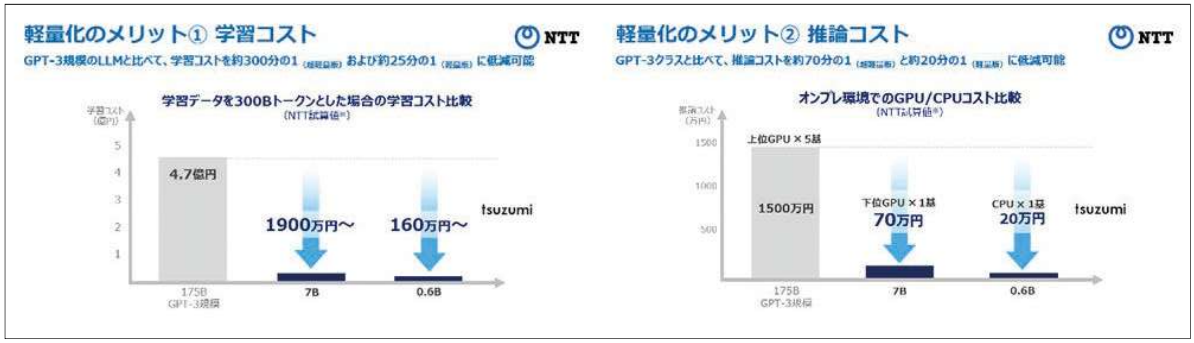
図表5-3-1 ▶tsuzumi　軽量化のメリット



出所：NTT「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」（2023年11月1日）

稼働させるために高性能なGPUが複数必要となり、学習にかかる電力量が膨大になるという弱点も持ち合わせていた。そこでNTTはパラメーター・サイズを抑制した軽量のLLMを開発した。

tsuzumiには2024年度末にサービス提供済みのLLMとして「軽量版（tsuzumi-7B）」と「超軽量版（tsuzumi-0.6B）」の2種類がある。そのパラメーター・サイズは発表当時にOpenAI社が開発していたGPT-3と比べて、軽量版が約1/25、超軽量版は約1/300に抑えられている。その結果、軽量版tsuzumiはGPU一基、超軽量版tsuzumiはCPUのみで動かすことが可能なAIとなった（**図表5-3-1**）。

すなわち、tsuzumiであれば、データセンターなどに用意された巨大な計算資源にアクセスしなくとも、オンプレミス[4]での運用が可能になる。このことはLLMの運用コストを低減させる効果を持つとともに、オンライン環境で利用しづらいような機微情報を取り扱う利用者にとっては非常に重要な特徴となっており、事業会社各社には実際に多くの案件相談が届いている。

2つ目の特徴は、「言語性能（特に日本語）が高い」ことである。

LLMはこれまで、パラメーター数を増やすことでその性能を高めてきた。例えば、ChatGPTの基盤モデルであるGPTシリーズのパラメーター数は、GPT-1（2018年）が約1億1,700万、GPT-2（2019年）が15億、GPT-3（2020年）が1,750億と増加してきており、GPT-4（2023年）以降のLLMについては、公表されていないが、数兆規模と推定されている。パラメーター数の増加に伴い、LLMの性能を向上させてきたのである。そのため、LLMにおけるパラメーター数は、一般的には性能を表す指標と捉えられていた側面があり、単に「軽量」であるだけで、その代償として性能が低下してしまうようでは、実用的なニーズに対応できなくなってしまう。

そこで意味を持ってくるのが長年にわたる研究成果の蓄積である。NTTはLLMという言葉が出てくるよりもはるか前の1980年に「自然言語処理」の研究開発を始めており、さまざまな言語処理技術の研究を続けてきた。その成果の一部は、検索サービス「goo」（1997年）、音声UI「しゃべってコンシェル」（2012年）、翻訳サービス「COTOHA」（2018年）などの形で、具体的なサービスとしても提供されてきている。

こうした言語処理分野における長年の蓄積の成果もありtsuzumiは、とりわけ日本語の言語性能に関しては、パラメーター数の大きな他のLLMと遜色のないパフォーマンスを示すことができている。パラメーター数が70億である「軽量版（tsuzumi-7B）」は、リリース当時の他のLLMとの比較において、軽量であっても、十分な性能を実現することが可能であることを証明している（**図表5-3-2**）。

GPU一基で動作する範囲における更なる精度向上をめざしてきた30B規模モデルは、「tsuzumi 2」として2025年10月にリリース予定となった。社内トライアルではRAG[5]型の問い合わせ回答精度が前モデル比約4倍に向上し、日本語の文脈・文意理解に関する評価でも同サイズ帯トップクラスの結果を示している（**図表5-3-3**）。

3つ目の特徴は「柔軟なカスタマイズ」ができることだ。

tsuzumiは3つのチューニング方法を提供している。コストを抑えたい利用者向けの「プロンプトエンジニアリング」、精度を高めたいときに有効な「フルファインチューニング」、そして、コストと精度のバランスの取れた「アダプタチューニング」という選択肢である。利用者、あるいは、利用シーンによって、LLMに求められる要件は変わってくる。異なる特徴を有する3種類の方法を用意することで多様なニーズに対応できるようにしているのである（**図表**

4 「オンプレミス（on-premises）」とは、企業や組織が自社内に設置したサーバーや設備を使って、ITシステムやソフトウェアを運用する形態を指す。クラウドとは対照的な概念。「オンプレ」と略されることも多い。

5 RAG（Retrieval-Augmented Generation）とは、質問に関連する社内文書やDBを検索で取り出し、その内容を文脈としてLLMに与えてから回答を生成させる方式。