D-05

生成 AI と雅に連歌を嗜もう
~ Amazon SageMaker で簡単に LLM を Fine Tune ~

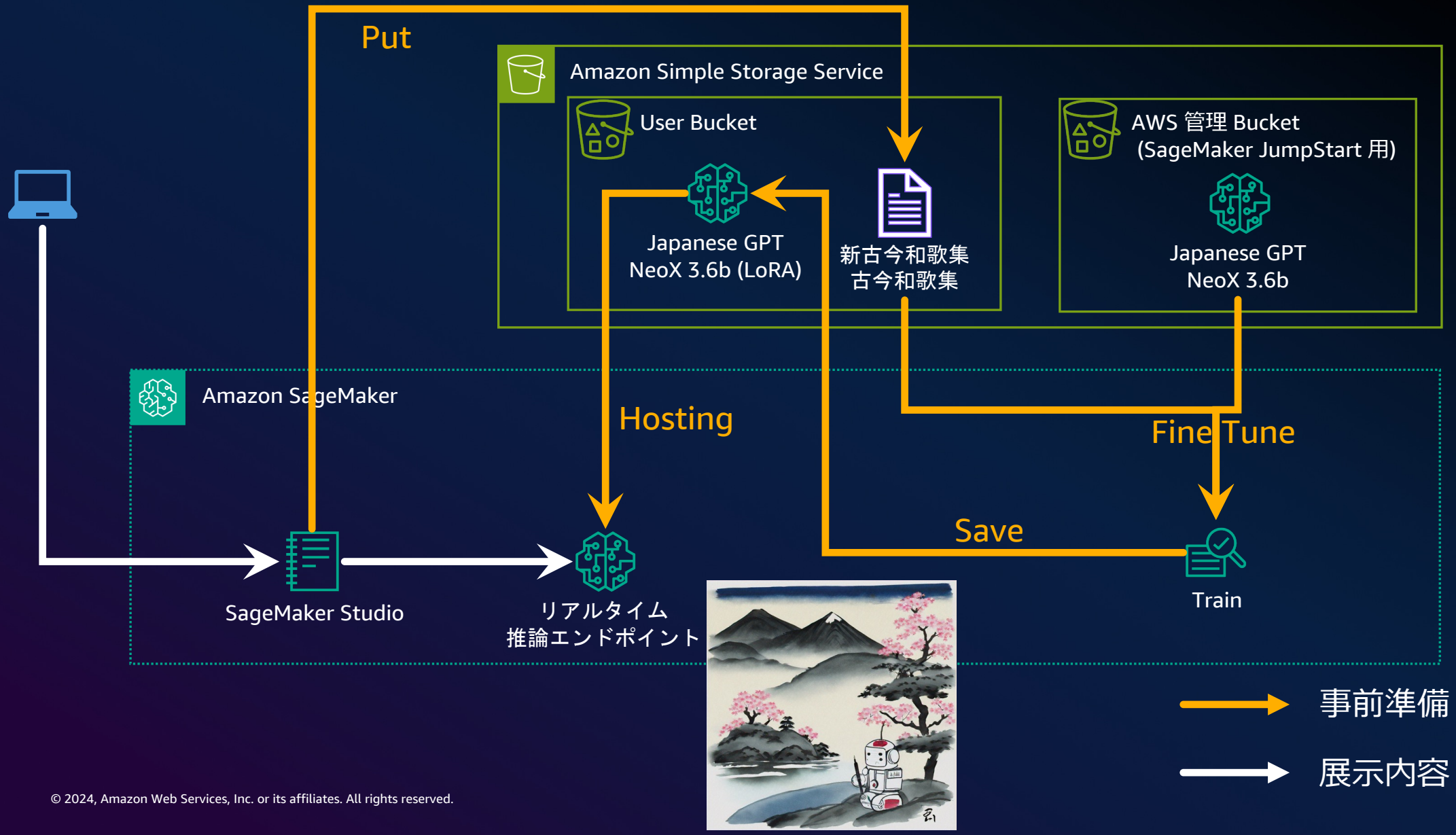
ネタ元：https://aws.amazon.com/jp/builders-flash/202308/generative-ai-renga/

このブースは何？

- 連歌を読める生成 AI を用意
  - 連歌を読めるように生成 AI を Fine Tune
  - トレーニングには Amazon SageMaker を利用
  - 推論には Amazon SageMaker Inference (Realtime Endpoint)を使用

Titan Image Generator で生成

データを準備して SageMaker で Fine Tune



Fine Tune ... の前に Text to Text の生成 AI とは

次のトークン（≒単語）を予測するタスクを繰り返し解くことで
テキストを生成する AI のこと

Input	Output
吾輩は	猫
吾輩は猫	で
我輩は猫で	ある
我輩は猫である	。
吾輩は猫である。	名前
吾輩は猫である。名前	は
吾輩は猫である。名前は	まだ
吾輩は猫である。名前はまだ	ない
吾輩は猫である。名前はまだない	。

Text to Text の生成 AI ってどう作るの？

学習でも次のトークン（≒単語）を予測するが予測する文章はフェーズによって違う

支離滅裂な文章
しか生成しない

文章を与えると
その続きを一方的に
ひたすら語る

対話したりタスクを
解いたりできる

未学習モデル

Pre Training

Base Model

Instruction Tuning

Intruction
Model

※追加で不適切なことを出力させない、文体の矯正などの調整を行うことも多い

大量のテキストデータ
(wikipedia など)

- Amazon SageMaker is a fully managed service that brings together a broad set …
- ある日のことでございます。お釈迦様は…
- 子曰学而時習之不亦説乎有朋自遠方来不…

最近のモデルは数10兆のトークンのデータを使う

Amazon SageMaker ⇒ **is**
ある日のことでござ ⇒ **い**
子曰学而時習之不亦 ⇒ **説**

を予測できるように
学ぶことで知識を獲得

Instruction Data
(AI に解かせたい教師データ)

ユーザー: 以下に与えるテキストを英語に訳せ
これはペンです
システム: This is a pen.
※数10万件の対話データを使うことが多い

ユーザー: 以下に与えるテキストを英語に訳せ
これはペンです
システム:
↓
This

ユーザー: 以下に与えるテキストを英語に訳せ
これはペンです
システム: This
↓
is

のように各タスクを解けるように学習
（要約や相談、品詞分解、等多様なタスク）

しかし・・・

Instruction Data には
連歌タスクなどない！（と思う）
ので、モデルによっては
あまりいい回答を得られない

既存モデルに対して欲しい出力を引き出す方法

手法	手法の概要	モデルの更新	可能なこと	トレーニングコスト
Continual Pre-Train	基盤モデルを追加で学習	そのものを更新	新しい知識を追加	高
LoRA をはじめとする PEFT など	パラメータをわずかに追加して学習	元のモデルのパラメータは変えずにパラメータをわずかに追加	知識獲得は期待できないが新タスクを解ける	
RAG (Retrieval-Augmented Generation)	結果出力の際にモデル外の情報を参照	変更しない	日々更新される知識をモデル外から追加	無
プロンプトの工夫	Few Shot や CoT など		LLM の知識を最大限引き出す	

Continual Pre-train が有効なケース

業界特化の専門知識の獲得（≒ AI が正確な知識を持っていない）

- **医療・ヘルスケア**：医学用語、診断プロトコル、治療法など、医療
- **法務**：法律用語、判例、法的文書の作成や解釈
- **金融**：金融用語、投資戦略、リスク管理、規制遵守
- **メーカー**：特定の製品やサービスに関する詳細な知識を持ち、正確に答えられる

LoRA をはじめとする Fine Tune が有効なケース

- 連歌を詠む
- 回答の好みの調整
  - 要約タスクは１行で、など
- 特定のキャラづけ
  - 口調など
- 与えたコンテキストのみから回答をさせるように調整

Fine Tuning (LoRA)

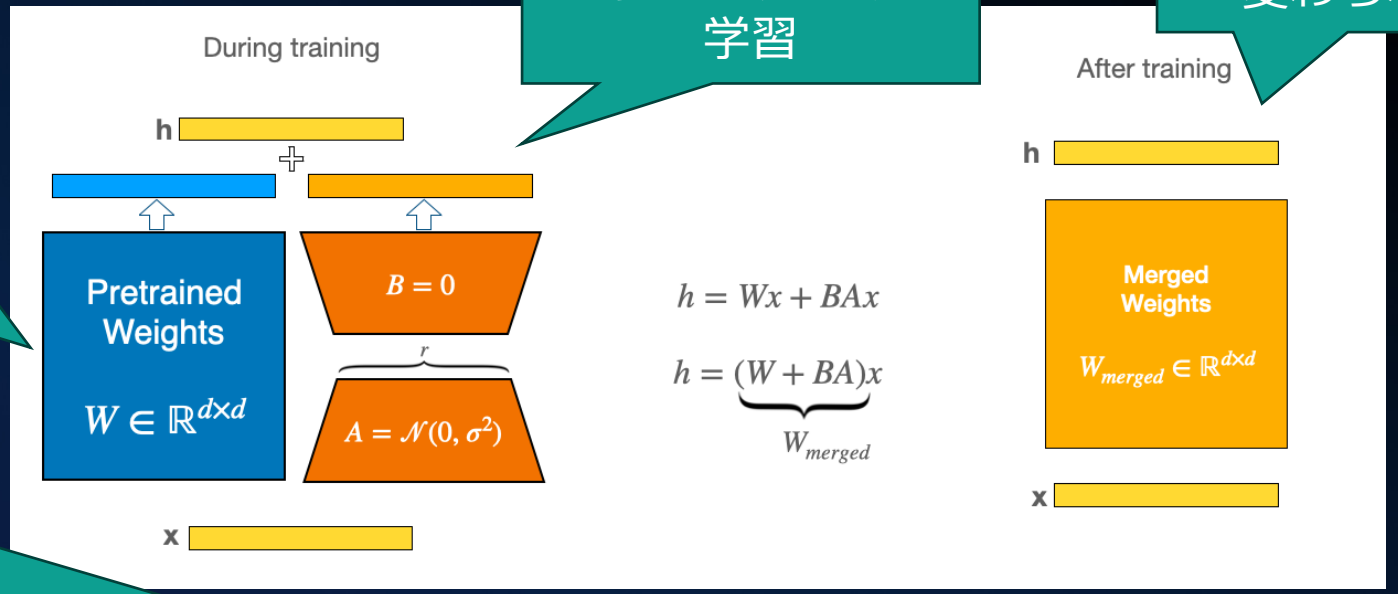
https://huggingface.co/docs/peft/main/en/conceptual_guides/lora

横付けした小さいパラメータのみを学習

推論時は合体（＝推論速度は変わらない）

元の重みは不変なので知識の破壊的忘却はない

横付けしたパラメータの数は少量



https://arxiv.org/pdf/2311.06243

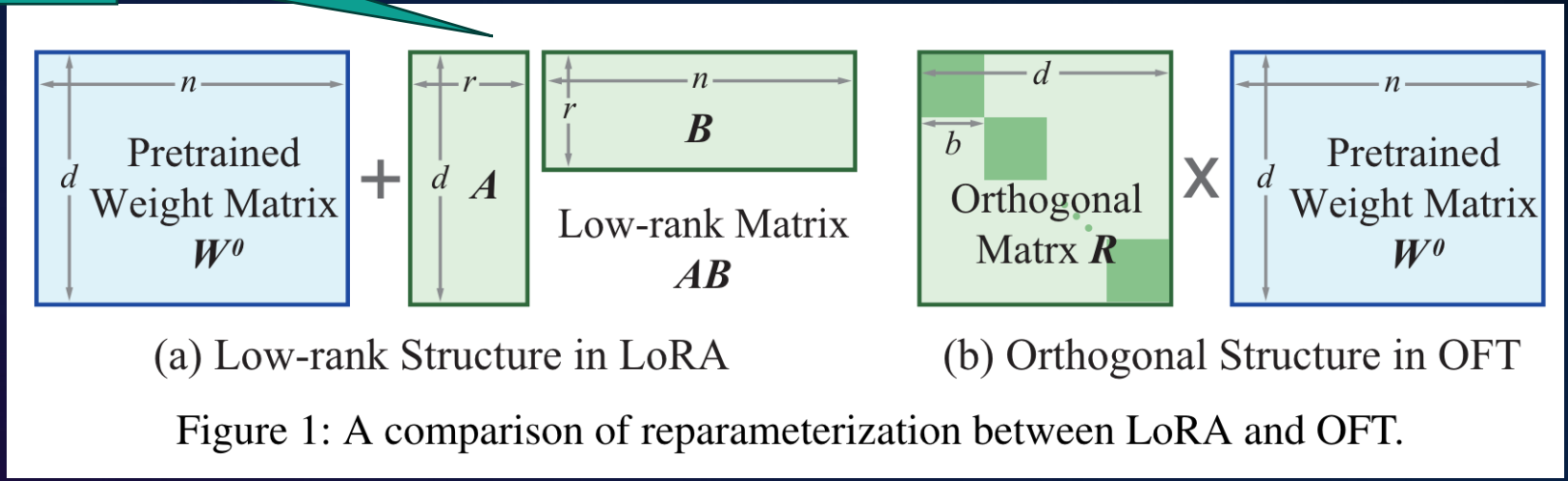


Figure 1: A comparison of reparameterization between LoRA and OFT.

デモで利用できるモデル

- Japanese GPT NeoX 3.6b (with LoRA)
  連歌タスクを LoRA で学ぶ
- Japanese GPT NeoX 3.6b
  LoRA 学習前のモデルで比較
  SageMaker JumpStart で用意
- ELYZA japanese Llama 2 7b chat
  パラメータ数が倍の大きいモデルで比較
  SageMaker JumpStart で用意
- Claude 3 Haiku
  API で提供しているモデルで比較(パラメータ数は非公開)
  Amazon Bedrock で用意

LoRA のトレーニングは SageMaker Training

Amazon SageMaker Training が提供するサービスを 1 文 で表すならば

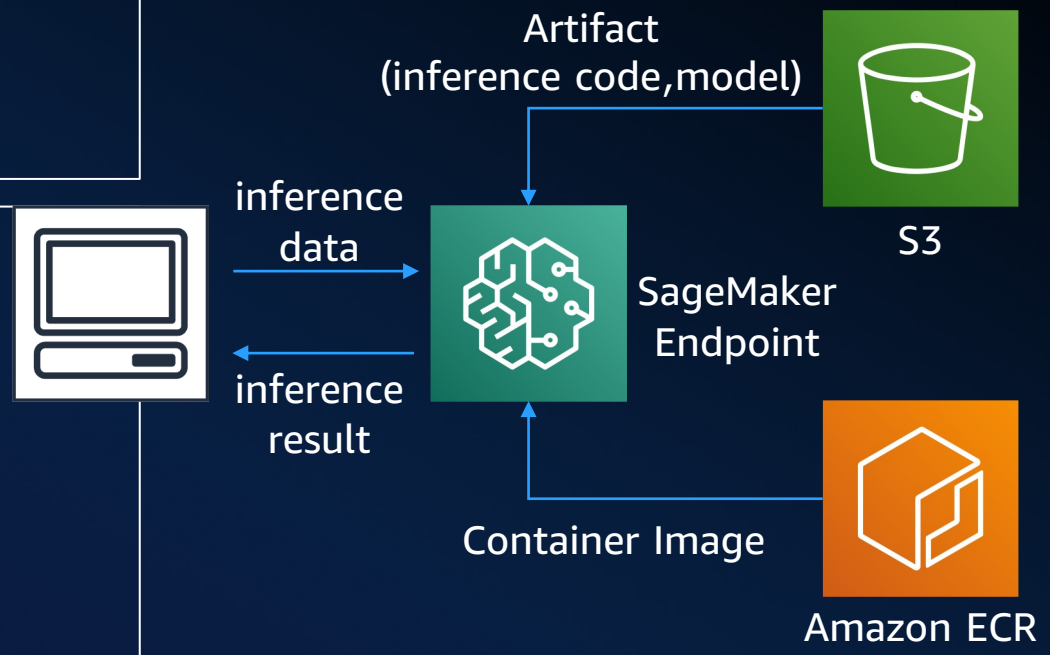
「用意したコード」と「用意したデータ」を
「指定したコンピューティングリソース」と「用意した環境」で実行し、
「実行履歴を自動記録」して「アーティファクトを自動で保存」する機能
を提供する

※ただし全てをユーザで用意する必要はない ⇒ e.g. 後述のSageMaker JumpStart
今回はコードとデータを準備

SageMaker を用いた連歌モデルのデプロイ

```
# デプロイ
response = sm_client.create_endpoint(
    EndpointName=endpoint_name,
    EndpointConfigName=endpoint_config_name,
)
```

```
# リクエスト
response = smr_client.invoke_endpoint(
    EndpointName=endpoint_name,
    ContentType='application/json',
    Accept='application/json',
    Body='上の句を与えるので・・・'
)
# レスポンス確認
predictions = json.loads(response['Body'].read().decode('utf-8'))
print(predictions)
```



SageMaker のリアルタイム推論の特徴

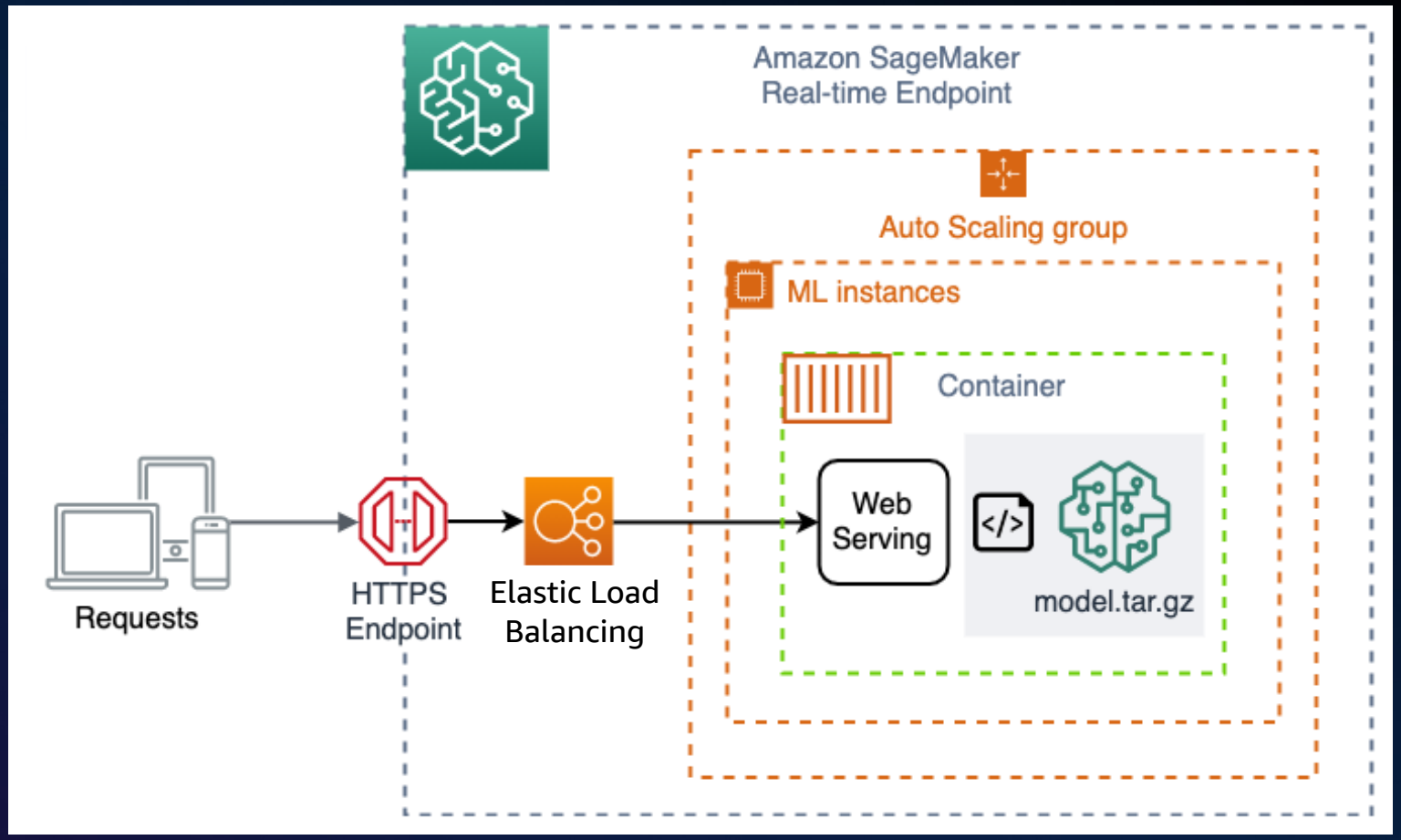
SageMaker
Real-time
Inference

永続的なマイクロサービスを作成

高速なレスポンス (ペイロード上限6MB)

外部アプリケーションからアクセス可能

自動スケーリング



SageMaker JumpStart は UI でも API でも Deploy と Finu Tune ができる

1
モデルカタログから基盤モデルを選択

Meta AI
AI21 labs
LightOn
We bring Light to AI
stability.ai
co:here
alexa

2
モデルをデプロイ

SageMaker の機能で
推論用にモデルをデプロイ

3
ファインチューニングと ML ワークフローの自動化

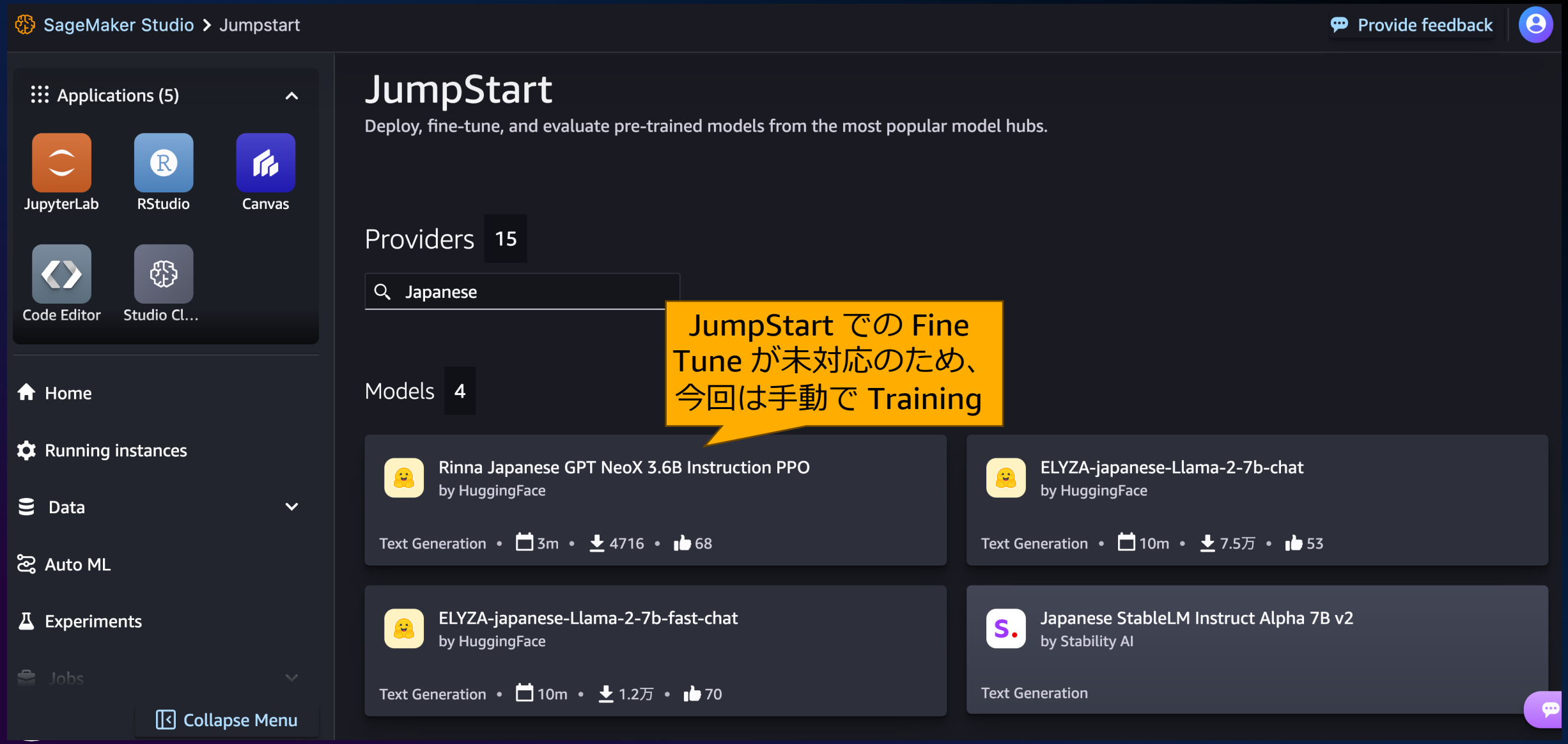
GUI ベースの
ファインチューニング

ML ワークフローの自動化

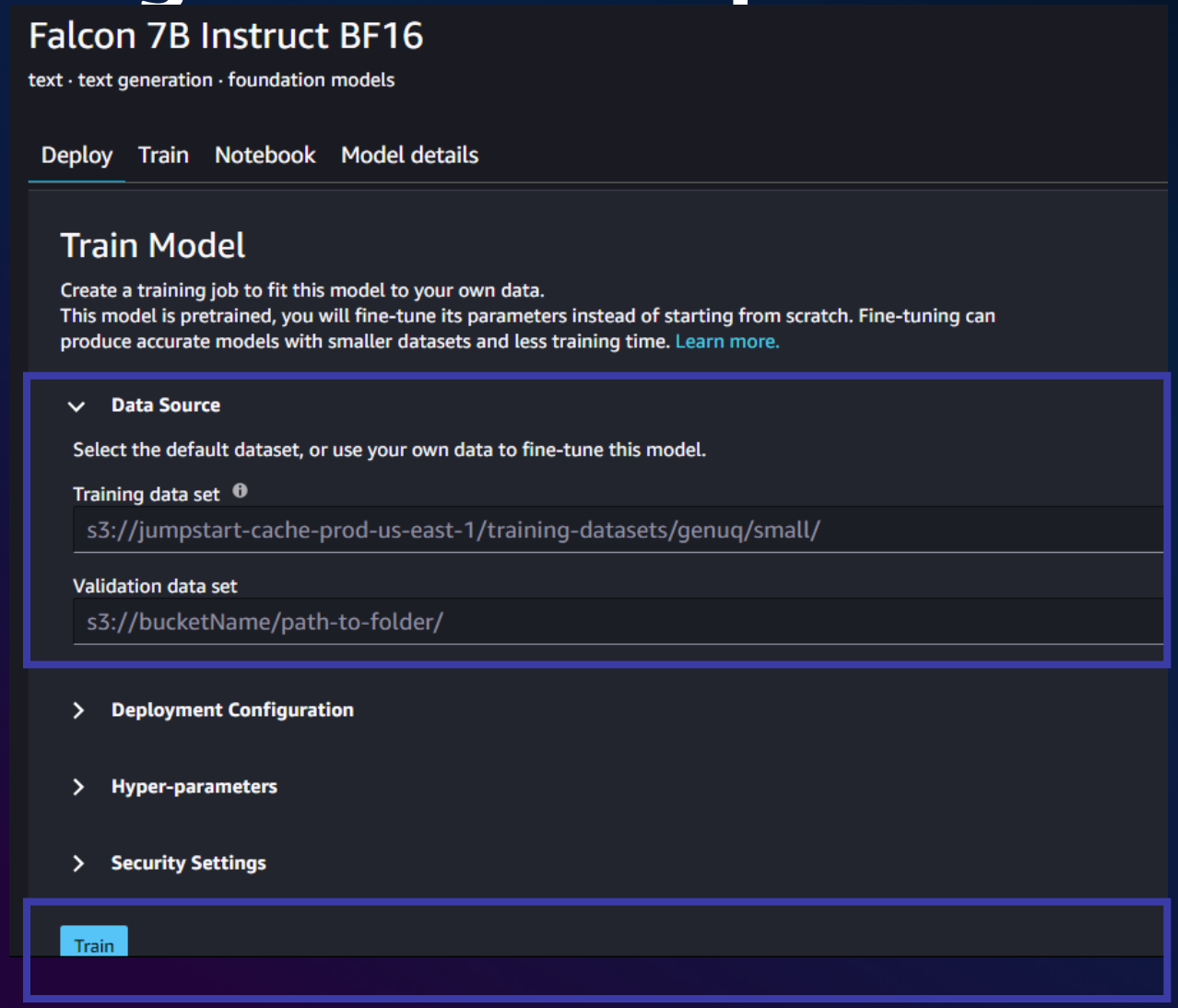
データはアカウント外に出さない
- モデル
- ログ
- インスタンス
- モデル入出力

Amazon SageMaker 機能群と**完全に統合**

日本語対応の大規模言語モデルも続々追加



SageMaker JumpStart でファインチューニング



学習データセットを
Amazon S3 に置いて「Train」
するだけ！

※ ファインチューニングに対応している
モデルとしていないモデルがあります

Fine Tune に関する FAQ

LLMのFine Tuneとは何ですか？

既存の大規模言語モデル（例えば、LlamaやMistral）を特定のタスクやドメインに適応させるために追加のデータで再訓練するプロセスです。

これにより、モデルは特定の専門知識や業務要件に合わせた応答や生成が可能になります。

Fine Tuneはなぜ必要ですか？

一般的なモデルが持つ広範な知識を特定の業務やドメインのニーズに適応させるために必要です。

例えば、医療、法務、金融などの専門領域や、企業固有のプロセスや用語に合わせた応答が求められる場合に有効です。

Fine Tuneにはどのくらいのデータが必要ですか？

タスクの複雑さやモデルの元の性能に依存しますが、一般的には千から数万の例が推奨されます。

データの質と多様性も重要です。

Fine Tuneにどのくらいの時間がかかりますか？

モデルのサイズ、データ量、計算資源の量に依存します。

小規模なデータセットであれば数時間、大規模なデータセットであれば数日かかることがあります。

Fine Tuneにはどのような技術やツールが必要ですか？

機械学習のフレームワーク（例：TensorFlow、PyTorch）、GPUなどの高性能計算資源、そして特定のドメインに関するデータセットが必要です。

多くの場合、Amazon SageMakerを使用して計算資源を提供します。

Fine Tuneのコストはどのくらいですか？

計算資源（特にGPU）の使用量、データ準備の手間、そして必要なトレーニング時間によって変動します。

Amazon SageMakerのリソースを使用する場合、リソースの使用時間に応じた料金が発生します。

Fine Tune後のモデルのメンテナンスはどのように行いますか？

Fine Tune後のモデルは、定期的に新しいデータやフィードバックを取り入れて再トレーニングすることが推奨されます。これにより、モデルは常に最新の情報やニーズに適応することができます。

また、モデルのパフォーマンスをモニタリングし、必要に応じて調整を行います。

Fine Tuneされたモデルの評価はどのように行いますか？

特定の業務やタスクに対するパフォーマンス指標（例：精度、再現率、F1スコア）を使用して行います。

ユーザーテストやフィードバックも重要な評価方法です。

Fine Tuneはどのようなビジネスに適していますか？

医療、法務、金融、カスタマーサポート、教育、マーケティング、製品開発など、特定の専門知識やカスタマイズされた応答が必要なビジネスに適しています。

Fine Tuneに失敗した場合、どのように対処すればよいですか？

まずデータの質や多様性を見直し、適切な前処理が行われているか確認します。

またモデルの設定やハイパーパラメータの調整も検討します。