



tsuzumi 2 on Azure ユーザーガイド v1.0

tsuzumi 2
user guide

1 tsuzumi 2ってどんなもの? 3

- 分類 4
- 特長 5
- 活用ポイント 6
- 実施できるタスクの例 7
- 実施できるタスクの具体的なイメージ 8
- tsuzumi 2 on Azure の提供仕様 10

2 tsuzumi 2を使おう 12

- チュートリアル 13
- クォータのリクエスト 14
- tsuzumi 2 のデプロイ 19
- 実行確認 25

3 推論を使いこなす 27

- 推論の仕組み 28
- 推論パラメータの意味 29

4 外部アプリケーションと連携する 31

- 外部アプリケーションとの連携方法 32
- フロントエンド実行手順 33

5 お問い合わせ 37

- お問い合わせ先 38

1 tsuzumi 2ってどんなもの？

分類

tsuzumi 2はGPT-5のような大規模なLLMとは異なり、軽量なSLM（Small Language Model）という分類です。チューニングを行わない状態ではLLMと比較して汎用的な精度は劣る点をご理解いただいた上で、用途に応じたカスタマイズで特定のタスクに適応することをご検討ください。

特性	SLM	LLM
パラメータ数	数億～数十億	数百億～数兆
学習データ	特定分野に特化	幅広い範囲
学習時間	数日程度	数十日～数か月
ファインチューニングコスト	比較的低い	非常に高い
適応性	特定タスクに強い	汎用的

※tsuzumi 2 on Azureではチューニング機能は利用できません。



tsuzumi

特定分野、特定タスク向けにチューニングしてご利用いただくのが効果的です。

特長

特長
1

軽量

パラメタサイズ 30B

高いコストパフォーマンス
大規模クラスと比べて
推論コストを約10～20分の1に低減可能
クラウドデータをお客さま環境で
セキュアに活用可能

特長
3

高カスタマイズ性

用途に応じたチューニングを 低コストで実現

1. アダプタチューニング
2. フルファインチューニング
3. プロンプトエンジニアリングを提供

※ tsuzumi 2 on Azureではチューニング機能は利用できません。

特長
2

高い言語性能

日本語性能 世界トップクラス

40年以上の自然言語処理研究による蓄積
GPT-5にも引けを取らない性能
同規模の国産LLMを大きく上回る
(日本語・英語対応)

特長
4

マルチモーダル性

言語に加え 図表読解等様々な形式に対応

文書画像の解析、聴覚等への拡張可能

※ tsuzumi 2 ではマルチモーダルはサービス対象外となります。
今後のバージョンでの追加を検討中です。

その他 tsuzumi 2の特徴については[こちら](#)をご参照ください。

活用ポイント

tsuzumi 2は軽量のモデルであり、GPT-5のような大規模LLMと同様の広範囲・汎用的な精度を期待した利用は困難です。特性を理解し、コストやセキュリティとのバランスを踏まえて活用を考えることがポイントです。

A クローズド環境でのセキュア利用

機微なデータを扱うなどのセキュリティ面の理由から、パブリッククラウド上での利用ではなく、自社オンプレ上やプライベートクラウド上で生成AIを利用したい。

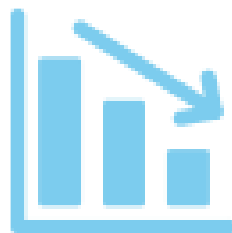


セキュリティ重視
特定の用途に必要な精度があればよい

B コスト抑制・レスポンス改善

特定の用途のみに活用するが、GPT-5のような汎用大規模LLMを利用した場合、継続利用コストがかかり、遅延も発生する。

軽量のtsuzumi 2を利用することでコストを抑制し、レスポンスを改善したい。



コストorレスポンス重視
特定の用途に必要な精度があればよい

C マルチLLMによる補完利用

GPT-5のような汎用大規模LLM + RAGを利用しているが、一部で日本語表現などに課題がある。

マルチLLMのオーケストレーションの一部にtsuzumi 2を活用したい (GPT-5で出力した結果をtsuzumi 2にインプットするなど)。



特定目的に対する精度重視
汎用大規模LLMと使い分けたい

実施できるタスクの例

以下 tsuzumi 2 が可能なタスクの例となります。
アダプタチューニングによって、タスクの精度向上、出力形式のコントロールが可能となります。

タスク	タスク説明	アダプタチューニングでできること
1.抽出	文章の中から固有名詞や固有表現を抽出する。 (人名、地域名などの一般ドメイン、病名などの専門ドメイン) 【活用例】 アンケート結果（自然言語項目）の集計など	抽出精度の向上 抽出フォーマットの固定
2.分類	文章中に含まれているワードをインプットに、内容を分類する。 【活用例】 カスタマーサポートにおける問合せ内容の カテゴリ設定など	分類精度の向上
3.要約	日本語の文章について、要約を生成出力する。 【活用例】 カスタマーサポートにおける問合せを記録 フォーマットの形式に沿って要約・出力するなど	要約結果のコントロール (フォーマットの固定化、テキスト長固定など)
4.翻訳	英語→日本語、日本語→英語の翻訳が可能。 【活用例】 カスタマーサポートにおける英語の問合せ内容を 日本語に翻訳して登録するなど	翻訳結果のコントロール (出力が必要な内容の過不足を小さくする)

※tsuzumi 2 on Azureではチューニング機能は利用できません。

実施できるタスクの具体的なイメージ

1.抽出

入力テキスト

安倍首相は15日首相官邸にて、台風19号の被害について「激甚災害」に指定する考えを明らかにしました。

抽出

安倍首相人名は15日日付表現首相官邸場所にて、台風19号イベントの被害について「激甚災害イベント」に指定する考えを明らかにしました。

2.要約

入力テキスト

このワクチンは複数の動物実験で、安全性や、効果的な免疫反応を引き起こすことが示されている。 今回の第1段階の後には、6000人を対象とした別の臨床試験が今年10月に予定されている。 インペリアル・コレッジ・ロンドンのチームは、2021年の早い時期からイギリスや海外でワクチンを配布できるようにしたいとしている。 世界中では約120のワクチンの開発が進められている。 英オックスフォード大学の専門家たちはすでに臨床試験を開始している。 新しいアプローチ 多くの従来のワクチンは、弱体化させたウイルスや改変したウイルスなどがもとになっている。 しかし今回のワクチンは新しいアプローチに基づいたもので、……

要約し、タイトルを抽出してください

出力

新型ウイルスワクチン、英大学が臨床試験を開始 300人対象

実施できるタスクの具体的なイメージ

3. 翻訳

テキスト

NTT DATA, a global digital business and IT services leader, today announced the launch of the tsuzumi 2 through the Microsoft Azure AI Models-as-a-Service (MaaS) offering. The development marks a fresh milestone in a 25-year collaboration committed to pioneering technological solutions that drive sustainability and innovation.

tsuzumi 2 – a Large Language Model (LLM) with robust capabilities in Japanese and English – is designed to ease environmental and financial challenges typically associated with LLMs. By adjusting model size without compromising performance, tsuzumi 2 makes advanced AI technologies, including Generative AI, accessible to a wider range of users and applications.

日本語に翻訳してください

出力

グローバルなデジタルビジネスおよびITサービスのリーダーであるNTTデータは、本日、Microsoft Azure AIモデルサービス(MaaS)の提供を発表しました。この開発は、25年にわたる持続可能性とイノベーションを推進するための技術ソリューションを先導するコミットメントの新たなマイルストーンです。

tsuzumi 2は、日本語と英語に強力な能力を持つ大規模言語モデル(LLM)であり、環境や財務の課題を軽減するために設計されています。tsuzumi 2は、パフォーマンスを犠牲にすることなくモデルサイズを調整することで、Generative AIなどの高度なAI技術をより多くのユーザーやアプリケーションにアクセス可能にします。

4. 分類

テキスト

ちょっと、近くの図書館に行ってみようかな。

この音量、ちょっと耳が痛いから下げてください。

機能に分類してください

出力

機能：行先検索
場所：図書館

機能：
ボリュームコントロール
ボリュームサイズ：
少し小さく

tsuzumi 2 on Azure の提供仕様

Azureにおけるtsuzumi 2の提供仕様は以下の通りとなります。

利用可能地域と提供リージョン

- 利用可能地域は日本に限定してサービスを提供します。
(Azureテナントの居住地に基づいて利用を制限します。)
- 本番ワークロードで利用する場合は、
レイテンシー及び国外へのデータ通信について考慮いただけるようお願い致します。

提供機能

- 推論機能を提供します。(Azureではチューニング機能は対象外)
- tsuzumi 2の機能アップデートに合わせて、アップデートバージョンのリリースは随時実施させていただきます。

料金仕様

- tsuzumi 2 on Azureはマネージドコンピューティングを採用しています。GPU起動時間単位で課金が発生します。利用価格については以下をご確認ください。
[NTT DATA tsuzumi 2 Instruct](#)
※マネージドコンピューティング：[Microsoft Foundry Models \(クラシック\) のデプロイ オプション - Microsoft Foundry \(classic\) portal | Microsoft Learn](#)

tsuzumi 2 on Azure の提供仕様

以下の注意事項を必ずお読みいただいた上でご利用ください。

- Azure AI Content Safety等ガードレール機能を持つサービスを利用することで、不適切な表現の出力を抑制することが可能です。倫理リスク抑制のため、**ガードレールの実装を推奨**します。
- Azure CSPサブスクリプションでは、CSPプロバイダによってtsuzumi 2 on Azureをご利用いただくことができない場合があります。利用希望される場合はご利用のCSPプロバイダをサポート窓口までお知らせください。または、EA契約のサブスクリプションをご利用ください。
- tsuzumi 2 on AzureはMicrosoft Foundryモデルではなく、Azure AI Hubベースのモデルとして提供されます。Azure AI Hubプロジェクトを作成の上、ご利用ください。
- 推論結果についてはテナント内のモデルにのみ適用され、他テナントのモデルに利用されることはありません。
- 大規模LLMとの汎用的な用途での精度比較・評価を行うことはお控えください。
- tsuzumi 2 on AzureのSLAは、Azure Machine Learningサービスに準拠します。
[Azure Machine Learning - サービスとしての ML | Microsoft Azure](#)
- 出力品質を担保するため、tsuzumi 2 on Azureに搭載しているシステムプロンプトをご利用いただくことを推奨します。
- tsuzumi 2 on Azureはマネージドコンピューティングを採用しています。本モデルをホストするにはご利用のサブスクリプションにGPUクォータが必要であり、事前にクォータ申請が必要となります。また、**GPU起動時間単位で課金が発生します**のでご注意ください。利用価格については以下をご確認ください。また、発生した料金の返金は実施できませんのでご注意ください。
[NTT DATA tsuzumi 2 Instruct](#)
※[Microsoft Foundry Models \(クラシック\) のデプロイ オプション - Microsoft Foundry \(classic\) portal | Microsoft Learn](#)

2 tsuzumi 2を使おう

チュートリアル

tsuzumi 2の利用手順をイメージしてもらうために、tsuzumi 2をデプロイしてからローカルから呼び出して回答を出すまでの手順を、画面操作を交えて説明します。
※課金が発生する点についてご了承ください。

チュートリアルの流れ



※AzureではGPU付き仮想マシンを使う前に、サブスクリプションごとのクォータ（上限）申請が必要です。
クォータはリージョンおよびVMファミリー単位で、GPU数ではなくvCPU数として管理されます。
Azure Portalの「使用量 + クォータ」から申請し、承認後にGPUリソースを作成できるようになります。
申請が必要な場合は数日の待ち時間が発生する可能性があります。

クォータのリクエスト

1. クォータサービスを開く
2. Machine learning を開く
3. マイ クォータを確認
4. クォータをリクエスト
5. リクエスト結果の確認

1. Azure ポータルの ① 検索欄に「クォータ」と入力して、表示された候補から ②クォータをクリック

クォータの
リクエスト

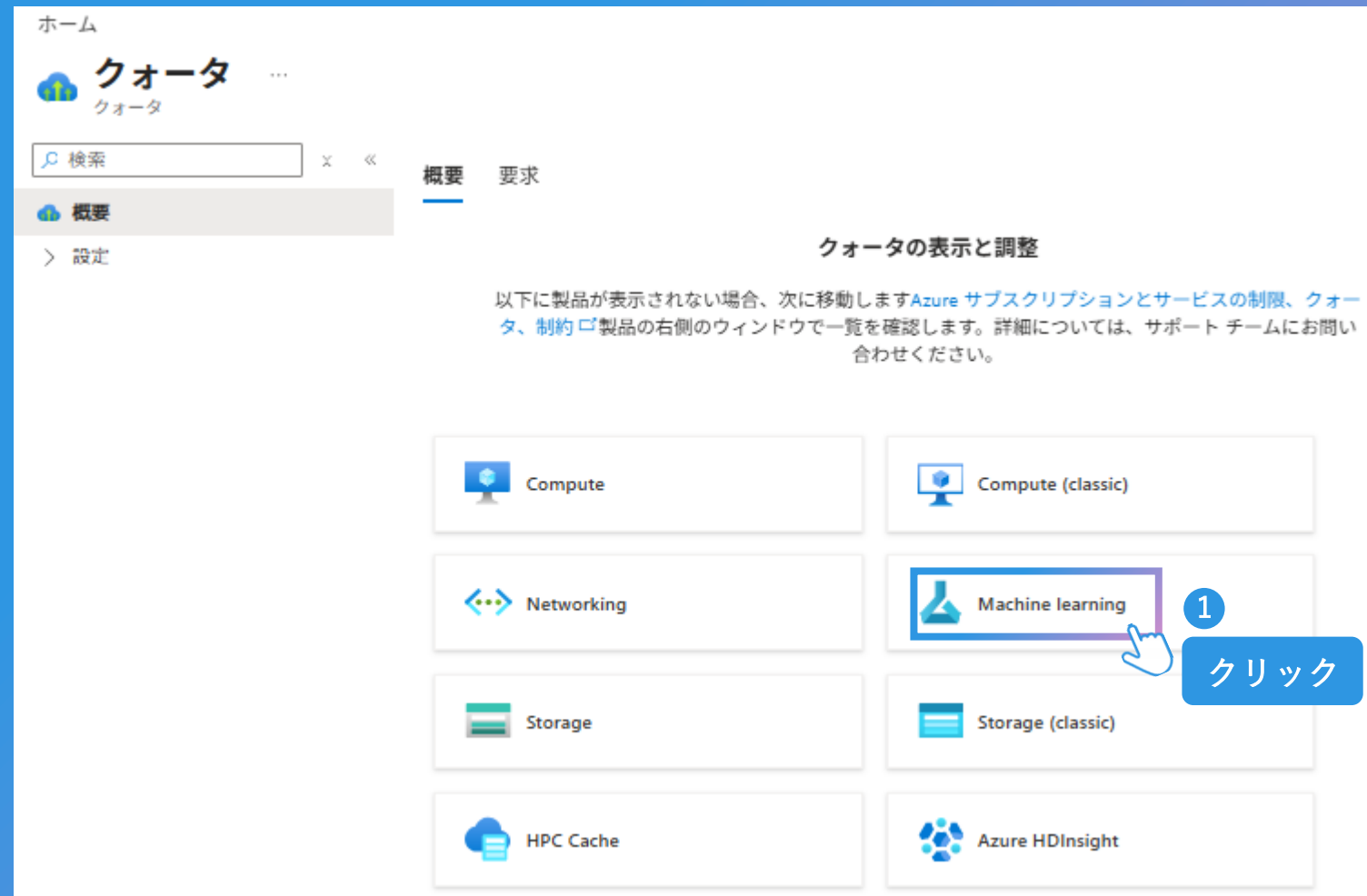
tsuzumi 2の
デプロイ

実行確認

クォータのリクエスト

1. クォータサービスを開く
2. Machine learning を開く
3. マイクォータを確認
4. クォータをリクエスト
5. リクエスト結果の確認

2. クォータの ①Machine learning をクリック



クォータの
リクエスト

tsuzumi 2の
デプロイ

実行確認

クォータのリクエスト

1. クォータサービスを開く
2. Machine learning を開く
3. マイクォータを確認
4. クォータをリクエスト
5. リクエスト結果の確認

3. 以下を入力あるいは選択しクォータを検索

- ① クォータ名
tsuzumi 2では以下のクォータが使用可能です。必要なクォータをリクエストして下さい。
「Standard NCADSA100v4 Family Cluster Dedicated vCPUs」
「Standard NCADSH100v5 Family Cluster Dedicated vCPUs」
- ② サブスクリプション
- ③ リージョン
「チェック」でクォータが無ければ④をクリック（クォータリクエスト）

クォータのリクエスト

クォータ名を入力

サブスクリプションを選択

リージョンを選択

チェック

クリック

クォータ | マイクォータ ...

クォータ

検索

新しいクォータ要求 最新の情報に更新 ダウンロード

Standard NCADSA* x プロバイダー: Machine learning サブスクリプション: リージョン: Japan East

マイクォータ

クォータグループ

発生したアラート (プレビュー)

アラートルール (プレビュー)

クォータ名	リージョン	サブスクリプション	現在の使用量 ↓	調整の要求
利用されていない (1)				
Standard NCADSA100v4 Family ...	Japan East		0% 0/0	チェック

クォータの
リクエストtsuzumi 2の
デプロイ

実行確認

クォータのリクエスト

1. クォータサービスを開く
2. Machine learning を開く
3. マイクォータを確認
4. クォータをリクエスト
5. リクエスト結果の確認

4. ① 要求数を入力

tsuzumi 2デプロイには最低(マシン1台)でも以下のクォータが必要になります
Standard NCADSA100v4 Family Cluster Dedicated vCPUs : 24
Standard NCADSH100v5 Family Cluster Dedicated vCPUs : 40

要求数入力後②をクリック

クォータのリクエスト

tsuzumi 2のデプロイ

実行確認

クォータのリクエスト

1. クォータサービスを開く
2. Machine learning を開く
3. マイクォータを確認
4. クォータをリクエスト
5. リクエスト結果の確認

5. クォータリクエストが成功し、クォータ上限が上がっていることをチェック

受取済み	新しい制限
Standard NCADSA100v4 Family Cluster Dedicated vCPUs	24 of 24

クォータ名	リージョン	サブスクリプション...	現在の使用量 ↓
Standard NCADSA100v4 Fa...	Japan East		0% 0/24

クォータの
リクエスト

tsuzumi 2の
デプロイ

実行確認

tsuzumi 2 のデプロイ

1. モデルカタログを選択

- 2. tsuzumi 2 を検索
- 3. tsuzumi 2 のデプロイ開始
- 4. デプロイ名、仮想マシンの選択
- 5. 価格と利用規約の同意と展開
- 6. デプロイ完了

1. Microsoft Foundryの①モデルカタログをクリック ※本手順はAzure AI hub プロジェクトを作成後に実施ください。



クォータのリクエスト

tsuzumi 2 のデプロイ

実行確認

tsuzumi 2のデプロイ

1. モデルカタログを選択
2. tsuzumi 2を検索
3. tsuzumi 2のデプロイ開始
4. デプロイ名、仮想マシンの選択
5. 価格と利用規約の同意と展開
6. デプロイ完了

2. ① tsuzumi 2を検索し、② tsuzumi 2をクリック

The screenshot shows a web interface for an AI model catalog. On the left is a navigation menu with items like '概要', 'モデルカタログ', 'プレイグラウンド', 'AI サービス', 'ビルドとカスタマイズ', 'エージェント', 'テンプレート', '微調整', 'コンテンツの解釈', 'プロンプトフロー', '監視と最適化', 'トレース', '監視', '保護と管理', '評価', 'ガードレールとコントロール', and 'リスク+ア'. The main content area is titled 'カスタム AI ソリューションを構築するための適切なモデル' and includes a 'モデル ランキング' section with two columns: '品質' (Quality) and '安全性' (Safety). Below the ranking is a search bar containing 'tsuzumi' and a list of search results. The first result is 'tsuzumi-7b チャットの完了' and the second is 'tsuzumi2 チャットの完了'. A red box highlights the search bar with a '1' and a red box highlights the 'tsuzumi2' result with a '2' and a 'クリック' (Click) button with a hand cursor pointing to it.

品質	安全性
1 claude-opus-4-6	1 gpt-5.4-mini
2 gpt-5.4	2 gpt-5.2-codex
3 claude-opus-4-5	3 gpt-5.1-codex-mini

検索: tsuzumi

- tsuzumi-7b
チャットの完了
- tsuzumi2
チャットの完了

クォータのリクエスト

tsuzumi 2のデプロイ

実行確認

tsuzumi 2のデプロイ

1. モデルカタログを選択
2. tsuzumi 2を検索
3. tsuzumi 2のデプロイ開始
4. デプロイ名、仮想マシンの選択
5. 価格と利用規約の同意と展開
6. デプロイ完了

3. ① このモデルを使用するをクリック

このスクリーンショットは、Azure AI Studioの「tsuzumi2」モデルの詳細ページを示しています。左側のナビゲーションメニューには「モデルカタログ」が選択されています。中央のメインコンテンツには、「このモデルを使用する」ボタンが赤い枠で囲まれ、赤い矢印と「1 クリック」というラベルで強調されています。右側の「クイックファクト」セクションには、モデルID「azureml://registries/azureml-nttdatacorp-p/models/tsuzumi2/versions/1」が示されています。

クイックファクト (NTT Data Corporation)

tsuzumi2

チャットの完了

前回トレーニングデータが更新済み
使用できません

価格 >
単価はデプロイの種類によって異なります。モデルのより詳しい説明を見る

モデル ID
コードでモデルを配置するときにこのモデル ID を参照する

azureml://registries/azureml-nttdatacorp-p/models/tsuzumi2/versions/1

データ、メディア、言語

Property	説明
サポートされるデータ型	入力 text 出力 text
サポートされている言語	ja, en

モデル プロバイダーからの詳細

To understand the capabilities, we compare tsuzumi 2 with a set of models over a variety of benchmarks using our internal benchmark platform. At the high-level overview of the model quality on representative benchmarks:

Category	Benchmark	tsuzumi2-28B-instruct (8bit)	gemma-3-27b-it	gemma-2-27b-it	Qwen2.5-32B-Instruct	Qwen2.5-72B-Instruct	Llama-3-ELYZA-JP-8B	Llama-3.3-70B-Instruct	calm3-22b-chat	sarashin-3b-instr v0.1
汎用ベンチマーク	Japanese MT-bench (turn-1)	8.78	9.14	8.22	8.70	9.06	6.60	8.06	7.42	7.67

クォータのリクエスト

tsuzumi 2のデプロイ

実行確認

tsuzumi 2のデプロイ

1. モデルカタログを選択
2. tsuzumi 2を検索
3. tsuzumi 2のデプロイ開始
4. デプロイ名、仮想マシンの選択
5. 価格と利用規約の同意と展開
6. デプロイ完了

4. ①デプロイ名を入力して②仮想マシンを選択して③次へ進む

展開 tsuzumi2

デプロイ名* ①
tsuzumi2-1

仮想マシン* ②
Standard_NC24ads_A10... 24 コア, 220 GB (RAM), 64 GB (ディスク), \$5.33/時間

展開のサマリー カスタマイズ

環境	仮想マシンのインスタンス数
既定の環境	1
プロジェクト名	エンドポイント
proj-hub-japan-east-tsuzumi2	proj-hub-japan-east-tsuzu-aefxd
推論データ収集 (プレビュー)	
無効	

③ 次へ キャンセル

デプロイ後は利用時間に応じた課金が発生します。
(選択したGPUインスタンス利用料とtsuzumi 2 利用料)
使用するときだけデプロイし、使用後は削除することをお勧めします。

クォータの
リクエスト

tsuzumi 2の
デプロイ

実行確認

クリック

tsuzumi 2のデプロイ

1. モデルカタログを選択
2. tsuzumi 2を検索
3. tsuzumi 2のデプロイ開始
4. デプロイ名、仮想マシンの選択
- 5. 価格と利用規約の同意と展開**
6. デプロイ完了

5. ①価格・利用規約を確認・同意して②展開をクリック

展開 tsuzumi2

【ご利用に関する注意事項】

NTT DATA 本モデルは現在プレビュー版のため、動作保証しておりません。
また、利用後の返金は承ることができませんのであらかじめご了承ください。
[詳細を読む](#)

価格の内訳

モデルの価格	¥118.8602 /時間	詳細
Azure Compute	この価格は、コンピューティングの構成によって異なります。	詳細

利用規約

私は (a) 上記の各 Marketplace サービスに関連する法律条項とプライバシーに関する声明に同意し、(b) サービスの利用を停止するまでの間、Microsoft より、その利用に伴う料金 (適用される税を含む) が、現在の支払い方法において Azure サブスクリプションと同じ請求頻度で課金または請求されることを認め、(c) Microsoft が私の連絡先情報および取引の詳細 (サービスに関連付けられた使用量を含む) を、サービスを販売する業者と共有し、販売業者がこの製品に関して私に連絡できるようにする可能性があること

[Azure Marketplace の使用条件](#) [プライバシーポリシー](#) [ライセンス契約](#)

私は価格について理解し、利用規約に同意します

戻る 展開 キャンセル

クォータのリクエスト

tsuzumi 2のデプロイ

実行確認

tsuzumi 2 のデプロイ

1. モデルカタログを選択
2. tsuzumi 2 を検索
3. tsuzumi 2 のデプロイ開始
4. デプロイ名、仮想マシンの選択
5. 価格と利用規約の同意と展開

6. デプロイ完了

6. プロビジョニングが成功していることを確認

The screenshot displays the Azure ML portal interface for a deployment named 'tsuzumi2-1'. The 'プロビジョニングの状態' (Provisioning Status) is highlighted with a red box and a magnifying glass, showing a green checkmark and the text '成功' (Success). A purple callout box with the text 'チェック' (Check) points to this status. The 'エンドポイント' (Endpoint) section shows the target URI as 'https://japaneast.inference.ml.azure.co...' and the Swagger URI as 'https://project-dev-japan-east-20-dnfus.japaneast.inference.ml.azure.co...'. The 'パブリック ネットワーク アクセス' (Public Network Access) is set to 'Enabled'.

配置情報	
名前	tsuzumi2-1
プロビジョニングの状態	成功
最終更新日時	Apr 13, 2026 10:16 AM
作成者	作成日
	Apr 13, 2026 10:16 AM
トラフィックの割り当て	インスタンス数
100%	1
コンピューティングの種類	SKU
専用	Standard_NC24ads_A100_v4
デプロイ時のモデルの追加料金	
¥118.8602 1 時間あたり	
モデル ID	
azureml://registries/azureml-nttdatacorp-p/models/tsuzumi2/versions/1	
作成日	モデルバージョン
03/26/2026	1
更新日	
04/09/2026	

エンドポイント	
ターゲット URI	https://japaneast.inference.ml.azure.co...
認証の種類	Key
主キー	再生成
Swagger URI	https://project-dev-japan-east-20-dnfus.japaneast.inference.ml.azure.co...
パブリック ネットワーク アクセス	コンピューティングの種類
Enabled	マネージド

クォータのリクエスト

tsuzumi 2 のデプロイ

実行確認

実行確認

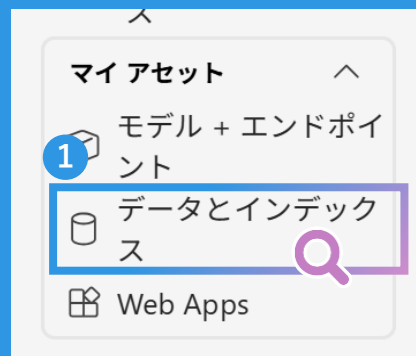
1. テストによるチャット確認

2. curlによるチャット確認

1. テストを使用したチャット確認

- ① 左ペイン「モデル+エンドポイント」
- ② 右ペインから作成したモデルを選択
- ③ 「テスト」タブを選択
- ④ サンプル推論の Input 欄に以下の JSON を貼付
- ⑤ 下部の「テスト」ボタンを押下

```
{"messages":[{"role":"user","content":"こんにちは"}]}
```



名前	モデル名
project	エンドポイント
tsuzumi2-1	tsuzumi2



クォータの
リクエスト

tsuzumi 2の
デプロイ

実行確認

実行確認

1. テストによるチャット確認

2. curlによるチャット確認

2. curlによるチャット確認 コマンドの ①エンドポイント、②主キーを実際の値に置換して実行

The screenshot shows the configuration page for 'tsuzumi2'. On the left, under '配置情報' (Configuration Information), the name is 'tsuzumi2', the deployment status is '成功' (Success), and the creator is '友克 浦'. On the right, under 'エンドポイント' (Endpoint), the 'ターゲット URI' (Target URI) is highlighted with a blue box and labeled '1', and the '主キー' (Main Key) is highlighted with a blue box and labeled '2'. The main key is represented by a series of dots.

```
■ Windows (コマンドプロンプト)  
curl -X POST ①エンドポイント ^  
-H "Authorization: Bearer ②主キー" ^  
-H "Content-Type: application/json" ^  
-d "{\"messages\": [{\"role\": \"user\", \"content\": \"こんにちは\"}]}"
```

```
■ Linux (bash)  
curl -X POST ①エンドポイント ¥  
-H 'Authorization: Bearer ②主キー' ¥  
-H 'Content-Type: application/json' ¥  
-d '{"messages": [{"role": "user", "content": "こんにちは"}]}'
```

クォータの
リクエスト

tsuzumi 2の
デプロイ

実行確認

3 推論を使いこなす

推論の仕組み

一般的に推論は、入力文のトークナイズ、モデルによる推論、推論結果のデコードという流れで処理されます。

	処理概要	処理例
入力	LLMへの指示 (=プロンプト)	日本で一番高い山は?
トークナイズ	プロンプトを、モデルが解釈できるように トークンと呼ばれるIDの列に変換	日本で一番高い山は? → [1,143,54,28,160,500,378,3]
推論	入力ID列をもとに、 次に続く文章を 確率的に推論	[1,143,54,28,160,500,378,3] → [1,143,54,28,160,500,378,3,432,39,9,0]
デコード	推論結果のID列を自然文に変換	[432,39,9,0] → 富士山です。
出力	推論結果	富士山です。

推論パラメータの意味

推論時の主なパラメータとその意味は以下の通りです。

パラメータ名	デフォルト値	解説
max_tokens	32768	一回の推論時にOutput可能な最大トークン数（プロンプトやシステムのトークン数を含む） 設定可能な値は最大で32768となります。
temperature	0.6	応答のランダム性を決める[0.0, 1.0]の値 0.0に近づけるほど出力が一意に定まり、 1.0に近いほどランダムになります。
top_k	-1 (制限なし)	予測する次トークンの候補の数（ ≥ 1 ） -1を設定すると制限なしとなり、1を設定すると出力が一意に定まります。
repetition_penalty	1.0	同じ文章や言い回しを繰り返し出力することを抑制するパラメータ

【参考】図解：temperature / top-k

Temperatureとtop-kは、どちらも次トークンの選び方を制御するパラメータです。LLMは、入力の次に続く単語（トークン）を確率的に生成しますが、この確率的生成の際の、**確率分布や次単語候補数を調整するために用いるのがこれらのパラメータとなります。**

パラメータ	例1 「私が好きなのは」の次トークン	例2 「私が好きなのは」の次トークン
temperature 次トークンの確率分布を決める 値が大きいほどなだらかな分布になる	temperature=0.01の確率分布 	temperature=0.7の確率分布
top-k 次トークンとして考慮する候補数を決定	temperature=0.7の確率分布 top-k=3 	temperature=0.7の確率分布 top-k=5

4 外部アプリケーションと連携する

外部アプリケーションとの連携方法

外部アプリケーション連携として、サンプルアプリを使ったご利用方法をご紹介します。

フロントエンド

- Chainlit※ を利用した、シンプルなフロントエンドを用意しました。
- ストリーミング応答およびマルチターン対話に対応しています。
- 手軽にデプロイした tsuzumi2 の動作確認が可能です。
- 開発・検証用途のサンプルとしてご利用ください。
- 以下よりダウンロードできます。

<https://www.nttdata.com/jp/ja/-/media/nttdatajapan/files/lineup/tsuzumi/tsuzumi2.zip>

※Chainlitについて

- チャットAIアプリを簡単に作成できるPythonのOSS(オープンソースソフトウェア)です。
- OSSのため、公式なサポートや品質保証はございません。

フロントエンド実行手順

1. エンドポイント情報の確認

2. サンプルアプリの準備

3. アプリ起動

4. 推論実行

1. モデル+エンドポイントタブから①ターゲット URI と②主キーを保存

The screenshot shows the configuration page for a model endpoint named 'tsuzumi2'. The page is divided into two main sections: 'Configuration Information' and 'Endpoint'. The 'Endpoint' section is highlighted with a red box, and the 'Target URI' and 'Key' fields are marked with red circles and numbered 1 and 2 respectively, indicating they should be saved.

Configuration Information

名前	最終更新日時
tsuzumi2	Apr 13, 2026 10:16 AM
プロビジョニングの状態	作成日
成功	Apr 13, 2026 10:16 AM
作成者	インスタンス数
	1
トラフィックの割り当て	
100%	

Endpoint

ターゲット URI
https://project-dev-japan-east-20-dnfus.japaneast.inference.ml.azure.com/v...

認証の種類
Key

主キー
..... 再生成

フロントエンド実行手順

1. エンドポイント情報の確認
2. サンプルアプリの準備
3. アプリ起動
4. 推論実行

2.
 - ① サンプルプログラムをダウンロードして展開
 - ② README.md に従いセットアップを実施
 - ③ env.sample ファイルを .env という名前のファイルとして複製
 - ④ .env ファイル中の ENDPOINT と API_KEY に 1. で取得した ターゲット URI と主 キーを転記、保存

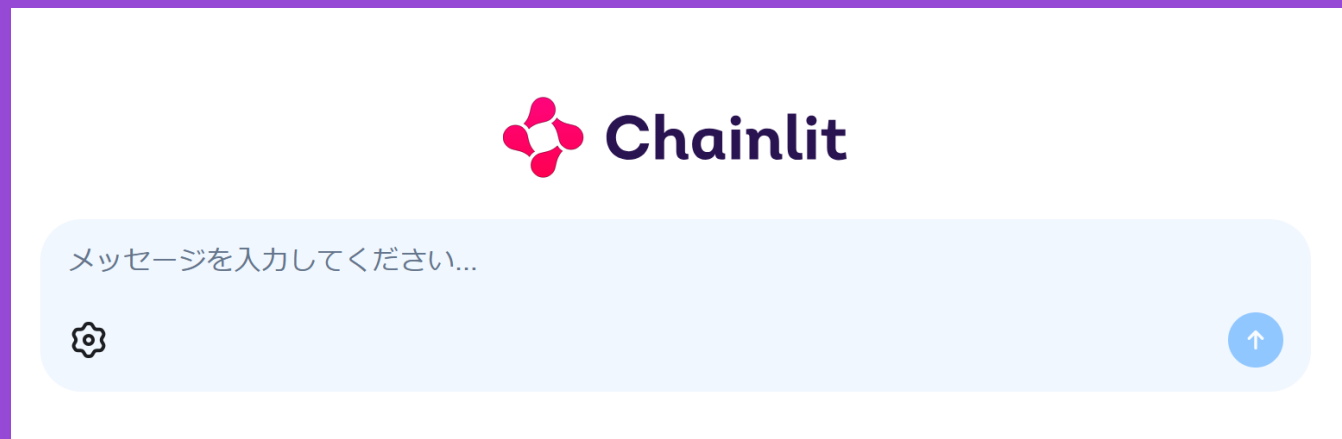
```
ENDPOINT="ターゲット URIの値"  
API_KEY="主キー の値"
```

フロントエンド実行手順

1. エンドポイント情報の確認
2. サンプルアプリの準備
3. アプリ起動
4. 推論実行

3. フロントエンドアプリ起動

README.md に従い python の仮想環境を設定したのち、アプリを起動してください。ブラウザで <http://127.0.0.1:8000> にアクセスすると以下の画面が表示されます。



フロントエンド実行手順

1. エンドポイント情報の確認
2. サンプルアプリの準備
3. アプリ起動
4. 推論実行

4. ①入力欄にメッセージを入力し ②送信してください。推論結果がストリーム表示されます。メッセージは履歴に残り、会話が継続できます。③を押すとメッセージ履歴を破棄して新しい会話が始まります。

The screenshot shows a chat interface with a white background and a blue header. In the top right corner, there are links for '説明書' (Documentation) and a settings icon. A blue message bubble on the right contains the text 'tusuzumi2 について5行'. Below it, a red flower icon is followed by the text 'もちろんです。以下に tusuzumi2 についての概要を5行でまとめました：'. A numbered list follows: 1. **開発元**：NTT株式会社が開発した日本語向けAIアシスタントです。 2. **特徴**：高度な自然言語処理能力と多言語対応が可能で、ユーザーのニーズに合わせた柔軟な応答ができます。 3. **利用シーン**：カスタマーサポート、コンテンツ作成、学習支援など、幅広い分野で活用できます。 4. **セキュリティ**：プライバシー保護とデータセキュリティに重点を置き、安全な環境で利用できます。 5. **更新情報**：定期的にアップデートされ、最新の技術とトレンドに対応しています。 Below the list, it says 'この情報は2025年4月までの知識に基づいています。'. At the bottom, there is a blue input field with a placeholder 'メッセージを入力してください...' and a blue send button with an upward arrow. A red flower icon is to the left of the input field. A blue box highlights the input field and the send button, with a '1' in a blue circle to the left and a '2' in a blue circle to the right. A blue box highlights the send button, with a '3' in a blue circle to the left. At the very bottom, a small disclaimer reads '大規模言語モデルは間違いを犯す可能性があります。重要な情報については確認を検討してください。'

①

②

③

説明書

tusuzumi2 について5行

もちろんです。以下に tusuzumi2 についての概要を5行でまとめました：

1. **開発元**：NTT株式会社が開発した日本語向けAIアシスタントです。
2. **特徴**：高度な自然言語処理能力と多言語対応が可能で、ユーザーのニーズに合わせた柔軟な応答ができます。
3. **利用シーン**：カスタマーサポート、コンテンツ作成、学習支援など、幅広い分野で活用できます。
4. **セキュリティ**：プライバシー保護とデータセキュリティに重点を置き、安全な環境で利用できます。
5. **更新情報**：定期的にアップデートされ、最新の技術とトレンドに対応しています。

この情報は2025年4月までの知識に基づいています。

メッセージを入力してください...

↑

大規模言語モデルは間違いを犯す可能性があります。重要な情報については確認を検討してください。

5 お問い合わせ

お問い合わせ先

tsuzumi 2に関するお問い合わせ

tsuzumi 2 on Azure版に関するお問い合わせは、以下のメールアドレス宛にご連絡ください。
ただし、モデル以外の内容に関するお問い合わせについては、[Azureのサポートリクエスト](#)を起票いただけるようお願い致します。
連絡先：support_tsuzumi_maas@hml.nttdata.co.jp

構築・チューニング支援サービス

法人・団体のお客様向けに有償の支援サービスもご相談可能です。
ご希望の方は、以下の連絡先にお問い合わせください。
連絡先：nttd_tsuzumi@hml.nttdata.co.jp

更新履歴

版	更新日付	変更箇所	変更内容
1.0	2026/5/20	初版	初版

