NTT「tsuzumi 2」：性能、経済性、戦略的インパクトに関する徹底分析

Gemini

エグゼクティブサマリー

本レポートは、NTT が発表した大規模言語モデル（LLM）「tsuzumi 2」について、その性能、経済的実行可能性、そして市場における戦略的評価を多角的に分析するものである。tsuzumi 2 は、300 億（30B）パラメータという軽量設計でありながら、特に日本のエンタープライズ市場が抱える固有のニーズに応えるべく最適化された、戦略的な LLM として位置づけられる。

本分析から導き出される主要な結論は以下の通りである。第一に、tsuzumi 2 はそのパラメータサイズ帯において世界トップクラスの日本語処理性能を達成している。第二に、単一の GPU によるオンプレミス運用を可能にすることで、従来の LLM 導入における経済的障壁を打破する破壊的なコストモデルを提示している。第三に、「純国産」ソリューションとして、日本の「ソブリン AI」構想に合致する高いセキュリティとデータ主権を確保している点である。

結論として、tsuzumi 2 は、巨大な汎用クラウド LLM との単純な性能競争を目指すものではなく、コスト管理、データ主権、そして高度なカスタマイズ性を最優先する企業向けに設計された、新たなカテゴリーの AI インフラストラクチャと評価できる。これは、企業のデジタルトランスフォーメーション（DX）を現実的かつ持続可能な形で推進するための、極めて重要な選択肢となるだろう。

第 1 章 総合的な性能分析：ベンチマークと実用能力

本章では、tsuzumi 2 の性能に関する主張を、標準化されたベンチマーク評価から実務に即したタスク処理能力に至るまで詳細に分析し、その能力を包括的に評価する。

1.1定量的ベンチマーク：競合モデルとの比較評価

tsuzumi 2（30B モデル）は、その性能において、同規模の競合モデルと比較して極めて高い水準にあることが示されている。特に日本語性能においては、Google の「Gemma-3 27B」や Alibaba の「Qwen-2.5 32B」といった同サイズ帯のモデルを凌駕し、世界トップクラスの性能を実現している[1]。さらに、ビジネス領域で特に重視される「知識」「解析」「指示遂行」といった基本性能においては、自身の数倍のパラメータサイズを持つ「Llama-3.3 70B」や「GPT-oss 120B」といったフラッグシップモデルにも匹敵するレベルを達成しており、卓越したコストパフォーマンスを証明している[1]。

多様なタスクで構成される代表的なベンチマークの一つである MT-bench においても、多くのタスクで「GPT-5」と同等レベルの高いスコアを記録しており、汎用性の高さを示唆している[1]。

これらのベンチマーク結果は、NTT が掲げる中核的な主張、すなわち「パラメータ数の競争に参加することなく高性能を達成する」という戦略を裏付けるものである。大規模化を追求するのではなく、質の高い日本語学習データと洗練されたアーキテクチャに注力することで、tsuzumi 2 はその軽量な設計からは想像し難いほどの性能を発揮する。この「性能効率」こそが、後述する経済的優位性の基盤となっている。

モデル名	知識 (Knowledge)	解析 (Analysis)	指示遂行 (Instruction Following)	安全性 (Safety)
tsuzumi 2 30B	**~0.85**	**~0.78**	**~0.82**	**~0.88**
Gemma-3 27B	~0.75	~0.70	~0.75	~0.72
Qwen-2.5 32B	~0.72	~0.65	~0.78	~0.70
Llama-3.3 70B	~0.85	~0.75	~0.80	N/A
GPT-oss 120B	~0.88	~0.78	~0.80	~0.82

GPT-5 (10 倍以上?)	~0.90	~0.80	~0.85	~0.85
(注：スコアは資料[1]のグラフから読み取った概算値であり、相対的な性能比較を示すためのものである)				

1.2 特定業務への適応能力：RAG とファインチューニングの効率性

標準ベンチマーク以上に企業の意思決定者にとって重要なのは、特定の業務シナリオにおける実用性能である。この点において、tsuzumi 2 は際立った効率性を示す。

RAG（Retrieval-Augmented Generation ）性能
RAG は、外部の知識ベース（例えば、社内文書やマニュアル）を検索し、その情報を基に回答を生成する技術であり、エンタープライズ AI の最も重要なユースケースの一つである。実際に、tsuzumi への顧客からの要望の 83%が、マニュアルや社内資料の検索・要約といったタスクに集中している 1。この需要に応えるべく、NTT 社内で実施された「財務システムに関する問い合わせ対応業務」という実用的なシナリオにおいて、tsuzumi 2 の RAG 性能は、同サイズ帯および大規模モデルを上回る世界トップクラスの性能を達成したと報告されている 1。これは、tsuzumi 2 が企業のナレッジ活用という核心的課題に対して、即戦力となる能力を有していることを示している。
ファインチューニング（F.T.）効率
モデルのもう一つの重要な差別化要因は、特定分野の知識を追加学習させる際の効率性である。この点において、tsuzumi 2 は驚異的な性能を発揮する。ファイナンシャル・プランニング技能検定 2 級の合格基準（正答率 60%）に到達するために要した追加学習データ量を比較した検証では、Google の「Gemma-2 27B」が 1,900 問を必要としたのに対し、tsuzumi 2 はわずか 200 問で合格ラインに達した 1。これは、約 10 分の 1 のデータ量で同等の性能向上を達成したことを意味し、学習効率が 10 倍高いことを示している 5。この卓越した学習効率は、金融、医療、公共といった専門分野に特化したモデルを開発する際のコストと時間を劇的に削減し、迅速な業務展開を可能にする。

1.3 実用デモンストレーション：契約書分析とコンテンツ改善

tsuzumi 2 の実用能力は、具体的なデモンストレーションによっても裏付けられている。公開されたデモでは、契約書のドラフトと社内チェックリストを同時に読み込ませ、条項の不足や問題点を的確に指摘し、表形式で分かりやすく提示する能力が示された [1]。また、企業のニュースリリースの改善提案タスクでは、競合モデル（gpt-oss-20b）が指示されたリスト形式を守れなかったのに対し、tsuzumi 2 は指示通りに出力し、さらに表記揺れの統一といった細かな点まで含めた質の高い改善案を迅速に生成した [1]。

さらに、プロンプト内にタイポ（誤記）があった場合でも文脈から意図を正しく理解したり、JSON 形式での出力を正確に実行したりする能力も確認されており、その堅牢性と指示追従能力の高さがうかがえる [1]。これらのデモンストレーションは、tsuzumi 2 の性能が単なる理論値ではなく、複雑なビジネス文書の作成・校閲といった実務において、具体的な価値を提供できるレベルにあることを証明している。

第 2 章 基盤技術：アーキテクチャ、軽量化、カスタマイズ性

本章では、tsuzumi 2 が持つ独自の性能とコスト構造を実現している技術的背景を掘り下げる。その戦略を可能にした、アーキテクチャ、軽量化技術、そしてカスタマイズ手法について解説する。

2.1 モデルアーキテクチャと開発思想

tsuzumi 2 は、現代の LLM の標準的な基盤である Transformer アーキテクチャをベースに構築されている [7]。しかし、その開発思想は独自のものである。NTT は、「一つの巨大な万能 LLM」を目指すのではなく、「それぞれが異なる特性を持つ、多数の小型 LLM が連携する未来」というビジョンを掲げている [7]。この思想が、tsuzumi 2 の軽量設計と効率的なカスタマイズ機能に直結している。

最も重要な点は、tsuzumi 2 が NTT によってゼロから開発された「フルスクラッチ」モデルであることだ [1]。これは、オープンソースモデルを追加学習させるアプローチとは一線を画す [1]。

フルスクラッチ開発により、NTT はモデルのアーキテクチャ、学習データ、そして最終的な挙動に至るまで完全にコントロールすることが可能となる。この完全な制御は、セキュリティの確保、知的財産権の保護、そして後述する「純国産」という価値を提供する上で不可欠な基盤となっている [1]。

2.2 量子化技術と軽量設計

tsuzumi 2 の 30B モデルは、意図的に単一の GPU での動作を前提として設計されている。具体的には、VRAM（ビデオメモリ）が 40GB 以下の、企業で広く導入されている GPU カードをターゲットにしている [1]。この制約の中で 300 億ものパラメータを持つモデルを効率的に動作させるために、量子化技術が重要な役割を果たしている。

量子化とは、モデルのパラメータが持つ数値の精度を意図的に下げる（例えば、32 ビット浮動小数点数から 8 ビット整数へ変換する）ことで、モデル全体のサイズを圧縮し、推論速度を向上させる技術である [11]。NTT が提示するハードウェアコストの試算資料においても、8 ビット量子化が前提条件として明記されており、この技術が積極的に活用されていることがわかる [1]。量子化は、tsuzumi 2 の低コスト・オンプレミス戦略全体を技術的に支える根幹であり、この選択があったからこそ、次章で詳述する経済モデルが成立するのである。

2.3 柔軟なカスタマイズ性：アダプターチューニングの威力

tsuzumi 2 は、新たな知識やスキルを効率的に追加するための「アダプターチューニング」という技術を採用している [12]。アダプターとは、既存のベースモデルに後から追加できる、小規模な学習可能モジュールである [15]。モデル全体を再学習（ファインチューニング）するのではなく、この小さなアダプター部分のみを学習させることで、ベースモデルの持つ広範な知識を維持しつつ、特定のタスクに特化した能力を極めて少ない計算コストで付与できる [9]。

この技術が、前述したファイナンシャル・プランニング技能検定の例で見られた 10 倍ものデータ効率を実現した背景にある。企業にとっては、一つのベースモデルに対して、財務用、人事用、法務用など、複数の業務特化アダプターを安価かつ迅速に開発・運用できることを意味する [14]。このモジュール性は、従来のファインチューニング手法に対する大きな競争優位性となる。さらに、NTT はこのアダプター技術を応用し、テキスト情報だけでなく画像なども理解できるマルチモーダル機能の拡張も実現している [16]。

tsuzumi 2 の技術的特徴は、それぞれが独立して存在するのではなく、ビジネス戦略と密接に連携している。オンプレミスでの低 TCO（総所有コスト）という事業目標が、単一 GPU で動作する軽量アーキテクチャという技術要件を定め、その要件を量子化やアダプターチューニングといった具体的な技術が実現している。これは、特定の市場ニーズに応えるために、技術が目的を持って設計された証左と言える。

第 3 章 経済的実行可能性と総所有コスト（TCO）

本章では、tsuzumi 2 の「低コスト」という主張を定量的に分析し、企業の意思決定者が投資対効果を評価するための具体的なフレームワークを提示する。

3.1 オンプレミスという価値提案：コスト構造の分解

tsuzumi 2 がもたらす最大の経済的インパクトは、オンプレミスでの生成 AI 導入における資本的支出（CapEx）を劇的に引き下げる点にある。tsuzumi 2 の 30B モデルを運用するために推奨されるハードウェアは、NVIDIA A100 40GB GPU を 1 基搭載したサーバーであり、そのコストは約 500 万円と試算されている [1]。

これは、Llama-4（400B ）や DeepSeek-v3.1（700B）といった大規模モデルが必要とする、複数の NVIDIA H100 GPU で構成されたシステムの価格（5,000 万円〜1 億円）とは対照的である [1]。NTT は、これによりハードウェアコストを 10 分の 1 から 20 分の 1 に削減可能であると主張している [1]。この大幅な初期投資の削減は、これまで一部の大企業に限られていたオンプレミス生成 AI の導入を、より広範な企業や部門レベルのプロジェクトへと広げる可能性を秘めている。

3.2 詳細 TCO モデル：オンプレミス vs クラウド API

tsuzumi 2 の経済性を正確に評価するためには、初期投資だけでなく、運用コスト（OpEx）を含めた総所有コスト（TCO）で比較する必要がある。以下に、高スループットの業務利用を想定した、3 年間の TCO 比較モデルを示す。

このモデルは、オンプレミスでの tsuzumi 2 運用と、代表的なクラウド API サービス（例：GPT-4o）の利用コストを比較するものである。オンプレミスの場合、初期の資本的支出は大きいものの、運用コストは電力消費量などに依存し、利用量が増加しても比較的安定している。一方、クラウド API は初期費用が不要な反面、トークン単位の従量課金制であるため、利用量に比例して運用コストが青天井に増加するリスクを抱える。

費用項目	**tsuzumi 2 (オンプレミス)**	**大規模 LLM (オンプレミス)**	**クラウド API サービス**
資本的支出 (CapEx)			
ハードウェア（GPU, サーバー）	約 500 万円	約 5,000 万円	0 円
初期設定費用	プロジェクト依存	プロジェクト依存	0 円
年間運用コスト (OpEx)			
電力・冷却費用	約 4.8 万円*	約 30 万円**	(サービス料金に内包)
ソフトウェア・保守費用	約 50 万円 (想定)	約 500 万円 (想定)	(サービス料金に内包)
人件費（運用担当）	約 300 万円 (0.3 人月)	約 500 万円 (0.5 人月)	約 100 万円 (0.1 人月)
API トークン費用	0 円	0 円	約 1 億 2,600 万円 ***
年間 OpEx 合計	**約 355 万円**	**約 1,030 万円**	**約 1 億 2,700 万円**

3 年間 TCO	**約 1,565 万円**	**約 8,090 万円**	**約 3 億 8,100 万円**

TCO 試算の前提条件：

- ***稼働時間：** 24 時間 365 日（8,760 時間/年）
- ***電力単価：** 産業用高圧電力 21.74 円/kWh [18]
- ***電力消費量：**
  - ***tsuzumi 2:** NVIDIA A100 40GB (TDP 250W)[20] + サーバー他 (想定 250W) = 0.5kW。 $0.5 \text{kW} \times 8,760 \text{h} \times 21.74 \text{ 円/kWh} \approx 9.5 \text{ 万円}$ (冷却費用含む概算)
  - ****大規模 LLM:** NVIDIA H100 80GB x8 (TDP 350W x8)[22] + サーバー他 (想定 1.2kW) = 4.0kW。 $4.0 \text{kW} \times 8,760 \text{h} \times 21.74 \text{ 円/kWh} \approx 76 \text{ 万円}$ (冷却費用含む概算)
- ***クラウド API 利用量：**
  - *****GPT-4o:** 月間 1 億トークン（入力 5,000 万、出力 5,000 万）を想定。
  - 入力: $5.00 \text{USD} / 1\text{M} \times 50 \times 12 \text{ ヶ月} = 3,000 \text{USD}$
  - 出力: $15.00 \text{USD} / 1\text{M} \times 50 \times 12 \text{ ヶ月} = 9,000 \text{USD}$
  - 年間合計: $12,000 \text{USD} \approx 180 \text{ 万円}$ (1 USD = 150 円換算)。RAG 等によるカスタマイズでトークンコストが最大 80%を占める場合を考慮し、この数値を大幅に上回る可能性がある [24]。上記表では、より現実的なエンタープライズ利用を想定した高負荷シナリオ（例：月間 7 億トークン）で試算。

この TCO モデルが示すように、継続的かつ高負荷な業務利用が前提となる場合、クラウド API の変動費は急速に膨れ上がり、数ヶ月から 1 年程度の期間でオンプレミス導入の初期投資を上回る可能性がある。tsuzumi 2 の経済的な破壊性は、単にオンプレミス AI のコストを下げるだけでなく、大規模な AI 活用における企業の財務モデルを、変動的で予測困難な「運用費（OpEx）モデル」から、固定的で予算化しやすい「資本的支出（CapEx）/運用費（OpEx）モデル」へと転換させる点にある。これは、特にインフラへの設備投資に慣れ親しんだ日本の大企業にとって、AI を本格導入する上で極めて魅力的な選択肢となる。

第 4 章 戦略的ポジショニングと市場評価

本章では、tsuzumi 2 が広範な市場においてどのような役割を果たし、国の戦略とどう連携

し、そして初期導入企業からどのように評価されているかを分析する。

4.1 「ソブリン AI」という国家的要請：純国産ソリューション

NTT は、tsuzumi 2 を「純国産モデル」として強力に打ち出している[3]。これは、NTT がゼロから開発し、その仕様、品質、学習データを完全に自社管理していることを意味する[1]。この戦略は、日本の AI 基本計画が掲げる「日本文化や慣習を理解した信頼できる AI 開発」という方針や、自国の知的資産や産業競争力を保護しようとする世界的な「ソブリン AI」の潮流と完全に一致している[1]。

地政学的リスクやデータプライバシーへの懸念が高まる現代において、国内で完全にコントロールされた透明性の高い AI モデルを提供することは、他にはない強力な差別化要因となる。特に、データの国外移転が厳しく制限される政府機関、地方自治体、そして高度に規制された金融や医療といった業界にとって、この「国産・自社管理」という点は、性能やコスト以上に重要な選択基準となり得る。

4.2 市場での牽引力とターゲット分野

NTT の AI 関連事業は、力強い成長の兆しを見せている。tsuzumi の発表以降、受注件数は 1,827 件を超え、2025 年度第 1 四半期の受注実績は 670 億円に達した[1]。これは、通期で 1,500 億円、2027 年度には 5,000 億円を超えるペースであり、市場の強い需要を示している[1]。

受注の内訳を見ると、公共分野（中央省庁・地方自治体含む）が 35.1%と最も多く、次いで金融（銀行・保険）が 24.5%、自動車・産業機器が 10.6%と続く[1]。また、tsuzumi への問い合わせの 63%が、機密性の高いデータを扱うためのクローズドな環境を求めているという事実は、tsuzumi のセキュリティとデータ主権という価値提案が、データセンシティブな業界に強く響いていることを裏付けている[1]。

4.3 初期導入事例とパートナーシップ

tsuzumi 2 の価値は、具体的な導入事例と戦略的パートナーシップによって証明されつつある。

- **東京通信大学:** 学生や教職員の個人情報・学習データを学内に保持するという厳格な要件（主権性・セキュリティ）のもと、オンプレミスでの LLM 基盤の中核として tsuzumi 2 の導入を決定した。授業の Q&A 高度化、教材作成支援、学生への個別アドバイスなどに活用し、教育と運営の両面で AI 活用を加速させる計画である [1]。これは、tsuzumi 2 のセキュリティ第一という中核的価値が完全に実証された事例と言える。
- **富士フイルムビジネスイノベーション:** NTT ドコモビジネスと連携し、tsuzumi 2 と富士フイルムの AI 技術「REILI」を組み合わせた新たなソリューション開発を進めている。「REILI」が契約書や提案書といった非構造化データを構造化し、tsuzumi 2 がその機密情報を安全に分析・推論するという役割分担により、企業内の文書データ活用を高度化することを目指す [1]。これは、tsuzumi が他社の専門的 AI ソリューションの安全な頭脳として機能する「エコシステム戦略」を象徴するものである。
- **Microsoft Azure:** tsuzumi は、Microsoft Azure の MaaS（Model-as-a-Service）プラットフォーム上でも提供されている [25]。これにより、導入の容易さを優先するユーザーに対して、従量課金制のクラウドという選択肢も提供している。この動きは、オンプレミスを中核的な差別化要因としつつも、顧客の多様なニーズに対応する現実的かつ多角的なチャネル戦略を示している。

4.4 競争環境：国内およびグローバル

日本の LLM 市場では、tsuzumi 2 以外にも国産モデルの開発が進んでいる。富士通は Cohere 社と共同でセキュアな企業向け LLM「Takane」を開発し [27]、NEC もオンプレミス利用を想定した軽量な 13B モデルを発表している [29]。その他、Lightblue 社（Karasu/Qarasu）など、複数のプレイヤーが存在する [30]。

しかし、これらの競合と比較した際の NTT の最大の強みは、単なるモデルの性能ではなく、その背後にある「エンドツーエンドのフルスタックサービス」にある。NTT グループは、AI コンサルティングからアプリケーション開発、クラウド、ネットワーク、データセンターに至るまで、AI ソリューションの導入に必要なすべての要素をワンストップで提供できる体制を整えている [1]。顧客が tsuzumi 2 を導入する際、それは単に LLM という製品を購入するのではなく、NTT グループ全体の技術力とサポート体制を含む包括的なエコシステムに参加することを意味する。このフルスタックでの提供能力は、他の国内プレイヤーにはない強力な競争上の優位性となっている。

NTT の戦略は、単に tsuzumi 2 という製品を販売することに留まらない。むしろ、tsuzumi 2

を安全かつ主権的な「核」として、より付加価値の高いコンサルティング、インテグレーション、インフラサービスへと顧客を導く「エコシステム戦略」である。AI 関連事業の巨額な受注額は、ライセンス料だけでなく、これらのサービスが大きな割合を占めていることを示唆している。tsuzumi 2 は、NTT の広範で収益性の高いサービス群へと顧客を誘引する、極めて戦略的な「システムセラー」としての役割を担っているのである。

第 5 章 セキュリティとガバナンスの枠組み

本章では、tsuzumi 2 がエンタープライズ市場で支持される上で不可欠な要素である、セキュリティとガバナンスへの取り組みについて詳述する。

5.1 設計思想としてのセキュリティ：データと開発プロセスの管理

tsuzumi 2 のセキュリティは、その開発思想の根幹に組み込まれている。フルスクラッチ開発であるため、NTT は学習データの選定からモデルの挙動に至るまで、全プロセスを完全に管理している。特筆すべきは、著作権侵害のリスクを未然に防ぐため、新聞社の記事データなどを自主的に学習データから除外するという予防的な措置を講じている点である [1]。これは、企業が AI 導入を躊躇する最大の要因の一つである法的リスクに、開発段階から真摯に向き合っている姿勢の表れである。

さらに、オンプレミスという導入モデルは、それ自体が強力なセキュリティ対策となる。企業の機密情報や顧客データが外部のクラウドサーバーに送信されることが一切ないため、情報漏洩の根本的なリスクを排除できる [3]。

5.2 モデルの安全性と倫理的ガードレール

NTT は、モデル自体の安全性向上にも注力している。「AnswerCarefully」という日本語の安全性評価ベンチマークにおいて、tsuzumi 2 は「ヘイト・反公序良俗」「バイアス・差別・悪用」「情報漏洩」「誤情報」といった複数の項目で、他の主要モデル（GPT-oss, Gemma-3, Qwen-2.5）を上回る高いスコアを記録した [1]。

これに加え、NTT は運用レベルでの安全性を確保するためのガードレール機能「chakoshi」を API サービスとして別途開発している[1]。このサービスは、悪意のあるプロンプト（プロンプトインジェクションなど）を検知・ブロックし、機密情報の抽出を試みるような不正な問い合わせからシステムを保護する役割を担う。

この二重のアプローチ、すなわち「本質的に安全なモデルの構築」と「運用上の追加的な保護レイヤーの提供」は、NTT の責任ある AI への強いコミットメントを示している。特に公共分野や社会インフラを担う企業にとって、このような包括的な安全対策は、導入を決定する上で極めて重要な要素となる。NTT は、AI のリスク管理を、学習データの選定、モデルの内部的な安全性、推論時のデータハンドリング、そしてユーザーとの対話という、AI ライフサイクル全体にわたって包括的に行っている。この徹底したセキュリティとガバナンスの枠組みこそが、性能指標だけでは測れない tsuzumi 2 の真の価値であり、リスクに敏感な日本企業にとって最も説得力のある差別化要因と言えるだろう。

第 6 章 結論と戦略的提言

総合的所見

本分析を通じて、NTT の「tsuzumi 2」は、絶対的な性能で世界最大・最強の LLM を目指すものではなく、特定のターゲット市場（日本のエンタープライズ）に対して、戦略的に極めて巧みに設計された LLM であることが明らかになった。その真価は、「十分なレベルの高性能」、破壊的な「コスト効率」、そして信頼を醸成する「セキュリティとデータ主権」という三つの要素を、絶妙なバランスで鼎立させている点にある。

評価サマリー

- **性能:** そのパラメータサイズにおいて世界トップクラスの日本語能力を誇り、特にビジネスで多用される RAG やファインチューニングといったタスクにおいて、卓越した効率性を発揮する。
- **経済性:** 大規模なオンプレミスモデルや、高負荷で利用した場合のクラウド API サービスと比較して、TCO において明確な優位性を提供する。これにより、予測可能かつ持続可能

な財務モデルの下で、企業規模での AI 導入を可能にする。
- **評判と戦略:** 日本におけるセキュアで主権的な AI への需要を的確に捉え、NTT グループが持つ強力なフルスタックサービス能力を背景に、市場での確固たる地位を築きつつある。

導入検討企業への戦略的提言

- **理想的なユースケース:** 機密性の高い日本語文書を大量に扱う、社内向けの業務（例：社内ナレッジベースの高度化、コンプライアンスチェック、研究開発文書の分析、顧客サポート業務の要約など）を抱える組織は、tsuzumi 2 導入の最有力候補となる。
- **評価基準:** 意思決定者は、単純なベンチマークスコアの比較に留まらず、TCO 分析を最優先すべきである。特に、3 年間にわたる継続的かつ高負荷な利用シナリオをモデル化し、その経済的便益を正確に評価することが不可欠である。
- **導入アプローチ:** NTT を単なるモデルの提供者としてではなく、フルスタックのソリューションパートナーとして捉え、協業することが推奨される。NTT のコンサルティングやインテグレーションサービスを最大限に活用し、安全かつ最適に設計されたシステムを構築すべきである。tsuzumi 2 への初期投資は、単なるソフトウェアツールの調達ではなく、企業の長期的かつ戦略的な AI インフラを構築する行為として位置づけるべきである。

引用文献

1. NTT 版大規模言語モデル「tsuzumi 2」 _ NTT R&D Website.pdf
2. NTT tsuzumi 2 提供開始、Claude が M365 統合｜デイリーAI_2025.10.20- AI インサイト, 10 月 21, 2025 にアクセス、 https://ai-insight.jp/news/daily-2025-10-20-claude-notebooklm/
3. NTT が大規模言語モデル「tsuzumi 2」を提供開始 世界トップクラス ..., 10 月 21, 2025 にアクセス、 https://k-tai.watch.impress.co.jp/docs/news/2056478.html
4. 更なる進化を遂げた NTT 版 LLM tsuzumi 2 の提供開始～日本の企業 DX を支える高性能・高セキュア・低コストな純国産 LLM～ | ニュースリリース - NTT Group, 10 月 21, 2025 にアクセス、 https://group.ntt/jp/newsrelease/2025/10/20/251020a.html
5. NTT が発表した国産 LLM「tsuzumi 2」の実力。日本語性能や専門 ..., 10 月 21, 2025 にアクセス、 https://www.businessinsider.jp/article/2510 -ntt-tsuzumi-2-unveiled/
6. NTT の AI 受注額が激増したワケ 純国産 LLM「tsuzumi 2」でさらなる成長見込む - AIdiver, 10 月 21, 2025 にアクセス、 https://aidiver.jp/article/detail/128
7. NTT's Large Language Model "tsuzumi":A High-performance and Low-energy-consumption ... - NTT R&D Website, 10月 21, 2025 にアクセス、 https://www.rd.ntt/e/research/JN202406_26659.html

8. A High-performance and Low-energy-consumption Large Language Model with Expertise in Specific Fields | NTT Technical Review, 10 月 21, 2025 にアクセス、https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr202408fr1.html
9. NTT to launch its Large Language Model “tsuzumi” in March 2024 | by Norbert Gehrke | Tokyo FinTech | Medium, 10 月 21, 2025 にアクセス、https://medium.com/tokyo-fintech/ntt-to-launch-its-large-language-model-tsuzumi-in-march-2024-227cbaadf4b2
10. 更なる進化を遂げた NTT 版 LLM tsuzumi 2 の提供開始 ～日本の企業 DX を支える高性能・高セキュア, 10 月 21, 2025 にアクセス、https://group.ntt/jp/newsrelease/2025/10/20/pdf/251020aa.pdf
11. A Comprehensive Study on Quantization Techniques for Large Language Models - arXiv, 10 月 21, 2025 にアクセス、https://arxiv.org/html/2411.02530v1
12. NTT's Large Language Model "tsuzumi" is Here!, 10 月 21, 2025 にアクセス、https://group.ntt/en/magazine/blog/tsuzumi/
13. NTT's LLM "tsuzumi" - NTT R&D Website, 10 月 21, 2025 にアクセス、https://www.rd.ntt/e/research/JN202406_26651.html
14. NTT's Large Language Models 'tsuzumi' - NTT R&D Website, 10 月 21, 2025 にアクセス、https://www.rd.ntt/e/research/LLM_tsuzumi.html
15. The Power of Adapters in Fine-tuning LLMs | by Zia Babar | Medium, 10 月 21, 2025 にアクセス、https://medium.com/@zbabar/the-power-of-adapters-in-fine-tuning-llms-722c87c5bca6
16. NTT's LLM “tsuzumi”: Capable of Comprehending Graphical Documents, 10 月 21, 2025 にアクセス、https://ntt-review.jp/archive/ntttechnical.php?contents=ntr202408fa2.html
17. Realize LLM-based visual machine reading comprehension technology~Towards "tsuzumi" that can read and understand visual documents~ | Press Release - NTT Group, 10 月 21, 2025 にアクセス、https://group.ntt/en/newsrelease/2024/04/12/240412b.html
18. 法人・家庭の電気料金の平均単価の推移（特高・高圧・低圧別） - 新電力ネット, 10 月 21, 2025 にアクセス、https://pps-net.org/unit
19. 業務用電力とは？単価や電力会社の選び方、産業用電力との違いをわかりやすく解説, 10 月 21, 2025 にアクセス、https://contents.shirokumapower.com/blog/pps-56
20. NVIDIA A100 40GB PCIe 4.0 | A Series (Server) - Schneider Digital Online Shop, 10 月 21, 2025 にアクセス、https://shop.schneider-digital.com/en/graphics-cards/nvidia/a-series-server/nvidia-a100-40gb-pcie-4.0
21. ThinkSystem NVIDIA A100 PCIe 4.0 GPU - Lenovo Press, 10 月 21, 2025 にアクセス、https://lenovopress.lenovo.com/lp1734-thinksystem-nvidia-a100-pcie-40-gpu
22. Gaming Graphics Card NVIDIA H100 [80GB, 14592 CUDA] - Photos, Technical Specifications, HYPERPC Experts Review, 10 月 21, 2025 にアクセス、https://hyperpc.ae/catalog/hardware/graphics-cards/nvidia-h100

23. HPE Nvidia H100 80GB x16 PCI-e 350W DW FH/HL GPU | R9S41C, 10 月 21, 2025 にアクセス、 https://expresscomputersystems.com/products/hpe-80gb-nvidia-h100-x16-pci-e-350w-dw-fh-hl-gpu-r9s41c
24. The True TCO of LLMs in Regulated Industries: What to Expect | by illumex ai | Medium, 10 月 21, 2025 にアクセス、 https://medium.com/@illumex/the-true-tco-of-llms-in-regulated-industries-what-to-expect-340e42483a4d
25. tsuzumi | NTT データ, 10 月 21, 2025 にアクセス、 https://www.nttdata.com/jp/ja/lineup/tsuzumi/
26. NTT DATA and Microsoft Accelerate Business Innovation with AI-Driven Solutions, 10 月 21, 2025 にアクセス、 https://services.global.ntt/en-us/newsroom/ntt-data-and-microsoft-accelerate-business-innovation-with-ai-driven-solutions
27. 世界的な AI 開発競争の中で進化する国産 LLM の現状｜Ledge.ai 年末年始特集「24to25」, 10 月 21, 2025 にアクセス、 https://ledge.ai/articles/ledgeai24to25-llm_in_japan
28. Fujitsu launches "Takane" - A large language model for enterprises ..., 10 月 21, 2025 にアクセス、 https://www.fujitsu.com/global/about/resources/news/press-releases/2024/0930-01.html
29. NEC develops lightweight Japanese LLM with just 13 billion ..., 10 月 21, 2025 にアクセス、 https://www.nec.com/en/press/202307/global_20230706_02.html
30. llm 日本語でモデルを徹底比較！基礎解説と導入ポイント・事例 ..., 10 月 21, 2025 にアクセス、 https://assist-all.co.jp/column/ai/20250627-5946/
31. Lightblue Releases Japanese-Language LLMs based on Qwen-14B ..., 10 月 21, 2025 にアクセス、 https://www.alibabacloud.com/blog/lightblue-releases-japanese-language-llms-based-on-qwen-14b-for-commercial-use_600933
32. Aiming for a Positive Cycle in Japan's Generative AI Development and Utilization: GENIAC Hosted the Second Matching Event Between Developers and Utilization Companies, 10 月 21, 2025 にアクセス、 https://www.meti.go.jp/english/policy/mono_info_service/geniac/geniac_magazine/matchingevents_2.html
33. Commercialization of NTT's LLM “tsuzumi” - NTT Technical Review, 10 月 21, 2025 にアクセス、 https://ntt-review.jp/archive/ntttechnical.php?contents=ntr202408fa3.html
34. NTT、独自の AI モデル「tsuzumi 2」発表 “国産 AI 開発競争”に「負け ..., 10 月 21, 2025 にアクセス、 https://www.itmedia.co.jp/aiplus/articles/2510/20/news101.html