（報道発表資料）

2023 年 11 月 1 日
日本電信電話株式会社

NTT 独自の大規模言語モデル「tsuzumi」を用いた商用サービスを 2024 年 3 月提供開始

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、大規模言語モデル（LLM※1）の普及に伴い課題となっている電力やコスト増加などの課題解決に向け、軽量でありながら世界トップレベルの日本語処理性能を持つ大規模言語モデル「tsuzumi※2」を活用し、NTT グループ発の商用サービスとして 2024 年 3 月に提供開始いたします。

「tsuzumi」は、2023 年 11 月 14 日～17 日に開催する「NTT R&D FORUM 2023 ― IOWN ACCELERATION※3」の IOWN Pickup の展示ブースにてご覧いただけます。また、本フォーラム内の基調講演・特別セッションにおいても具体的な取り組み・展望などを紹介いたします。

1. 提供の背景

近年、ChatGPT を始めとする大規模言語モデル（LLM）に大きな注目が集まっておりますが、これらは膨大な知識をモデル内に有することで高い言語処理性能を示す一方、学習に要するエネルギーは原発1基1時間分の電力量が必要（GPT-3 のケース）とも言われており、また、運用には大規模な GPU クラスタを必要とし様々な業界に特化するためのチューニングや推論にかかるコストが膨大であることから、サステナビリティおよび企業が学習環境を準備するための経済的負担面で課題があります。

こうした課題を踏まえ、NTTでは、研究所が保有する 40 年以上に及ぶ自然言語処理研究の蓄積、世界トップレベルの AI 分野の研究力を活かし、軽量でありながら世界トップレベルの日本語処理性能を持つ大規模言語モデル「tsuzumi」を活用し、NTT グループ発の商用サービスとして 2024 年 3 月に開始します。

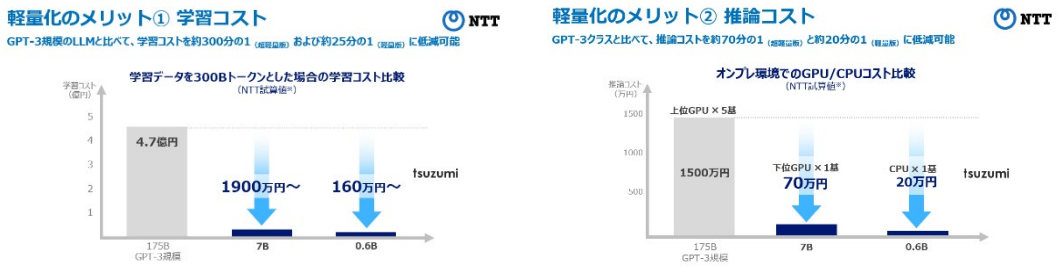
提供開始に向け 2023 年 10 月から先行して、メディカル分野の京都大学医学部附属病院様やコンタクトセンタ分野の東京海上日動火災保険株式会社様などのパートナー様とトライアルを開始しています。

2. tsuzumi の特長

<1> 軽量な LLM

- パラメタサイズが 6 億の超軽量版と 70 億の軽量版の「tsuzumi」を開発。
- Open AI 社の GPT-3 の 1750 億パラメタと比べ、およそ約 300 分の1（超軽量版）および 25 分の1（軽量版）と軽量。

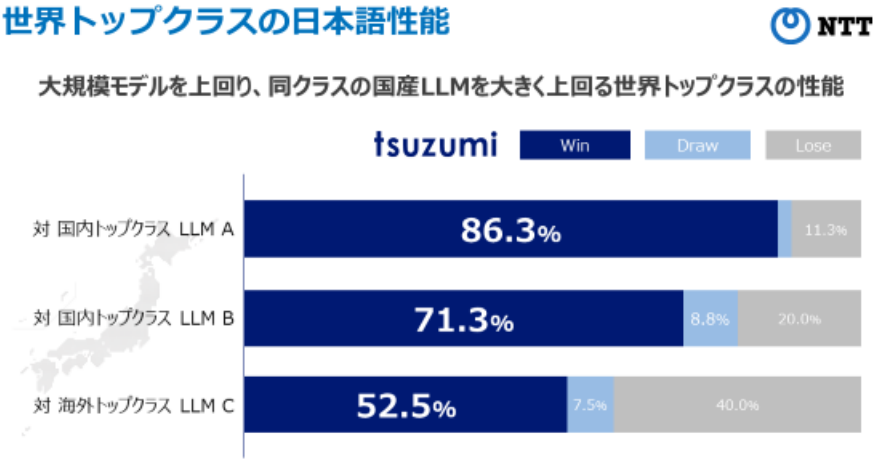
・　軽量版は 1GPU で、超軽量版は CPU で高速に推論動作可能であり、チューニングや推論に必要なコストを抑えることが可能(図 1)。
GPUクラウドの利用料金に換算すると、学習コストを約 300 分の1（超軽量版）および 25 分の1（軽量版）、推論コストを約 70 分の1（超軽量版）および 20 分の1（軽量版）に低減可能。　＊NTT試算



（図 1）軽量によるコストメリットの例※4

<2> 日本語と英語に対応 ～特に日本語が得意な LLM～

・　「tsuzumi」は日本語と英語に対応しており、特に日本語処理性能については長年の研究で得た知見を活かすことで、高い性能を具備。
生成 AI 向けのベンチマークである Rakuda では GPT-3.5 や国産トップの LLM 群を上回ることを確認（図 2）。
英語でも、世界トップクラスと同程度の性能を実現しており、多言語にも今後対応。



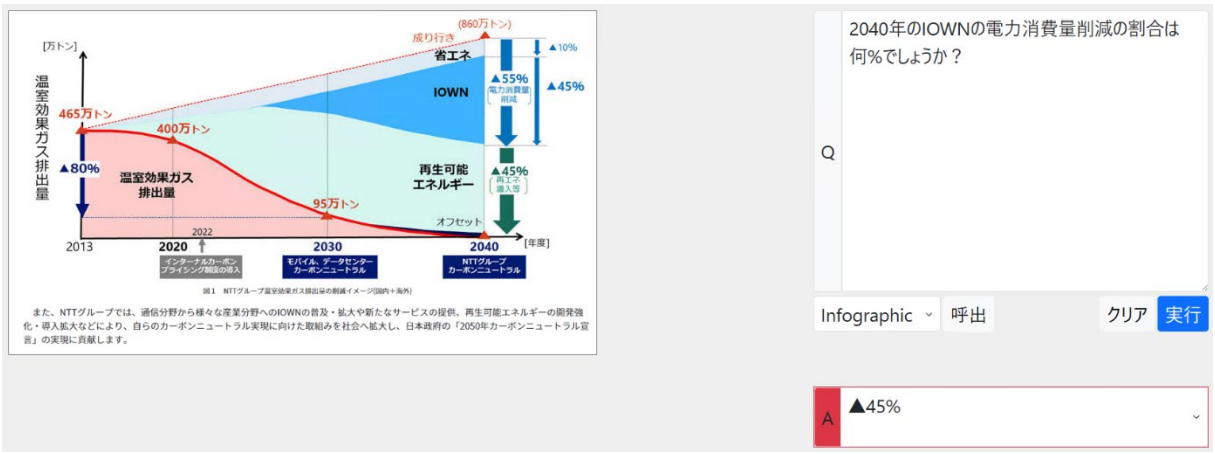
（図 2）日本語性能に関する tsuzumi と他 LLM との対戦結果※5

<3> 柔軟なチューニング　～基盤モデル＋アダプタ～

・　効率的に知識を学習させることのできるアダプタ※6 により、例えば特定の業界に特有の言語表現や知識に対応するようなチューニングを、少ない追加学習量で実現可能。

<4> マルチモーダル※7 ～言語＋視覚・聴覚・ユーザ状況理解～

・　言語化されていないグラフィカルな表示や音声のニュアンス、顔の表情などを理解し、現実世界での人との協調作業を可能とするような、マルチモーダルへの対応を予定(図 3)。



（図 3）視覚的読解技術の実施例

3. tsuzumi が志向する方向性

まずは業界に固有なデータを柔軟・セキュアに学習することが可能となる点を生かし、業界に特化した領域にフォーカスしてまいります。

全ての知識を集約した 1 つの巨大な LLM が存在するのではなく、専門性や個性をもった小さな LLM の集合知が多種多様な AI 群と連携してリアルワールドの社会課題を解決する世界をめざします。

こうした大量の LLM の連携基盤にはローカル環境と遜色のない安全かつ低遅延環境が必要になりますが、NTT では tsuzumi の学習のために IOWN APN(All Photonics Network)を利用した環境を構築しました。これにより数百 km 離れたデータセンタ間で GPU とストレージを接続し、安全かつ性能低下の非常に少ない LLM 学習環境を実現しています。

4. 今後の展開

商用サービス提供後もチューニング機能の充実やマルチモーダルの実装についても順次展開してまいります。また、サイバーセキュリティ分野への応用、自律的に連携し議論する AI コンステレーション等の開発を進めます

これらにより NTT は新たな価値創造、お客さま体験の高度化に向けた取り組みをより一層加速してまいります。

【用語解説】

※1 LLM:　大規模言語モデル（Large Language Models）：大量のテキストデータを使って学習された言語モデルで、言語の理解や文章の生成に優れた能力をもつもの。

※2「tsuzumi」は商標出願中です。日本語の処理性能を重視し、産業の発展を牽引する言語モデル技術への期待を、雅楽の合奏の開始の切っ掛けを担う鼓に寄せました。

※3 「NTT R&D FORUM 2023 -IOWN ACCELERATION」では、2023 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 技術を中心とした 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わかりやすくご紹介します。

URL：https://www.rd.ntt/forum/

※4 コスト試算条件

（学習コスト）

- Llama-1 7B の学習時 82,432 GPU-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

（推論コスト）

- 量子化: 16 ビット
- 必要 GPU メモリサイズ：パラメタ数 x 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

※5 日本語性能の評価方法

rakuda ベンチマーク：https://yuzuai.jp/benchmark
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

※6 アダプタ:　事前学習済みモデルの外部に追加されるサブモジュール。ファインチューニングの際に事前学習済みモデルのパラメタを固定したままアダプタのパラメタのみを更新することで、計算コストの高いベースモデルの再学習を行わずに知識を学習することができる。

※7 マルチモーダル：AI への入力情報の種類（テキスト、画像、音声など）をモーダルと言い、これらの異なる入力情報を組み合わせて使う能力をもった人工知能の特性を指す。

■NTT 版大規模言語モデル「tsuzumi」の解説情報
https://www.rd.ntt/research/LLM_tsuzumi.html

■本件に関する報道機関からのお問い合わせ先
日本電信電話株式会社
広報部門
ntt-pr@ntt.com