URL https://journal.ntt.co.jp/article/33345
DOI https://doi.org/10.60249/25041201

NTT西日本
デジタル革新本部　技術革新部

福田航平　Kohei Fukuda

tsuzumiの活用に向けて検証，お客さま案件支援，そして社内各部門の自力活用をめざした研修に取り組む

2023年11月にNTT研究所が大規模言語モデル（LLM）「**tsuzumi**」を発表しました．**tsuzumi**には，「軽量」「世界トップレベルの日本語処理性能」「柔軟なチューニング」「マルチモーダル」という特徴があり，2024年３月に商用サービスが開始されました．こうした中で，**tsuzumi**は各方面から注目を集めており，NTTグループ各社においても，**tsuzumi**を活用した事業展開，お客さまにおける導入・活用等に向けてさまざまな活動が活発化しています．NTT西日本では，デジタル革新本部技術革新部が中心となって**tsuzumi**の展開に取り組んでおり，そのキーパーソンである福田航平氏に，**tsuzumi**の活用に向けた取り組みと導入事例，**tsuzumi**の可能性を確認する検証，そして発想の転換と自らの手で実施することの重要性を伺いました．

tsuzumiを自ら活用するための研修と，お客さま案件における導入支援

現在，手掛けている業務の概要をお聞かせいただけますか．

NTT研究所の大規模言語モデル（LLM）「tsuzumi」をNTT西日本グループのビジネスへの活用をめざして，tsuzumi研修の実施，tsuzumiの社外案件の支援，tsuzumiに関する各種検証に取り組んでいます．

tsuzumi研修については，NTT西日本グループ内組織横断で生成AI（人工知能）に関する取り組む仕組みを構築してその中で行っています．NTT西日本では，tsuzumiをはじめとする生成AIに関する取り組みを技術革新部が中心に行ってきたのですが，生成AIのお客さまへの提案を含む事業での活用を推進していくためには，それぞれの部門において自ら生成AIを扱うことができるような環境構築と人材育成を行う必要があり，そのために組織横断で生成AI活用に取り組む体制を構築し，活動をしています．研修は，生成AIやtsuzumi，LLM等の基本的仕組みのほか，ユースケースの事例，導入手順等に関する研修，そして実環境におけるファインチューニング実習まで実施しています．その後の精度向上に向けたチューニングについては，学習のためのデータ等がユースケースにより異なるため，各部門において研修終了後の受講生が行うこととしています．

tsuzumiの社外案件の支援は，山口県様と三重大学医学部附属病院（三重大病院）様の案件の支援のほか，法人営業部門等からの個別案件に対して，活用支援を行っています．山口県様の案件は，行政DX（デジタルトランスフォーメーション）や生成AIに関する技術・ノウハウを活かし，機微データを扱う自治体業務への本格的な展開も見据えた実証実験です．機微なデータを扱うためにオンプレミス環境において小型のGPU（Graphics Processing Unit）サーバにtsuzumiを実装し，業務に特化したチューニングを行い動作させ，業務上の機微なデータを扱う業務の対応記録の要約・校正，各種業務マニュアルの検索・要約等，庁内の実データ活用を想定した実証を行っています．

三重大病院様の案件は，病院での業務におけるtsuzumiの実用性と適用可能性の評価です．医師の事務作業の１つである，患者の入院期間中の診療経過をまとめた「退院サマリ」（年間約1.5万件）の作成過程において，tsuzumiが生成した電子カルテの要約文章を活用することで，作成業務の効率化を検証しています．電子カルテデータを基に医師自らが文章を作成するのではなく，tsuzumiが生成した文章を医師が確認・修正することで，作成に要する時間を短縮しつつ，最終的なアウトプットの質を落とさない業務フローの実現をめざしています．

このほかにも，人事業務において業務実施状況から社員のスキ

ルレベルを可視化するなど，多様な活用方法についてご相談をいただいており，実用的なユースケースは社内活用を通じて技術を磨くことにも取り組んでいます．

tsuzumiの可能性を確認するための検証

tsuzumiの検証に関して具体的にどのように検証を行ってきたのでしょうか．

検証は，外部情報を検索して生成を行う技術である「RAG（Retrieval-Augmented Generation）の精度向上」と新しい技術としての「視覚読解RAG」「MoE（Mixture of Experts）」について実施してきました．

RAGは，「情報を検索・取得するRetrieve」「Retrieveで取得した情報と質問を組み合わせてLLMに入力する指示文であるプロンプトを作成するAugment」「Augmentで生成したプロンプトから回答を生成するGenerate」の3つのフェーズで構成されています．RAGの精度向上はこれまで継続して実施してきていますが，新たに文書検索手法の改善とプロンプトエンジニアリングの工夫を行いました（**図1**）．文書検索手法には主として，質問文の単語と参照文書内の単語の一致から検索するもっとも基本的な手法である「キーワード検索」，参照文書の内容をベクトル化して保存しておき，質問文をベクトル化した値の類似度を算出する手法である「ベクトル検索」，全文検索，ベクトル検索を両方実施し，共に一致度が高い文書が上位にくる検索手法である「ハイブリッド検索」，そして，AIモデルを用いて直接類似度を計算する手法である「リランキング」の4タイプがあります．それぞれの検索手法を組み合わせた特徴を評価し，案件ごとに最適な手法を選択してきました．

プロンプトエンジニアリングにおいては，NTT研究所推奨の検索結果入力方法（プロンプト作成方法）を参考にNTT西日本独自のプロンプトを作成しました．さらに，検索結果を上位のものだけに絞って参照することで，要約精度が大幅に向上しました．今回の結果から，tsuzumiが参照する文書を，いかに回答に必要な情報だけに絞れるかが精度向上のカギとなるのではないかと考えています．

新しい技術としての「視覚読解RAG」「MoE」の検証はいかがでしょうか．

今までのRAGユースケースでは基本的にテキストだけを扱い，図表やフローチャートは対象外としています．しかし，実際の文書においては図表等のほうが重要な情報が入っているケースが多く，RAGを実用化するにはそれらの解釈は必須となります．さらに，テキスト情報を取得したとしても，フォーマットが崩れるため使い物になりません．「視覚読解RAG」は，tsuzumiの視覚読解機能を活用して，図表を対象にしてもRAGによる情報取得がどこまでできるかを試してみました．

検証では，一般的なRAGと同様，①情報検索に使用するDB（データベース）作成，②RAGの推論の2ステップで構築しました．文書データを画像として認識してベクトル化を行うことで図表処理をする「ColPali」では，独自Embeddingモデル（テキストや単語などの言語データを数値ベクトルに変換する方法）でベクトル化し，画像を分割してメタデータを付加し，ベクトルデータの類似度を基に検索し，もっとも適した画像を取得しました．そして，質問内容と取得した画像を用いて視覚読解機能を持ったtsuzumiが回答を生成する，というアプローチで視覚読解RAGを構築し，

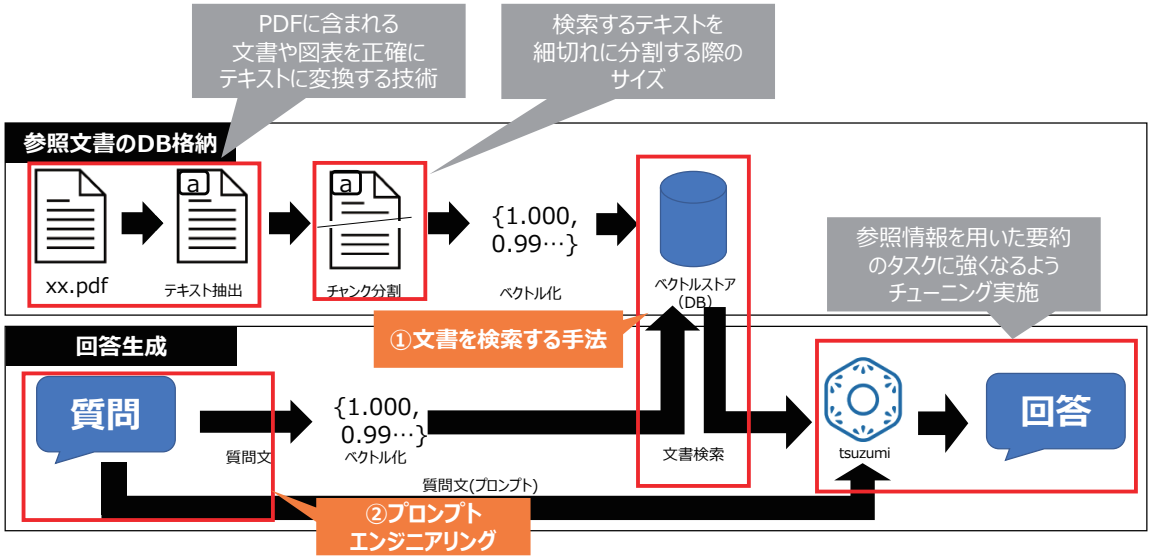


図1　RAGの精度向上

- 一般的なRAGと同様，①DB作成　②RAGの推論　の２ステップで構築．
- 図表処理をする「ColPali」では，独自Embeddingモデルでベクトル化し，画像を分割してメタデータを付加．
  ベクトルデータの類似度をもとに検索し，もっとも適した画像を取得．
- 質問内容と取得した画像を用いて視覚読解機能を持ったtsuzumiが回答を生成する．

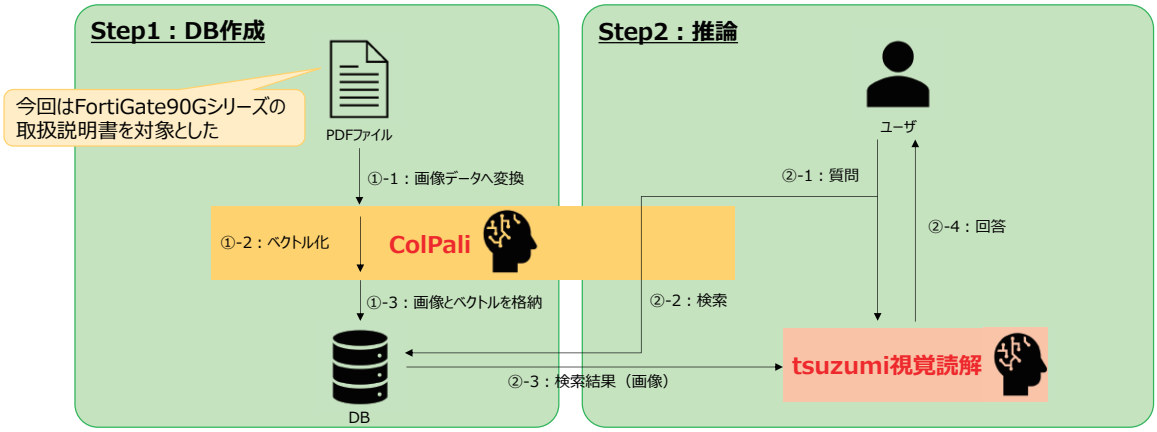


図２　視覚読解RAGの検証

評価しました（**図２**）．

図を使用した質問に対する回答を生成した結果，ColPaliによる検索が高精度で成功する一方，視覚読解tsuzumiによる文書化精度は改善の工夫が必要です．tsuzumiそのものの性能改善についても，今後もNTTグループ各社と連携して取り組んでいこうと思います．

「MoE」は，機械学習モデルの中で特定のタスクに適した専門家を選んで働かせる仕組みで，質問内容に対してゲーティングがどの専門家に聞けば良いかを判断し，必要な専門家にだけ回答を生成させる技術です．MoEを使用するメリットは，大きく４点あります．１点目は，モデル全体ではなく選択されたエキスパートだけが活性化されるため，計算資源を効率的に使用できることです．これにより，大規模モデルを用いても一部のパラメータだけを活用することでトレーニングや推論時のコストを削減可能となります（計算効率の向上）．２点目は，同じ計算コスト内でより多くのパラメータを持つモデルを構築可能であるため，複雑なタスクに対しても高いモデル容量の適応可能性を向上させることができます（モデル容量の拡大）．３点目は，各エキスパートが特定のタスクやデータ特性に特化するため，モデル全体の性能が向上する可能性があるとともに，モジュールごとに異なるタスクやドメインに対応できる柔軟性もあります（専門性の向上）．４点目は，新しいエキスパートを追加することで，タスクやデータの規模が増えても対応可能となります（スケーラビリティ）．

検証は，コンタクトセンタ業務で方言が分からず対応しにくいといったニーズがあるため，対象とする方言を関西弁・博多弁・沖縄弁の３つとし，方言を標準語に変換するタスクを考えました（**図３**）．評価は，①インプットに対して専門家を正しく選択できたか，②専門家が正しく回答できたか，の２つの観点で精度を確認し，以下の結果となりました．

・沖縄弁：他の方言と比べて独自性が強いため①の精度は高かった一方で，標準語への変換タスクの難易度も高く，②の精度が低い結果となりました．
・関西弁：tsuzumiのベースモデルが一定量の学習を行っているため，①および②の両項目において高い精度を達成しました．
・博多弁：①で全く選択が行えなかったため，それに伴い②の精度も低下しました．

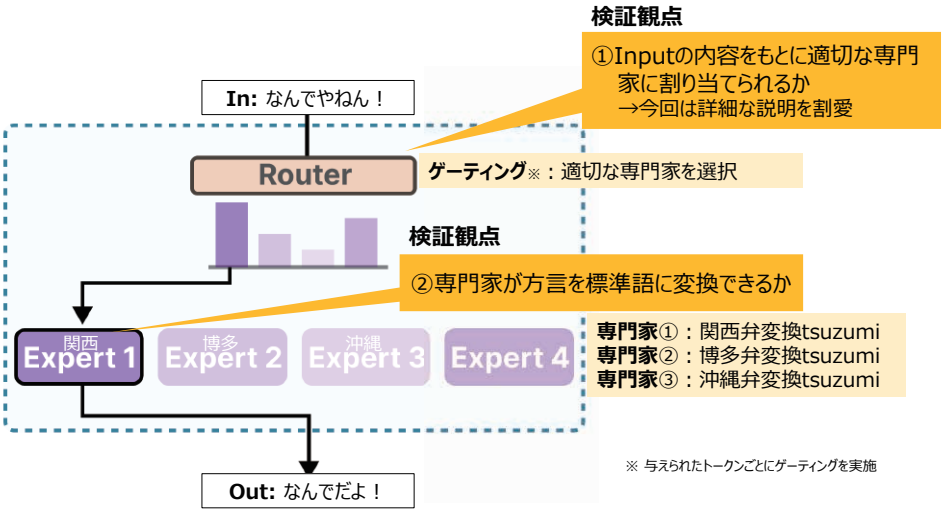
この結果に対する原因を解析しているところで，その状況により今後の対応を検討していくつもりです．

生成AIは分からないこともまだ多く，それを検証していくために発想の転換，実際に手を動かすことが重要

開発者としてのスキルはどのように磨いているのでしょうか．

私は，2022年12月にNTT西日本に入社したのですが，それ以前は気象・防災関連のソフトウェア開発を行いつつ，AIを勉強してシステムから取得した画像を用いたAIの開発を行ってきました．実案件をとおしてCNN（Convolutional Neural Network）による機械学習を施したAIを画像処理に適用・検証を行うことでAIシステムの開発を行いました．それ以外にも人流予測等の時系列予測AI・故障検知AIの開発・技術検証も行ってきました．そして，tsuzumiが発表された以降は，tsuzumiをはじめとした生成AIの技術検証をしています．

AIは，その種類によって機械学習の方法や，処理方式等が異なるため，画像処理関連から生成AIに対象を移した際には，生成AIの自然言語処理等は未知の分野であり，最初から勉強しなけれ



https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-mixture-of-experts

図 3　MoEの検証

ばならない状況でした．しかし，そのおかげで現在では生成AIを普通に使うことができるまで，そして人にレクチャーするまでスキルが向上しました．こうした過程を振り返ってみると，学習の方法や処理方式はそれぞれ異なってはいるものの，AIの学習から動作まで，基本的な考え方は同じであること，精度は入力データの質に大きく依存することが，これまで扱ったすべてのAIにおいて共通であるということに気付きました．したがって，データをしっかり見ることにより，そのデータの質，処理の方法を確認したうえでAIのつくり込みをしていくことが重要であり，これは日ごろから意識している点でもあります．

さて，あるお客さまの案件でRAGのシステムを構築し，展示会でデモを行いました．お客さまのデータをお借りして短期間で検証を行ったのですが，パラメータを変えつつ試行錯誤してもなかなか精度が上がらず悪戦苦闘しました．展示会の1週間前になり，発想，視点を変えて今までイメージしていなかった方法で取り組んだところ，これまで40%程度だった精度が一気に70%程度まで向上し，展示会のデモは好評を博しました．これまでお客さまとの接点が少なかったので，非常に達成感のある仕事の経験をさせていただきました．それとともに，発想の転換の重要性，机上の理論ではなく実際に試してみることの重要性を実感しました．併せて，生成AIが注目を集めてさまざまなところで研究・勉強・実証されてはいるものの，まだまだ分からないことが多くある，ということも気付きました．

AI関連の技術に携わってこられましたが，今後はどのようなことに取り組んでいきたいですか．

NTT西日本に入社して日が浅いこともあり，しばらくは技術革新部でAIの技術検証を続けていきたいと思います．特に，生成AIやRAGが注目されている中で，キーワードとして登場している技術として，マルチモーダルやAIエージェント等があり，こういった分野における開発・検証に取り組んでみたいと思います．

これとは別に，生成AIに世の中の関心が集まっている中で，従来からある画像処理関連のAIや機械翻訳等といった生成AIとは異なるAIもあり，現実の場で活用されています．このような既存のAIによるシステムを構築し，DXの効果を生む取り組みで世の中に貢献していきたいと思います．

AIを活用し，社会課題解決に向けた貢献のための連携

後進や読者へのメッセージをお願いします．

少子高齢化に伴う就労人口の減少による人手不足が社会問題になっています．その結果，特に技術者や専門家にとって雑用に奔走し，その専門性を活かす業務の時間が割かれる，といった問題も実際に起こっています．

生成AIに限らずAIにより，人手不足となっている業務の肩代わりや，それにより専門性を活かした業務への集中に寄与することで，社会課題解決につながると考えています．私たちはそこに貢献するためにtsuzumiをはじめとするAIの導入，そしてそのための検証に取り組んでいます．同様な課題意識をお持ちの方がいらっしゃれば，連携してAIの活用に取り組んでみませんか．