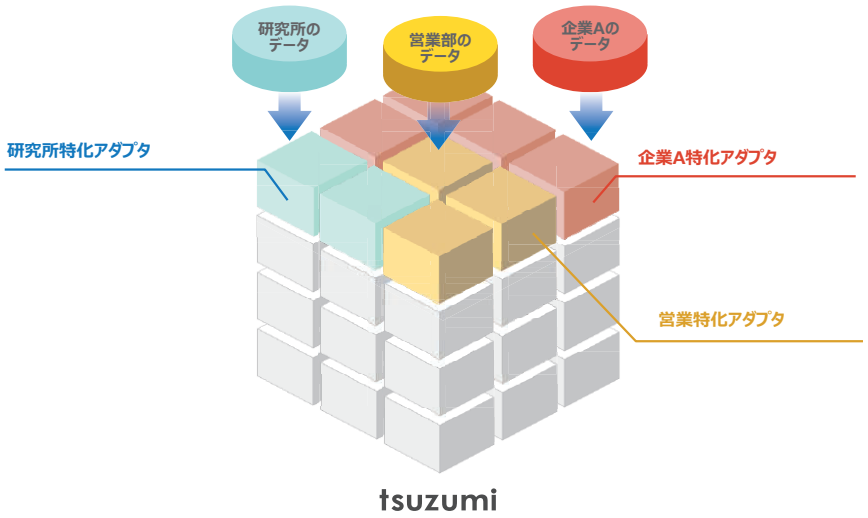


図 5　マルチアダプタ：利用シーンに応じてアダプタの切替や組合せが可能

ニングのイメージを**図 3** に示します．

チューニングにはほかに，推論時に質問と同時に関連情報を与えるプロンプトエンジニアリング，モデル全体を更新するフルファインチューニングがあります．これらの特徴をまとめたものを**図 4** に示します．

コストと精度の面でそれぞれのチューニング方法に特徴がありますが，アダプタチューニングはその両面を一定レベルで両立できるメリットがあります．また，企業内でのユースケースを考えた場合，組織ごとに異なるアダプタを用意することで，簡単にそれぞれの組織に特化したモデルの作成が可能です．

将来的には，**図 5** に示すようなマルチアダプタに対応し，1 つの tsuzumi の上で複数のアダプタが連携し，さまざまなシーンでの要求に対応できるようにすることを想定しています．

マルチモーダルへの対応

LLM はその名のとおり言語を操るモデルのため，画像や音声の入出力は想定していません．しかしながら，一般的な文書はテキストだけでなく図表を含むことが多く，またその図表が大きな意味を持つ場合が多々あります．tsuzumi は，こうした図表理解[2, 3]をはじめ，音声や状況を理解する新しい LLM をめざしており（**図 6**），その事例を2023年の NTT R&D フォーラムで展示しました．

本特集記事『グラフィカルな文書を理解できる「tsuzumi」』[4]にて，マルチモーダルの 1 つである視覚読解技術について解説していますので，そちらも併せてご覧ください．

今後の展望

ここまで述べてきたように，tsuzumi は小型軽量でありながら，優れた日本語処理能力やチューニングの柔軟性，マルチモーダルといった多様な特長を備えていますが，これからも以下に述べるような観点での研究開発を進め，さらに発展していく予定です．

■多言語対応

現状の tsuzumi は，日本語以外に英語にも対応しています．今後，日英の言語処理性能のさらなる向上をめざすとともに，英語以外の言語（例えば，中国語，韓国語，フランス語，ドイツ語など）への対応にも取り組み，世界中に tsuzumi を発信していきます．

また，人間が操る言語だけでなく，プログラミング言語への対応にも取り組んでいます．例えば，仕様書の内容を理解し，仕様どおりのソースコードを指定された言語で出力できるようになれば，ソフトウェア開発の稼働削減に大きく貢献します．現状でも一定程度のソースコード出力は可能ですが，さらなる学習用データ収集とモデル学習を進め，性能向上をめざします．

■中型版

2023年11月の発表では，超軽量版，軽量版のほかに，中型版もリリース予定であることを明らかにしました．中型版は，パラメータサイズは軽量版 7 B のおよそ倍となる13 B で，多言語対応はもちろんより多くの知識を蓄え，LLM としての性能向上を図る予定です．

パラメータサイズが大きくなるため，軽量版よりも要求するハードウェアリソース等は大きくなりますが，LLM を限られたハードウェアリソースで効率的に動作させるための技術（量子化）についても検証を進め，要求リソースを低減していく予定です．

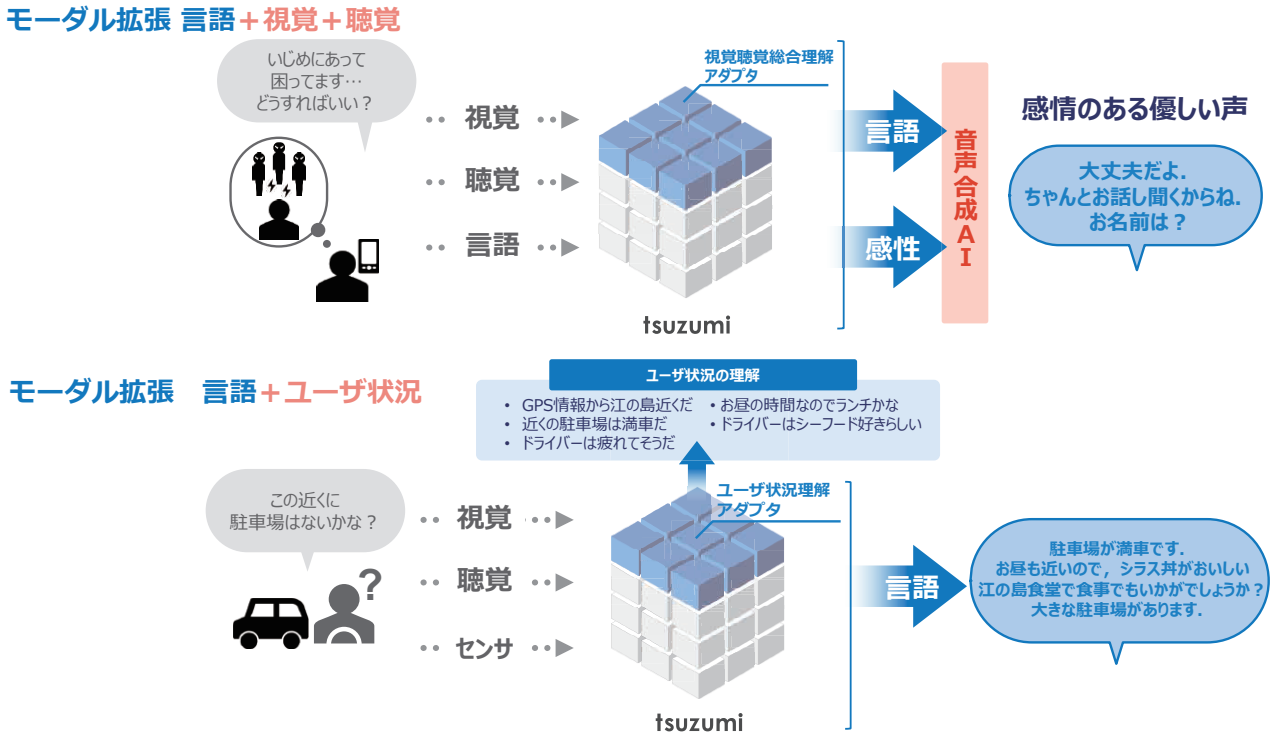


図6　マルチモーダルのイメージ

■安心・安全への対応

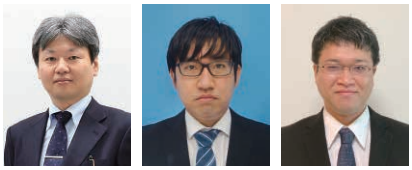
現在，LLMをはじめとする生成AI（人工知能）について，安全面や倫理面での議論が非常に多くかわされています．制度的な面では，学習データに著作物や個人情報が入っていた場合どうするのかといったことで，海外では訴訟に発展したケースもあります．日本国内でも活発に議論されていますが，まだ明確な結論は出ていません．技術的な面でも，例えば人権侵害にあたるような回答をさせようとする不適切な質問に対して，しっかりと拒否する仕組み等が必要になります．

これらの課題に対して，より質の高い学習データの整備はもちろんですが，制度面，技術面での検討を進めている他研究所や他組織とも連携しながら，より安心・安全にtsuzumiを利用できるよう検討を進めていく予定です．

tsuzumiはこれからも発展していきます．今後のtsuzumiにご期待ください．

■参考文献

(1) 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 AAAI-24, Feb. 2024.
(3) T. Hasegawa, K. Nishida, K. Maeda, and K. Saito：“DueT：Image-Text Contrastive Transfer Learning with Dual-adapter Tuning,” Proc. of EMNLP 2023, pp. 13607-13624, Dec. 2023.
(4) 田中・壹岐・長谷川・西田：“グラフィカルな文書を理解できる「tsuzumi」,” NTT技術ジャーナル，Vol. 36, No. 6, pp. 14-17, 2024.



（左から）　清水　健太郎／　西田　光甫／
西田　京介

「tsuzumi」という名称は，雅楽で用いられる鼓が演奏の流れを統率する役割を担っていることになぞらえ，これからの自然言語処理技術の発展をリードしていく存在をめざす，という意味を込めて付けました．

◆問い合わせ先

NTT人間情報研究所
思考処理研究プロジェクト
E-mail　tsuzumi-nttlab @ ntt.com