



# 生成AIをエッジAIで使うための勘所

東京エレクトロン デバイス株式会社

クラウドIoTカンパニー

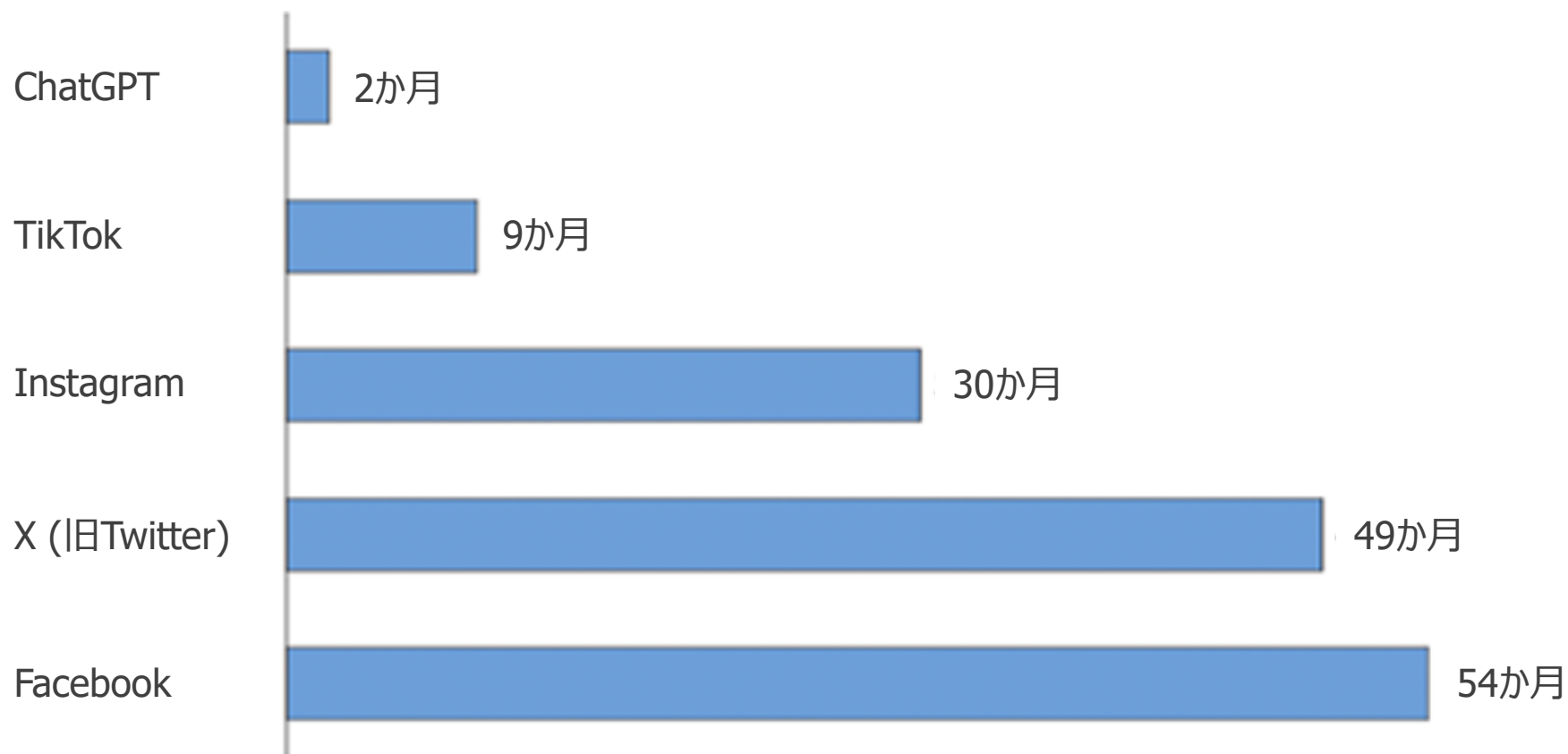
エッジクラウドソリューション部

2024年12月

## 想像を超える速度で、生成AIは社会に浸透。

2022年11月に、OpenAI社がChatGPT（GPT 3.5ベース）をリリースしてから、まだ2年。

図表 I-3-1-2 各種サービスにおける1億ユーザー達成までにかかった期間



(出典) Reuters等を基に作成

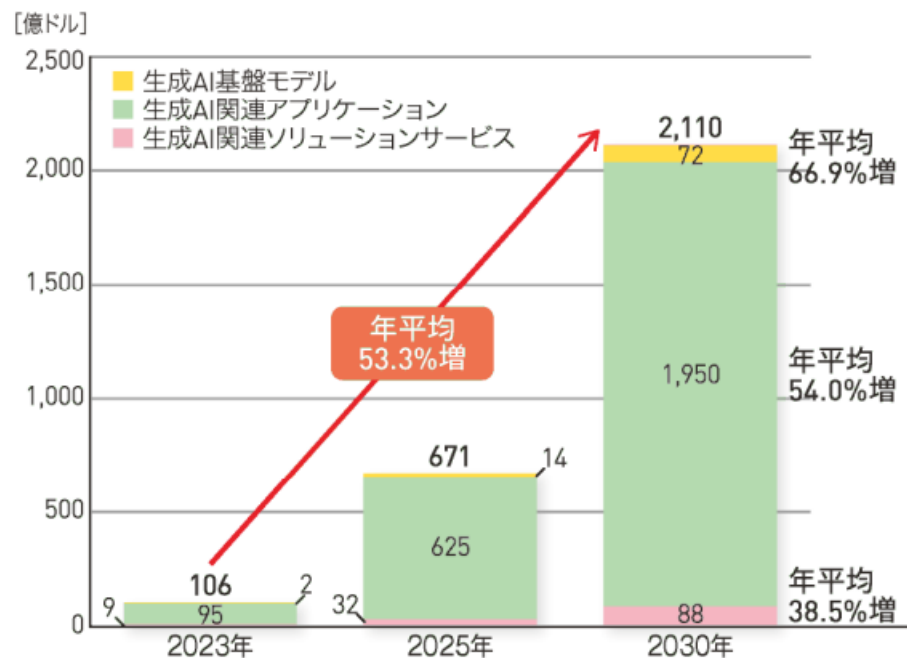
情報通信白書令和6年版より抜粋

# 製造業にも大きな影響を与える生成AI

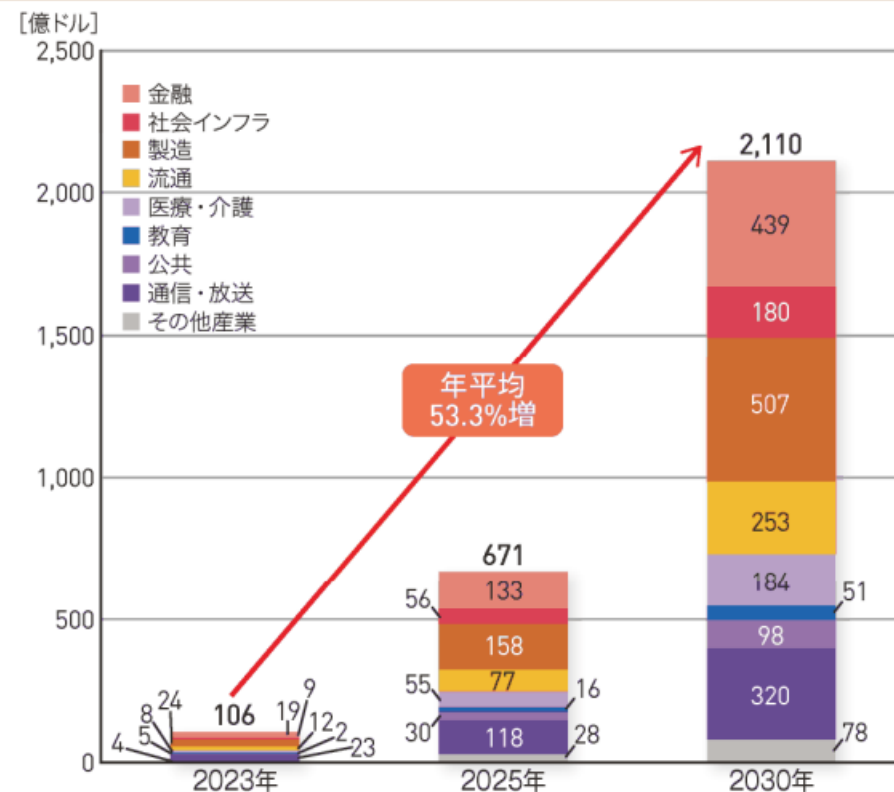
生成AI市場は今後も爆発的に成長。生成AI関連アプリケーションに大きな需要。

製造業は、生成AIが最も活用される産業となる。

生成AI市場の需要額見通し(世界)



利活用分野別 需要額見通し(世界)



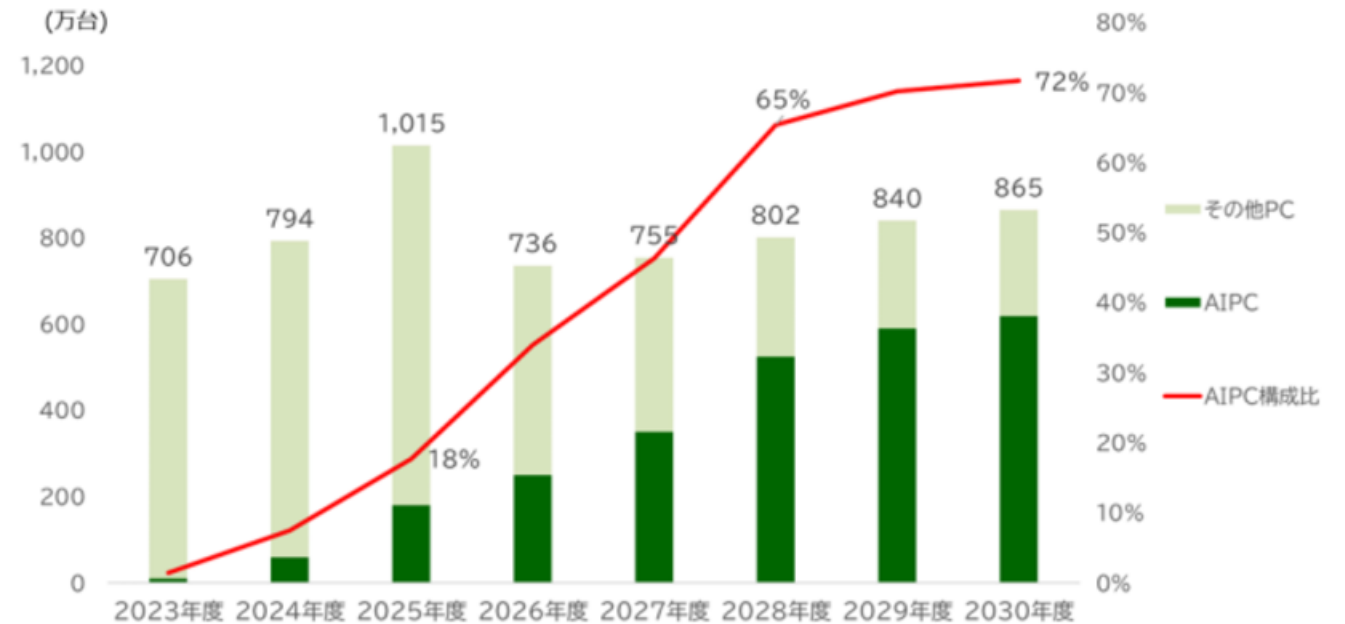
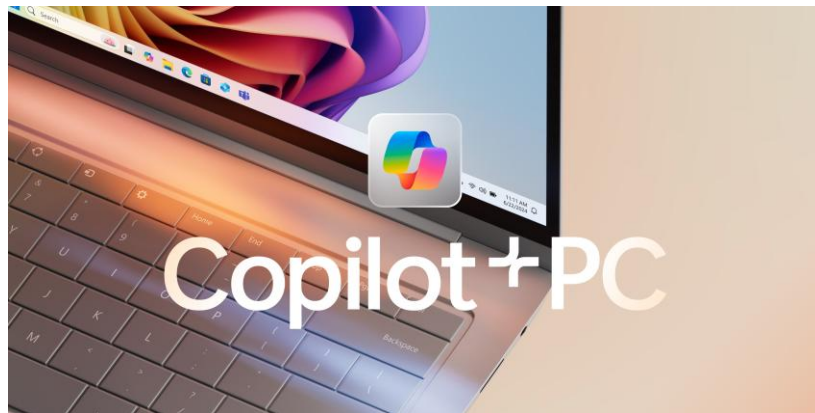
2023年12月21日「JEITA、生成AI市場の世界需要額見通しを発表」一般社団法人電子情報技術産業協会

## Copilot+ PC の登場により、エッジ分野に生成AIの浸透が加速

2030年度には、国内法人PC市場の72%がAI PCに

マイクロソフトが提供するWindows + Phiシリーズ（Phi Shilica）が、**最も使われるエッジでの生成AI構成**に

【データ1】日本のAIパソコン 法人市場規模予測

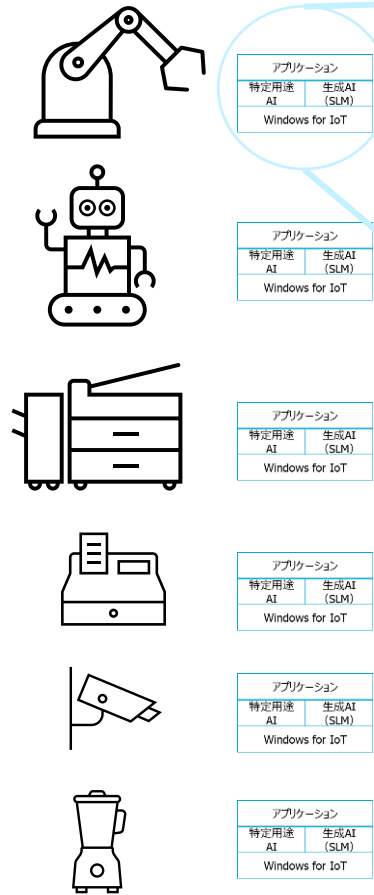


2024年07月01日 [今後5年間で法人市場にAIパソコンが普及](#) < プレスリリース | 株式会社MM総研

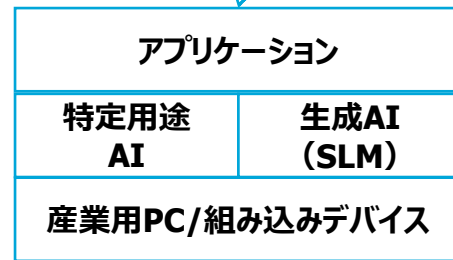
# エッジでの生成AI SLMとは

エッジAIを活用することで、これまでよりもインテリジェントな製品を開発することが可能です。  
**生成AIにより、創造性や柔軟性を持った多様な処理能力**を活用できる時代に入りました。

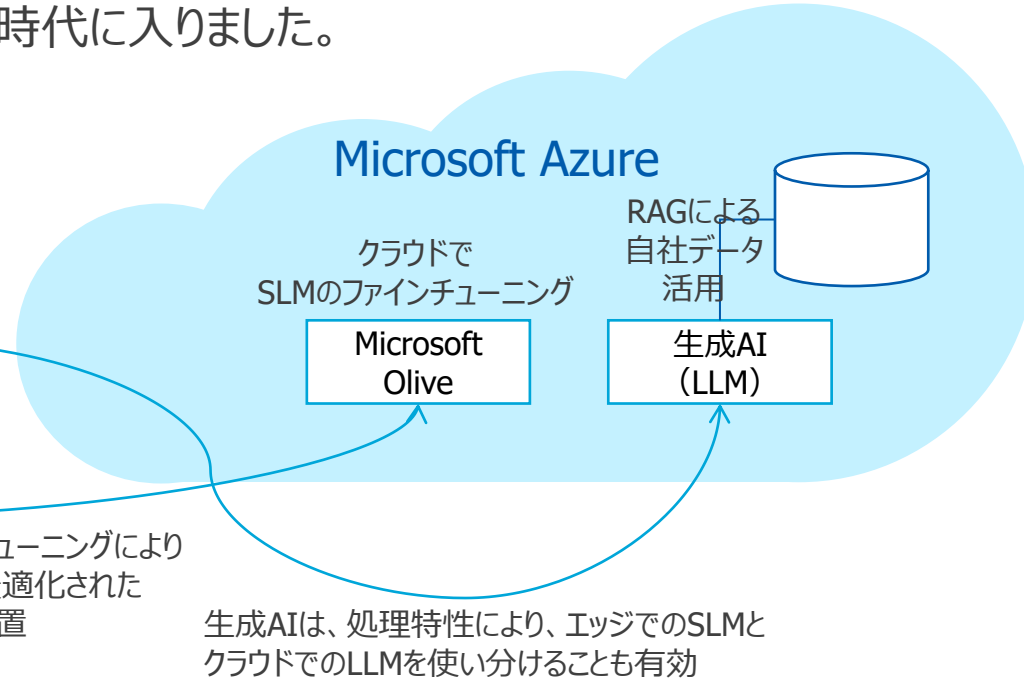
エッジAIを  
内蔵あるいは連携した  
インテリジェント製品



- エッジAIの特徴**
- ・ オフライン利用
  - ・ プライバシーデータのローカル処理
  - ・ 低遅延
  - ・ 帯域幅の節約



**エッジでの特定用途AI・生成AIを組み合わせた  
アプリケーションで付加価値提供**

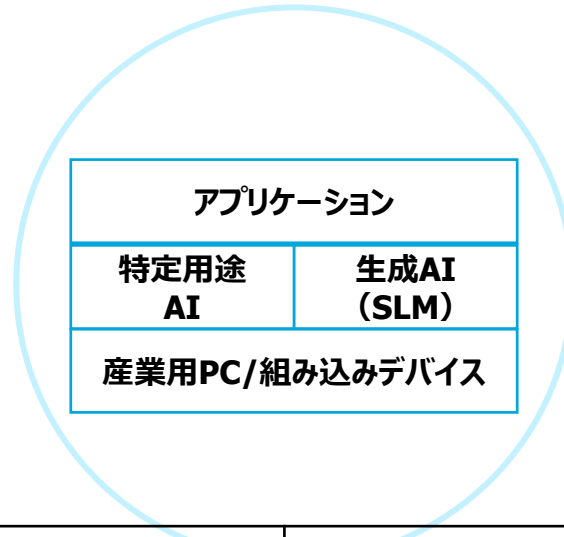


ファインチューニングにより  
業務に最適化された  
SLMを配置

SLM（小規模言語モデル）	LLM（大規模言語モデル）
SLM は、より小さいデータセットを使ってトレーニングされます。パラメーターの数は LLM よりも少なくなります。	LLM は、広い範囲の一般的な膨大な量のテキストでトレーニングされます。トレーニング時、LLM には数十億（または数兆もの）パラメーターがあります。
SLM のサイズが小さいほど、デバイスやオンプレミス コンピューターへのローカル展開など、展開のオプションが増える。	サイズが大きいため、パフォーマンスに影響し、デバイスやコンピューターにローカルに展開することは困難で、クラウドでの実行が一般的。

# エッジでの特定用途AI・生成AIを組み合わせたアプリケーションで付加価値提供

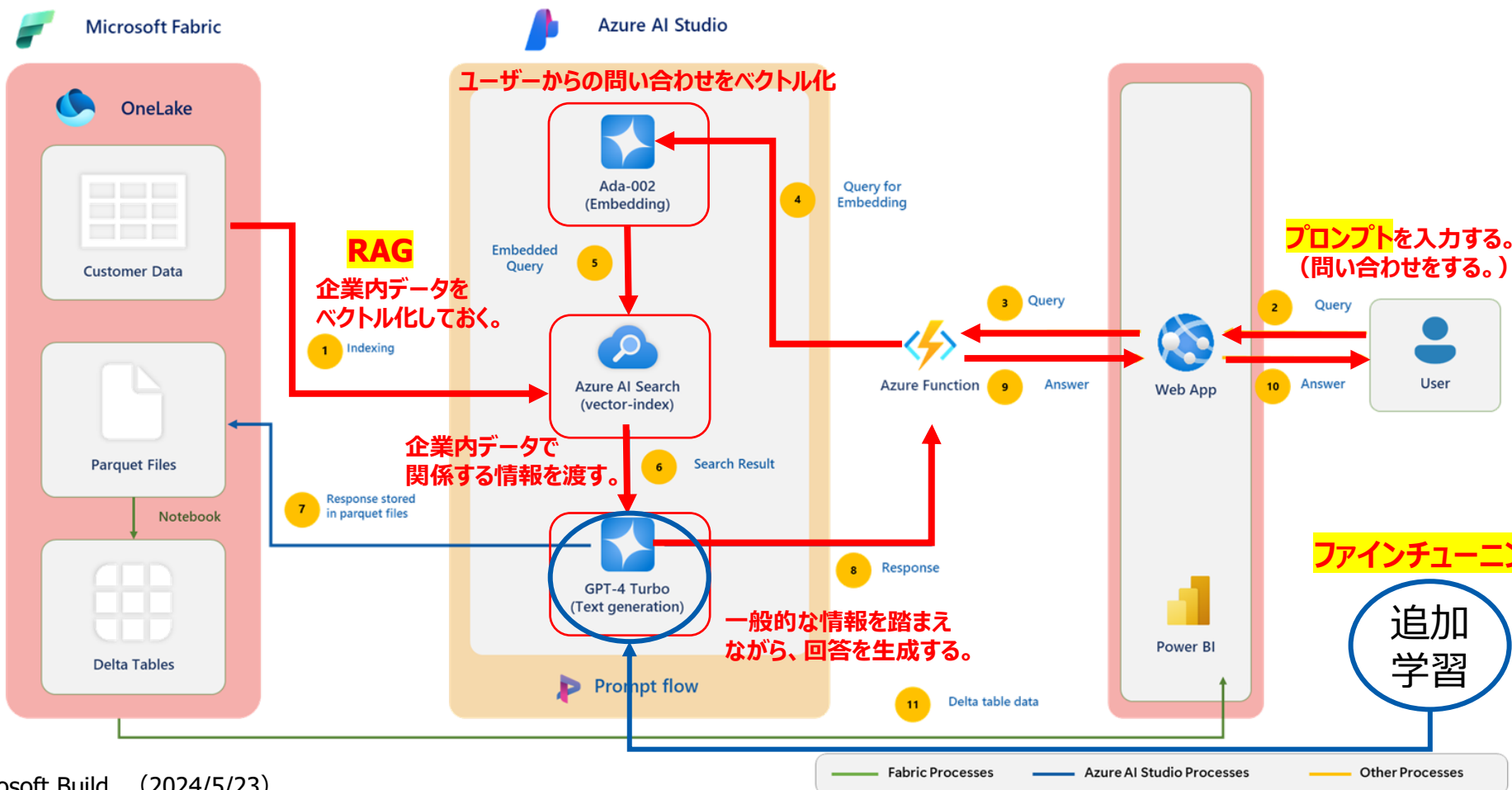
特定用途AIと生成AIにはそれぞれの特徴があり、組み合わせて活用することも可能です。  
生成AIは急速に進化し、対応領域を広めています。



	特定用途AI	生成AI (SLM)
主な用途	物体検出、予測、分類など	テキスト生成、会話、要約など
メリット	高精度、低遅延	柔軟性、創造性
デメリット	汎用性が低い、タスク限定	計算資源の制約、業務に応じたファインチューニングが必要
応答速度	高速	処理内容、HWスペックにより応答速度は異なり特性理解が必要
ユースケース例	製造業の品質管理・故障予知、医療画像診断	コンテンツ生成、ナレッジ支援、音声ガイド
	スマートファクトリー、医療アシスタント、接客ロボット	

# 独自データを含めた生成AIシステムを構築する手法

エッジでの独自データを含めた生成システムでは、エッジでのリソース制約から、ファインチューニングは有力な手段となります。エッジにおいても小規模なRAGは活用出来ます。また、プロンプトの工夫（プロンプトエンジニアリング）は重要な要素となります。



**プロンプト**とは  
プロンプトは、言語モデルにタスクを実行するよう指示する自然言語命令です。適切なプロンプトを入力しないと、生成AIは、期待される回答を返すことが出来ません。

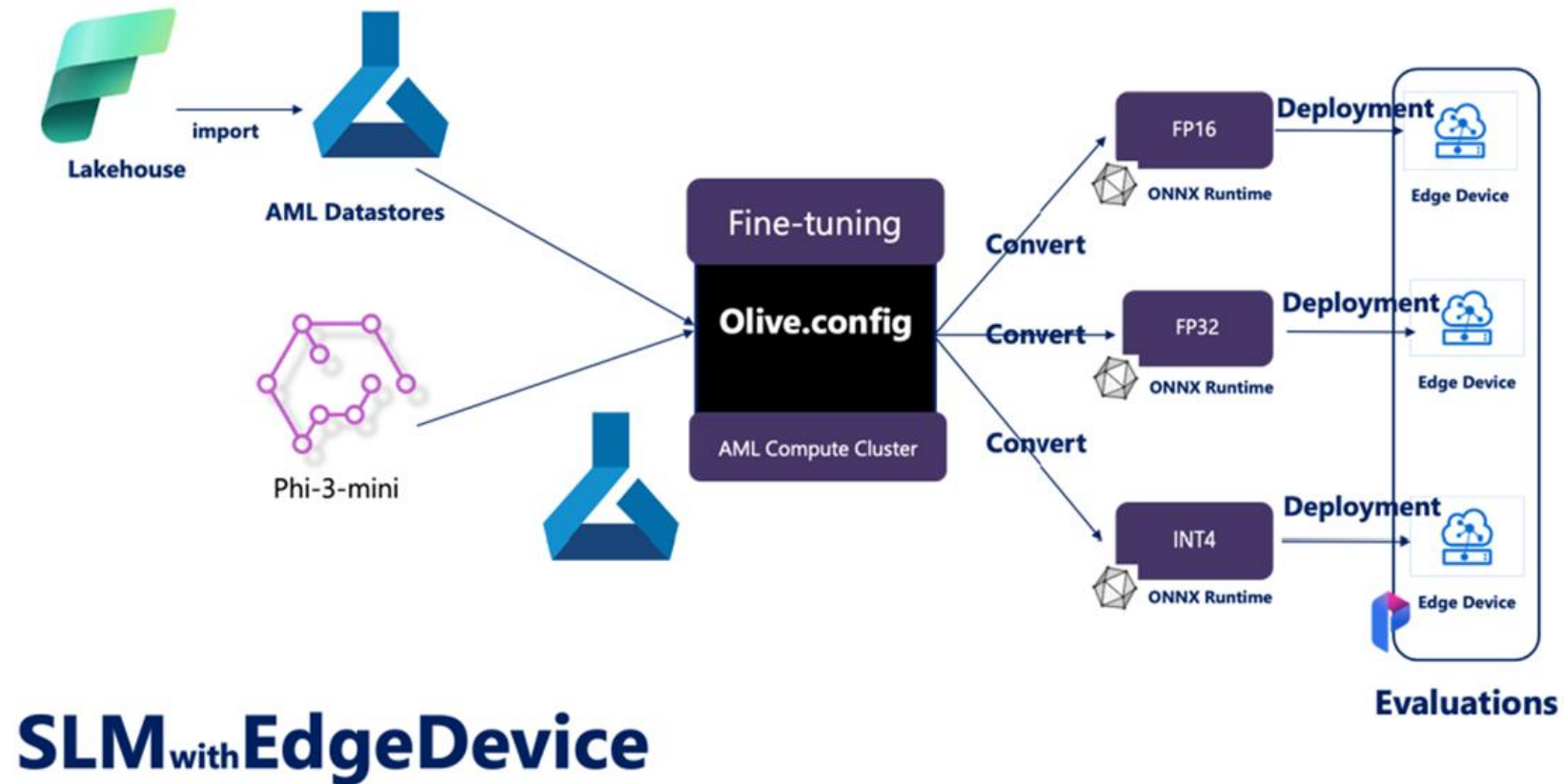
**RAG**とは  
(Retrieval Augmented Generation: 取得拡張生成)  
RAGは、独自データを提供する情報取得システムを追加することで、言語モデルの機能を拡張するアーキテクチャです。

**ファインチューニング**  
**追加学習**  
追加学習とは、専門データセットで生成AIを追加的にトレーニングすることで、より正確で関連性の高い応答を導くアーキテクチャです。大規模な計算リソースが必要。



# ファインチューニングしたSLMのエッジ展開の最適環境

ファインチューニングから展開まで、マイクロソフトのフレームワーク（**Microsoft Olive**）に基づいて運用が可能  
Microsoft OliveをSLMのファインチューニングを行い、エッジプラットフォームの種類に応じて、モデルの最適化を行います。



[Phi-3CookBook/md/04.Fine-tuning/FineTuning\\_MicrosoftOlive.md at main · microsoft/Phi-3CookBook · GitHub](#)



## 医療機器



プライバシーデータのローカル端末での処理  
診断結果のレポート自動生成  
検査から診断までのリードタイム短縮

## 店舗端末



自然言語での会話を通じた接客・お薦めの案内

- お客様の好みに応じたお薦めの提示
- 食材に関する知識の披露

## 射出成形機における生成AIを使ったエッジAI デモ

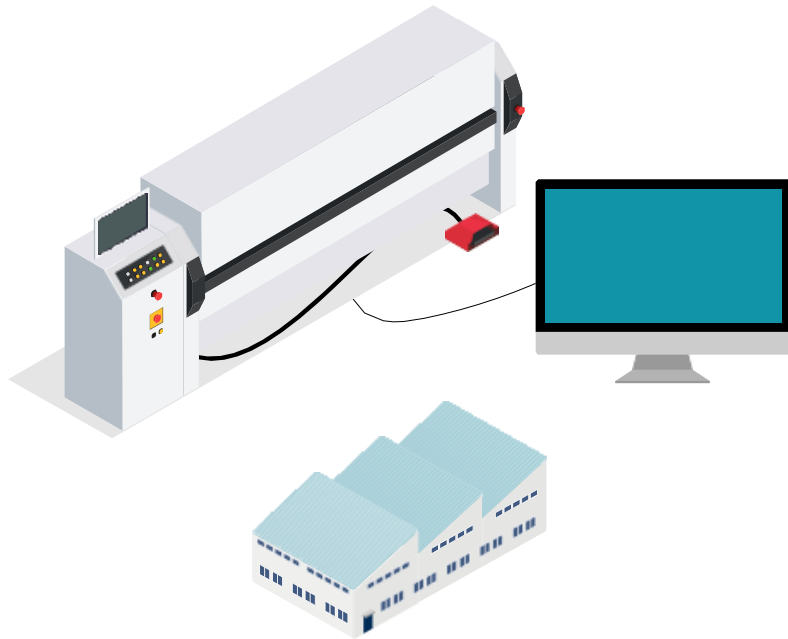
射出成形機に付属したアプリケーションでは、従来より、機械の制御や基本的な統計データの確認を行える仕組みが提供されていることを想定。

エッジAIとして生成AI（SLM）を活用し、新たな付加価値を提供する。

**生成AI（SLM）が以下の対応を行います。**

**デモシナリオ 1：柔軟性に富んだ保全レポート作成**

**デモシナリオ 2：保全作業の支援**



# デモで使用したミニPC

## エッジPC NUC 14 Pro+ NUC14RVSU7 (ASUS)

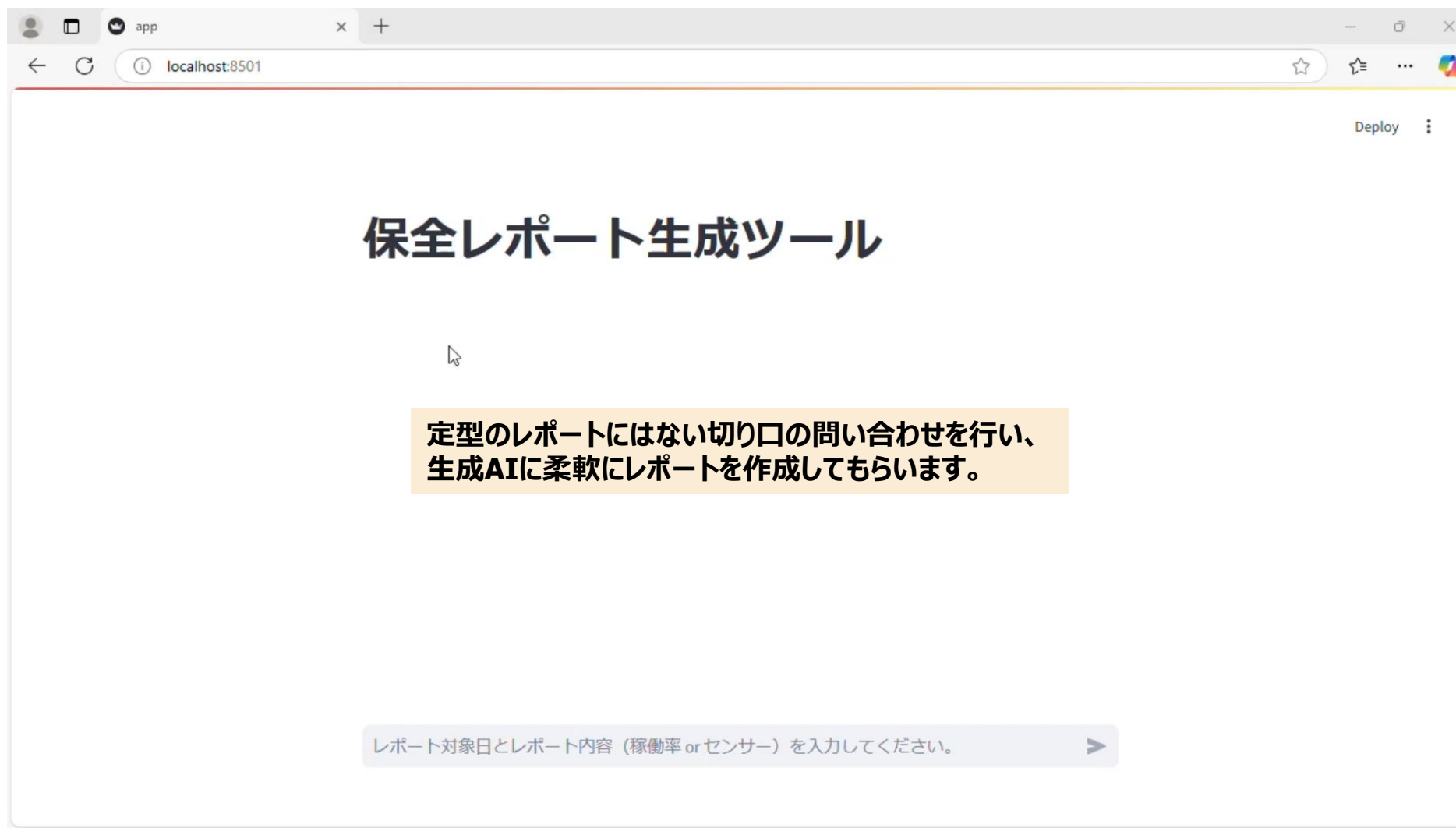
AI 対応: インテル® Core™ Ultra 7 プロセッサの能力を活用して、  
生成的な AI ワークロードをローカルで実行します。



## Phi 3 mini on Windows 11 IoT Enterprise

# デモシナリオ 1 動画：柔軟性に富んだ保全レポート作成

## 自然言語に対応したレポート作成



# デモシナリオ 1 動画：柔軟性に富んだ保全レポート作成

## 自然言語に対応したレポート作成



自然言語で、「2日前のセンサー」と問い合わせます。

2日前のセンサー



異常あり との回答が返ってきました。

# デモシナリオ 1 動画：柔軟性に富んだ保全レポート作成

## 自然言語に対応したレポート作成

更に、主要指標の推移をグラフで表示してくれています。視覚的に閾値との差異の発生を確認出来ます。

レポート対象日とレポート内容（稼働率 or センサー）を入力

異常のあった時間帯に関しては、数値の詳細レポートを報告してくれます。

### 詳細

- <MoldTemperature(Celsius)> : 異常なし
- <ClosingForce(N)> : 異常なし
- <TorqueMean(Nm)> : 2件異常あり

判定理由：センサー閾値 (80.0) を下回っています。

- 2024/10/23 09:00 - 76.5
- 2024/10/23 15:00 - 65.5

- <ShotVolume(cc)> : 異常なし

レポート対象日とレポート内容（稼働率 or センサー）を入力してください

# デモシナリオ 1 動画：柔軟性に富んだ保全レポート作成

## 自然言語に対応したレポート作成

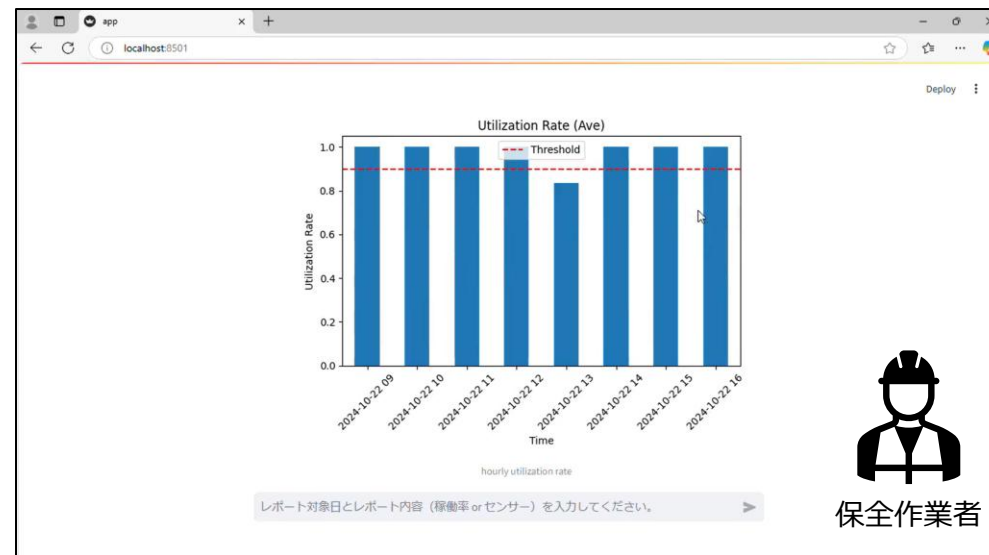
### 概要

- エッジデバイス内のIoT Edgeモジュールで実行されるアプリ
- 産業機器の**保全レポート**を生成する
- 保全作業者がエッジデバイスで**オフライン実行**する
- モジュール内の**SLM (Phi3-mini)** を利用
- チャットGUI上でユーザーの自然言語での**あいまいな指示でも、SLMが日付とレポート項目を読み取る**
- **生成AIがグラフ出力コードを生成**し、そのコードで出力されたグラフがレポート内容に加わる
- **モジュールの開発/管理/デプロイはAzure上で完結**

### 使用データ

#### [Machine Learning for Quality Prediction in Plastic Injection Molding](#)

- エッジデバイス内に本アプリとは別に存在する、射出成型機の保全アプリが取得するデータを想定
- 保全アプリがこのデータから、さらに保全集計データ（センサー/稼働率、日毎/時間毎）を出力することを想定
- オリジナル：日付、時刻、射出成型のセンサーデータ値、サイクルタイム、品質ラベルの列がある
- カスタマイズ：修理時間、総生産時間、稼働率の列を追加



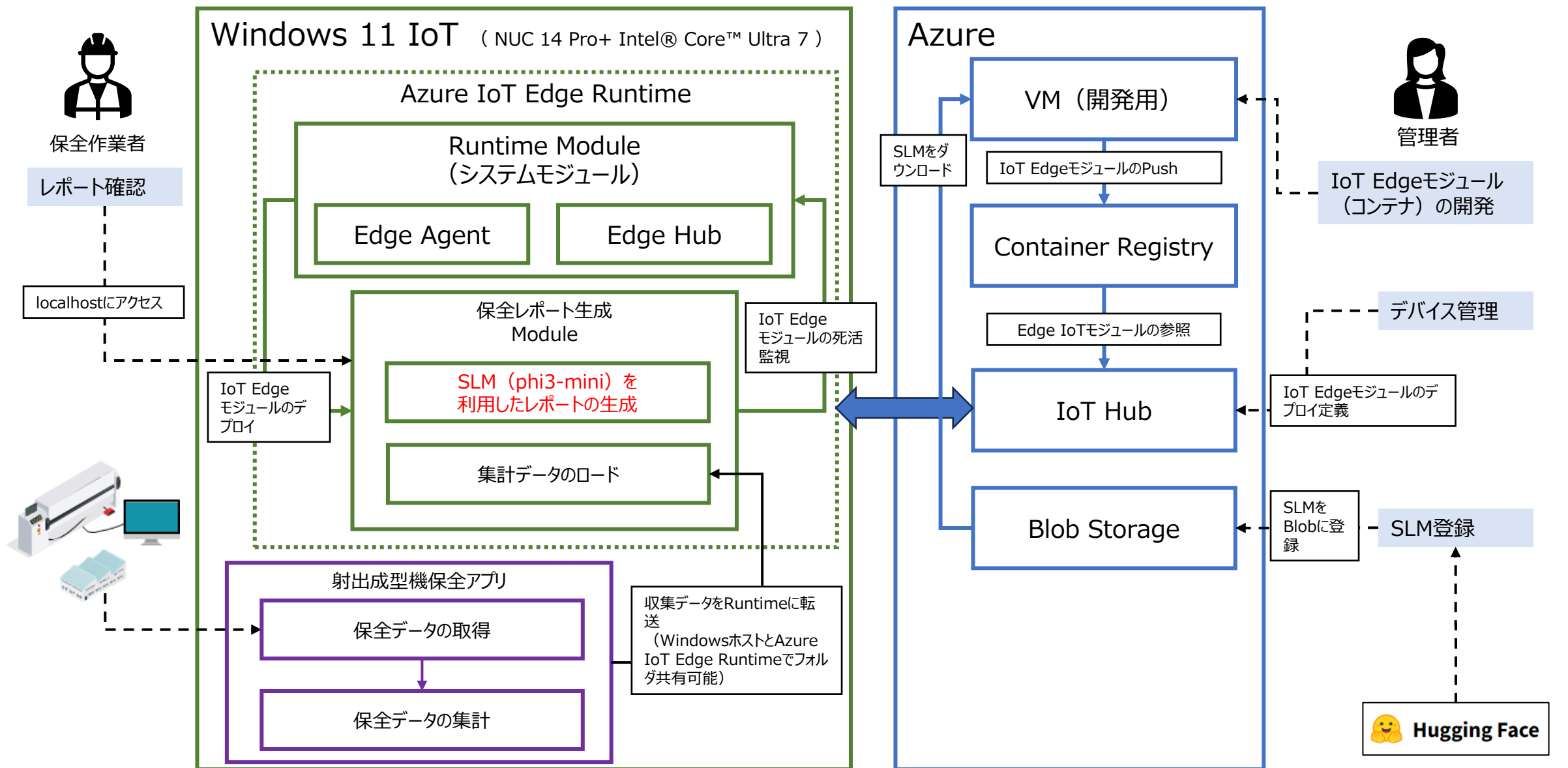
保全作業者

Date	Time	Datetime	Mold Temperature (Celsius)	Closing Force(N)	Torque Mean(Nm)	Shot Volume(cc)	Cycle Time(sec)	Maintenance Time(s)	Total Production Time	Utilization Rate	Quality
2024/10/21	9:00	2024-10-21 09:00	80.0	895.2	76.5	18.7	75	0	75	1.00	Target
2024/10/21	9:01	2024-10-21 09:01	78.9	890.4	105.8	19.1	75	0	75	1.00	Target
2024/10/21	9:02	2024-10-21 09:02	80.7	893.5	81.1	19.0	76	0	76	1.00	Target
2024/10/21	9:03	2024-10-21 09:03	80.7	896.6	82.6	18.9	76	0	76	1.00	Target
2024/10/21	9:05	2024-10-21 09:05	81.0	894.8	89.3	19.1	76	0	76	1.00	Target
2024/10/21	9:06	2024-10-21 09:06	80.9	895.0	87.8	19.0	76	0	76	1.00	Target
2024/10/21	9:07	2024-10-21 09:07	81.0	901.2	87.1	18.9	76	0	76	1.00	Target
2024/10/21	9:08	2024-10-21 09:08	80.9	899.4	90.1	19.1	76	0	76	1.00	Target
2024/10/21	9:10	2024-10-21 09:10	80.9	899.6	89.7	19.1	76	0	76	1.00	Target
2024/10/21	9:11	2024-10-21 09:11	80.9	902.0	88.7	19.0	76	0	76	1.00	Target
2024/10/21	9:12	2024-10-21 09:12	81.0	899.4	93.5	19.1	76	0	76	1.00	Target
2024/10/21	9:13	2024-10-21 09:13	80.8	900.7	87.7	19.0	76	0	76	1.00	Target
2024/10/21	9:15	2024-10-21 09:15	81.0	901.1	93.1	19.0	76	0	76	1.00	Target
2024/10/21	9:16	2024-10-21 09:16	81.0	900.1	95.6	19.0	76	0	76	1.00	Target
2024/10/21	9:17	2024-10-21 09:17	81.0	898.9	91.8	18.9	76	0	76	1.00	Target
2024/10/21	9:18	2024-10-21 09:18	70.6	886.9	104.3	18.7	0	75	75	0.00	Waste
2024/10/21	9:20	2024-10-21 09:20	70.6	886.9	104.3	18.7	0	75	75	0.00	Waste
2024/10/21	9:21	2024-10-21 09:21	70.6	886.9	104.3	18.7	0	75	75	0.00	Waste
2024/10/21	9:22	2024-10-21 09:22	70.6	886.9	104.3	18.7	0	75	75	0.00	Waste
2024/10/21	9:23	2024-10-21 09:23	70.6	886.9	104.3	18.7	0	75	75	0.00	Waste
2024/10/21	9:25	2024-10-21 09:25	70.6	886.9	104.3	18.7	0	75	75	0.00	Waste
2024/10/21	9:26	2024-10-21 09:26	70.6	886.9	104.3	18.7	0	75	75	0.00	Waste
2024/10/21	9:27	2024-10-21 09:27	70.6	886.9	104.3	18.7	0	75	75	0.00	Waste
2024/10/21	9:28	2024-10-21 09:28	70.6	886.9	104.3	18.7	0	75	75	0.00	Waste
2024/10/21	9:30	2024-10-21 09:30	70.6	886.9	104.3	18.7	0	75	75	0.00	Waste
2024/10/21	9:31	2024-10-21 09:31	70.6	886.9	104.3	18.7	0	75	75	0.00	Waste
2024/10/21	9:32	2024-10-21 09:32	70.6	886.9	104.3	18.7	0	75	75	0.00	Waste



# デモシナリオ 1 動画：柔軟性に富んだ保全レポート作成

## 自然言語に対応したレポート作成 アーキテクチャ図



# デモシナリオ 2 動画：保全作業の支援 ファインチューニングの効果



設備保全 Q&Aチャット

☆☆☆何でも聞いてくださいね☆☆☆

質問する

トレーニング前のモデル  
(microsoft/Phi-3-mini-4k-instruct)

トレーニング後のモデル  
(microsoft/Phi-3-mini-4k-instruct)

同じ質問をファインチューニング前後の生成AIの回答を比較します。

# デモシナリオ 2 動画：保全作業の支援 ファインチューニングの効果

☆☆☆何でも聞いてくださいね☆☆☆

「エア圧不足検出」で現在の課題を入力

エア圧不足検出

質問する

### トレーニング前のモデル (microsoft/Phi-3-mini-4k-instruct)

エア圧不足検出

エア圧不足検出は、空気の圧力が予想値より低かった状態を指します。これは、気象条件や機材の問題、または温度変化によるものかもしれません。エア圧温度計や圧力計を使用して、実際の空気圧を測定し、予測値との差異を分析することで、エア圧不足を検出することができます。この問題を解決するためには、気象条件の再評価、機材のメンテナンス、または温度変化の影響を考慮に入れる必要があります。

ファインチューニング前だと、気象条件など一般情報から、回答に関係のない情報も追加してしまっている。

### トレーニング後のモデル (microsoft/Phi-3-mini-4k-instruct)

エア圧不足検出

原因は『エア圧が不足しています。』です。対処方法は『エア供給元とエア圧計を確認し、適切なエア圧を確保します。』です。

ファインチューニング後は、今回の設備保全に関する回答のみを行っている。

# デモシナリオ 2 動画：保全作業の支援 ファインチューニングの効果

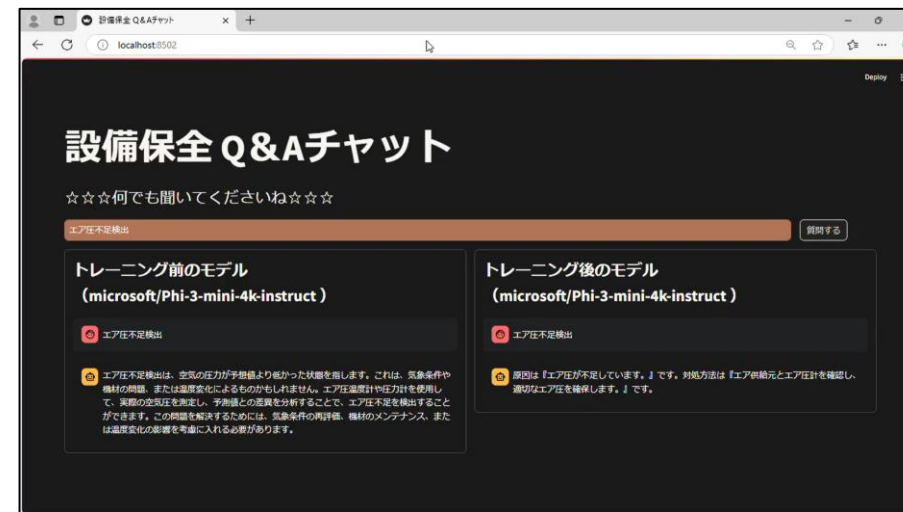
## 概要

- エッジデバイス内のIoT Edgeモジュールで実行されるアプリ
- 産業機器に関するマニュアルのQAチャットボット
- 保全作業者がエッジデバイスでオフライン実行する
- モジュール内の学習済みのSLM (Phi3-mini) を利用
- 学習により、SLMが生産現場のドメイン知識を元に回答を生成
- SLM学習、モジュールの開発/管理/デプロイはAzure上で完結

## 使用データ

トラブルシューティングガイド (.pdf)

産業機器のトラブルシューティングガイドより学習データを作成  
アラームNo.、エラーメッセージ、原因、対処方法の列がある  
このドキュメントから.json形式に変換する (自動変換可能)



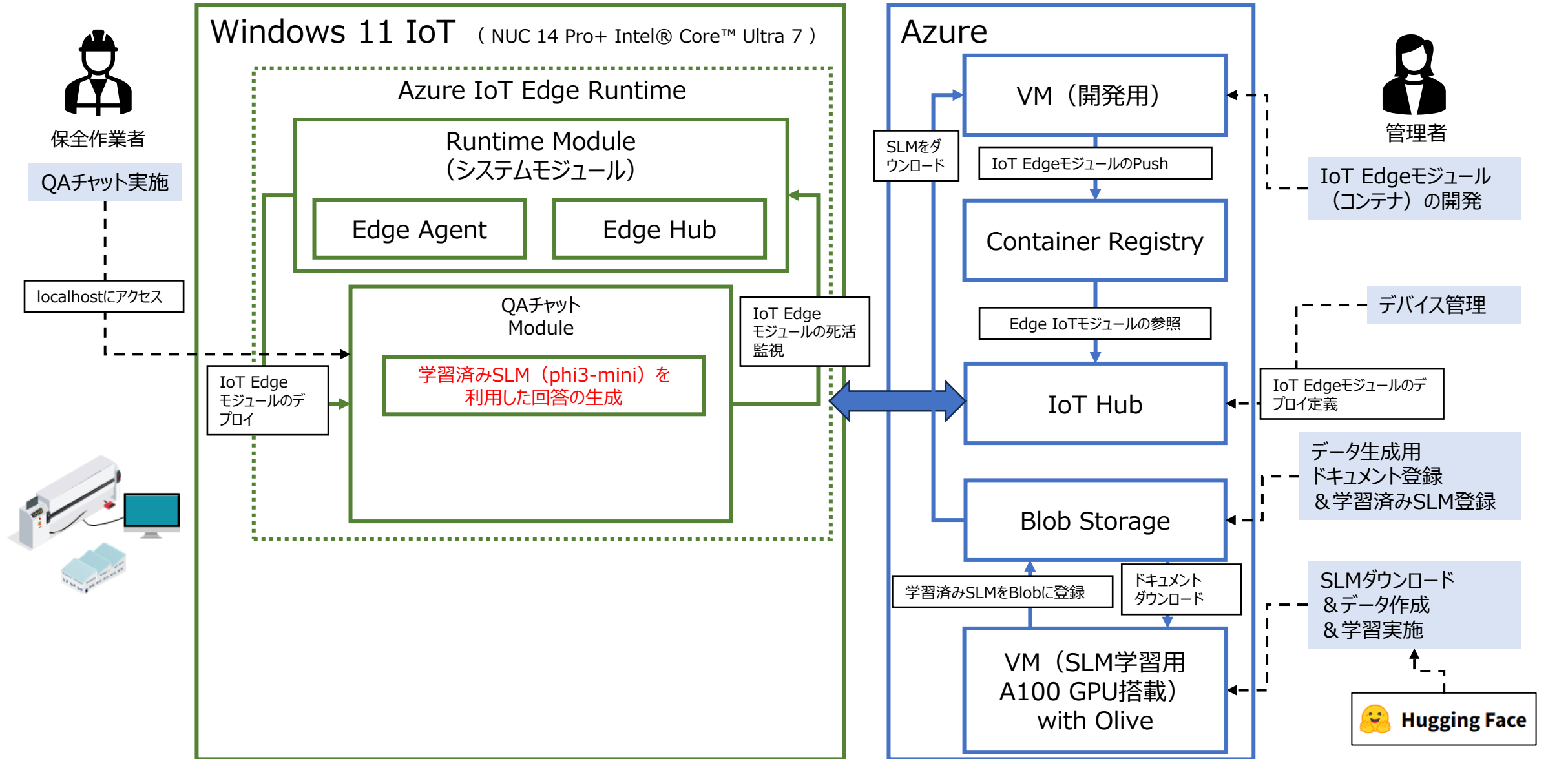
## トラブルシューティングガイド

対象機種：TED2000c

更新日：2024/6/1

アラーム No	エラーメッセージ	原因	対処方法
001	非常停止が押されています	非常停止ボタンが押されています。	搭載中の基板を手動で取り除き、非常停止ボタンを解除します。その後、原点復帰を行い、搭載前の基板を基板搬入センサーの上に配置し、生産開始ボタンを押します。
002	原点復帰が行われていません	原点復帰が行われていません。	原点復帰を手動で実行し、正常に完了するまで待ちます。
003	部品切れ	指定されたフィーダ番号のフィーダに部品が不足しています。	フィーダに部品を補充し、正常に供給されているか確認します。

# デモシナリオ 2 動画：保全作業の支援 ファインチューニングの効果 アーキテクチャー図





共に創る 新たな価値を



東京エレクトロン デバイス株式会社

クラウドIoTカンパニー エッジクラウドソリューション部

<https://esg.teldevice.co.jp/iot/azure/>



東京エレクトロンデバイス  
IoTソリューションポータルサイト