NTT DATA

tsuzumi

tsuzumi 2 on Azure ユーザーガイド v1.3

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
- 応用パラメータ紹介 31

4 外部アプリケーションと連携する 36

- 外部アプリケーションとの連携方法 37
- フロントエンド実行手順 38

5 お問い合わせ 42

- お問い合わせ先 43

© 2026 NTT DATA Japan Corporation

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は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

※tsuzmi 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。

© 2026 NTT DATA Japan Corporation

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
クローズドデータをお客さま環境で
セキュアに活用可能

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 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

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

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。

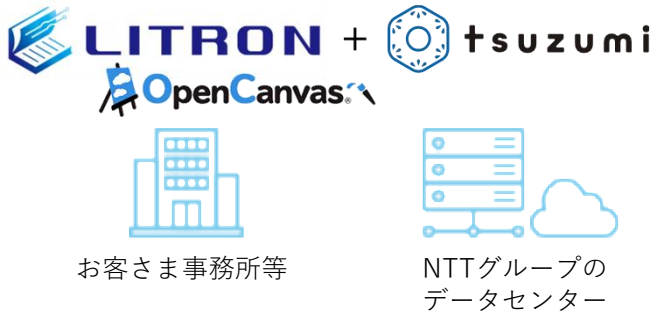
© 2026 NTT DATA Japan Corporation

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。

軽量な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。

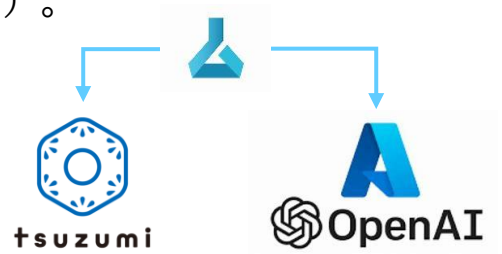


コストorレスポンス重視
特定の用途に必要な精度があればよい

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

© 2026 NTT DATA Japan Corporation

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）

※tsuzmi 2 on Azureではチューニング機能は利用できません。

© 2026 NTT DATA Japan Corporation

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

© 2026 NTT DATA Japan Corporation

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れる？

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

© 2026 NTT DATA Japan Corporation

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にお願い致します。
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。 NTTDATA 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

© 2026 NTT DATA Japan Corporation

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。
  NTTDATA 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

© 2026 NTT DATA Japan Corporation

tsuzumi

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



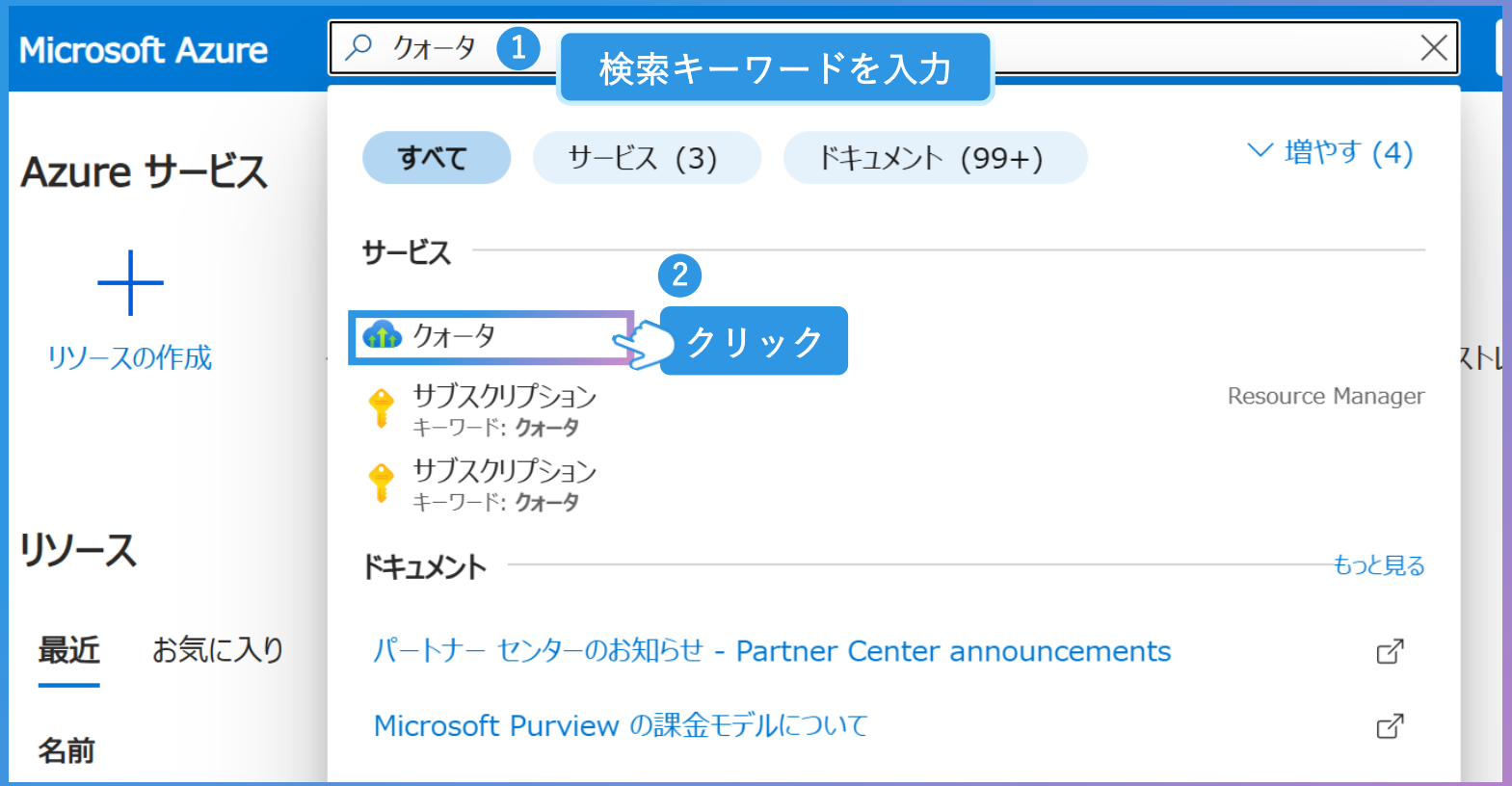
※AzureではGPU付き仮想マシンを使う前に、サブスクリプションごとのクォータ（上限）申請が必要です。
クォータはリージョンおよびVMファミリー単位で、GPU数ではなくvCPU数として管理されます。
Azure Portalの「使用量 + クォータ」から申請し、承認後にGPUリソースを作成できるようになります。
申請が必要な場合は数日の待ち時間が発生する可能性があります。

© 2026 NTT DATA Japan Corporation

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 ❶ 検索欄に「クォータ」と入力して、表示された候補から ❷クォータをクリック

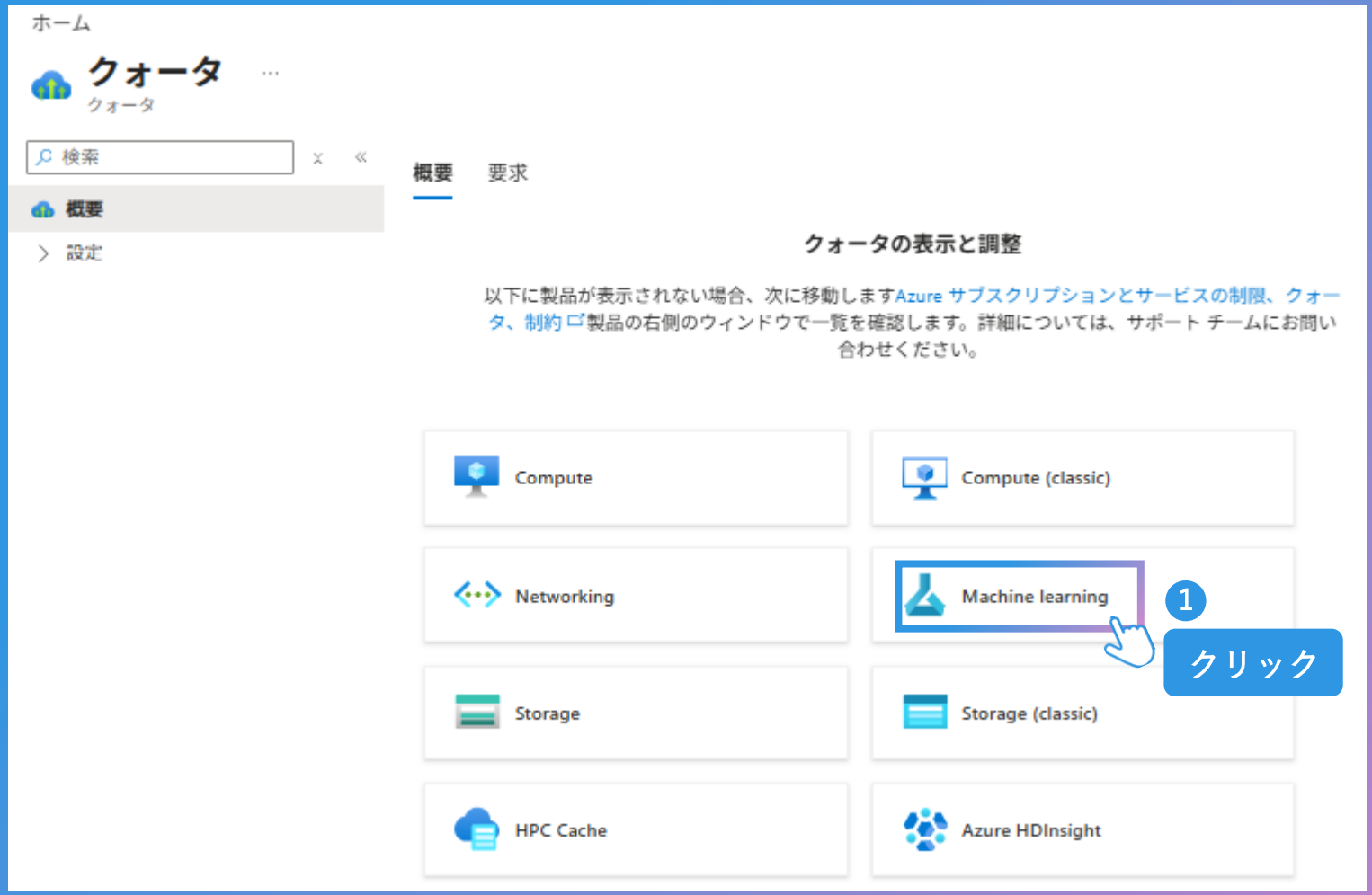


© 2026 NTT DATA Japan Corporation

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



© 2026 NTT DATA Japan Corporation

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）

※クォータ申請において、Microsoft側のGPUリソースの在庫状況により、より小さいGPUサイズを推奨される場合があります。
その際は、tsuzumi 2の利用にはA100またはH100が必要である旨をサポートへお伝えください。



© 2026 NTT DATA Japan Corporation

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. ❶ 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❷をクリック



© 2026 NTT DATA Japan Corporation

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





クォータのリクエスト
tsuzumi 2のデプロイ
実行確認

© 2026 NTT DATA Japan Corporation

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❶ 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。



© 2026 NTT DATA Japan Corporation

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. ❶ tsuzumi 2を検索し、❷ tsuzumi 2をクリック



© 2026 NTT DATA Japan Corporation

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. ❶このモデルを使用するをクリック



© 2026 NTT DATA Japan Corporation

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. ❶デプロイ名を入力して❷仮想マシンを選択して❸次へ進む



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。

© 2026 NTT DATA Japan Corporation

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. ❶価格・利用規約を確認・同意して❷展開をクリック



© 2026 NTT DATA Japan Corporation

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



© 2026 NTT DATA Japan Corporation

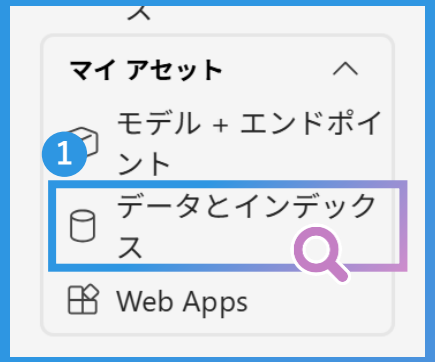
実行確認

1. テスト によるチャット確認
2. curl によるチャット確認

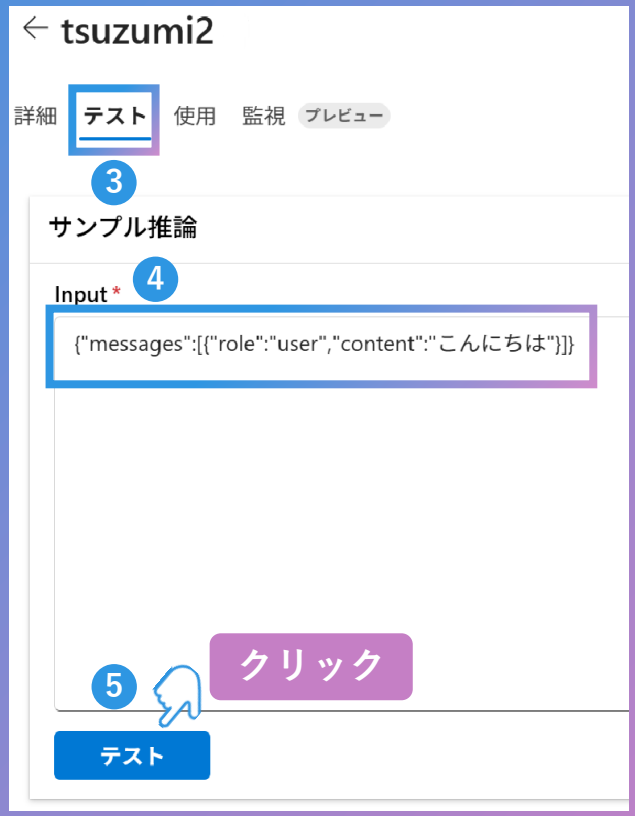
1. テストを使用したチャット確認

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

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







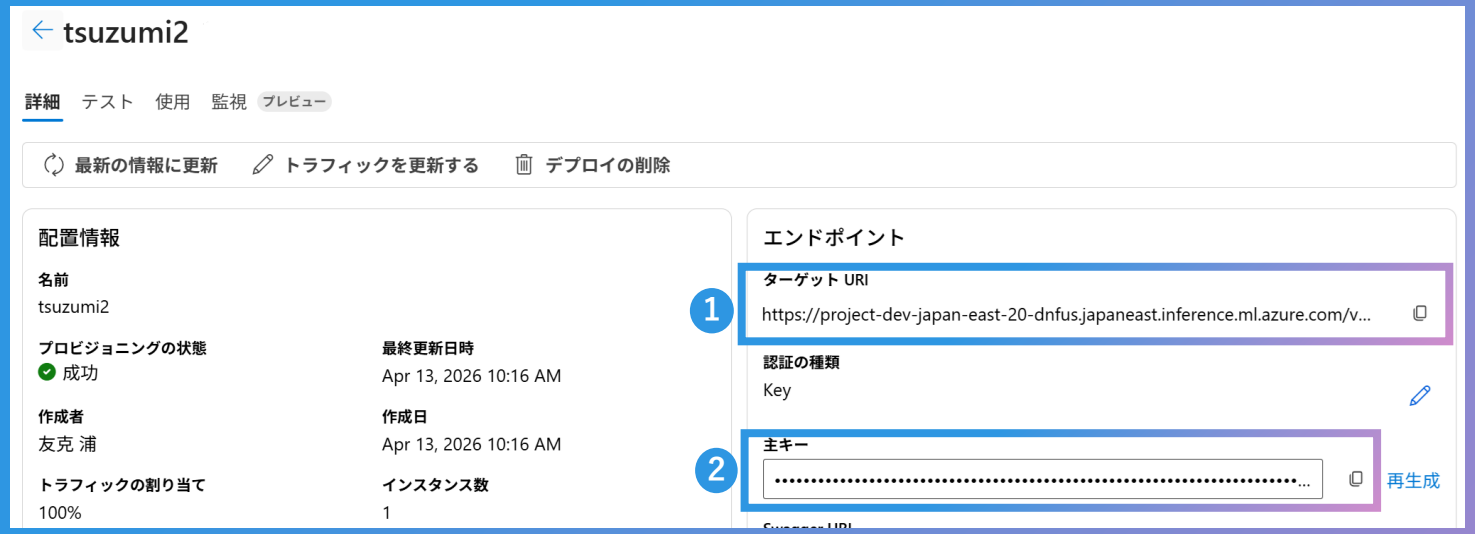
クォータのリクエスト
tsuzumi 2のデプロイ
実行確認

© 2026 NTT DATA Japan Corporation

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 ❶エンドポイント、❷主キーを実際の値に置換して実行



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

■ Linux (bash)
```
curl -X POST ❶エンドポイント ¥
  -H 'Authorization: Bearer ❷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

© 2026 NTT DATA Japan Corporation

tsuzumi

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。

© 2026 NTT DATA Japan Corporation

推論パラメータの意味

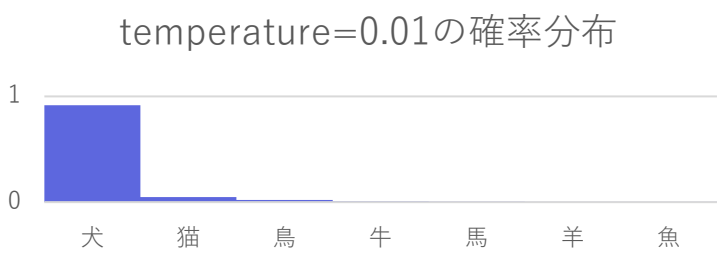
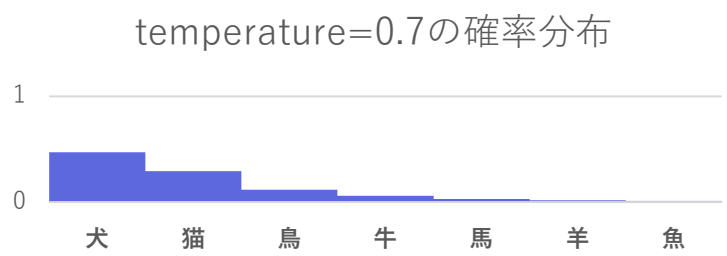
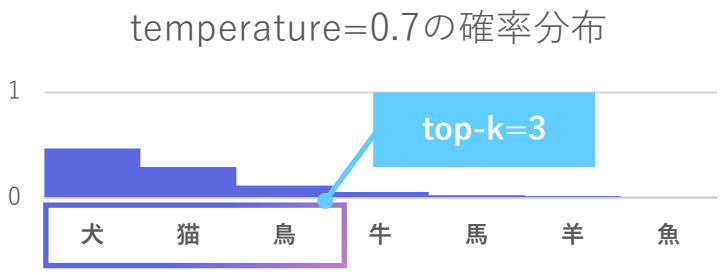
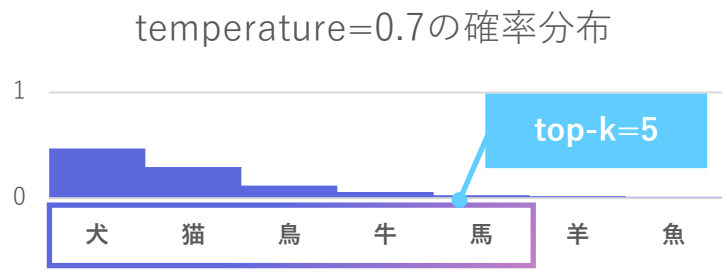
推論時の主なパラメータとその意味は以下の通りです。

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

© 2026 NTT DATA Japan Corporation

【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		
top-k 次トークンとして考慮する**候補数**を決定		

© 2026 NTT DATA Japan Corporation

応用パラメータ紹介

推論時に使用できる応用的なパラメータを紹介します。

パラメータ名	設定値	解説
tools	"function": { “name”: “ファンクション名", “description”: “ファンクションの説明", “parameters”: {パラメータ}, “required”: [必須パラメータ]}	外部ツールの呼び出し(function calling)ができます。 呼び出し対象のツールの定義をtoolsに json形式で与えることで使用し、モデルが与えられた外部ツールを呼び出すことが適切と判断した場合、レスポンスに呼び出すべきfunction名とその引数が含まれます。
enable_thinking	True / False	reasoningを行います。かつモデルの思考過程（<think> </think> タグ内の推論）をレスポンスに含めて出力します。数学的推論や論理的判断の根拠を確認する場合にも有用です。
documents	“id”: 数値, “title”: “タイトル", “text”: “内容” “date”: “日付” “url”: “URL”	RAG（検索結果に基づく生成）を実施する際に、根拠となる情報を指定します。 文書ID（id）、タイトル（title）、本文（text）、日付（date）、URL（url）をキーに持つ辞書のリストの形式で、検索結果となる文書集合を与えます。text以外は省略が可能です。
today_date	“YYYY-MM-DD” 例：”2026-05-15”	本日の日付を与えます。日付を加味して推論を行います。

© 2026 NTT DATA Japan Corporation

tools 具体例

toolsを使う場合の具体的なソースコード例を示します。

リクエスト

```
{
  "messages": [
    {
      "role": "user",
      "content": "明日の東京の天気を教えて"
    }
  ],
  "max_tokens": 1024,
  "tools": [
    {
      "type": "function",
      "function": {
        "name": "getWeather",
        "description": "指定した地域の天気予報を取得する",
        "parameters": {
          "type": "object",
          "properties": {
            "location": {
              "type": "string",
              "description": "地域名"
            },
            "date": {
              "type": "string",
              "description": "日付(YYYYMMDD)"
            }
          },
          "required": [
            "location",
            "date"
          ]
        }
      }
    }
  ],
  "tool_choice": "auto"
}
```

レスポンス

```
{
  "role": "assistant",
  "content": null,
  "refusal": null,
  "annotations": null,
  "audio": null,
  "function_call": null,
  "tool_calls": [
    {
      "id": "chatcmpl-tool-92e6946b0015194b",
      "type": "function",
      "function": {
        "name": "getWeather",
        "arguments": "{¥"location¥": ¥"東京¥", ¥"date¥": ¥"20260528¥"}"
      }
    }
  ],
  "reasoning": null
}
```

© 2026 NTT DATA Japan Corporation

enable_thinking 具体例

enable_thinkingを使う場合の具体的なソースコード例を示します。

リクエスト	レスポンス

{
"messages": [
{"role": "user", "content": "2次方程式 x^2 - 5x + 6 = 0 を解いてください"}
],
"max_tokens": 2048,
"chat_template_kwargs":{"enable_thinking": True}
}
	{
	"role": "assistant",
	"content": "<think>¥nまず、与えられた2次方程式は x² - 5x + 6 = 0 です。この方程式を解くためには、因数分解や平方完成、または判別式を用いた方法が考えられます。¥n¥nまず、因数分解を試みます。この方程式を (x - a)(x - b) = 0 の形に分解できれば、a と b は係数から求められます。つまり、a + b = 5 であり、ab = 6 です。この条件を満たす整数を探します。例えば、2 と 3 です。2 + 3 = 5 であり、2 × 3 = 6 です。したがって、方程式は (x - 2)(x - 3) = 0 と分解できます。したがって、解は x = 2 または x = 3 です。¥n¥n別の方法として、判別式を用いる方法もあります。判別式 D = b² - 4ac を計算します。ここでは a = 1、b = -5、c = 6 です。D = (-5)² - 4*1*6 = 25 - 24 = 1 です。判別式が正の場合、実数の2つの解があります。解は x = [5 ± √1]/2 = [5 ± 1]/2 で、x = (5+1)/2 = 3 または x = (5-1)/2 = 2 となります。¥n¥nどちらの方法でも、解は x = 2 と x = 3 です。¥n</think>¥n¥n与えられた2次方程式 ¥¥(x^2 - 5x + 6 = 0¥¥) を解きます。¥n¥n### 方法1：因数分解¥n方程式を ¥¥((x - a)(x - b) = 0¥¥) の形に分解します。¥n条件：¥¥(a + b = 5¥¥) および ¥¥(ab = 6¥¥)¥nこの条件を満たす整数は ¥¥(a = 2¥¥)、¥¥(b = 3¥¥) です。¥nしたがって、方程式は ¥¥((x - 2)(x - 3) = 0¥¥) と分解できます。¥n¥n解：¥n¥¥[¥nx = 2 ¥¥quad ¥¥text{または} ¥¥quad x = 3¥n¥¥]¥n¥n### 方法2：判別式¥n判別式 ¥¥(D = b^2 - 4ac¥¥) を計算します。¥n¥¥(a = 1¥¥)、¥¥(b = -5¥¥)、¥¥(c = 6¥¥) なので¥n¥¥[¥nD = (-5)^2 - 4 ¥¥times 1 ¥¥times 6 = 25 - 24 = 1¥n¥¥]¥n判別式が正の場合、実数の2つの解があります。¥n解：¥n¥¥[¥nx = ¥¥frac{5 ¥¥pm ¥¥sqrt{1}}{2} = ¥¥frac{5 ¥¥pm 1}{2}¥n¥¥]¥nしたがって、¥n¥¥[¥nx = ¥¥frac{5 + 1}{2} = 3 ¥¥quad ¥¥text{または} ¥¥quad x = ¥¥frac{5 - 1}{2} = 2¥n¥¥]¥n¥n### 結論¥n方程式の解は ¥¥(x = 2¥¥) と ¥¥(x = 3¥¥) です。",
	（以下略）
	}

© 2026 NTT DATA Japan Corporation

documents 具体例

documentsを使う場合の具体的なソースコード例を示します。

リクエスト

```
{
   "messages": [
      {"role": "user", "content": "年次有給休暇は何日付与されますか？"}
   ],
   "max_tokens": 1024,
   "chat_template_kwargs": {
      "documents": [
         {
            "id": 1,
            "title": "就業規則 第20条",
            "text": "勤続年数が6ヶ月以上の社員には年次有給休暇として10日を付与する。勤続年数が1年6ヶ月以上の場合は11日とし、以降1年ごとに2日を加算する。"
         }
      ]
   }
}
```

レスポンス

```
{
  "role": "assistant",
  "content": "年次有給休暇の付与日数は、勤続年数に応じて次のように決まります。¥n¥n- 勤続年数が6ヶ月以上の場合、10日が付与されます[[1]]。¥n- 勤続年数が1年6ヶ月以上になると、11日が付与されます[[1]]。¥n- 以降、勤続年数が1年増えるごとに2日ずつ加算されます[[1]]。¥n¥n[1]: about:blank ¥"就業規則 第20条¥"",
  （以下略）
}
```

© 2026 NTT DATA Japan Corporation

today_date 具体例

todat_dateを使う場合の具体的なソースコード例を示します。

リクエスト	レスポンス

{
"messages": [
{"role": "user", "content": "今日から1週間後は何日ですか？"}
],
"max_tokens": 1024,
"chat_template_kwargs":{"today_date": "2026-05-15"}
}
	{
	"role": "assistant",
	“content”: “現在の日付は **2026年5月15日** です。¥n1週間後は **2026年5月22日** になります。¥n¥n（もし別の日付を基準に計算したい場合は、その日付を教えていただければ再計算します。）”,
	（以下略）
	}

© 2026 NTT DATA Japan Corporation

tsuzumi

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

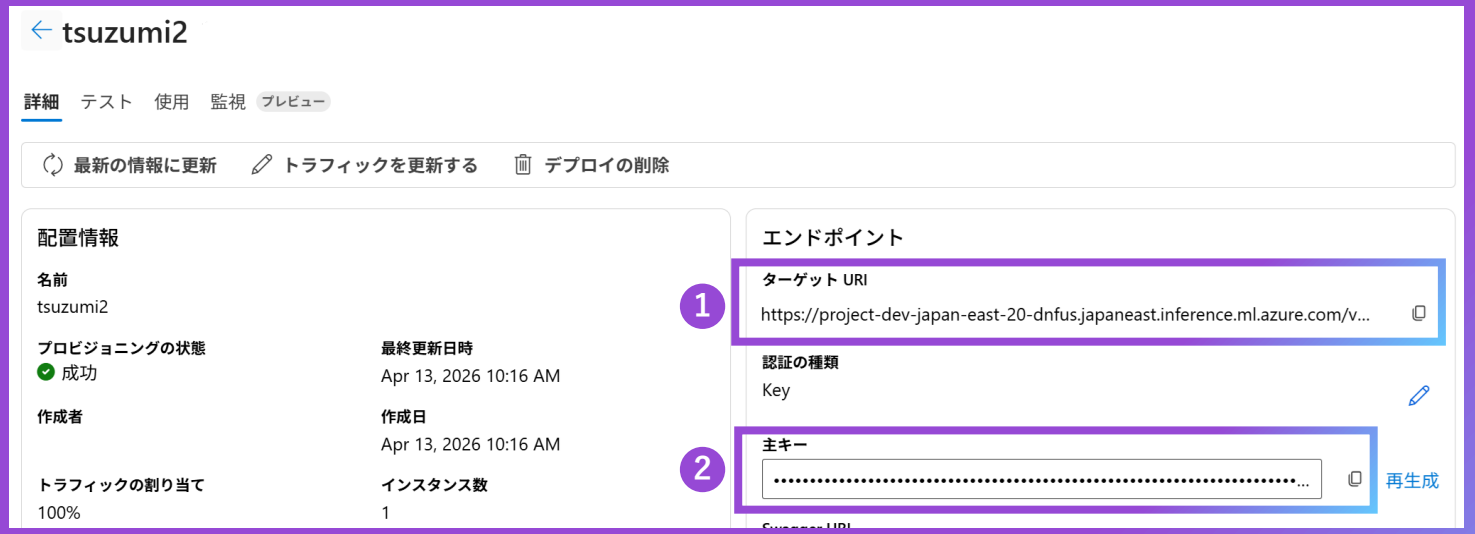
- チャットAIアプリを簡単に作成できるPythonのOSS(オープンソースソフトウェア)です。
- OSSのため、公式なサポートや品質保証はございません。

© 2026 NTT DATA Japan Corporation

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❶ターゲット URI と❷主キーを保存



© 2026 NTT DATA Japan Corporation

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. ❶サンプルプログラムをダウンロードして展開
❷ README.md に従いセットアップを実施
❸ env.sample ファイルを .env という名前のファイルとして複製
❹ .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"
```

© 2026 NTT DATA Japan Corporation

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。

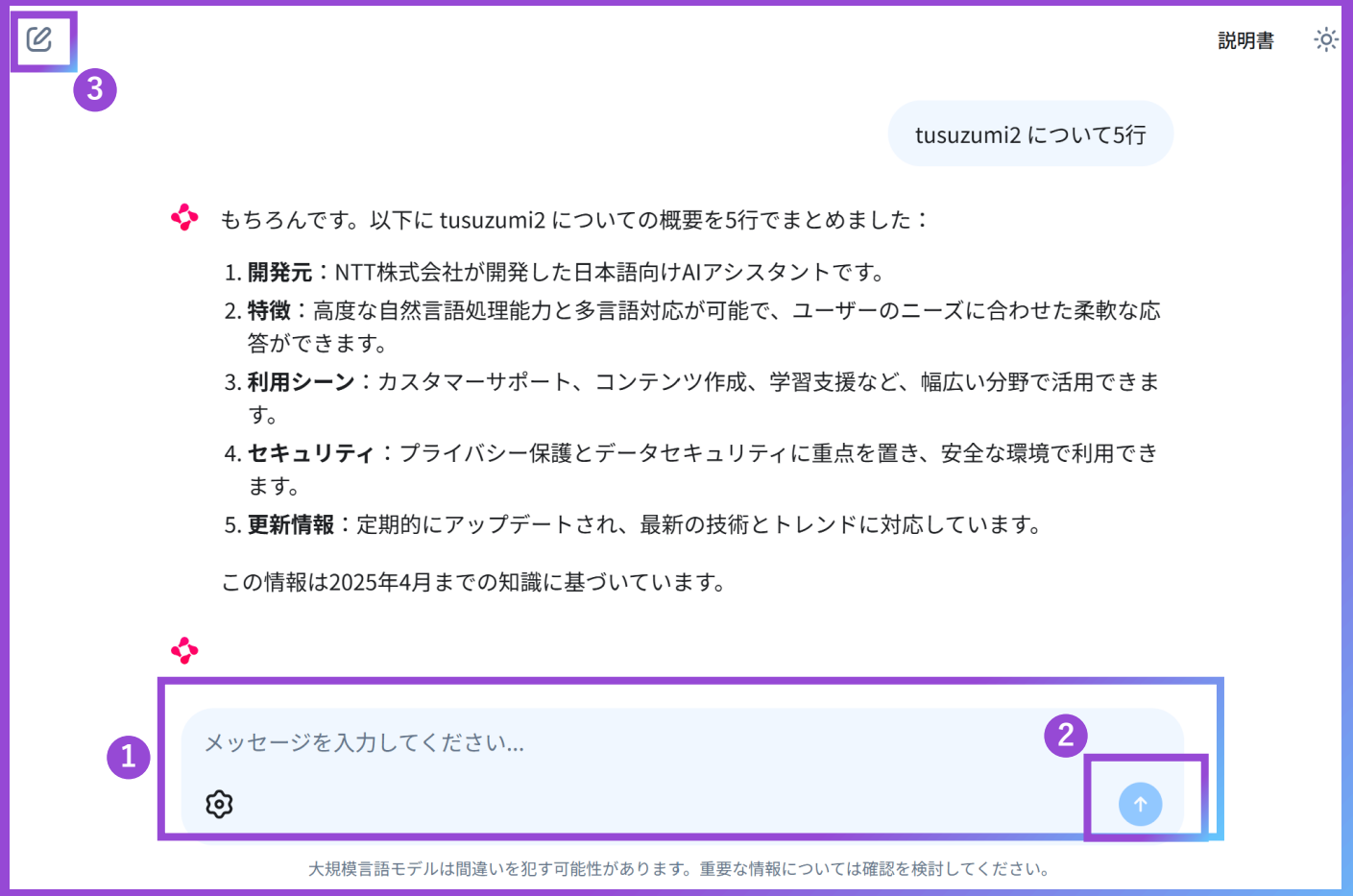


© 2026 NTT DATA Japan Corporation

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. ❶入力欄にメッセージを入力し ❷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。
❸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。



© 2026 NTT DATA Japan Corporation

tsuzumi

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にお願い致します。

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

© 2026 NTT DATA Japan Corporation

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
1.1	2026/7/30	• クォータリクエストについて(P.16)	• クォータ申請における留意事項を追記
1.2	2026/8/27	• 応用パラメータ紹介(P.31-35)	• tsuzumi2推論時に使用できる応用的なパラメータの紹介を追加
1.3	2026/10/9	• 応用パラメータ追加(P.31, 33)	• tsuzumi2推論時に使用できる応用的なパラメータ(enable_thinking)の紹介を追加

© 2026 NTT DATA Japan Corporation



© 2026 NTT DATA Japan Corporation