Microsoft

Fine-tune SLM on Azure

Alex Shen
Senior Solution Specialist
AI Global Black Belt
Microsoft

SLM Innovator Lab Workshop
2024/11/13

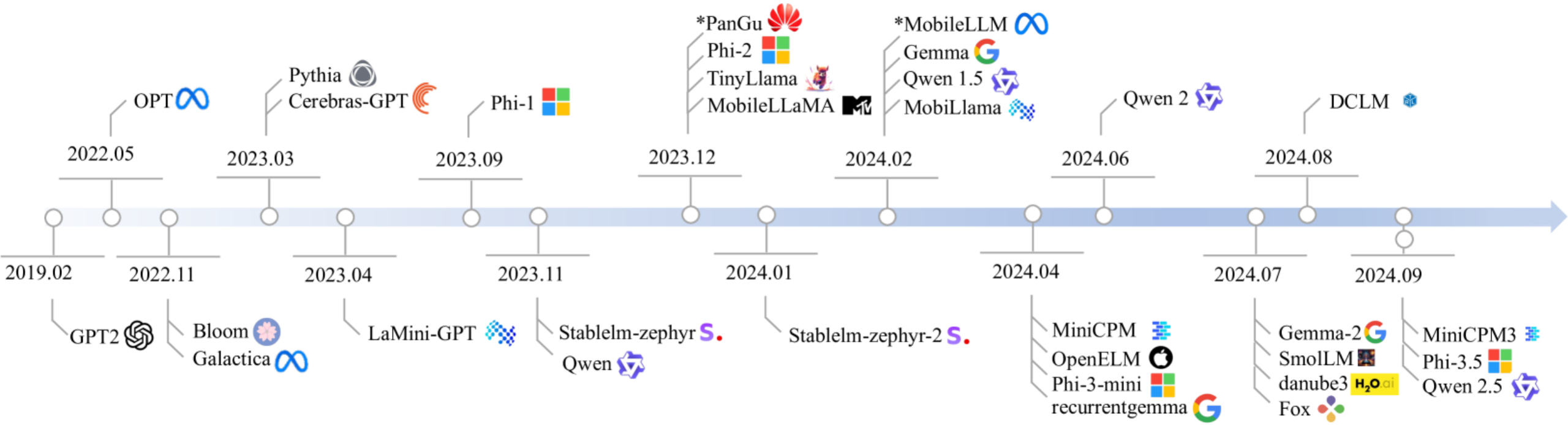
Agenda

- Small Language Models
  - Microsoft の SLM - Phi-3
- Fine-Tuning
  - Azure での Fine-Tuning

Small Language Models

Small Language Models (SLMs)とは

より小さく、計算負荷の低いモデルで、単純なタスクで優れたパフォーマンスを発揮します



Source: Small Language Models: Survey, Measurements, and Insights

Small Language Models (SLMs) のメリット

より小さく、計算負荷の低いモデルで、単純なタスクで優れたパフォーマンスを発揮します

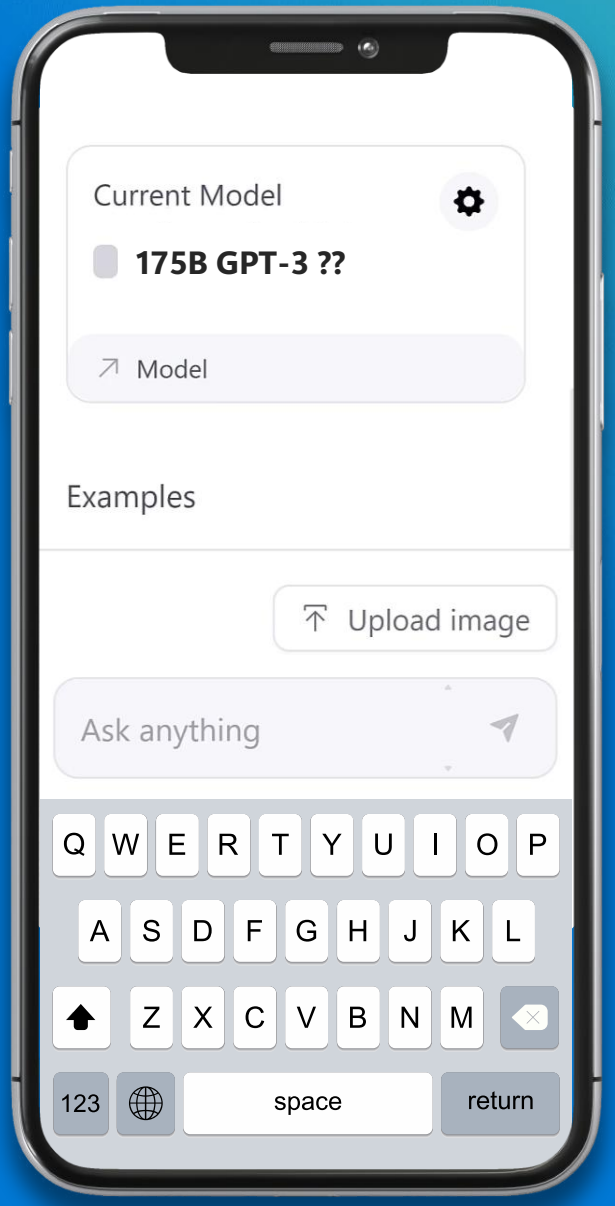
高いコスト
パフォーマンス

デプロイの
柔軟性

極めて低い
レーテンシー

容易な
カスタマイズ(FT)

LLM on your phone?



推論メモリのニーズ:

32-bit precision:

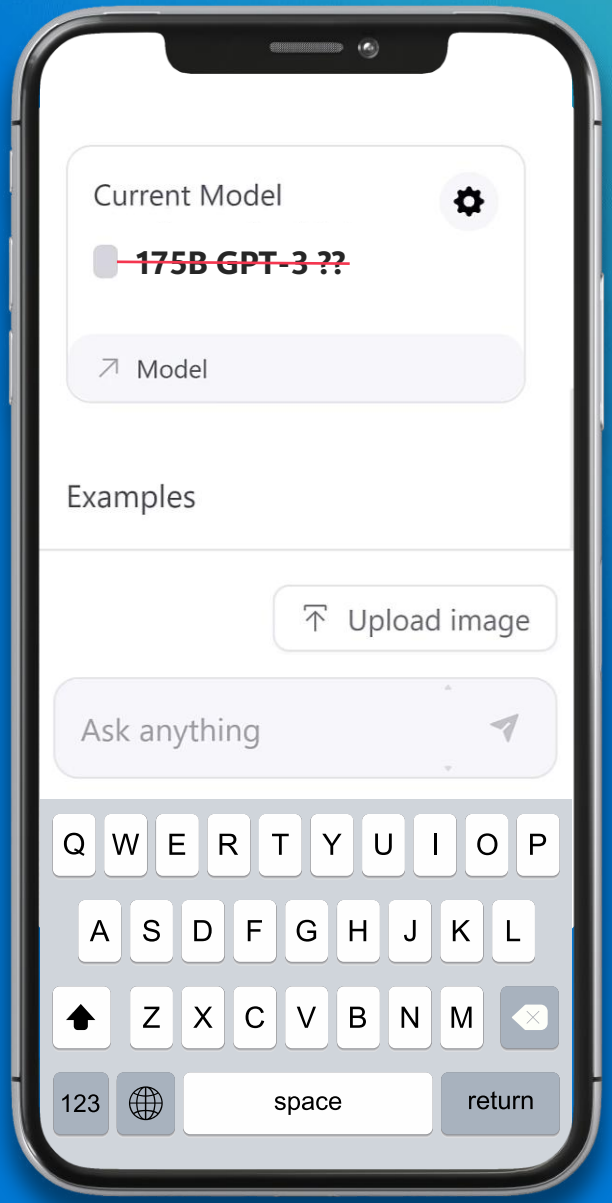
Model parameters: 175 bil x 4 bytes = 700 GB
Intermediate activations: 700-1400 GB
Overheads: 10-20 GB

Total: 1410-2120 GB

16-bit precision:

Model parameters: 175 bil x 2 bytes = 350 GB
Intermediate activations: 350-700 GB
Overheads: 10-20 GB

Total: 710-1070 GB

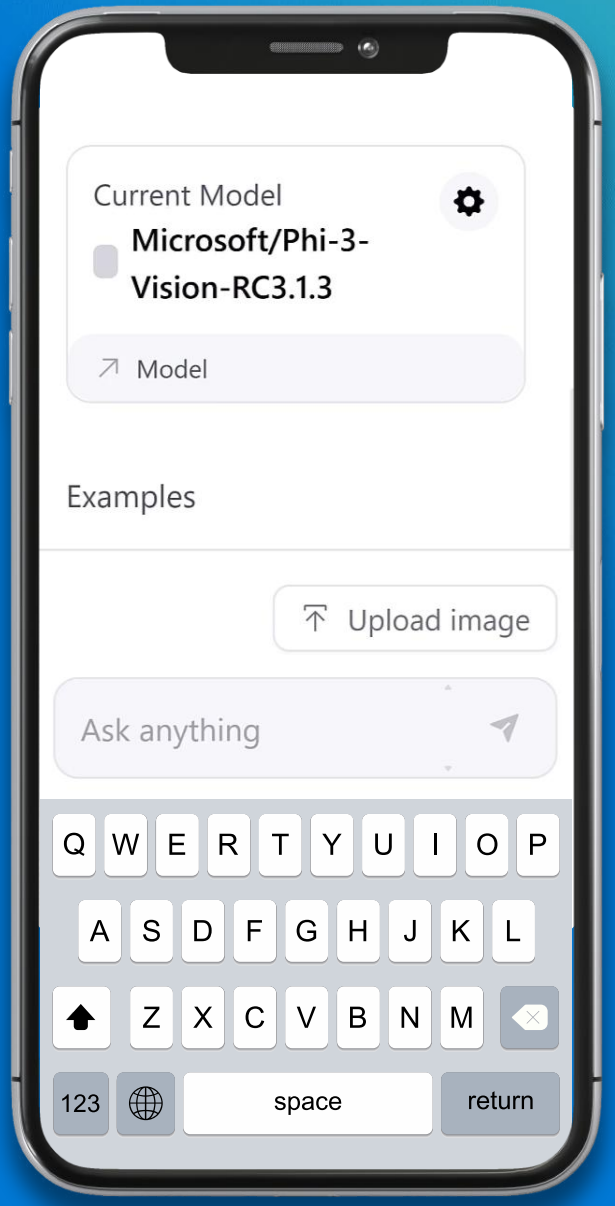


~~LLM~~

SLM on your phone!!

WebGPU Demo

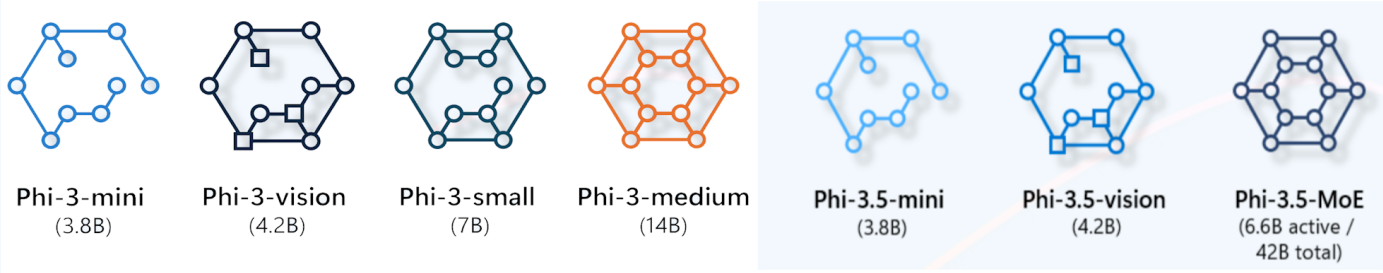
https://caniuse.com/webgpu



Introducing Phi-3

- 多言語マルチモーダルSOTA SLM ファミリー
- Microsoft Research から Open Weight リリース
- サイズに対する的画期的な性能

Open Source with MIT License



Instruction Tuned **RAI Safety Aligned**

Available on

Azure AI Model Catalog

GitHub Model Catalog

Hugging Face

ONNX Runtime

NVIDIA NIM

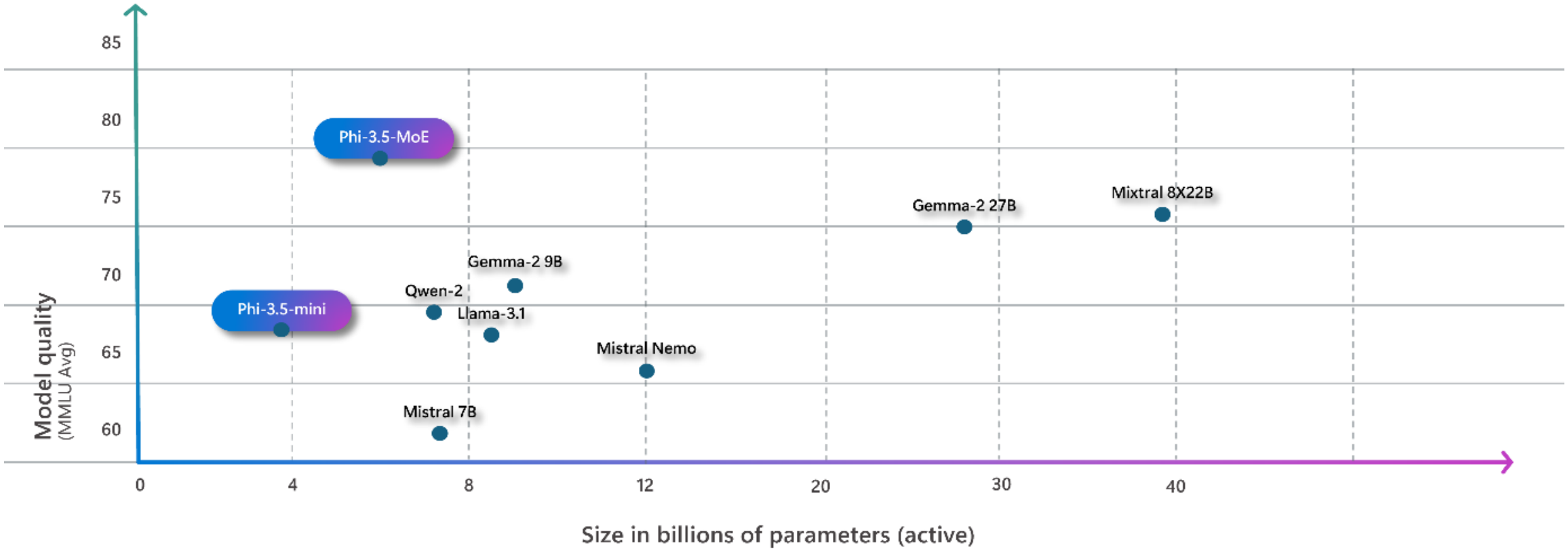
Ollama

LMStudio

VSCode AI Toolkit

Phi-3 サイズに対する画期的な性能

Phi-3.5 (2024/08)



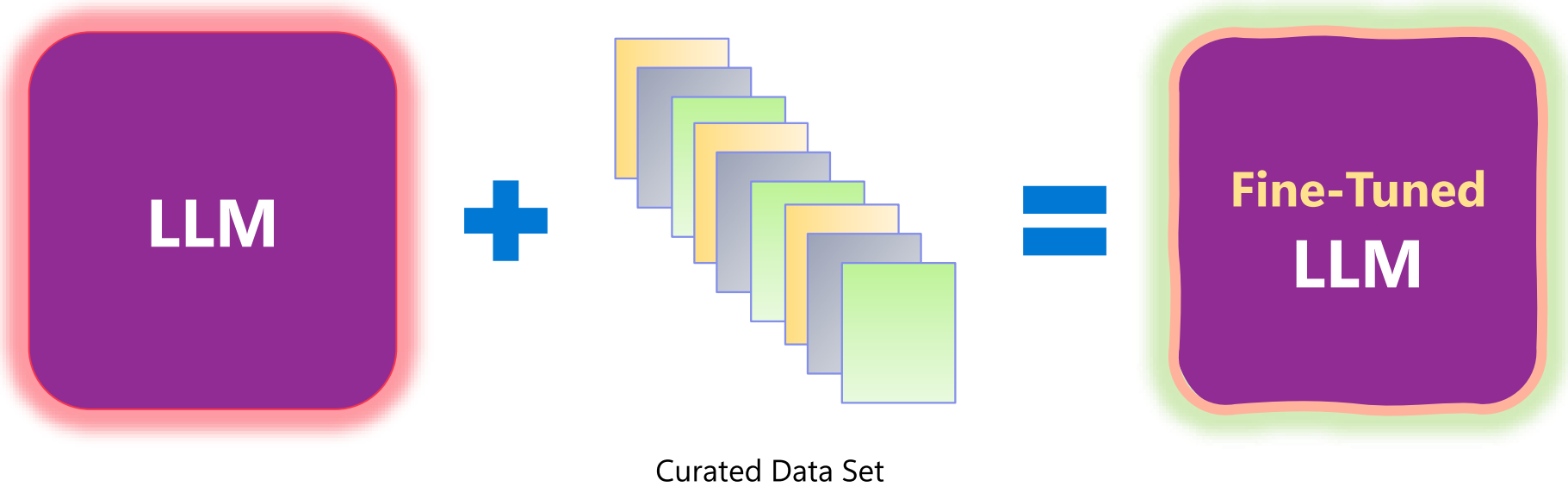
SLMが適したユースケース

- デバイス上またはオンプレミスでオフライン環境でローカル推論が必要な場合
- 応答の速さが重要なレーテンシー・クリティカルなシナリオ
- コストに制約のあるユースケース、特に単純なタスク
- リソースに制約のある環境
- Fine-tuneによってパフォーマンスの向上を確認できる (out-of-boxのLLMと比較して)

Fine-Tuningの基本

Fine-Tuningとは?

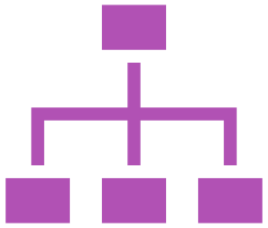
ファインチューニングとは、**特定のタスク**または**新しいデータセット**に対する追加のトレーニングを行い、事前トレーニング済みのLLM/SLMをカスタマイズすることを指します。



Fine-Tuningのメリット

パフォーマンスの向上

Fine-Tuningにより、対象タスクに対するモデルの精度と有効性が向上し、より正確で信頼性の高い出力が得られます。



ドメイン/タスクの特化

Fine-Tuningされたモデルは、タスク対応により適切な挙動を示します。

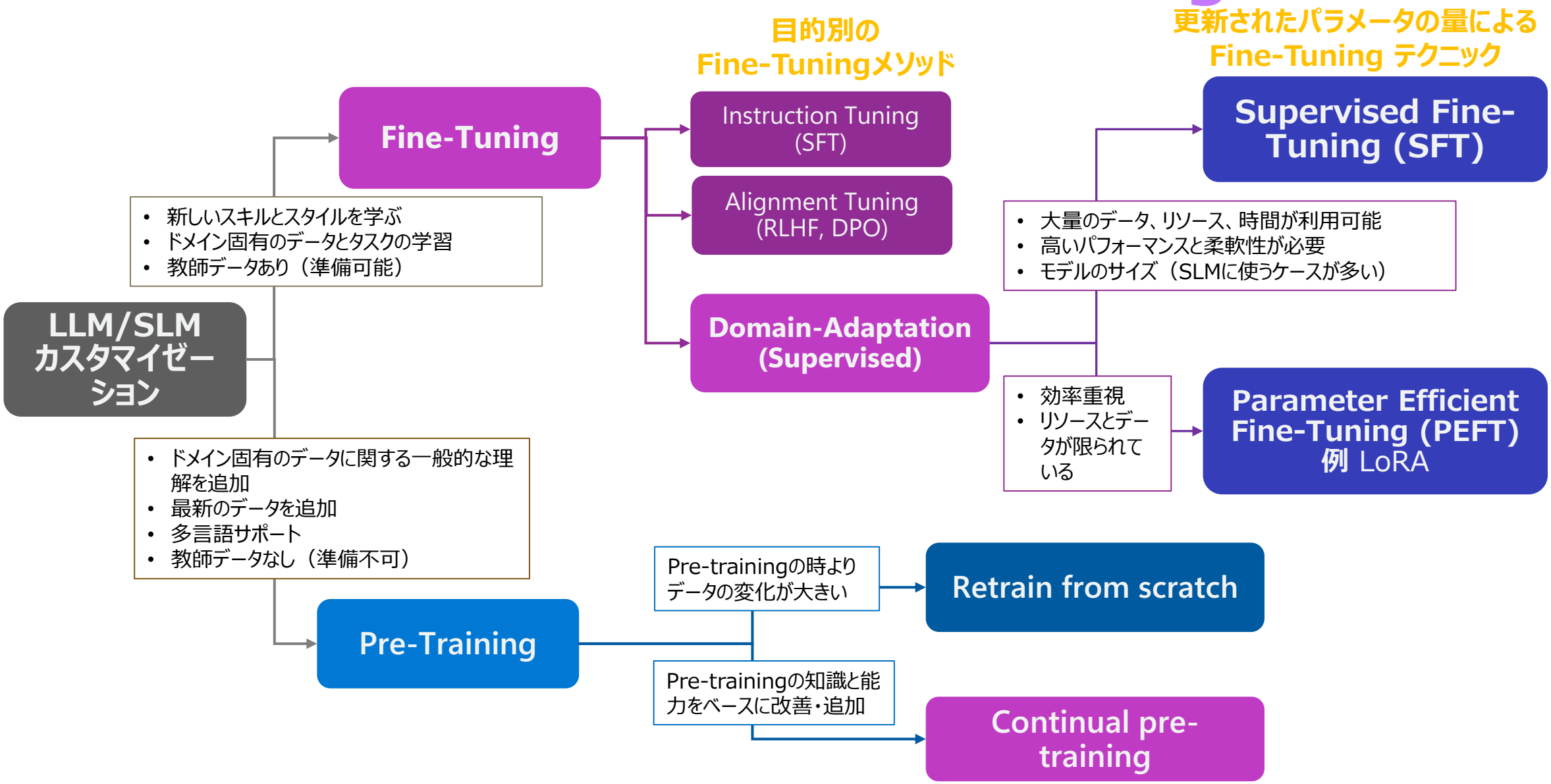
Fine-Tuningされたモデルは、ドメイン固有のコンテンツをよりよく理解し、生成します。

効率の向上

Fine-Tuningされたモデルを使用すると、Few-shotプロンプトや複雑な指示トークンを省け、プロンプトトークンを節約できます。

Fine-TuningされたSLMを使用すると、LLM を使用する場合よりレーテンシーの改善を見込めます。

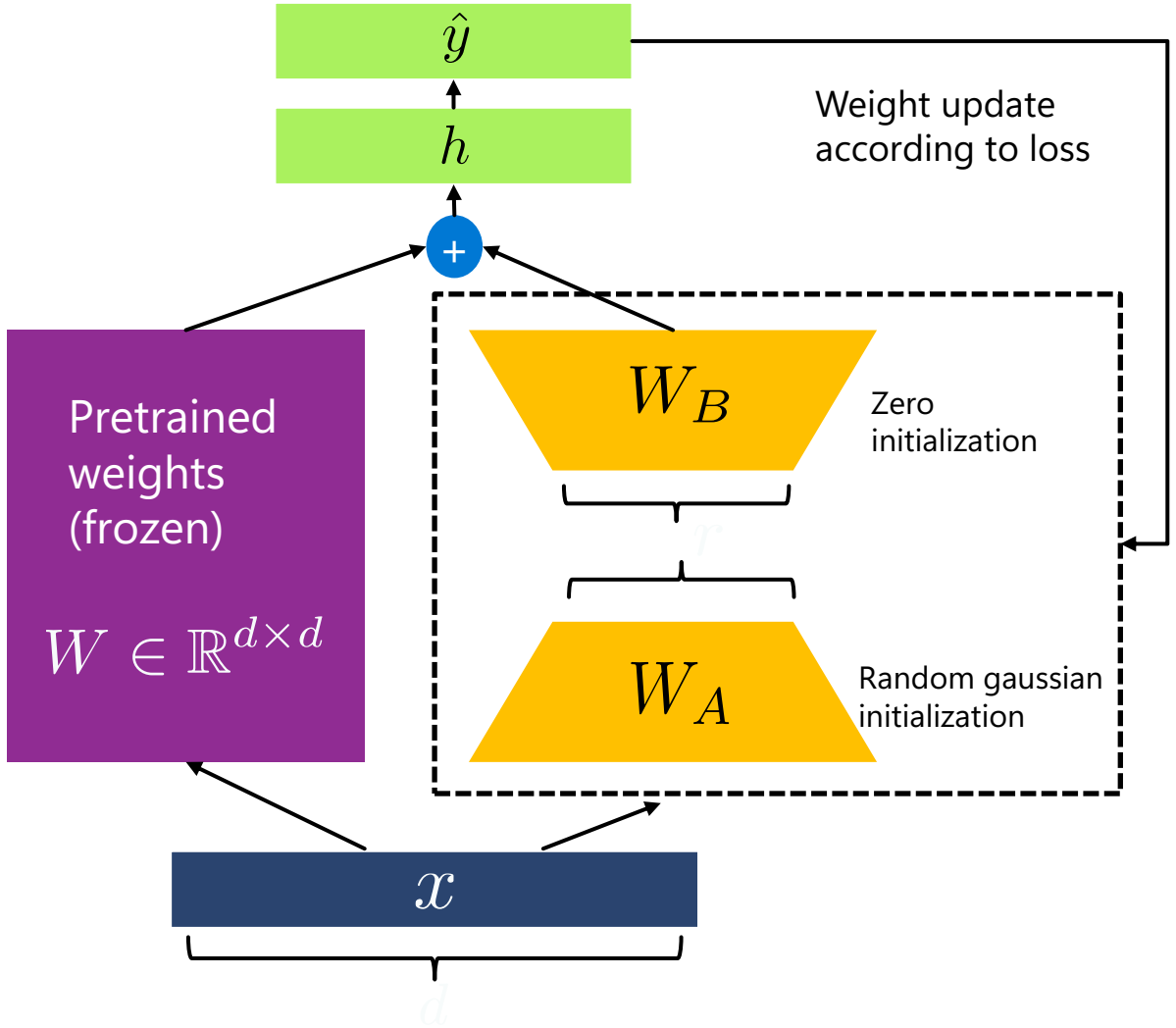
カスタマイゼーション目的別のFine-Tuningメソッド



[Ref.] A Comprehensive Overview of Large Language Models: https://arxiv.org/pdf/2307.06435.pdf

LoRA (Low-Rank Adaptation of LLMs)

- モデル全体をトレーニングする代わりに、小さなアダプターをモデルに追加してFine-Tuningします
- アップデートするパラメーター数はモデルのパラメーター数の0.1%-5%



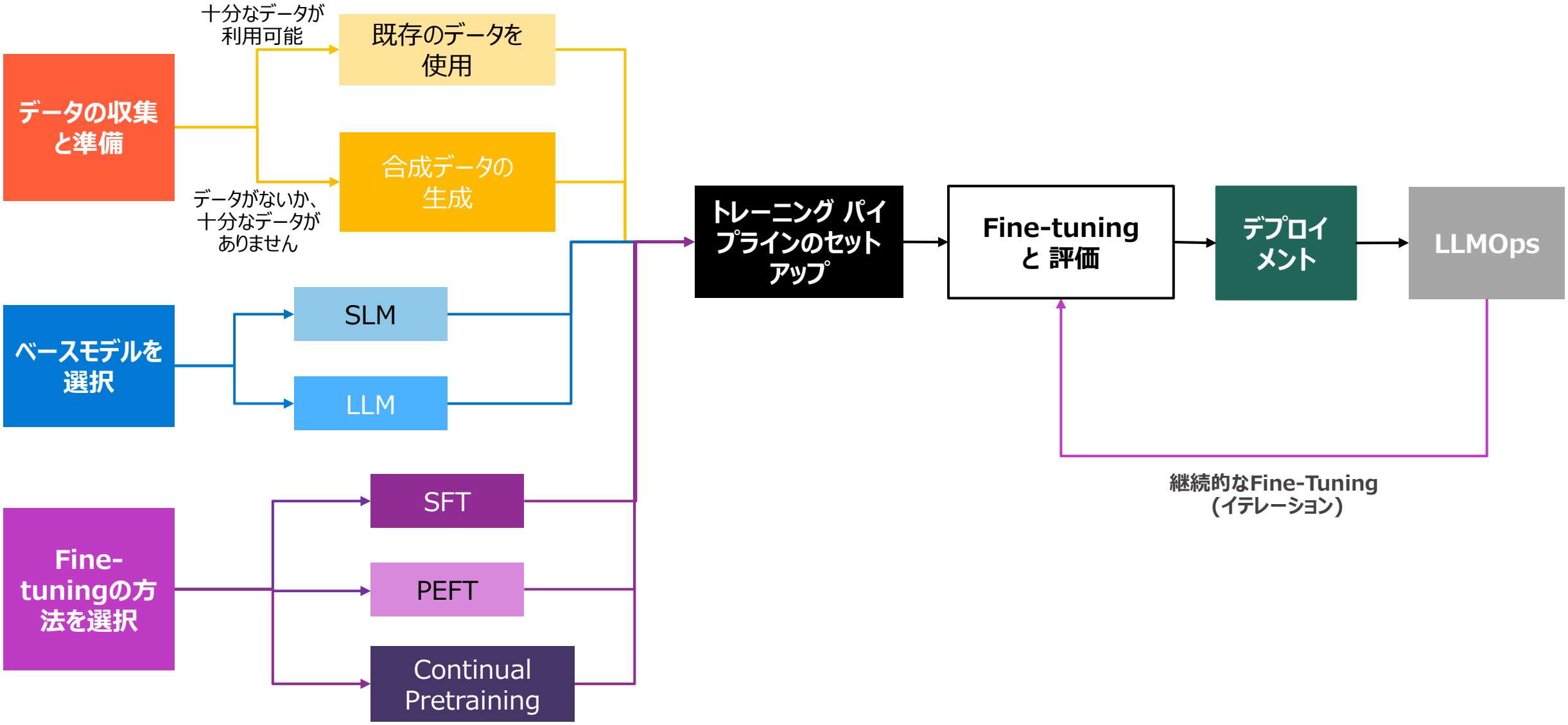
r: Low-rank dimension (smaller r makes model training faster and saves memory, but reduces accuracy)

e.g., If r = 4, d = 512,
W = 512*512*FP32(32bit) = 8,388,608

W_A = 512*4*FP32 = 65,536
W_B = 4*512*FP32 = 65,536

→ Only **1.56%** parameters

Fine-Tuningの流れ



Azure上でのFine-Tuning

Azure AI Stack: GenAI ライフサイクルの完全サポート

AI Models

- Microsoft Research Model Family
- Azure OpenAI Model Family
- Mistral AI Model Family
- Meta Llama Model Family
- Databricks Model Family
- Cohere Model Family
- Hugging Face Model Family
- NVIDIA Model Family
- Snowflake

Azure AI

Azure AI Services

インテリジェントアプリケーション向けの事前トレーニング済みのAIサービス

Azure Machine Learning

AIモデルを設計および管理するためのフルライフサイクルツール

Responsible AI Tooling

信頼できるAIアプリの構築と管理

Azure AI Studio

カスタムCopilotを開発および展開するための包括的なプラットフォーム

AI Infrastructure

- NC-series GPU (P100, V100, A100)
- ND-series GPU (A100, H100)
- MI300 GPU
- CPUs and FPGA

Azure AI または AOAI Studio での従量課金 Fine-Tuning

ベースモデルの選択

- AOAI models: Babbage-002, Davinci-002, GPT-3.5-Turbo, GPT-4o & GPT-4o mini, GPT-4
- Meta Llama 3.1/3/2
- **Microsoft Phi-3-mini/medium**

準備済みデータの提供

- 100以上のサンプルを収集・生成
- Chat または completionフォーマットに変換済み
- Training と validation セット指定
- データをアップロード

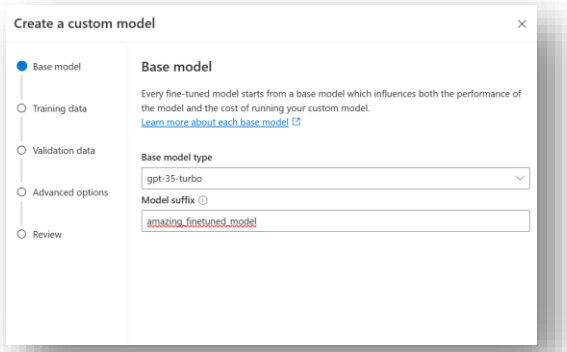
トレーニングと評価

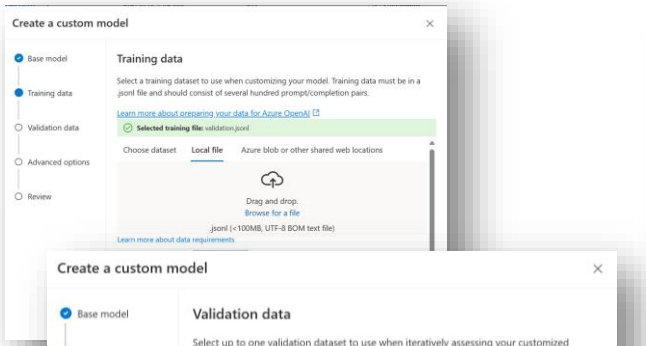
- Epoch、Batch size、学習率などのハイパーパラメータを指定
- トレーニングとLoss評価
- トレーニングと精度評価

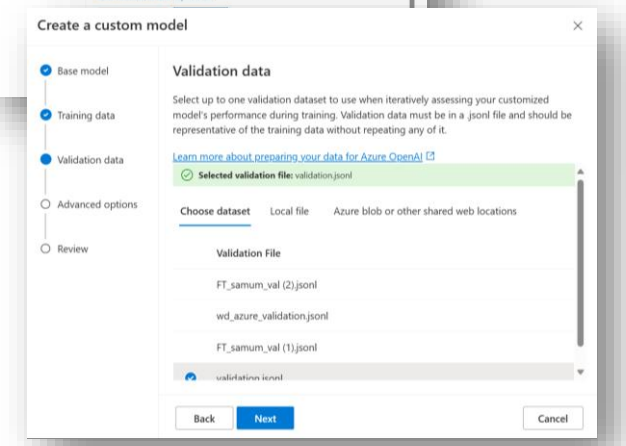
デプロイメント

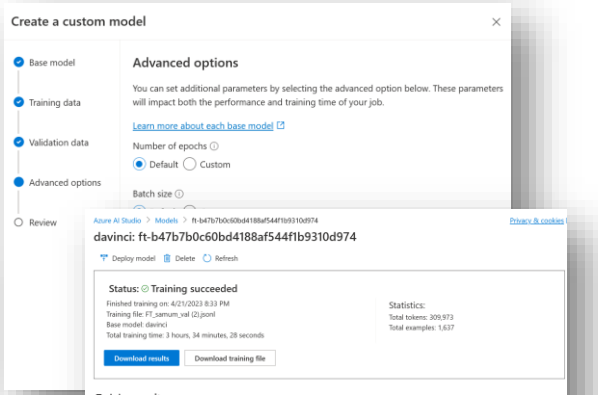
- Fine-tuning したモデルをエンドポイントにデプロイ

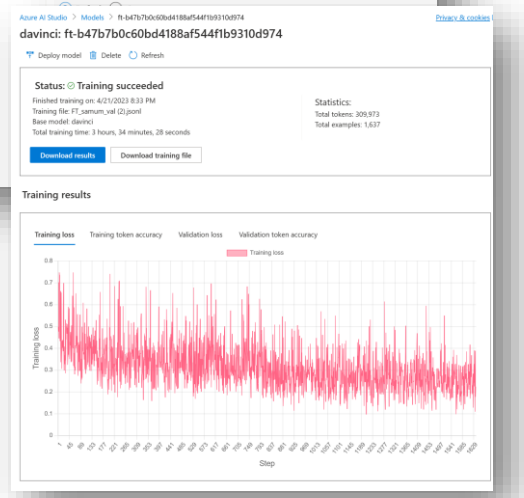
Azure AI Studio

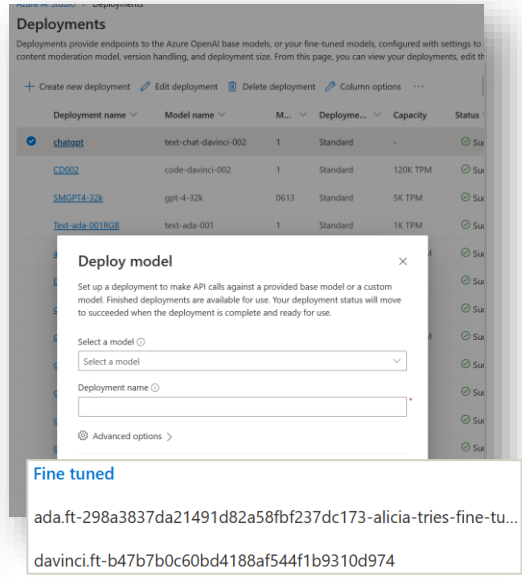






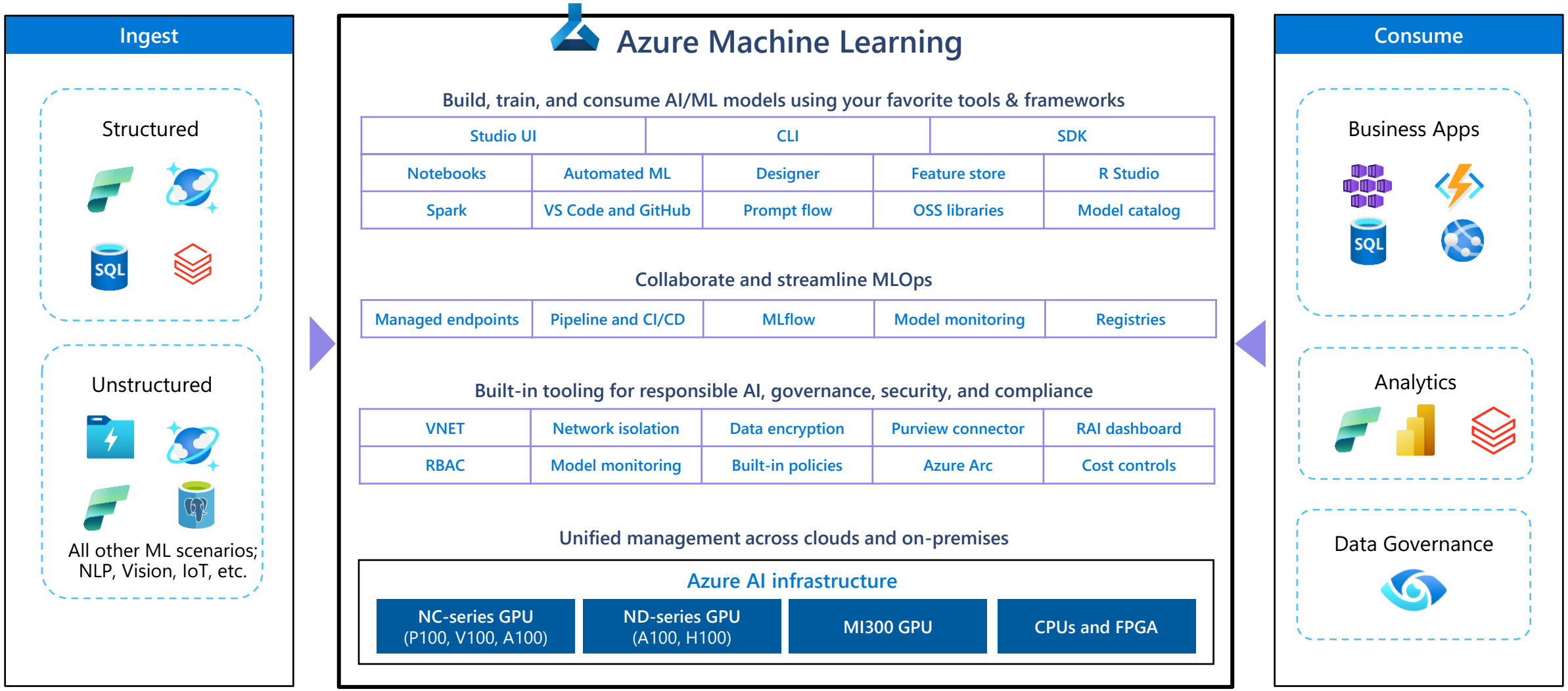






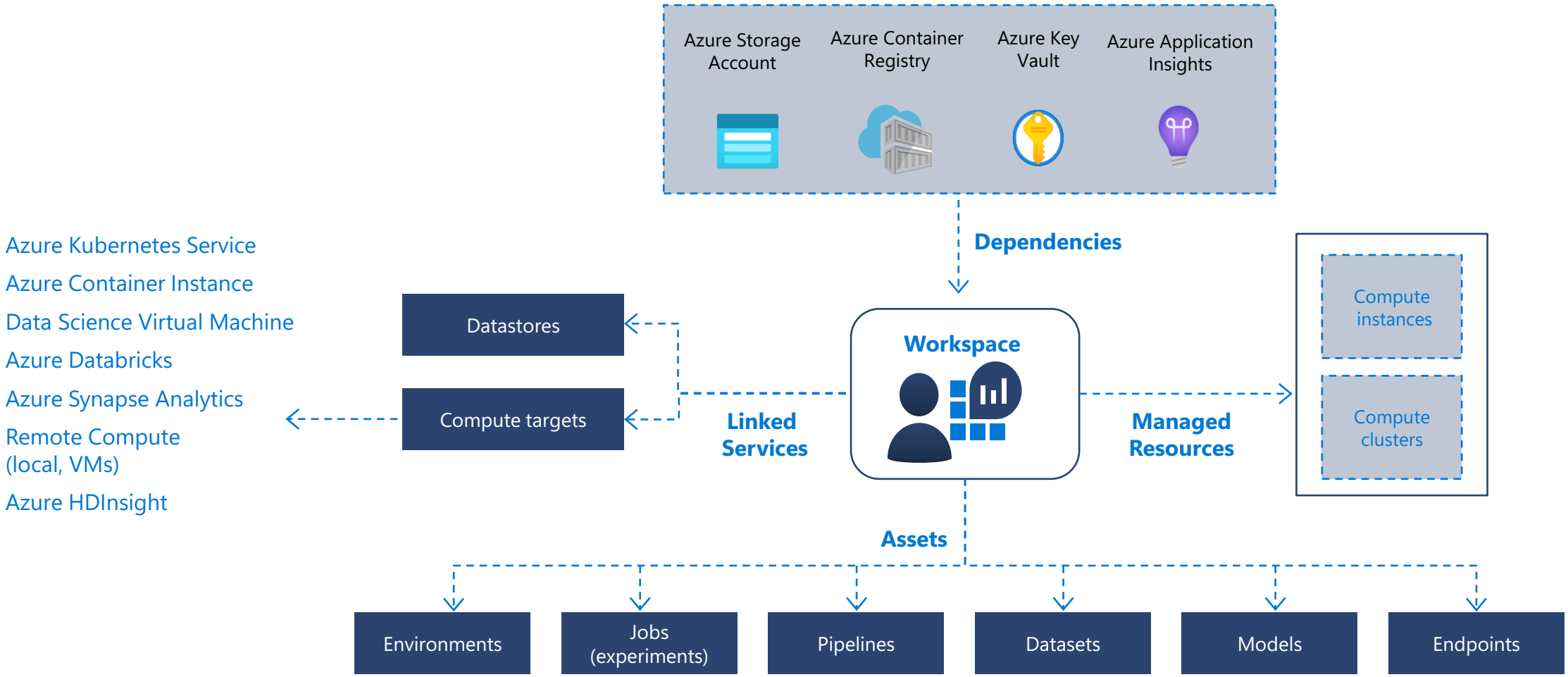
Azure ML での完全にカスタマイズ可能な Fine-Tuning

- ベースモデル、Fine-Tuningメソッド、Training/Inferenceパイプライン、リソースの最適化とスケーリングに最高の柔軟性を提供しています



Azure ML: End-to-end ML platform for ML professionals

Key Elements of Azure ML



Compute リソース: Azure AI VM ポートフォリオ

Azure Instance →	NCasT4_v3	NC A100 v4	NC H100 v5	NDm A100 v5	ND H100 v5
Cores	4, 8, 16, 64	24, 48, 96	40, 80	96	96
GPU	1-4x Tesla T4	1-4x A100 GPU, PCIe	1-2x H100NVL GPU, PCIe	8x A100 GPU, SXM	8x H100 GPU, SXM
Memory	28/56/110/440 GB	220/440/880 GB	320/640 GB	1900 GB	2048GB
Local Disk	180/360/2880 GB SSD	1123/2246/4492 GB SSD	4/8 TB SSD	6.4 TB SSD	36 TB SSD
Network	Azure Network	Azure Network + NVLink GPU Interconnect (pair)	Azure Network + NVLink GPU Interconnect (pair)	Azure Network + InfiniBand EDR + NVLink GPU Interconnect	Azure Network + InfiniBand NDR + NVLink GPU Interconnect
	NCv3 T4 1-4x T4	**NC A100v4** 1-4x A100 80GB	**NC H100v5** 1-2x H100 96GB	**NDm A100v4** 8x A100 80GB	**ND H100v5** 8x H100 80GB

Microsoft

Thank you