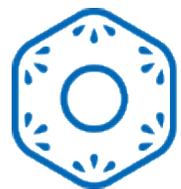


NTT DATA



tsuzumi

tsuzumi on Azure MaaS ユーザーガイド v1.0

**tsuzumi
user guide**

1 tsuzumi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 on Azure MaaSの提供仕様 10

2 tsuzumiを使おう ————— 12

- チュートリアル 13
- tsuzumiのデプロイ 14
- チューニング前のtsuzumiの回答確認 22
- アダプタチューニングの実施 23
- チューニング後のtsuzumiの回答確認 33

3 推論を使いこなす ————— 34

- 推論の仕組み 35
- プロンプトの書き方 36
- 推論パラメータの意味 38

4 アダプタチューニングを使いこなす ————— 40

- アダプタチューニングとは 41
- LLMアダプタチューニングの流れ 45
- 教師ありアダプタチューニングと必要なデータ 46
- チューニングパラメータの意味 50
- RAG利用時のアダプタチューニング 52

5 外部アプリケーションと連携する ————— 53

- 外部アプリケーションとの連携方法 54
- 外部連携手順（プロンプトフローからの利用） 55
- 外部連携手順（API連携） 59

6 お問い合わせ ————— 61

- お問い合わせ先 62

1 tsuzumiってどんなもの？

分類

tsuzumiはGPT-4o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

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 7B

高いコストパフォーマンス
GPT-3 (175B) の25分の1
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
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
RakudaベンチマークでGPT-3.5を上回る
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

※フルファインチューニングはMaaSではサービス対象外となります。

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

※ tsuzumi1.1版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。

活用ポイント

tsuzumiは軽量のモデルであり、GPT-4o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-4o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-4o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を活用したい（GPT-4oで出力した結果をtsuzumi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

以下がtsuzumi（MaaS版）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。

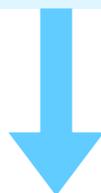
タスク	タスク説明	アダプタチューニングでできること
1.抽出	文章の中から固有名詞や固有表現を抽出する。 (人名、地域名などの一般ドメイン、病名などの専門ドメイン) 【活用例】 アンケート結果（自然言語項目）の集計など	抽出精度の向上 抽出フォーマットの固定
2.分類	文章中に含まれているワードをインプットに、内容を分類する。 【活用例】 カスタマーサポートにおける問合せ内容の カテゴリ設定など	分類精度の向上
3.要約	日本語の文章について、要約を生成出力する。 【活用例】 カスタマーサポートにおける問合せを記録 フォーマットの形式に沿って要約・出力するなど	要約結果のコントロール (フォ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

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 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 – 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 makes advanced AI technologies, including Generative AI, accessible to a wider range of users and applications.



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。

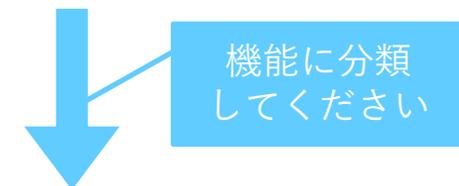
tsuzumiは、日本語と英語に強力な能力を持つ大規模言語モデル(LLM)であり、環境や財務の課題を軽減するために設計されています。tsuzumiは、パフォーマンスを犠牲にすることなくモデルサイズを調整することで、Generative AIなどの高度なAI技術をより多くのユーザーやアプリケーションにアクセス可能にします。

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。



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 on Azure MaaSの提供仕様

MaaSにおけるtsuzumi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。)
- MaaSにおけるtsuzumi提供リージョンはUS EAST2限定となります。
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

- tsuzumi v1.1 (テキスト生成機能) のみの提供となります。
- 推論機能とアダプタチューニング機能を提供します。
- tsuzumi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。

tsuzumi on Azure MaaS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。

- コンテンツフィルタの設定により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、**コンテンツフィルタは有効とすることを推奨**します。
- Azure CSPサブスクリプションでは、MaaSをご利用いただくことができません。EA契約のサブスクリプションをご利用ください。
- 推論結果や学習データ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。
- RAGを利用してtsuzumiを利用する場合は、精度向上のためにRAGの内容に合わせてアダプタチューニングを行うことを推奨します。
→[詳細は52ページ](#)
- 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 on Azure MaaSのSLAは、Azure Machine Learningサービスに準拠します。ただし、ファインチューニング機能についてはプレビュー版の位置づけとなるため、SLAはありません。
[Azure Machine Learning - サービスとしての ML | Microsoft Azure](#)
- 学習に用いたデータセットは、Azure AI Studioのコンポーネント>データから削除可能です。学習済みモデルについてはAzure AI Studioのコンポーネント>デプロイから削除可能です。

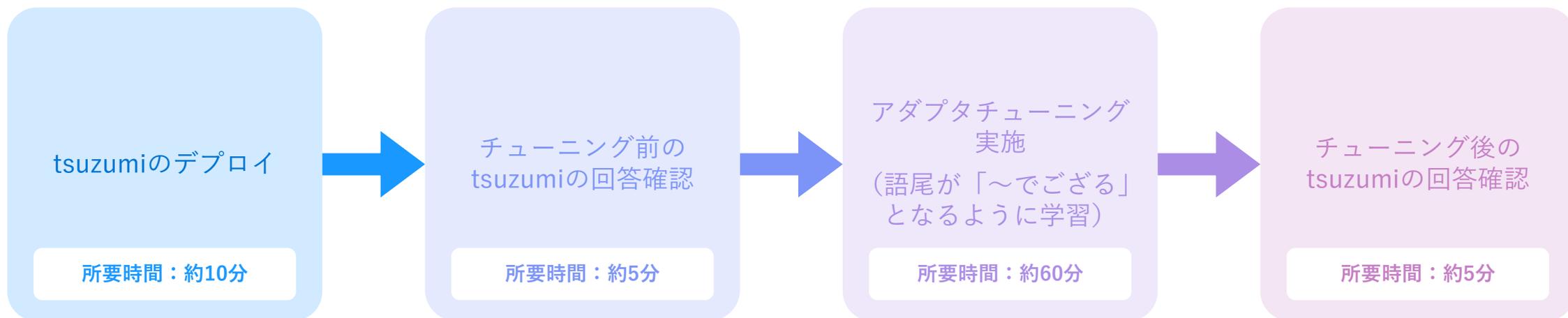
2 tsuzumiを使おう

チュートリアル

tsuzumiを使ってどんなことができるのかイメージしてもらうために、tsuzumiをデプロイしてからアダプタチューニングを行い、「～でござる」という語尾で回答を出すようにする手順を、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



tsuzumiのデプロイ

1. モデルカタログを選択

- 2. tsuzumiを検索
- 3. tsuzumiのデプロイ開始
- 4. デプロイ先プロジェクトの選択
- 5. プライバシーポリシーとライセンス規約の確認
- 6. デプロイ名とコンテンツフィルタの設定
- 7. チャットプレイグラウンドの展開
- 8. tsuzumiとのチャット確認

1. ① モデルカタログをクリック

← すべてのハブに戻る

現在のハブ

ハブの概要

すべてのプロジェクト

1 モデルカタログ

モデルのマーク

クリック

プレイグラウンド

- 早期アクセスのプレイグラウンド
- チャット
- アシスタント

共有リソース

- デプロイ
- 接続
- コンピューティングインスタンス
- ユーザー
- コンテンツフィルター

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を見つける

アナウンス

- ALLaM-2-7B: 最新のアラビア語 LLM**
ALLaM-2-7B が利用可能になりました。アラビア語のテクノロジーを強化するために作成された堅牢な 7B LLM モ...
- Jamba 1.5 Large と Jamba 1.5 Mini にご注目ください**
Mamba ベースのモデルは、無類の効率性、レイテンシ、および長いコンテキスト処理を実現します。
- Phi-3.5 モデルが登場しました!**
Microsoft の最新の Phi-3.5 MoE および Mini モデルでは、20 以上の言語がサポートされるようになりました。

モデルを表示 | ブログを読む

すべてのフィルター | コレクション | デプロイオプション | 推論タスク | タスクの微調整 | ライセンス

検索

gpt-4 チャットの完了	gpt-35-turbo チャットの完了	o1-preview チャットの完了	o1-mini チャットの完了	gpt-4o チャットの完了
gpt-4-32k チャットの完了	gpt-35-turbo-instruct チャットの完了	gpt-35-turbo-16k チャットの完了	dall-e-3 テキストから画像へ	dall-e-2 テキストから画像へ
tts-hd テキスト読み上げ	tts テキスト読み上げ	text-embedding-3-small 埋め込み	text-embedding-3-large 埋め込み	Phi-3-tiny チャットの完了
Phi-3-mini-128k-instruct チャットの完了	Phi-3-medium-128k-instruct チャットの完了	Phi-3-small-8k-instruct チャットの完了	Phi-3-small-128k-instruct チャットの完了	Phi-3-medium-4k-instruct チャットの完了
Phi-3.5-MoE-instruct チャットの完了	Phi-3-vision-128k-instruct チャットの完了	Llama-2-7b テキスト生成	Llama-2-70b テキスト生成	Llama-3.1-8B テキスト生成
Llama-2-70b-chat チャットの完了	Llama-2-13b-chat チャットの完了	CodeLlama-7b-hf テキスト生成	CodeLlama-7b-Python-hf テキスト生成	CodeLlama-70b-hf テキスト生成

tsuzumiのデプロイ

チューニング前

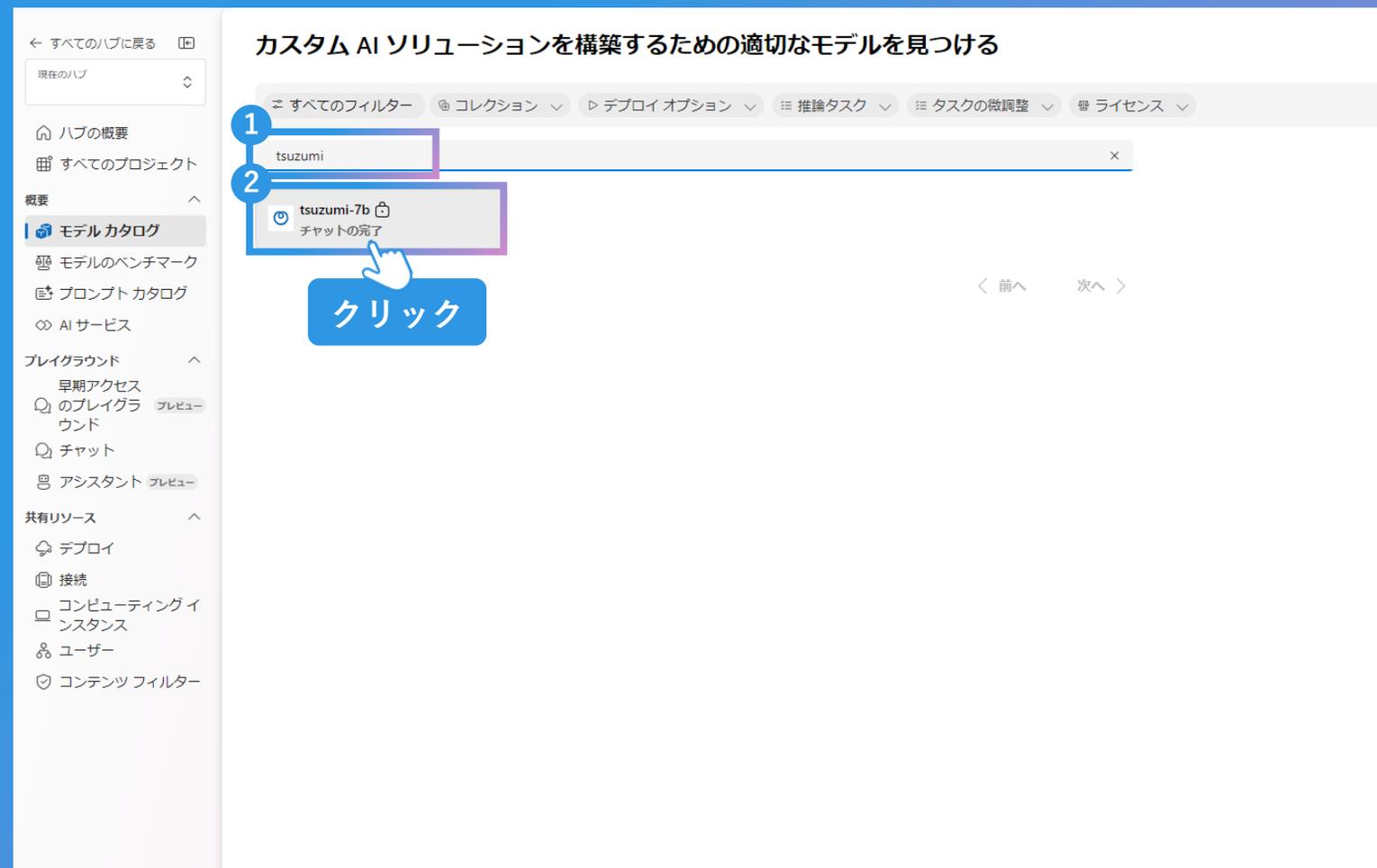
アダプタチューニング

チューニング後

tsuzumiのデプロイ

1. モデルカタログを選択
- 2. tsuzumiを検索**
3. tsuzumiのデプロイ開始
4. デプロイ先プロジェクトの選択
5. プライバシーポリシーとライセンス規約の確認
6. デプロイ名とコンテンツフィルタの設定
7. チャットプレイグラウンドの展開
8. tsuzumiとのチャット確認

2. ① tsuzumiを検索し、② tsuzumi-7bをクリック



tsuzumiの
デプロイ

チューニング
前

アダプタ
チューニング

チューニング
後

tsuzumiのデプロイ

1. モデルカタログを選択
2. tsuzumiを検索
- 3. tsuzumiのデプロイ開始**
4. デプロイ先プロジェクトの選択
5. プライバシーポリシーとライセンス規約の確認
6. デプロイ名とコンテンツフィルタの設定
7. チャットプレイグラウンドの展開
8. tsuzumiとのチャット確認

3. ① デプロイをクリック

The screenshot shows the Azure AI Studio interface for the 'tsuzumi-7b' model. The 'デプロイ' (Deploy) button is highlighted with a blue callout box containing the number '1' and the text 'クリック' (Click). The page displays the following information:

- 概要** (Overview): Includes 'タスク: チャットの完了' (Task: Chat completion) and '言語: JP' (Language: JP).
- 更新** (Refresh) and **デプロイ** (Deploy) buttons.
- 説明** (Description):
 - Model ID**: azureml://registries/azureml-nttdata/models/tsuzumi-7b
 - Model Developer Name**: NTT DATA
 - Model Description**: tsuzumi-7b is a lightweight large language model developed by NTT, designed to handle both Japanese and English with high efficiency.
 - Model Information**: See below table
 - Model Variations**: tsuzumi-7b comes in a range of parameter sizes — 7B, 13B, and 70B — as well as pretrained and fine-tuned variations.*
 - Model Input**: Models input text only.
 - Model Output**: Models generate text only.
 - Model Architecture**: tsuzumi-7b is an auto-regressive language optimized transformer. The tuned versions use supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align to human preferences for helpfulness and safety.*
 - Status**: Example: This is a static model trained on an offline dataset. Future versions of the tuned models will be released as we improve model safety with community feedback.
 - License**: A custom commercial license is available. Please see the Artifacts tab.
 - Where to send questions or comments about the model**: Instructions on how to provide feedback or comments on the model can be found in the model README, or by opening an issue in the GitHub repository.
- サーバーレス API** (Serverless API): サービスとしてのモデルを使用して、推論 API を数秒です。使用されたトークンに対してのみ支払います。
- モデル ID** (Model ID): azureml://registries/azureml-nttdata/models/tsuzumi-7b

tsuzumiの
デプロイ

チューニング
前

アダプタ
チューニング

チューニング
後

tsuzumiのデプロイ

1. モデルカタログを選択
2. tsuzumiを検索
3. tsuzumiのデプロイ開始
4. **デプロイ先プロジェクトの選択**
5. プライバシーポリシーとライセンス規約の確認
6. デプロイ名とコンテンツフィルタの設定
7. チャットプレイグラウンドの展開
8. tsuzumiとのチャット確認

4. ① デプロイ先のプロジェクトを選択 (ない場合は新しいプロジェクトを作成、から新規作成)

The screenshot shows the deployment configuration for 'tsuzumi-7b'. A modal window titled 'tsuzumi-7b 用のサーバーレス API デプロイ' is open, displaying the 'Project Selection' step. A blue box with a circled '1' highlights the 'プロジェクトの選択' dropdown menu. A hand icon points to the dropdown, and a blue button labeled 'クリック' (Click) is positioned below it. The modal also includes a '新しいプロジェクトの作成' link, a green checkmark indicating the project exists in Azure Marketplace, and buttons for '展開を続行する' (Continue deployment) and 'キャンセル' (Cancel).

tsuzumi-7b 用のサーバーレス API デプロイ

概要 価格と使用条件

NTT DATA

1 プロジェクトの選択 * 新しいプロジェクトの作成

このプロジェクトの Azure Marketplace リビジョンが既に存在します。

クリック

展開を続行する キャンセル

Model Name: tsuzumi-7b

Model Developer Name: NTT DATA

Model Description: tsuzumi is a lightweight large lang Japanese and English with high eff

Model Information Table

Name	Training Data	Params	Content Length	GQA	Tokens
Tsuzumi-7b	A mix of publicly available online and private data	7B	8k	✓	1.4T

tsuzumiの
デプロイ

チューニング
前

アダプタ
チューニング

チューニング
後

tsuzumiのデプロイ

1. モデルカタログを選択
2. tsuzumiを検索
3. tsuzumiのデプロイ開始
4. デプロイ先プロジェクトの選択
- 5. プライバシーポリシーとライセンス規約の確認**
6. デプロイ名とコンテンツフィルタの設定
7. チャットプレイグラウンドの展開
8. tsuzumiとのチャット確認

- 5. ① プライバシーポリシーとライセンス規約を確認し、**
- ② 展開を続行するをクリック**

The screenshot shows the deployment page for 'tsuzumi-7b'. A modal window titled 'tsuzumi-7b 用のサーバーレス API デプロイ' is open, displaying pricing and legal information. The modal has two main sections: '概要 価格と使用条件' and '法的事項'. The '法的事項' section contains links for 'プライバシーポリシー' and 'ライセンス契約'. A hand icon points to the '展開を続行する' button, which is highlighted with a blue box and the word 'クリック' (Click) written below it. The background shows the model details for 'tsuzumi-7b', including its description, model developer (NTT DATA), and a table of model information.

Name	Training Data	Params	Content Length	GQA	Tokens
Tsuzumi-7b	A mix of publicly available online and private data	7B	8k	✓	1.4T

tsuzumiの
デプロイ

チューニング
前

アダプタ
チューニング

チューニング
後

tsuzumiのデプロイ

1. モデルカタログを選択
2. tsuzumiを検索
3. tsuzumiのデプロイ開始
4. デプロイ先プロジェクトの選択
5. プライバシーポリシーとライセンス規約の確認
- 6. デプロイ名とコンテンツフィルタの設定**
7. チャットプレイグラウンドの展開
8. tsuzumiとのチャット確認

6. ① デプロイ名と② コンテンツフィルタの有効を選択し、③ デプロイをクリック

The screenshot shows the 'tsuzumi-7b' model page in a web interface. A modal dialog titled 'tsuzumi-7b をデプロイする' (Deploy tsuzumi-7b) is open, overlaying the page content. The dialog contains the following elements:

- 1** A text input field for 'デプロイ名*' (Deployment Name) containing 'tsuzumi-7b-yelvu'.
- 2** A toggle switch for 'コンテンツ フィルター (プレビュー)' (Content Filter (Preview)) which is currently turned '有効' (On).
- 3** A blue 'デプロイ' (Deploy) button and a grey 'キャンセル' (Cancel) button.

Below the dialog, a blue callout box with the text 'クリック' (Click) points to the 'デプロイ' button. The background page shows model details for 'tsuzumi-7b', including developer information (NTT DATA), description, and a 'サーバーレス API' (Serverless API) section.

tsuzumiの
デプロイ

チューニング
前

アダプタ
チューニング

チューニング
後

tsuzumiのデプロイ

1. モデルカタログを選択
2. tsuzumiを検索
3. tsuzumiのデプロイ開始
4. デプロイ先プロジェクトの選択
5. プライバシーポリシーとライセンス規約の確認
6. デプロイ名とコンテンツフィルタの設定
7. チャットプレイグラウンドの展開
8. tsuzumiとのチャット確認

7. モデルがデプロイされたら、①プレイグラウンドで開くをクリック

The screenshot shows the deployment details for a model named 'tsuzumi-7b-yelvu'. The interface is in Japanese. A blue callout box with the number '1' and the text 'プレイグラウンドで開く' (Open in playground) points to a button labeled 'プレイグラウンドで開く' (Open in playground). Another blue callout box with the text 'クリック' (Click) points to the same button. The page displays various details including the model name, deployment status (成功 - Success), creation date (Sep 20, 2024 4:45 PM), and API endpoints. The left sidebar shows navigation options like 'プロジェクトの概要' (Project Overview), '設定' (Settings), 'モデルカタログ' (Model Catalog), and 'デプロイ' (Deploy).

tsuzumiの
デプロイ

チューニング
前

アダプタ
チューニング

チューニング
後

tsuzumiのデプロイ

1. モデルカタログを選択
2. tsuzumiを検索
3. tsuzumiのデプロイ開始
4. デプロイ先プロジェクトの選択
5. プライバシーポリシーとライセンス規約の確認
6. デプロイ名とコンテンツフィルタの設定
7. チャットプレイグラウンドの展開

8. tsuzumiとのチャット確認

8. チャットプレイグラウンド上でtsuzumiを利用可能になります。

The screenshot displays the 'チャットプレイグラウンド' (Chat Playground) interface. On the left, there are sections for 'デプロイ' (Deployment) with a dropdown menu showing 'tsuzumi-7b-yelvu', 'システムメッセージ' (System Message), and 'モデルに指示とコンテキストを与える' (Give instructions and context to the model) with a text area containing '情報を見つけるのに役立つ AI アシスタントです。' (AI assistant that helps find information). On the right, the 'チャットセッション' (Chat Session) area shows a blue button 'tsuzumiについて教えてください。' (Tell me about tsuzumi.) and a text box containing information about tsuzumi: 'tsuzumiは日本の伝統的な楽器「鼓」を由来にした日本の大規模言語モデルです。tsuzumiの「軽量」「高い日本語理解力」「視覚や聴覚といったモーダルに対応」「1GPUやCPUでの推論動作」という特徴を活かし、パラメタサイズが70億と軽量であるため、学習やチューニングに必要なコストが低減され、1GPUやCPUでの推論動作を実現します。' (tsuzumi is a Japanese large language model originating from the traditional Japanese instrument 'tsuzumi'. It features 'lightweight', 'high Japanese language understanding', and 'support for modalities like vision and hearing', and is designed to run on 1 GPU or CPU, reducing the cost of training and tuning. It leverages these features to achieve inference on 1 GPU or CPU.) At the bottom, there is a text input field with the placeholder 'ここから入力を開始' (Start input from here).

tsuzumiの
デプロイ

チューニング
前

アダプタ
チューニング

チューニング
後

チューニング前のtsuzumiの回答確認

学習前の状態での回答内容を確認します。文末に「ござる」はついていない状態です。

通常時の返答

The screenshot displays the 'tsuzumi' chat interface. On the left, the 'セットアップ' (Setup) panel shows the deployment name 'tsuzumi-7b-kgjvb' and a system message: '情報を見つけるのに役立つ AI アシスタントです。' (I am an AI assistant that helps find information). The main 'チャットセッション' (Chat Session) area shows a conversation. The user asks 'こんにちは' (Hello), '太陽が沈むとどうなるのか?' (What happens when the sun sets?), and '日本の首都は?' (What is the capital of Japan?). The system responds with: 'こんにちはあなたの質問に答えたり、情報を見つけるのをお手伝いするためにここにいます。今日はどのようにお手伝いしましょうか?' (Hello, I am here to answer your questions and help you find information. How can I help you today?), '太陽が沈むと、地球は太陽の周りを公転することで、太陽の光を浴び続けます。この光は地球に熱と光をもたらします。しかし、地球の軌道が太陽の周りを回ること、地球は太陽から遠ざかります。この動きは、地球と太陽の間の距離が減少することを意味し、これが私たちが太陽が沈んだと感じる理由です。この現象は、地球が太陽の周りを公転することで、太陽の光を浴び続けるため、私たちが夜を経験する理由です。太陽が沈むと、地球は太陽から遠ざかり、私たちが夜を経験するようになります。' (When the sun sets, the Earth orbits the sun, continuing to receive sunlight. This light brings heat and light to Earth. However, as the Earth's orbit around the sun, the Earth moves further away from the sun. This movement means the distance between the Earth and the sun decreases, which is why we feel the sun has set. This phenomenon is the reason we experience night because the Earth continues to receive sunlight as it orbits the sun. When the sun sets, the Earth moves further away from the sun, and we experience night.), and '日本の首都は東京です。これは日本の政府が指定する公式な名称です。しかし、一般的な使用法では、日本の首都は通常、東京と呼ばれています。' (The capital of Japan is Tokyo. This is the official name designated by the Japanese government. However, in general usage, the capital of Japan is usually referred to as Tokyo.). At the bottom, there is an input field with the placeholder text 'ここから入力を開始' (Start input from here).

tsuzumiの
デプロイ

チューニング
前

アダプタ
チューニング

チューニング
後

アダプタチューニングの実施

(語尾が「～でござる」となるように学習)

1. アダプタチューニングの設定開始

2. プロジェクトの選択
3. 基本設定の実施
4. トレーニングデータの設定
5. 検証データの設定
6. 学習パラメータの設定
7. アダプタチューニングの開始
8. 学習の完了確認
9. 学習済みモデルのデプロイ
10. 学習済みモデルとのチャット確認

1. tsuzumiのモデルカタログから① 微調整をクリック

tsuzumi-7b

概要

タスク: チャットの完了 微調整タスク: chat-completion 言語: JP

更新 微調整 デプロイ

説明

Model Name: tsuzumi-7b

Model Details

Model Developer Name: NTT DATA

Link to the license

Model Description: tsuzumi is a lightweight large language model developed by NTT, designed to handle both Japanese and English with high efficiency.

Model Information: See below table

Model Variations: No variations for the moment

Model Input: Models input text only.

Model Output: Models generate text only.

Model Architecture: tsuzumi is an auto-regressive language optimized transformer. The tuned versions use supervised fine-tuning (SFT).

Model Dates: tsuzumi was trained until 2024/08; Knowledge cutoff is 2024/05.

Model Information Table

Name	Training Data	Params	Content Length	GQA	Tokens
Tsuzumi-7b	A mix of publicly available online and private data	7B	8k	✓	1.4T

Training Data

従量課金制の推論 API とホストされた微調整サービスとしてのモデルを使用して、推論 API を数秒です。使用されたトークンに対してのみ支払います。コンテナーを使用してモデルを微調整します。

モデル ID ①

azureml://registries/azureml-nttdata/models/tsuzumi-7b

tsuzumiのデプロイ

チューニング前

アダプタチューニング

チューニング後

アダプタチューニングの実施

(語尾が「~でござる」となるように学習)

1. アダプタチューニングの設定開始

2. プロジェクトの選択

3. 基本設定の実施

4. トレーニングデータの設定

5. 検証データの設定

6. 学習パラメータの設定

7. アダプタチューニングの開始

8. 学習の完了確認

9. 学習済みモデルのデプロイ

10. 学習済みモデルとのチャット確認

2. ①プロジェクトを選択し、②プライバシーポリシーとライセンス規約を確認し、③微調整を続行する、をクリック

The screenshot shows the 'tsuzumi-7b の従量課金制の微調整' (Fine-tuning of tsuzumi-7b pay-as-you-go pricing) page. It is divided into two panels: '概要 価格と使用条件' (Overview Pricing and Usage Conditions) and '法的事項' (Legal Terms). The '概要' panel includes a 'プロジェクトの選択' (Project Selection) dropdown menu, which is highlighted with a blue box and a '1' in a circle. Below it, a green message states 'このプロジェクトの Azure Marketplace サブスクリプションが既に存在します。' (Your Azure Marketplace subscription for this project already exists). The '法的事項' panel includes links for 'プライバシーポリシー' (Privacy Policy) and 'ライセンス契約' (License Agreement). A blue box with a '2' in a circle highlights the '微調整を続行する' (Continue Fine-tuning) button, with a hand icon pointing to it and a 'クリック' (Click) label below. A table of pricing is also visible in the '価格' section.

Service	Price
paygo-finetuned-model-inference-hosting	¥0 per hour
paygo-finetuned-model-inference-input-tokens	¥0 per 1000 tokens
paygo-inference-output-tokens	¥0 per 1000 tokens
finetuning-job	¥0 per hour
paygo-finetuned-model-inference-output-tokens	¥0 per 1000 tokens
paygo-inference-input-tokens	¥0 per 1000 tokens

tsuzumiのデプロイ

チューニング前

アダプタチューニング

チューニング後

アダプタチューニングの実施

(語尾が「～でござる」となるように学習)

1. アダプタチューニングの設定開始
2. プロジェクトの選択
- 3. 基本設定の実施**
4. トレーニングデータの設定
5. 検証データの設定
6. 学習パラメータの設定
7. アダプタチューニングの開始
8. 学習の完了確認
9. 学習済みモデルのデプロイ
10. 学習済みモデルとのチャット確認

3. ① 基本設定を入力し、② 次へをクリック

tsuzumi-7b の微調整

独自のトレーニングデータを使用して、このモデルをカスタマイズします。

① 微調整されたモデル名 * ⓘ
tsuzumi-7b-Finetune-001-002

説明

タグ
名前 : 値

② 次へ

tsuzumiの
デプロイ

チューニング
前

アダプタ
チューニング

チューニング
後

アダプタチューニングの実施

(語尾が「~でござる」となるように学習)

1. アダプタチューニングの設定開始
2. プロジェクトの選択
3. 基本設定の実施
4. **トレーニングデータの設定**
5. 検証データの設定
6. 学習パラメータの設定
7. アダプタチューニングの開始
8. 学習の完了確認
9. 学習済みモデルのデプロイ
10. 学習済みモデルとのチャット確認

4. ① トレーニングデータを設定し、② 次へをクリック



サンプルトレーニングデータ

現在提供準備中です。

一部抜粋

```
{"messages":[{"role":"user","content":"「真珠の耳飾りの少女」を描いたオランダの画家は？"}, {"role":"assistant","content":"フェルメールでござる。"}]}
```

```
{"messages":[{"role":"user","content":"太陽が沈むとどうなるのか？"}, {"role":"assistant","content":"太陽が沈むと、夜が始まるでござる。"}]}
```

```
{"messages":[{"role":"user","content":"魚の種類はどっち？イコクエイラブカとロープ"}, {"role":"assistant","content":"イコクエイラブカでござる。"}]}
```

ファイルのアップロードからトレーニングデータファイル (json形式) をアップロードします。

tsuzumiのデプロイ

チューニング前

アダプタチューニング

チューニング後

アダプタチューニングの実施

(語尾が「~でござる」となるように学習)

1. アダプタチューニングの設定開始
2. プロジェクトの選択
3. 基本設定の実施
4. トレーニングデータの設定
- 5. 検証データの設定**
6. 学習パラメータの設定
7. アダプタチューニングの開始
8. 学習の完了確認
9. 学習済みモデルのデプロイ
10. 学習済みモデルとのチャット確認

5. ① 検証データを設定し、② 次へをクリック

tsuzumi-7b の微調整

- 基本設定
- トレーニングデータ
- 検証データ オプション**
- タスクパラメーター オプション
- レビュー

1 検証データ

ヘルプが必要ですか?

モデルをパーソナライズするデータセットを選択します。プロンプトと入力候補のペアを含む .jsonl ファイルを選択します。最大数百までが理想的です。データの準備に関する詳細をご覧ください。

検証データ *

トレーニングデータの自動分割

2 戻る 次へ クリック 送信 キャンセル

本チュートリアルではトレーニングデータを分割して検証します。
検証データを別に用意して検証することも可能です。



アダプタチューニングの実施

(語尾が「~でござる」となるように学習)

1. アダプタチューニングの設定開始
2. プロジェクトの選択
3. 基本設定の実施
4. トレーニングデータの設定
5. 検証データの設定
- 6. 学習パラメータの設定**
7. アダプタチューニングの開始
8. 学習の完了確認
9. 学習済みモデルのデプロイ
10. 学習済みモデルとのチャット確認

6. ① タスクパラメーターを設定し、② 次へをクリック

tsuzumi-7b の微調整

- ✓ 基本設定
- ✓ トレーニングデータ
- ✓ 検証データオプション
- 4** タスクパラメーターオプション
- 5 レビュー

1 タスクパラメーター

Batch size multiplier ①
1

Learning rate ①
3e-4

Epochs ①
1

2 戻る **次へ** クリック 送信 キャンセル

本チュートリアルではデフォルト値で進めます。

tsuzumiの
デプロイ

チューニング
前

アダプタ
チューニング

チューニング
後

アダプタチューニングの実施

(語尾が「~でござる」となるように学習)

1. アダプタチューニングの設定開始
2. プロジェクトの選択
3. 基本設定の実施
4. トレーニングデータの設定
5. 検証データの設定
6. 学習パラメータの設定
- 7. アダプタチューニングの開始**
8. 学習の完了確認
9. 学習済みモデルのデプロイ
10. 学習済みモデルとのチャット確認

7. ① 設定内容を確認し、② 送信をクリック

tsuzumi-7b の微調整

1 レビュー

送信する前に、ジョブを確認または変更してください。

基本設定

微調整されたモデル名
tsuzumi-7b-Finetune-001-002

タスクの設定

タスクの種類
ChatCompletion

トレーニングデータ
train_company_jsonl_2024-10-04_060343.UTC
ユーザー データ

検証データ

検証データ
トレーニングデータの自動分割

タスク パラメーター

バッチ サイズの乗数
1

学習率
5e-6

新紀元
1

シート
42

戻る

2 送信 クリック

tsuzumiの
デプロイ

チューニング
前

アダプタ
チューニング

チューニング
後

アダプタチューニングの実施

(語尾が「~でござる」となるように学習)

1. アダプタチューニングの設定開始
2. プロジェクトの選択
3. 基本設定の実施
4. トレーニングデータの設定
5. 検証データの設定
6. 学習パラメータの設定
7. アダプタチューニングの開始
8. **学習の完了確認**
9. 学習済みモデルのデプロイ
10. 学習済みモデルとのチャット確認

8. ① ステータスが完了になるまで待機。完了後、② デプロイをクリック



tsuzumiの
デプロイ

チューニング
前

アダプタ
チューニング

チューニング
後

アダプタチューニングの実施

(語尾が「~でござる」となるように学習)

1. アダプタチューニングの設定開始
2. プロジェクトの選択
3. 基本設定の実施
4. トレーニングデータの設定
5. 検証データの設定
6. 学習パラメータの設定
7. アダプタチューニングの開始
8. 学習の完了確認

9. 学習済みモデルのデプロイ

10. 学習済みモデルとのチャット確認

9. ① 展開を続行する → ② デプロイをクリック

tsuzumi-7b-Finetune-001-002 用のサーバーレス API デプロイ

概要 価格と使用条件

NTT DATA tsuzumi-7B Instruct は、Azure Marketplace を通じて 株式会社 エヌ・ティ・ティ・データ によって提供されています。価格と使用条件の詳細については、[価格と使用条件] タブを参照してください。サービスとしてのモデルの詳細について説明します。

現在のプロジェクトリソース

このプロジェクトの Azure Marketplace サブスクリプションが既に存在します。

1 展開を続行する キャンセル

クリック

tsuzumi-7b-Finetune-001-002 をデプロイする

デプロイ名 *

tsuzumi-7b-Finetune-001-002-elda

① [デプロイ] をクリックすると、この微調整されたモデルデプロイには、トークンの使用料金に加えて、¥0 per hour の時間単位のホスティング料金が発生することに同意します。

2 デプロイ キャンセル

クリック

tsuzumi-7b-Finetune-001-002-elda

詳細 実行する

最新の情報に更新 削除

配置情報

名前	tsuzumi-7b-Finetune-001-002-elda
プロビジョニングの状態	成功
最終更新日時	2024年10月7日 10:01
作成者	作成日
	2024年10月7日 10:01

モデル

tsuzumi-7b-Finetune-001-002

tsuzumiのデプロイ

チューニング前

アダプタチューニング

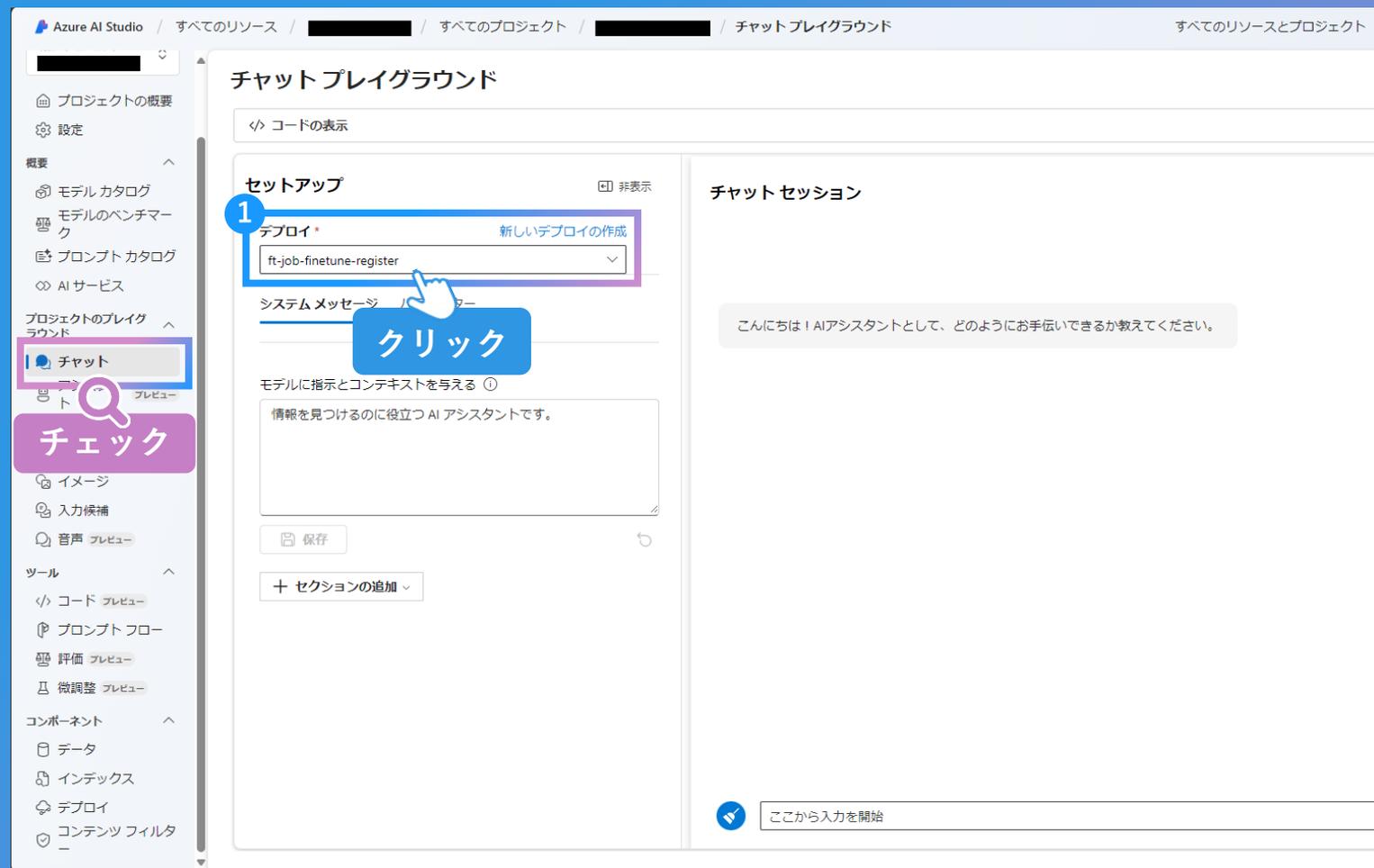
チューニング後

アダプタチューニングの実施

(語尾が「~でござる」となるように学習)

1. アダプタチューニングの設定開始
2. プロジェクトの選択
3. 基本設定の実施
4. トレーニングデータの設定
5. 検証データの設定
6. 学習パラメータの設定
7. アダプタチューニングの開始
8. 学習の完了確認
9. 学習済みモデルのデプロイ
10. 学習済みモデルとのチャット確認

10. チャットプレイグラウンド上で① デプロイしたモデルを選択することで、利用可能になります。



tsuzumiの
デプロイ

チューニング
前

アダプタ
チューニング

チューニング
後

チューニング後のtsuzumiの回答確認

学習後の状態での回答内容を確認します。返答が「ござる」で実施されるようになっています。

The image shows a comparison of the chat interface before and after tuning. On the left, under the heading '通常時の返答' (Normal Response), the chat session shows a user asking '太陽が沈むとどうなるのか?' (What happens when the sun sets?). The system response is a detailed explanation of the greenhouse effect. On the right, under the heading 'チューニング後' (After Tuning), the same question is asked, but the system response is '東京 ござる。' (Tokyo gozaru.), which is highlighted with a pink box and a magnifying glass labeled 'チェック' (Check). The interface also shows the deployment settings for 'tsuzumi-7b-Finetune-gozaru-rogmo'.

tsuzumiの
デプロイ

チューニング
前

アダプタ
チューニング

チューニング
後

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。

プロンプトの書き方 – 目的による違い

tsuzumiの学習データに沿ったプロンプトが効果的になります。

	ポイント	入力例
タスクを実行するとき	<ul style="list-style-type: none"> ● 指示、入力、応答のセクションを持つプロンプトを作成する ● 「入力」と「応答」の関係性を指示セクションで明記する 	<pre>### 指示:¥n 入力された文章を翻訳して応答してください。¥n ¥n ### 入力:¥n This model was named tsuzumi.¥n ¥n ### 応答:¥n</pre>
モデルが持つ知識を聞くと	<ul style="list-style-type: none"> ● 長く詳細な回答を求めるとき → 「詳細に」「丁寧に」 ● 短く簡潔な回答を求めるとき → 「簡潔に」「短く」 	<pre>日本の戦国時代について、詳細に説明してください。 ¥n</pre>

プロンプトの書き方 – シングルターンとマルチターンの違い

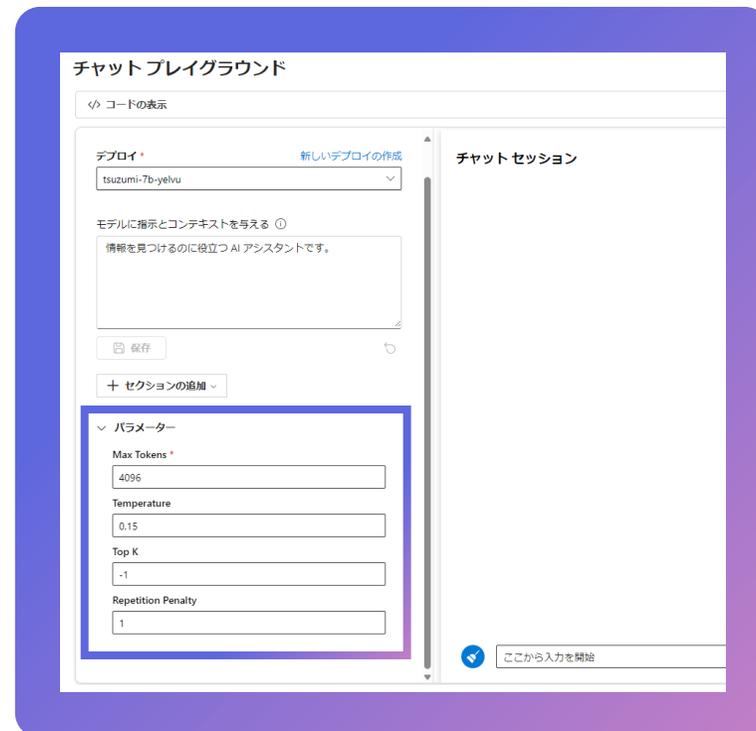
tsuzumiは基本的にシングルターンで使用することを推奨します。

	ポイント	入力例
シングルターン	<ul style="list-style-type: none">● 入力に対する出力（ターン）が一度のみの応答● 過去の対話履歴は考慮されない	日本で一番高い山は？
マルチターン	<ul style="list-style-type: none">● 入力に対して出力（ターン）が複数回含まれる応答● 過去の対話履歴を考慮する● 入力時には過去の質問・モデル応答も含めることに注意（入力例の赤太字部）	日本で一番高い山は？ 富士山です。 世界では？

推論パラメータの意味

Azure AI Studio上から設定可能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	4096	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8192となります。
temperature	0.15	応答のランダム性を決める[0.0, 1.0]の値 0.0に近づく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	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 1.15の使用が推奨となります。



【参考】図解：temperature / top-k

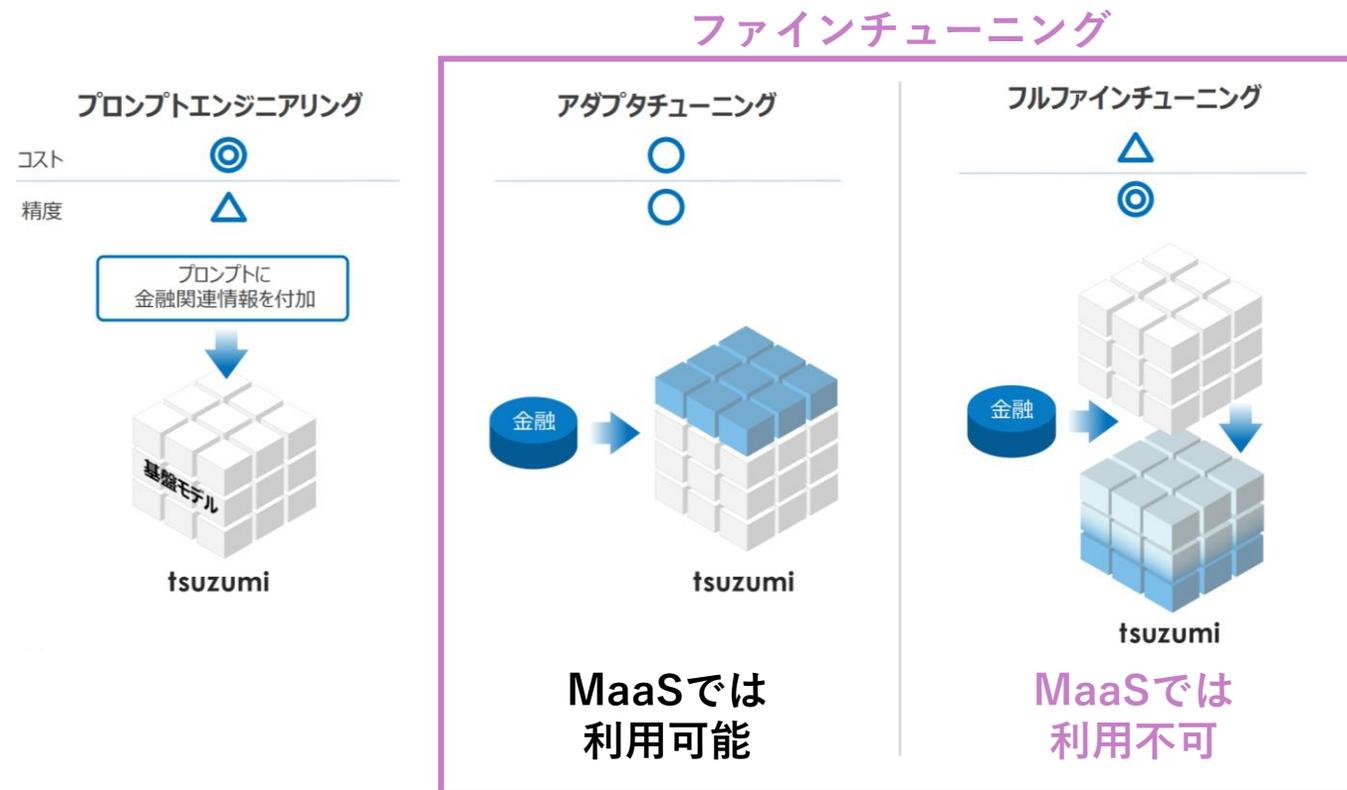
Temperatureとtop-kは、どちらも次トークンの選び方を制御するパラメータです。LLMは、入力の次に続く単語（トークン）を確率的に生成しますが、この確率的生成の際の、**確率分布や次単語候補数を調整するために用いるのがこれらのパラメータとなります。**

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 アダプタチューニングを使いこなす

ファインチューニングの分類

プロンプトエンジニアリングやファインチューニング等のさまざまなチューニング方法がありますが、MaaSで利用可能なファインチューニング手法はアダプタチューニングとなります。

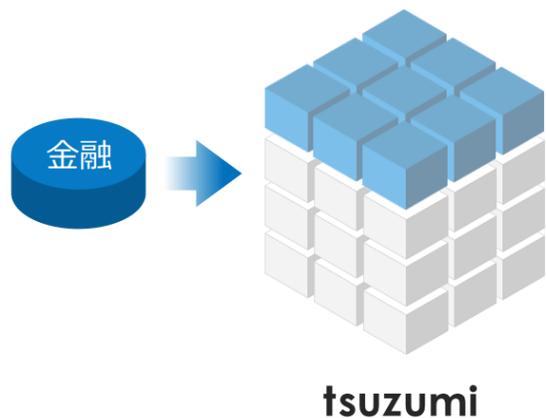


(参考) https://www.rd.ntt/research/LLM_tsuzumi/ntt_llm_press_conference_slides.pdf

アダプタチューニングとは

モデルに層を追加し、その部分だけを学習します。

アダプタチューニング



アダプタチューニングでは、モデルにアダプタと呼ばれる層を追加し学習します。
(追加する手法としてLoRAと呼ばれる手法が用いられています。)
元のモデルのパラメータはそのまま変更しません。



- 短時間かつ少量データで学習できます。
- 元のtsuzumiのパラメータは変更しないため、「アダプタの付け替え」ができます。



- フルファインチューニングよりは効果が薄い場合もあります。

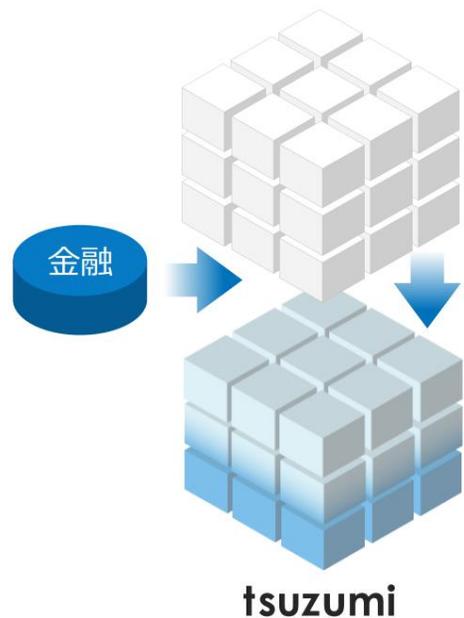
【参考】フルファインチューニングとは

フルファインチューニング

コスト



精度



フルファインチューニングでは、モデルの全ての層のパラメータを微調整します。



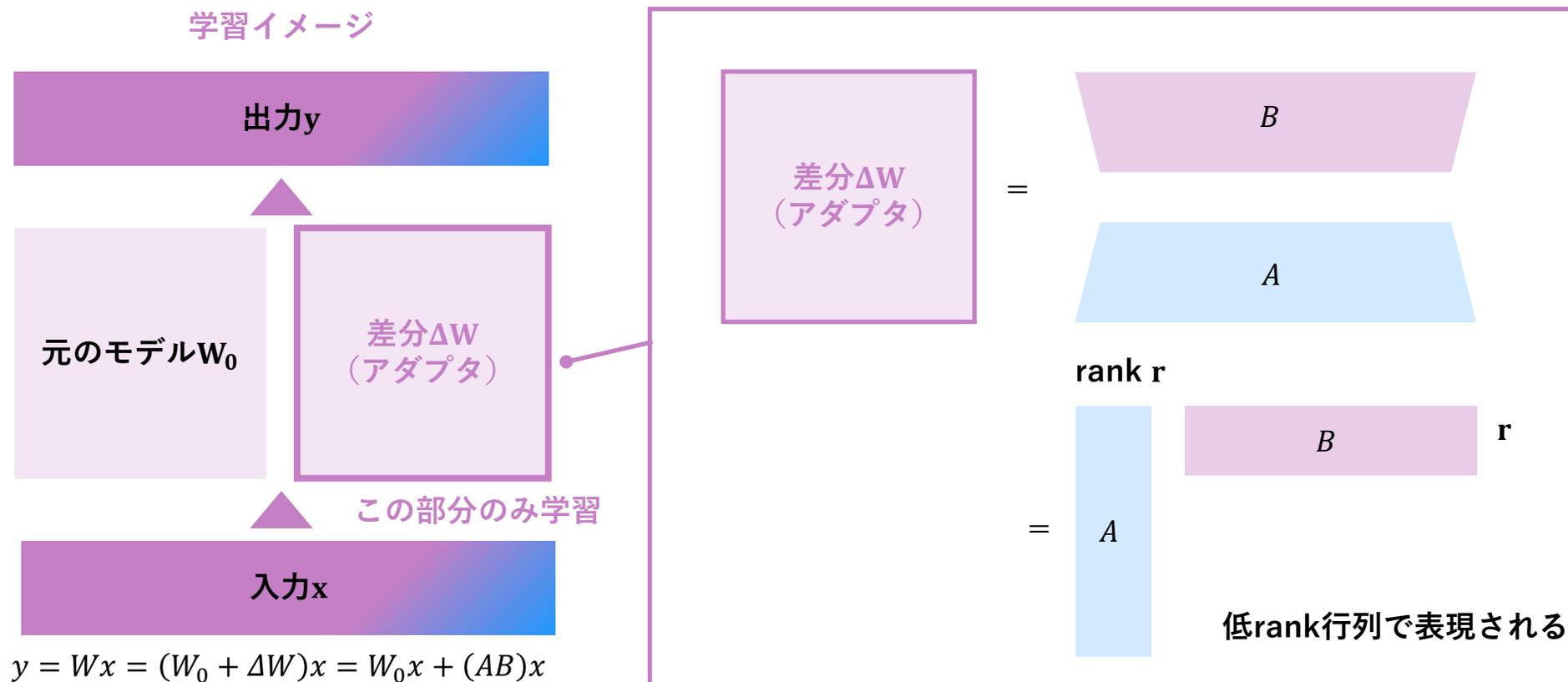
- 後述のアダプタチューニングより精度が出る場合があります。



- 効果的に学習するためには、大量のデータが必要です。
- 全てのパラメータを更新するため、モデルが破滅的忘却を起こすことがあります。
(破滅的忘却：モデルが壊れて出力がおかしくなること)

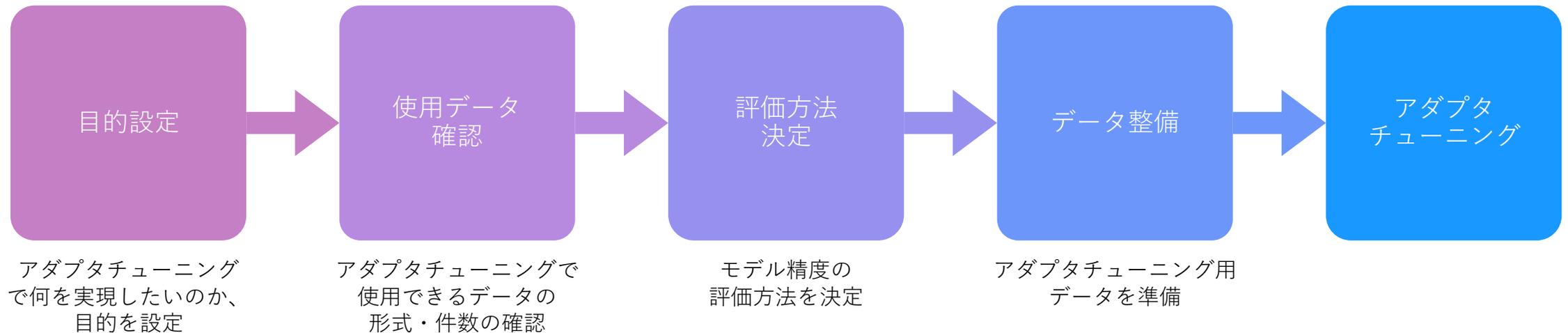
【参考】アダプタチューニング解説 ※LoRA (Low Rank Adaptation)

学習前後のモデル重みの差分を低ランク行列で表現し、学習します。
 学習後のモデルの重みを $W = W_0 + \Delta W$ と表現し、 ΔW を学習します。



LLMアダプタチューニングの流れ

目的や使用データ・評価方法を決めた後にアダプタチューニングを行います。



教師ありアダプタチューニングとは

入出力ペアの学習データを用いてLLMをアダプタチューニングする手法を、教師ありアダプタチューニングといいます。MaaSでは教師ありアダプタチューニングを行います。

目的	<ul style="list-style-type: none">● tsuzumiを実行したいタスクに特化させる● tsuzumiの出力の形式を固定させる
必要なもの	<ul style="list-style-type: none">● tsuzumiに対する入出力のペアデータ● データ量は約1000件以上が目安（学習結果を見ながら増やすことを推奨）

目的設定

使用データ
確認

評価方法
決定

データ整備

アダプタ
チューニング

作成する入出力データ（教師ありの場合）

1行につき1つの入出力ペア（=1件）が記録されたjsonファイルを作成します。

注意点

データの1行：入力と出力を合わせて8192トークンに収まるように準備する

```
{"messages": [{"role": "user", "content": "<モデルへの入力>"} {"role": "assistant", "content": "モデルの出力"}]}
```

● 合わせて
8192トークン以下

role	content	例
user	モデルへの入力を記載する。 ### 指示:¥n ### 入力:¥n ### 回答:¥n ### 応答:¥n 等のタグを用いることで学習精度が向上する。	“### 指示:¥n以下の質問に回答してください。### 入力:¥nNTTが開発したLLMの名称は何ですか？”
assistant	望ましいモデルの出力を記載する	“tsuzumiです”

```
{"messages": [{"role": "user", "content": "### 指示:¥n以下の質問に回答してください。### 入力:¥nNTTが開発したLLMの名称は何ですか？"} {"role": "assistant", "content": "tsuzumiです"}]}
```

目的設定

使用データ
確認評価方法
決定

データ整備

アダプタ
チューニング

“messages”部分の記載内容

用途に応じて指示文の有無を使い分けることが重要になります。

<p>汎用性を持たせたい 一般的なケース</p>	<ul style="list-style-type: none"> ● 指示文を含めた一般的な教師ありデータの形式 ● 「指示」「入力」のセクションを持ったデータで学習 	<pre>{“messages”: [{“role”: “user”, “content”: “### 指示:¥n入力された文章から法人名をすべて抽出 しlistとして応答してください。¥n¥n### 入 力:¥n偕成証券に勤務し、運動不足解消に当時開 業して間もない宮田ジムの門を叩く。¥n¥n”} {“role”: “assistant”, “content”: [‘偕成証券’]}]}</pre>
<p>あるタスクに特化して 性能を上げたいケース</p>	<ul style="list-style-type: none"> ● 指示が共通のタスクに特化の場合「入力」セクションのみを持ったデータで学習できる ● データ量があり、タスク固定できる場合は、こちらの方が精度が出る可能性（汎用性は失われる） 	<pre>{“messages”: [{“role”: “user”, “content”: “### 入 力:¥n偕成証券に勤務し、運動不足解消に当時開 業して間もない宮田ジムの門を叩く。¥n¥n”} {“role”: “assistant”, “content”: [‘偕成証券’]}]}</pre>

目的設定

使用データ
確認

評価方法
決定

データ整備

アダプタ
チューニング

【参考】“messages”部分の記載内容テクニック

以下の内容で学習データを作成すると効果が出やすい可能性があります。

学習データ	内容
一般	指示文やフォーマットは表現の揺らぎ（variation）を持たせるとよい ※表現の揺らぎ：内容を変えない範囲で文章・形式を変更すること
箇条書きや表を含むデータ	markdown形式が効果的
長いコンテキストを含む応答	質問／指示文を最後に置くことが効果的

目的設定

使用データ
確認評価方法
決定

データ整備

アダプタ
チューニング

チューニングパラメータの意味

Azure AI Studio上から設定可能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	チューニング優先度	チューニング方法
Batch size multiplier	1	バッチサイズの合計	1	一般に、大きいほど良いとされています。
Learning rate	3e-4	学習率 大きいほどモデルの重みを大きく更新します。 まずはデフォルト値の3e-4を設定することが推奨です。	3	epochを固定した後は、この値チューニングするのが、最も影響があります。 アダプタでは3e-4からスタートし、推論結果が壊れているようであれば下げるのが良いです。 ベースライン（学習無しモデル）での推論結果を基準として判断します。
Epochs	1	学習エポック数	2	長く設定し、最適なチェックポイントを学習済みモデルとします。

目的設定

使用データ
確認評価方法
決定

データ整備

アダプタ
チューニング

【参考】学習にかかる時間・料金について

学習手法、データ件数と1件当たり入出力長に影響されます。

手法	入力データ長	出力データ長	データ件数	epoch	バッチサイズ	時間
アダプタ	約100文字	約100文字	800	5	32	48分
	約100文字		2,450	3		52分

RAG利用時のアダプタチューニング

RAGを組み合わせてtsuzumiを利用する場合は、tsuzumiが自身の持っている知識優先で回答しないようにアダプタチューニングを行うことを推奨します。

チューニングを実施しない場合、RAGで与えた情報以外の出力を返す可能性が高まることが知られています。

【学習データ例】

role	content
user	<p>"あなたには検索結果に基づいて生成を行うRAGシステムに貢献するためのタスクが与えられます。検索で得られた情報を活用して、質問に答えてください。提供された検索結果を唯一の参照として、応答を行ってください。検索結果は¥"-----¥"で区切る形式です。回答が検索で確定しない場合、不正確な情報の伝達を控えてください。他の資料を参照せず、提供された検索結果だけで応答してください。</p> <p>検索結果： ----- [実際に参照すべき文章] -----</p> <p>質問： [実際の質問]"</p>
assistant	"[質問に対する理想的な応答例]"

5 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やAIオーケストレーションを行うためのツールとしてAzure AI Studioで利用可能な機能を2つご紹介します。

プロンプトフロー

- プロンプト フローは、大規模言語モデル (LLM) によって動作する AI アプリケーションの開発サイクル全体を合理化するために設計された開発ツールです。プロンプト フローは、AI アプリケーションのプロトタイプ作成、実験、反復、デプロイのプロセスを簡素化する包括的なソリューションを提供します。
- 参考サイト
[Azure AI Studio でのプロンプト フロー - Azure AI Studio | Microsoft Learn](#)

Azure AI Model Inference (API連携)

- Azure AI Model Inference は、基本モデルの共通の機能セットを公開する API であり、開発者がさまざまなモデルセットからの予測を一定かつ一貫した方法で利用するために使用できます。開発者は、使用している基になるコードを変更することなく、Azure AI Studio にデプロイされたさまざまなモデルと対話できます。
- 参考サイト
[Azure AI Model Inference API - Azure AI Studio | Microsoft Learn](#)
[Azure AI モデル推論チャット入力候補 - Azure AI Studio | Microsoft Learn](#)

外部連携手順

(プロンプトフローからの利用)

1. エンドポイント情報の確認

2. 接続の追加

3. エンドポイント情報の設定

4. プロンプトフローからの呼び出し確認

- 1 デプロイタブから2 追加したいモデルを選択し、エンドポイント情報を確認します。

1

2

tsuzumi-7b-Finetune-ttjex

最新の情報に更新 削除

配置情報

名前	tsuzumi-7b-Finetune-ttjex
プロビジョニングの状態	最終更新日時 成功 Oct 29, 2024 4:37 PM
作成者	作成日 Oct 29, 2024 4:37 PM

モデル

tsuzumi-7b-Finetune

アプリケーション開発に役立つリンク

[コードサンプルリポジトリ](#)

[チュートリアル](#)

エンドポイント

ターゲット URI

https://tsuzumi-7b-Finetune-ttjex.eastus2.models.ai.azure.com

キー

.....

エンドポイントの状態

成功

消費

Swagger URI

https://tsuzumi-7b-Finetune-ttjex.eastus2.models.ai.azure.com/swagger/v1/swagger.json

API ルート

Azure AI model inference: Chat Completion

https://tsuzumi-7b-Finetune-ttjex.eastus2.models.ai.azure.com/chat/completions

tsuzumi-7b-instruct: Chat Completion

https://tsuzumi-7b-Finetune-ttjex.eastus2.models.ai.azure.com/v1/chat/completions

チェック

外部連携手順

(プロンプトフローからの利用)

1. エンドポイント情報の確認
2. 接続の追加
3. エンドポイント情報の設定
4. プロンプトフローからの呼び出し確認

2. プロジェクトの① 設定タブから② 新しい接続をクリックし、③ サーバーレス モデルをクリックします。

外部資産への接続を追加する

接続するリソースの種類を選択してください。

Azure AI

- Azure OpenAI Service AI モデル
- Azure AI サービス AI モデル
- Azure AI 検索 プレビュー 検索とインデックス作成
- サーバーレス モデル プレビュー AI モデル
- Azure AI Content Safety プレビュー 責任ある AI
- Azure AI 音声 プレビュー 音声

データ

- Azure Data Lake Storage Ge... プレビュー データ
- Microsoft OneLake プレビュー データ

その他のリソースの種類

- Serp 検索
- Open AI AI モデル
- API キー カスタム
- カスタム キー カスタム

クリック

閉じる

総コスト
プロジェクトのコストは、消費されたコンピューティ
スト最適化の方法について調べる
リソースのコストを表示する

外部連携手順

(プロンプトフローからの利用)

1. エンドポイント情報の確認
2. 接続の追加
- 3. エンドポイント情報の設定**
4. プロンプトフローからの呼び出し確認

3. 1.で確認した① ターゲットURIと② APIキーを入力し、
③ 接続名を記載して④ 接続を追加するをクリック

サーバーレス モデルを接続する

← 資産の種類を選択に戻る

① ターゲット URI *

https://tsuzumi-7b-Finetune-ttjex.eastus2.models.ai.azure.com

認証 *

API キー

② キー *

.....

③ 接続名 * ①

tsuzumiFT

アクセス ①

すべてのプロジェクトで共有

④ 接続を追加する 閉じる

クリック

外部連携手順

(プロンプトフローからの利用)

1. エンドポイント情報の確認
2. 接続の追加
3. エンドポイント情報の設定
4. プロンプトフローからの呼び出し確認

4. 接続が追加されると、Azure AI Studioのプロジェクト内で利用することができます。①プロンプトフローから呼び出し、他LLMやpythonコードと連携することも可能です。

名前	種類	値
question	string	\$(inputs.question)
chat_history	string	\$(inputs.chat_history)

外部連携手順 (API連携)

1. エンドポイント情報、サンプルコードの確認

2. APIキーの確認

1. API連携情報は、①デプロイから②モデルを選択し、
③実行するをクリックすることで確認可能です。

② tsuzumi-7b-yelvu

③ 実行する

エンドポイント情報が確認可能です。

基本消費量に関する情報

REST エンドポイント
https://tsuzumi-7b-yelvu.eastus2.models.ai.azure.com/v1/chat/completions

消費オプション

使用の種類
Python C# JSON

```
1 # pip install azure-ai-inference
2 import os
3 from azure.ai.inference import ChatCompletionsClient
4 from azure.core.credentials import AzureKeyCredential
5
6 api_key = os.getenv("AZURE_INFERENCE_CREDENTIAL", '')
7 if not api_key:
8     raise Exception("A key should be provided to invoke the endpoint")
9
10 client = ChatCompletionsClient(
11     endpoint='https://tsuzumi-7b-yelvu.eastus2.models.ai.azure.com',
12     credential=AzureKeyCredential(api_key)
13 )
14
15 model_info = client.get_model_info()
16 print("Model name:", model_info.model_name)
17 print("Model type:", model_info.model_type)
18 print("Model provider name:", model_info.model_provider_name)
19
20 payload = {
21     "messages": [
22         {
23             "role": "user",
24             "content": "I am going to Tokyo, what should I s
```

チェック

外部連携手順 (API連携)

1. エンドポイント情報、サンプルコードの確認

2. APIキーの確認

2. APIキーは① 詳細タブのキーから確認可能です。

The screenshot displays the configuration page for a model named 'tsuzumi-7b-yelvu'. The interface includes a top navigation bar with a '詳細' (Details) tab selected, indicated by a red circle with the number '1'. Below the navigation bar, there are buttons for 'プレビューで開く' (Open in Preview), '最新の情報に更新' (Update latest information), '編集' (Edit), and '削除' (Delete). The main content area is divided into several sections:

- 配置情報 (Configuration Information):** A table showing the model's name ('tsuzumi-7b-yelvu'), deployment status ('成功' - Success), and creation date ('Sep 20, 2024 4:45 PM').
- モデル (Model):** A link to the model's page ('tsuzumi-7b').
- アプリケーション開発に役立つリンク (Links helpful for application development):** Links to 'コードサンプルリポジトリ' (Code sample repository) and 'チュートリアル' (Tutorial).
- エンドポイント (Endpoint):** Shows the target URI ('https://tsuzumi-7b-yelvu.eastus2.models.ai.azure.com').
- キー (Key):** A field containing the API key, highlighted with a red box. A callout bubble points to this field with the text 'APIキーが確認可能です。' (API key can be confirmed). A purple 'チェック' (Check) button is overlaid on the key field.
- エンドポイントの状態 (Endpoint status):** Shows '成功' (Success) and the Swagger URI ('https://tsuzumi-7b-yelvu.eastus2.models.ai.azure.com/swagger/').
- API ルート (API Routes):** Lists two routes: 'Azure AI model inference: Chat Completion' and 'tsuzumi-7b-instruct: Chat Completion', both with their respective URIs.
- 監視と安全性 (Monitoring and Security):** Shows 'Azure AI Content Safety' is '有効' (Enabled).

6 お問い合わせ

お問い合わせ先

tsuzumiに関するお問い合わせ

tsuzumi on Azure MaaS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

法人・団体のお客様向けに有償のチューニング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いただければと思います。
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	2024/11/20	初版	初版

NTT DATA