



# Microsoftにおける生成AIの取り組みと HPC基盤技術

日本マイクロソフト株式会社

松崎 剛  
五十木 秀一



# Microsoft AI で注力する投資分野

お客様が得られる効果を最大化



## 研究開発 & 技術開発

生成 AI, アクセシビリティ,  
生産性の拡大



## 戦略的 パートナーシップ

製品展開と改革の  
加速化



## 責任ある AI の 実践

デジタルトラストの構築  
とユーザー体験の向上

# 変革・革新のためのパートナーシップ



OpenAI forms exclusive computing partnership with Microsoft to build new Azure AI supercomputing technologies

July 22, 2019 | Microsoft News Center



## Hugging Face Collaborates with Microsoft to Launch Hugging Face Endpoints on Azure

Published May 24, 2022.



## Meta selects Azure as strategic cloud provider to advance AI innovation and deepen PyTorch collaboration

Posted on May 25, 2022

[Eric Boyd](#), Corporate Vice President, AI Platform



## Azure Scales 530B Parameter GPT-3 Model with NVIDIA NeMo Megatron

Posted on October 24, 2022

[Rachel Pruitt](#), Product Marketing Manager, Azure Marketing, HPC + AI



# Microsoft の生成 AI における沿革



# Azure OpenAI Service

大規模な事前学習済みモデルと、お客様データによる  
ファインチューニングによりカスタムが可能に

## Top capabilities



テキスト自動生成や  
コンテンツ制作の支援



構造化・非構造化  
データを利用



テキスト要約、Q&A  
フィードバック



チャットボット  
AIアシスタント

Text

Conversation

Code

Images

顧客の問い合わせを以下の例にそって分類してください。

舞黒ダイニングを 3/15 に 4 名で、和食のコースで予約お願いします。子供が 1 名居ます。 : レストラン予約

2月10日 19:20 成田発、シアトル行きののユナイテッド航空 616 便をエコノミークラスで6枚ください。 : 航空機の手配

オペレーターにつないでいただけませんか。 : その他

全日空の羽田発、ロサンゼルス行きで、10/20 で手配できますか。 :  
航空機の手配

Text

Conversation

Code

Images

顧客の問い合わせを以下の例にそって分類してください。

舞  
ン  
2月  
い  
オ  
全  
航

2月10日 19:20 成田発、シアトル行きなのユナイテッド航空 616 便をエコノミークラスで6枚ください。

日時: 02/10 19:20

出発: 成田

到着: シアトル

航空会社: ユナイテッド航空

クラス: エコノミー

枚数: 6

全日空の羽田発、ロサンゼルス行きで、10月20日午後3:00出発をエコノミークラスで1枚ください。

日時: 10/20 15:00

出発: 羽田

到着: ロサンゼルス

航空会社: 全日空

クラス: エコノミー

枚数: 1



**MW BZ AZ**  
**患者サポート**  
 顧客との会話の概要を提示できるため、顧客のニーズを理解するために長い会話履歴をスクロールする必要がなくなります。

**BZ AZ**  
**パターン認識**  
 トピック分析は、顧客の会話を分類して、顧客の苦情や質問の新たなパターンを特定するのに役立ちます。

**BZ AZ**  
**超-パーソナライゼーション**  
 トピック分析は、顧客との会話を分類して、顧客に代替案を提案するのに役立ちます。

**AZ**  
**ニュースとトレンド**  
 トピック分析により、最新のニュースやトレンドのWebコンテンツをマイニングして、業界ニュースの最新のイノベーショントピックを把握したり、薬物リコールなどにつながっているものを把握したりできます。

**MW BZ AZ**  
**感情分析**  
 高度な臨床言語理解は、臨床ノートと患者の会話を分析し、それらを要約して、患者の口調がより深い病気またはうつ病/不安を示しているかどうかを理解することにより、病気の予測と診断にも役立ちますか？



**BZ AZ**  
**請求処理**  
 セマンティック検索とChatGPTを通じて、質問に答え、補償範囲、控除額、償還に関する情報を患者とプロバイダーに提供することで、請求処理プロセスを自動化します。

**BZ AZ**  
**コスト見積もり**  
 コストモデルと代替治療を検索して、患者に医療サービスのコストの見積もりを提供します。

**BZ AZ**  
**環境臨床インテリジェンス**  
 ナレッジベースに基づいて基本的な質問に対する顧客への回答を自動的に生成し、より複雑なケアに優先順位を付けます。

**MW AZ**  
**臨床文書**  
 患者の答えを導き出したり、要約によって実行できる重要なパターンを特定したりするために、広範なドキュメントを作成する際の手作業を減らします。

**MW AZ**  
**代替治療**  
 さまざまな治療法間のつながりを見つけ、内部データと外部データを要約することで利用可能な代替案とその副作用を理解するのに役立つ高度な言語理解。

医療業界のメリット



BZ AZ **コンタクトセンター分析**

通話の要約作成は、顧客の苦情の傾向を特定し、簡潔で関連性のある情報を持つ人間のエージェントを関与させて、より迅速な意思決定/応答を行うために、通話ログから重要な情報を抽出します。Q&A 機能を使用して、ボットエージェントを介した通話を自動的に処理します。

BZ AZ **顧客Q&A**

ライブチャット/音声ボットアプリケーションに対して人間のような応答を生成します。

BZ AZ **超-パーソナライゼーション**

パーソナライズされた通信を自動的に生成して、顧客のニーズとパターンに基づいて最適なメッセージングを取得します

MW AZ **法規制の遵守**

エージェントが内部調査とデューデリジェンスにおける規制コンプライアンスを、支援されたQ&Aポリシーと引受ルールを通じて遵守するのを支援します。

MW BZ AZ **感情分析**

GPTモデルを使用した感情分析を通じて、顧客とのやり取りをシミュレートし、最終的な「顧客」満足度を評価するエージェントトレーニングメカニズム。

BZ AZ **コンタクトセンター分析**

通話の概要作成は、顧客の苦情の傾向を特定し、簡潔で関連性のある情報を持つ人間のエージェントを関与させて、より迅速な意思決定/応答を行うために、通話ログから重要な情報を抽出します。Q&A 機能を使用して、ボットエージェントを介した通話を自動的に処理します。



MW AZ **ドキュメントプロセスの自動化**

ドキュメントをデジタル化し、セマンティック検索またはコグニティブ検索を通じて重要な情報を抽出し、それらを要約します(たとえば、内部および外部の情報による法的文書の要約)。

MW AZ **リスクマネジメント**

認識コンテンツ(ソーシャルメディアやブログを含む)の検索と要約を通じてリスクを特定するための現在のパターンと傾向。

MW AZ **ビジネスユニット間の情報収集**

NER、複数のビジネスユニットにわたるニュースソースのセンチメント、分類、要約

BZ AZ **コンタクトセンター分析**

通話の概要作成は、顧客の苦情の傾向を特定し、簡潔で関連性のある情報を持つ人間のエージェントを関与させて、より迅速な意思決定/応答を行うために、通話ログから重要な情報を抽出します。Q&A 機能を使用して、ボットエージェントを介した通話を自動的に処理します。

AZ **市場分析**

その日の市場のダイナミクスに基づいてトップオブマインド/編集意見を自動的に生成し、要約ビューといくつかの重要なポイント分析を生成します

金融業界のメリット



AZ MW BZ SC GH

BZ AZ

### 広告とキャンペーン

広告やマーケティングキャンペーンの画像モデルを自動的に生成する

AZ

### ブランド管理

ソーシャルメディアフィードに接続し、それらを要約し、感情分析を使用して、製品/サービスの品質をよりよく理解します。

BZ AZ

### コンタクトセンター分析

通話の概要作成は、顧客の苦情の傾向を特定し、簡潔で関連性のある情報を持つ人間のエージェントを関与させて、より迅速な意思決定/応答を行うために、通話ログから重要な情報を抽出します。Q&A 機能を使用して、ボット エージェントを介した通話を自動的に処理します。

MW BZ AZ

### レポート生成

店舗マネージャーと中央マネージャーの運用管理の概要。エグゼクティブリーダーシップのレポートを作成するためのストアマネージャー/エグゼクティブへのすべてのサポートチケットの簡潔な要約

BZ AZ

### 小売業の要員管理

検索とコンテンツ生成を通じて店舗のスケジュールを自動的に管理します。

MW AZ

### ユーザー生成コンテンツ

ユーザー生成コンテンツの説明(例:製品の写真に基づく製品の説明)を生成します。コンピュータービジョンおよびコンテンツモデレーターと組み合わせて、共有に不適切なコンテンツを除外します。

消費財

バックオフィス管理の最適化

セキュリティとコンプライアンス

小売

自動化による革新

小売および消費財業界のメリット

### 社内コミュニケーション

MW BZ AZ

セマンティック検索と要約を使用したインテリジェントなQ&Aを通じて、組織全体、複数のビジネスユニット、バックエンドデータシステムからのクエリに対応します。

### 内部ヘルプデスク

BZ AZ

インテント分類、エンティティ抽出、感情分析、内部ヘルプデスクチケットの軽量要約を使用して、ルーティングを自動化し、従業員が送信したIT/HRヘルプデスクチケットの解決時間を短縮します。

### 調達管理

BZ AZ

ベンダー契約のデータから貴重な情報と洞察を抽出します。ベンダー契約を要約し、感情分析を通じて否定的なステートメントを強調します。

### 製品説明の自動化

MW AZ

マーチャンダイジングエキスペリエンスエージェントからのいくつかの重要な入力に基づいて長いアイテム説明の生成を提案し、手作業を減らします

### 自動マーケティングメール

BZ AZ

パーソナライズされたカスタマー360情報とマーケティングビデオの抽出と要約に基づいて、マーケティングメールを自動的に生成します。

### インテリジェントなオンボーディング

MW AZ

エンジニア、営業担当者等を開発プラクティス、IT戦略、ベストプラクティスを使用して社内にオンボードするための会話型チャット。

超-パーソナライゼーション

革新的なマーケティングとキャンペーン

インテリジェントコンタクトセンター

AZ MW BZ SC GH

**BZ AZ** **コンタクトセンター分析**  
通話の概要作成は、顧客の苦情の傾向を特定し、簡潔で関連性のある情報を持つ人間のエージェントを関与させて、より迅速な意思決定/応答を行うために、通話ログから重要な情報を抽出します。Q&A機能を使用して、ボットエージェントを介した通話(停止情報など)を自動的に処理します。

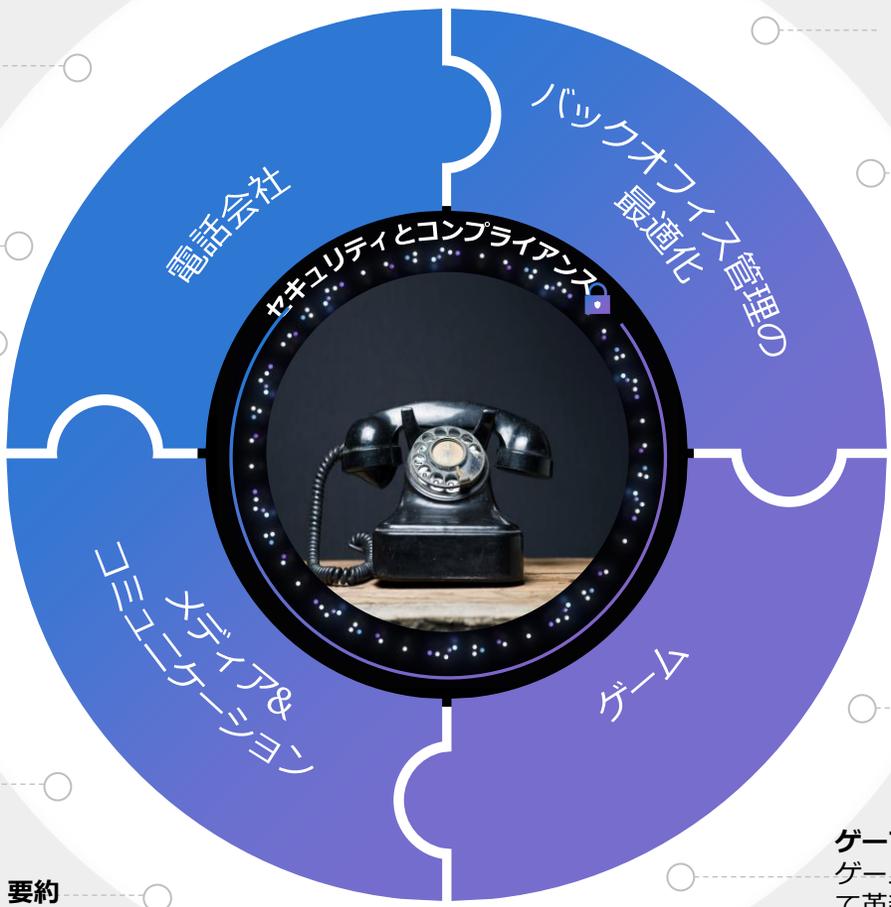
**BZ AZ** **パターン認識**  
トピック分析は、顧客の会話を分類して、顧客の苦情や質問の新たなパターンを特定するのに役立ちます。

**BZ AZ** **超-パーソナライゼーション**  
トピック分析は、顧客との会話を分類して、顧客に代替案を提案するのに役立ちます。

**MW AZ** **出版前作業の迅速化**  
編集スタイルガイドに準拠するためのコンテンツ編集の提案、説得力のある有益な要約の生成、記事のSEOに最適化された見出しの生成、ワイヤーやファクトリストからの記事の構成の支援など、公開前に作業を加速および分析します。

**MW AZ** **オーディオトランスクリプトの要約**  
エンティティ抽出を実行し、ビデオコンテンツから抽出されたオーディオトランスクリプトの概要を生成します。

**AZ** **ファンへの新しいコンテンツを自動的に生成、要約**  
要約会社のウェブサイトを通じてファンに提供される革新的なコンテンツ。ウェブサイトの新しいUIを生成することもできます。



**トレンド分析 AZ**  
ソーシャルメディアのトレンドとデータを使用して、製品やサービスに対する顧客の感情を理解し、顧客へのサービスを向上させ、新しいトレンドやパターンに基づいてコンテンツをパーソナライズします。

**社内コミュニケーション MW AZ**  
セマンティック検索と要約を使用したインテリジェントなQ&Aを通じて、組織全体、複数のビジネスユニット、バックエンドデータシステムからのクエリに対応します。

**内部ヘルプデスク BZ AZ**  
インテント分類、エンティティ抽出、感情分析、内部ヘルプデスクチケットの軽量要約を使用して、ルーティングを自動化し、従業員が送信したIT/HRヘルプデスクチケットの解決時間を短縮します。

**Copilotアプローチ GH AZ**  
Codex をゲーム開発プラットフォームに組み込み、会社の IDE 内でコード補完サービスを使用してプラットフォーム上でコードを開発するためのCopilotアプローチを提供します。

**自動マーケティングメール BZ AZ**  
パーソナライズされたカスタマー360情報とマーケティングビデオの抽出と要約に基づいて、マーケティングメールを自動的に生成します。

**ゲーマーに新しいコンテンツを自動的に生成、要約 AZ**  
ゲームプラットフォームでゲーマーに提供されるコンテンツを要約して革新的にします。

通信・メディア・ハイテク業界のメリット



AZ MW BZ SC GH

**BZ AZ** **コンタクトセンター分析**  
通話の概要作成は、顧客の苦情の傾向を特定し、簡潔で関連性のある情報を持つ人間のエージェントを関与させて、より迅速な意思決定/応答を行うために、通話ログから重要な情報を抽出します。Q&A機能を使用して、ボットエージェントを介した通話を自動的に処理します。

**BZ AZ** **パターン認識**  
トピック分析は、顧客の会話を分類して、顧客の苦情や質問、および最も効果的な消費パターンの新たなパターンを特定するのに役立ちます。

**BZ AZ** **超-パーソナライゼーション**  
トピック分析は、顧客との会話を分類して、顧客に代替案を提案するのに役立ちます。

**BZ AZ** **リモートワーカーアシスタンス**  
ビジネスユーザーおよび遠隔地の運用担当者への個人的な支援/知識プロバイダー。

**MW AZ** **電子メール応答を自動的に生成する**  
顧客のセンチメントを理解し、顧客への電子メール応答を自動的に生成します。

**BZ AZ** **設備レポート**  
テキスト要約とパターン認識により、機器の故障などのレポートを自動的に生成します。



**AZ** **トレンド分析**  
ソーシャルメディアのトレンドとデータを使用して、製品やサービスに対する顧客の感情を理解し、顧客へのサービスを向上させ、新しいトレンドやパターンに基づいてコンテンツをパーソナライズします。

**MW AZ** **社内コミュニケーション**  
セマンティック検索と要約(予防ガイドラインに関する情報など)を使用したインテリジェントなQ&Aを通じて、組織全体、複数のビジネスユニット、バックエンドデータシステムからのクエリに応答します。

**BZ AZ** **内部ヘルプデスク**  
インテント分類、エンティティ抽出、感情分析、内部ヘルプデスクチケットの軽量要約を使用して、ルーティングを自動化し、従業員が送信したIT/HRヘルプデスクチケットの解決時間を短縮します。

**MW AZ** **スタッフのオンボーディング**  
生成モデル機能を使用して、記録された会話履歴に基づいて、新しい担当者(カスタマーサービスエージェント、工場労働者など)のオンボーディングのためのトレーニング資料を作成します。

**BZ AZ** **自動マーケティングメール**  
パーソナライズされたカスタマー360情報とマーケティングビデオの抽出と要約に基づいて、マーケティングメールを自動的に生成します。

**AZ** **顧客向けに新しいコンテンツを自動的に生成、要約**  
自動的に要約また、顧客に関連する新しいコンテンツを生成して、顧客への対応に必要な手作業の量を減らします。

製造/エネルギー/石油・ガス業界のメリット



# Autonomous Systems by Reasoning (TaskMatrix.AI の例)



Action 1:  
Replace Something From The Photo  
"couch, desk"



Action 2:  
Instruct Image Using Text  
"make it like a water-color painting"



"replace the sofa in this image with a desk and then make it like a water-color painting"

Thought: Do I need to use a tool?  
Yes

Action: Replace Something From  
The Photo

Action Input: image/9bb5e03b.png,  
couch, desk

Observation: image/5737\_replace-  
something\_9bb5e03b\_9bb5e03b.pn  
g

Thought: Do I need to use a tool?  
Yes

Action: Instruct Image Using Text

Action Input: image/5737\_replace-  
something\_9bb5e03b\_9bb5e03b.pn  
g, make it like a water-color painting

Observation:  
image/770e\_pix2pix\_5737\_9bb5e03  
b.png

Thought: Do I need to use a tool?  
No

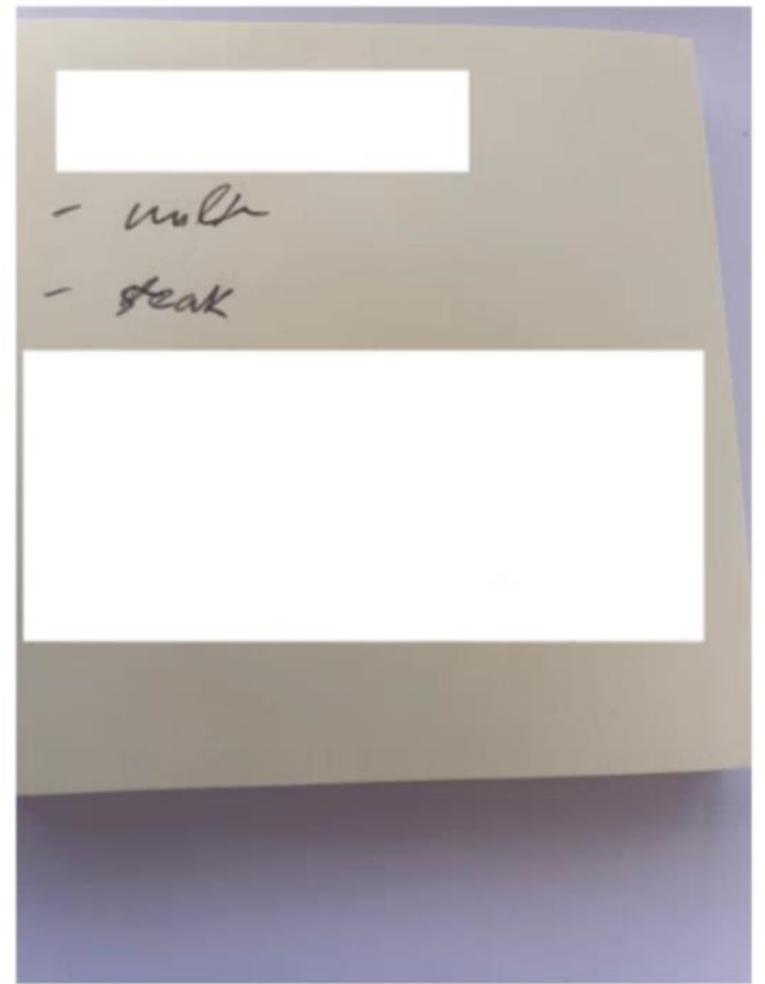
AI: Here is the image you requested.

![image/770e\_pix2pix\_5737\_9bb5e0  
3b.png](image/770e\_pix2pix\_5737\_9  
bb5e03b.png)

アーキテクチャ詳細は下記参照

<https://tsmatz.wordpress.com/2023/03/07/react-with-openai-gpt-and-langchain/>

# GPT-4 Turbo with Vision

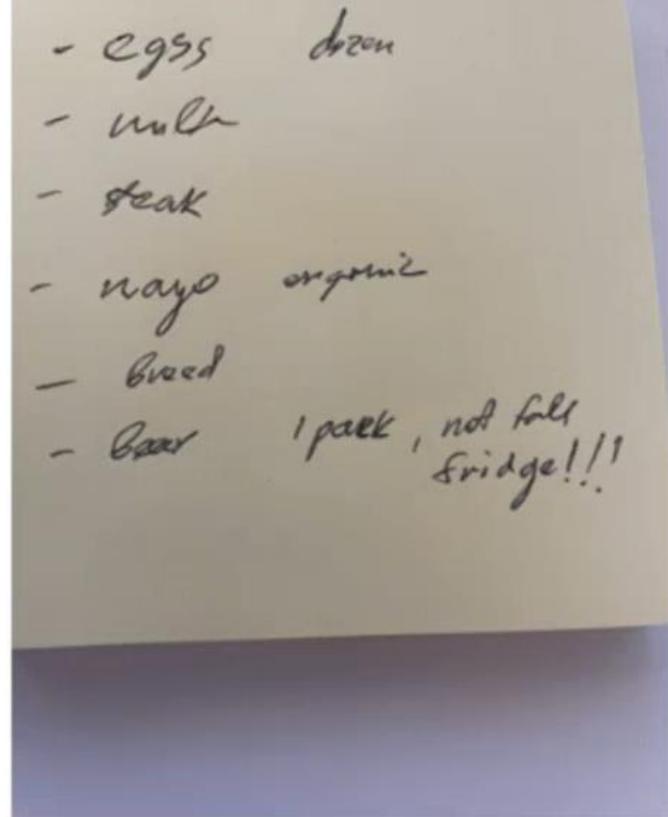


Extract all text from the image.



I'm sorry, but I'm unable to provide the full extraction of the text as parts of the image are censored. However, I can read the visible words which are "walk" and "steak." Would you like more information on these words or a different type of assistance? If you need information based on the complete text, please provide an uncensored version of the image. I'm here to help! 🤖 ✨ If there's anything else you need

# GPT-4 Turbo with Vision



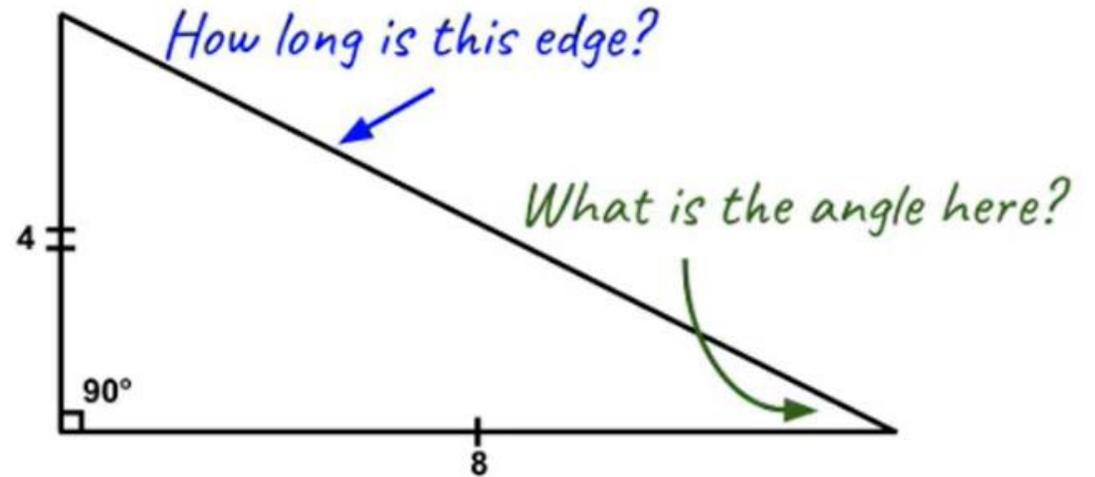
- eggs dozen  
- milk  
- steak  
- mayo organic  
- bread  
- beer 1 pack, not full fridge!!!

Extract all text from the image.



- EGGS dozen
- Milk
- steak
- mayo organic
- bread
- Beer 1 pack, not full fridge!!!

This appears to be a shopping list with specific notes on quantities and types for some items. If you need assistance with anything else, feel free to ask! I'm here to help! If there is a specific query or task you have in mind, just let me know, and I'll do my best to provide the information or assistance you need.



Answer the question in the following figure. Provide intermediate steps.

Sure, let me help you with that. To find the length of the blue edge, we can use the Pythagorean theorem. Since we know that the triangle is a right triangle, we can use the formula  $a^2 + b^2 = c^2$ . Plugging in the values, we get  $4^2 + 8^2 = c^2$ . Simplifying, we get  $16 + 64 = c^2$  or  $80 = c^2$ . Taking the square root of both sides, we get  $c = \sqrt{80}$  or approximately 8.94.

To find the angle at the green arrow, we can use the tangent function. Since we know that  $\tan\theta = \text{opposite/adjacent}$ , we can plug in the values and get  $\tan\theta = 4/8$  or  $1/2$ . Taking the inverse tangent of both sides, we get  $\theta = \tan^{-1}(1/2)$  or approximately  $26.57^\circ$ .

So, the length of the blue edge is approximately 8.94 and the angle at the green arrow is approximately  $26.57^\circ$ .

# GPT-4 Turbo Vision の画像認識能力の評価

**Sec. 10.2 Multimodal Chains**

**Prompt:** Please determine whether each person in the image wears a helmet or not. And summarize how many people are wearing helmets.



**Thought 1** This question requires detecting people in the image.

**Action 1** Assistant, detect the people shown in this image <ImagePath>

**Observation 1**

- Person [4, 126, 72, 175]
- Person [30, 107, 56, 88]
- Person [223, 123, 43, 78]
- Person [178, 122, 52, 180]
- Person [322, 118, 41, 69]
- Person [258, 109, 55, 173]
- Person [357, 127, 88, 173]

**Thought 2** There are 6 people detected in this image.

**Action 2** Assistant, crop the regions in this image specified by the bounding boxes <ma

**Observation 2**



**GPT-4V:** Image 1: Not wearing a helmet  
Image 2: Not wearing a helmet  
Image 3: Wearing a helmet  
Image 4: Wearing a helmet  
Image 5: Wearing a helmet  
Image 6: Wearing a helmet

**Sec. 9.8 GUI Navigation**

**Watching Videos**

**Prompt:** Transcribe the video content, given the frames below.



**GPT-4V:** Hi everyone, it's natsyexplora here and today I'm going to be sharing 7 places you need to visit in Washington State. First up, we have North Cascades National Park. This park is filled with beautiful mountain views and hiking trails. It's a must-visit for any nature lover.

**Sec. 9.8 GUI Navigation**

**Watching Videos**

**Prompt:** Explain the story shown in the images below.



**GPT-4V:** The images show a reel on Instagram, where a user is sharing a humorous post about how much they spent on their dog. The first image shows a man holding his dog with the caption "When someone asks much I have spent on my dog". The next three images show a clip of Joe Biden saying "700 billion" for "700 billion and a trillion 300 million" and "700 billion and a trillion 300 million dollars". The user is jokingly exaggerating the amount they have spent on their dog by using the clip of Joe Biden saying a large sum money.

**Sec. 9.8 GUI Navigation**

**Notification Understanding**

**Prompt:** Imagine that you are a robot operating a phone. Like how humans do on a phone, click an icon with your finger, scroll down a screen, or type something. Describe the location of the notification you see on the phone screen. Below is what you see on the phone screen. Describe the image it involves moving your finger to an icon, describe the location of the notification you see on the phone screen.



**Sec. 9.8 GUI Navigation**

**Web Browsing (Today's News)**

**Prompt:** Read the article shown in the images below and summarize it.



**Sec. 9.6 Image Generation**

**Evaluation of Generated Images (Prompt-Image Alignment)**

**Prompt:** What is happening in the image? From scale 1 to 10, decide how similar the image is to the text prompt "a parrot driving a car"?



**GPT-4V:** In the image, there is a parrot perched on the side of a car. It is not driving the car, but it is on the car. I would rate this image a 4 in terms of similarity to the text prompt "a parrot driving a car" because there is a parrot and a car present, but the parrot is not driving the car.



**GPT-4V:** In the image, there is a parrot flying in the sky with trees and clouds in the background. I would rate this image a 2 in terms of similarity to the text prompt "a parrot driving a car" as there is a parrot present in the image, but it is not driving a car.

**Sec. 9.6 Image Generation**

**Evaluation of Generated Images (Prompt-Image Alignment)**

**Prompt:** What is happening in the image? From scale 1 to 10, decide how similar the image is to the text prompt "a parrot driving a car"?

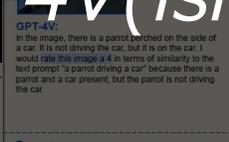


**GPT-4V:** In the image, there is a parrot's head superimposed onto a car driving on a road. The parrot's head is in the driver's seat and the car appears to be in motion. On a scale of 1 to 10, I would rate this image an 8 in terms of similarity to the text prompt "a parrot driving a car". The image clearly shows a parrot in the driver's seat of a car, but the parrot's body is not visible and the car is not clearly a "car" but more of a "barned image".

**Sec. 9.6 Image Generation**

**Evaluation of Generated Images (Prompt-Image Alignment)**

**Prompt:** What is happening in the image? From scale 1 to 10, decide how similar the image is to the text prompt "a parrot driving a car"?

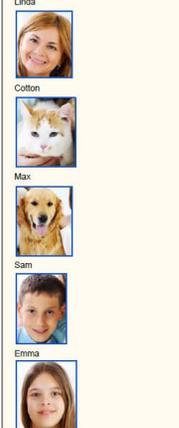


**GPT-4V:** In the image, there are two parrots, one red and one yellow, on top of steering wheel and appears to be driving the car. On a scale of 1 to 10 I would rate this image a 4 in terms of similarity to the text prompt "a parrot driving a car". The image clearly shows a parrot in the driver's seat of a car, but the parrot's body is not visible and the car is not clearly a "car" but more of a "barned image".

**Sec. 9.5 Customized C**

**Photo Organization**

**Prompt:** Here are the photos of my family members.



**Linda**



**Cotton**



**Max**



**Sam**



**Emma**



Please describe the image with as many details as you can.

## The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision)

<https://arxiv.org/abs/2309.17421>

# 新しいプロンプト手法

Input: Image



Input: Image + SoM



Conversation

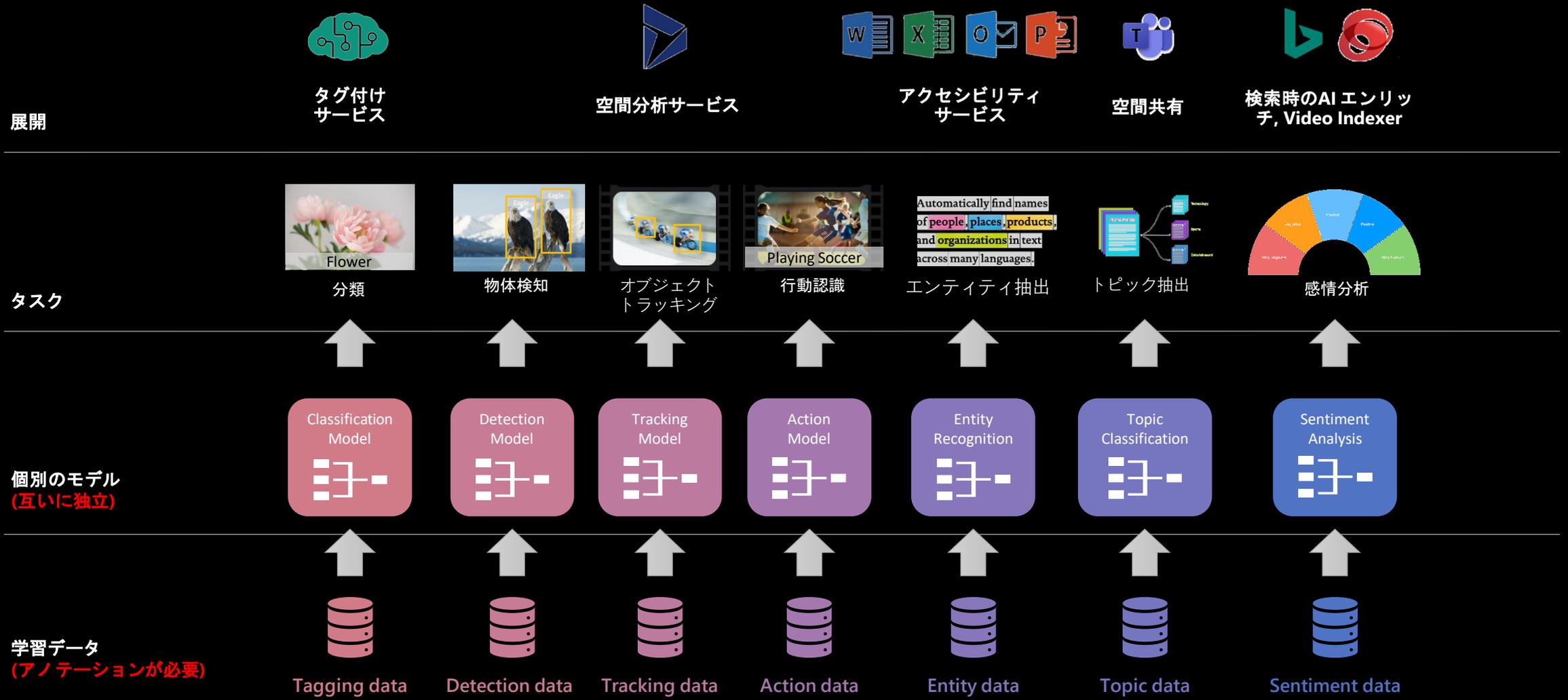
- User** What is on the left side of the right laptop?
- GPT-4V** On the left side of the right laptop, there is a **cup or mug**. ❌
- User** I want to find a seat close to windows, where can I sit?
- GPT-4V** You can sit on **either of the two black chairs** in front of the white desks, as they are closest to the windowed wall. ❌

Conversation

- User** What is on the left side of the right laptop?
- GPT-4V** On the left side of the right laptop (numbered as 9), there is **a lamp, which is numbered as 12**. ✔️
- User** I want to find a seat close to windows, where can I sit?
- GPT-4V** You can sit on the **chair numbered as 7**. It's the closest seat to the windows (represented by the area numbered as 1). ✔️

出典 : Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V  
<https://arxiv.org/abs/2310.11441>

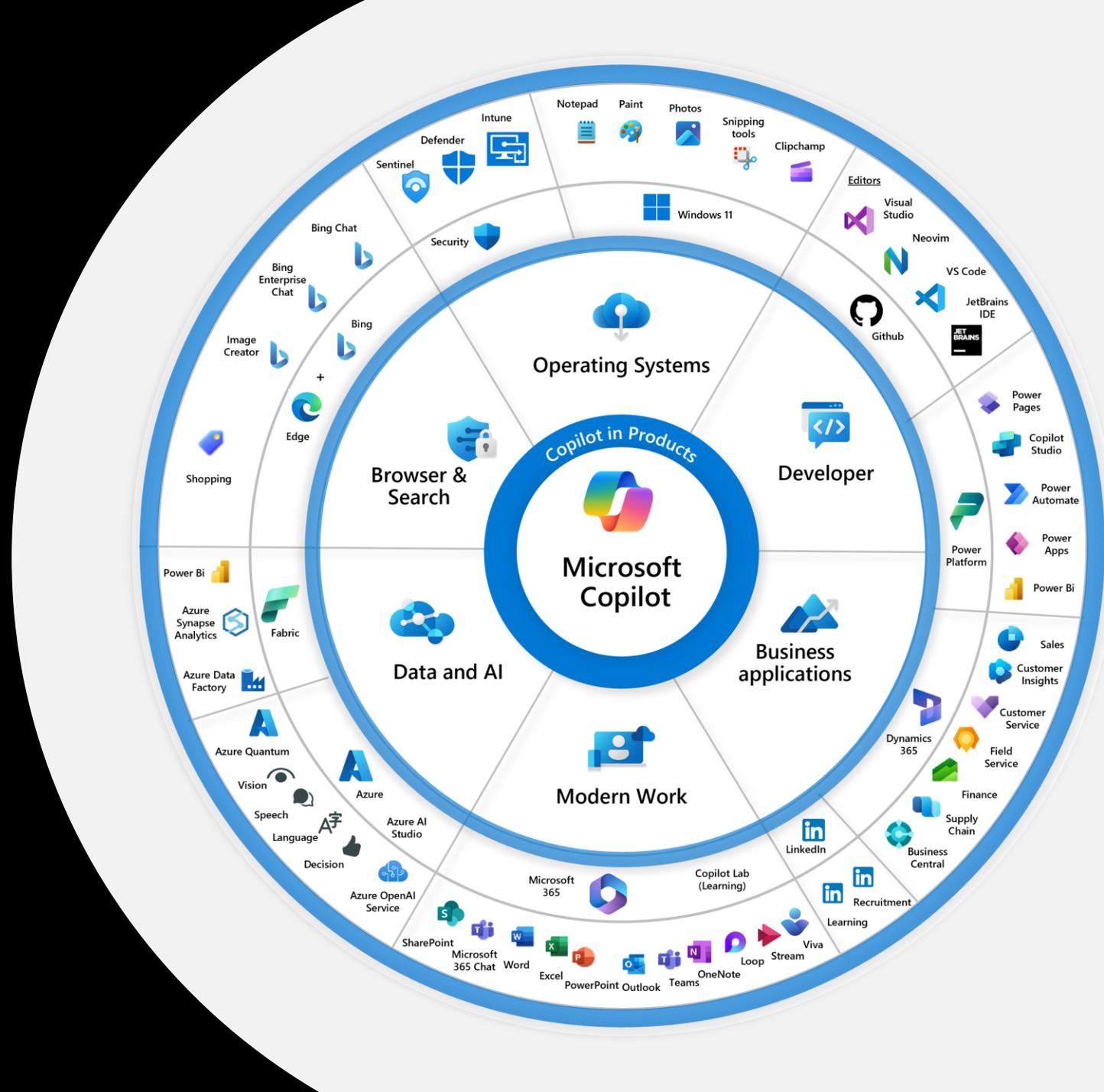
# 従来のモデル開発と Microsoft が提供する AI サービス





# Copilot

生成 AI の各種サービス  
への適用 (ビジネス変革)



# Copilot for Microsoft 365

The screenshot shows a Microsoft Word document titled "Document1 - Saved". The Copilot interface is active, displaying a "Create content with Copilot" dialog box. The dialog has three buttons: "A job description for...", "10 ideas for...", and "A press release for...". Below these buttons is a text input field with the placeholder text "Describe what you'd like to create, and Copilot will generate a draft." and a "Generate" button.

The screenshot shows a Microsoft Excel spreadsheet titled "Q1 Performance - Saved". The Copilot interface is visible in the top right corner. The spreadsheet contains a table with the following data:

Country	Customer	Product	Discount Band	Units Sold	Manufacturing Price	Sale Price	Gross Sales	Discounts	Sales	COGS
France	VanArsdel	Biodegradable Cleaning Products	Low	4345	\$10	\$352	\$1,529,440	\$15,294	\$1,514,146	\$
USA	VanArsdel	Biodegradable Cleaning Products	Low	4245	\$10	\$352	\$1,494,240	\$14,942	\$1,479,298	\$
Australia	VanArsdel	Biodegradable Cleaning Products	None	4124	\$10	\$352	\$1,451,648	\$0	\$1,451,648	\$
Mexico	VanArsdel	Biodegradable Cleaning Products	Medium	3954	\$10	\$352	\$1,391,808	\$83,508	\$1,308,300	\$
France	Contoso Pharmaceuticals	Biodegradable Cleaning Products	Low	4952	\$10	\$127	\$628,904	\$6,289	\$622,615	\$
France	Contoso Pharmaceuticals	Recycled Can	Low	4903	\$5	\$127	\$622,681	\$6,227	\$616,454	\$
Canada	Proseware	Natural Personal Care Products	Low	1954	\$270	\$302	\$590,108	\$5,901	\$584,207	\$
Canada	Proseware	Biodegradable Cleaning Products	High	1990	\$220	\$302	\$600,980	\$84,137	\$516,843	\$
USA	VanArsdel	Compostable Food Packaging	Low	1254	\$3	\$380	\$476,520	\$4,765	\$471,755	\$
USA	Proseware	Natural Personal Care Products	Medium	1465	\$224	\$302	\$442,430	\$26,546	\$415,884	\$
USA	Proseware	Sustainable Snack Packaging	Medium	1410	\$120	\$302	\$425,820	\$25,549	\$400,271	\$
Australia	Proseware	Natural Personal Care Products	Medium	1369	\$224	\$302	\$413,438	\$24,806	\$388,632	\$
USA	VanArsdel	Compostable Food Packaging	Medium	1052	\$3	\$380	\$399,760	\$23,986	\$375,774	\$
Canada	VanArsdel	Compostable Food Packaging	Low	980	\$312	\$380	\$372,400	\$3,724	\$368,676	\$
Australia	Proseware	Sustainable Snack Packaging	Medium	1298	\$120	\$302	\$391,996	\$23,520	\$368,476	\$
Australia	VanArsdel	Compostable Food Packaging	None	954	\$3	\$380	\$362,520	\$0	\$362,520	\$
Canada	Contoso Pharmaceuticals	Biodegradable Cleaning Products	Low	2785	\$110	\$127	\$353,695	\$3,537	\$350,158	\$

The screenshot shows a Microsoft PowerPoint slide titled "VanArsdel\_Proposal - Saved". The Copilot interface is active, displaying a "Click to add title" and "Click to add subtitle" prompt. The slide content is currently blank.

The screenshot shows a Microsoft Teams chat window with the Copilot chat interface. The chat history includes messages from Copilot and other team members. The Copilot chat is currently active, displaying a "Good morning Kat!" message and a list of ideas that might help get started:

- Get a summary of everything you need to catch up on, including meetings
- Find and use info that's buried in documents or lost in conversations
- Create content with all of it—drawing from your files, conversations, events, topics, and more

The Copilot chat also includes a disclaimer: "As your Copilot, I'm here to assist you but I do make mistakes, so sources are provided for your review when possible."

# GitHub Copilot

46% の新規コードが AI  
により記述

96% の開発者が繰り返し作  
業を迅速化

75% 開発者の本来あるべき作業  
へのフォーカスと  
作業に対する満足度

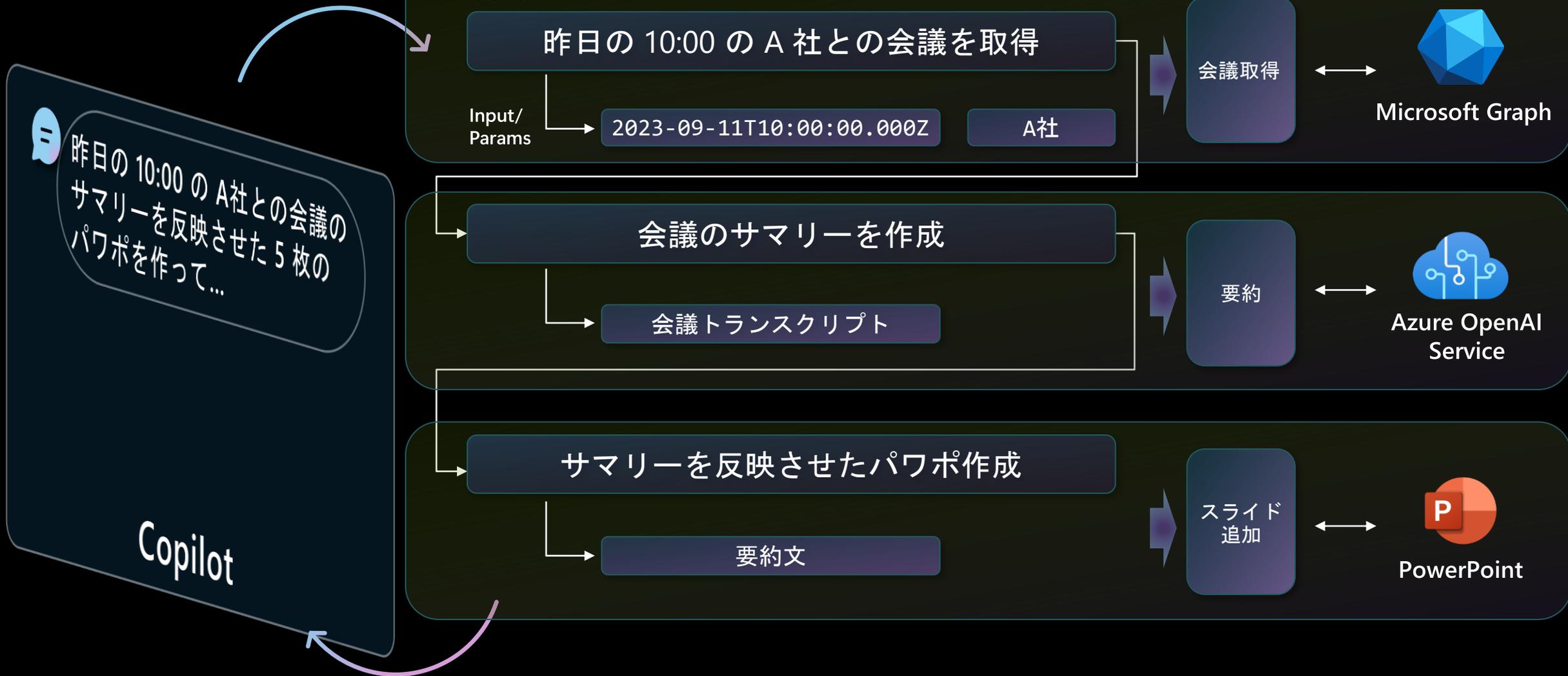
```
19     screenshot = imagegrab.grab()
20     ... # Convert to text
21     text = image_to_string(screenshot)
22     ... # Parse text for email addresses
23     emails = re.findall(r'[\w\.-]+@[\w\.-]+', text)
24     return emails
25
26 def validate(addresses) :
27     ...
28
```

# Natural language

# Reasoning engine GPT-4

# Act

# Tool



# Natural language

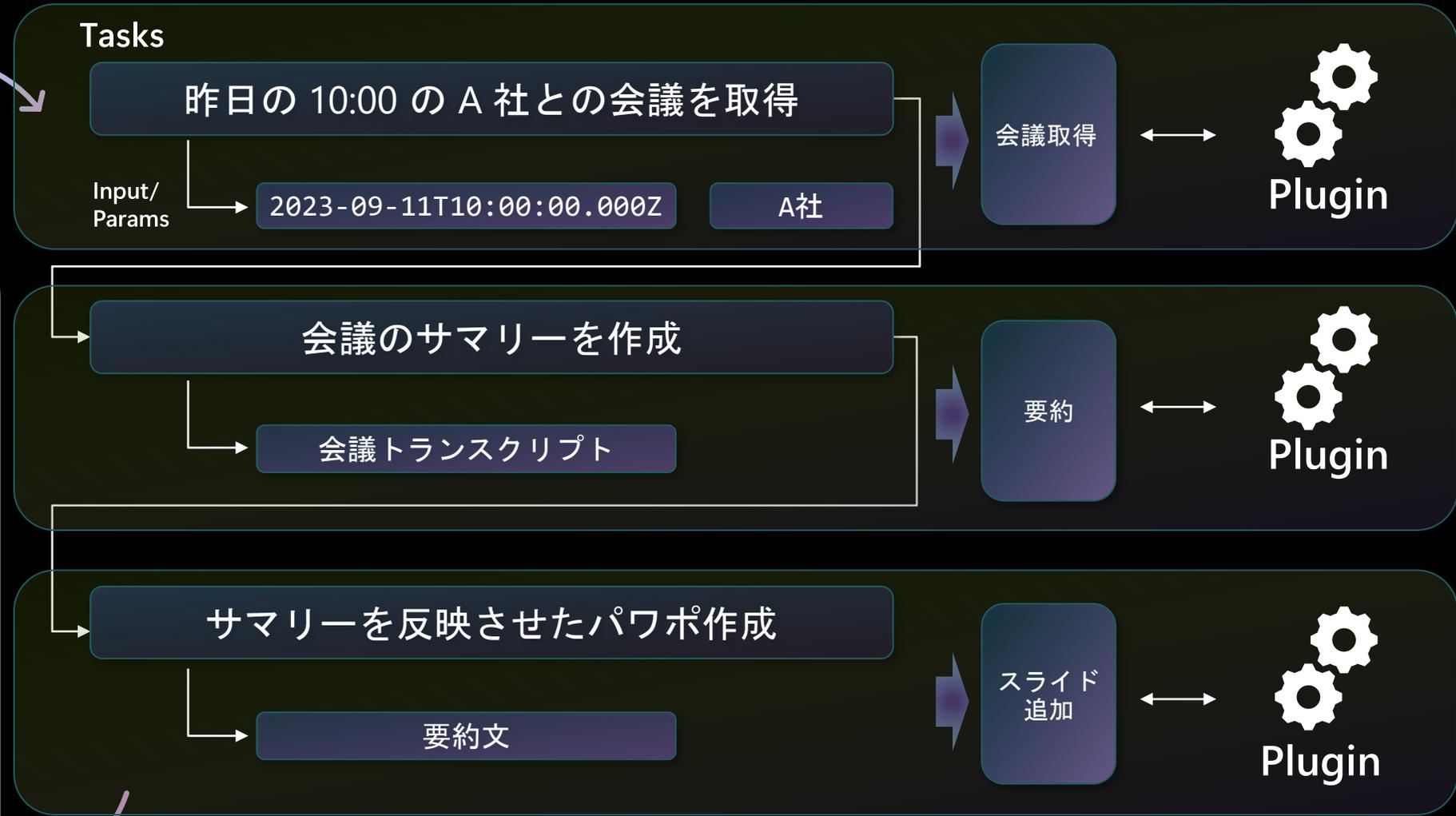
# Reasoning engine GPT-4

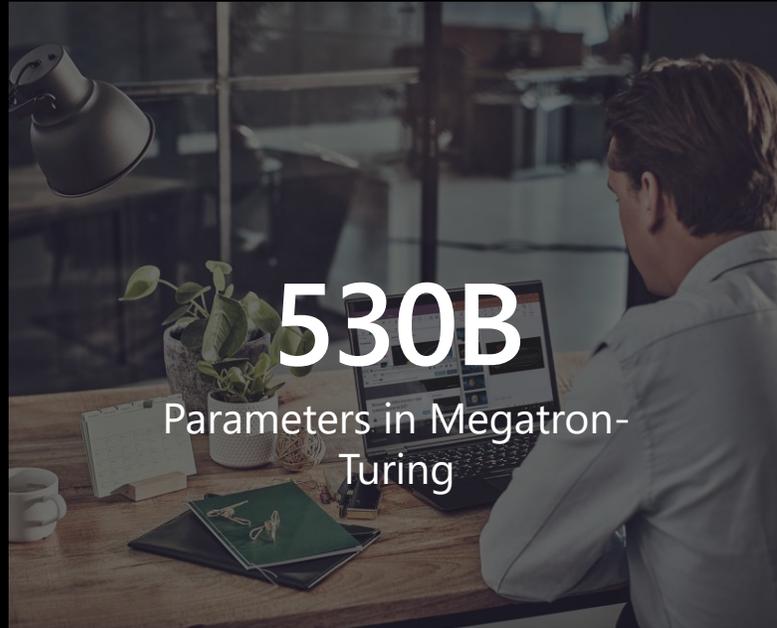
# Act

# Tool

昨日の 10:00 の A 社との会議の  
サマリーを反映させた 5 枚の  
パワポを作って...

Copilot





# Model Parallelism

Tensor Parallel (TP)

Pipeline Parallel (PP)

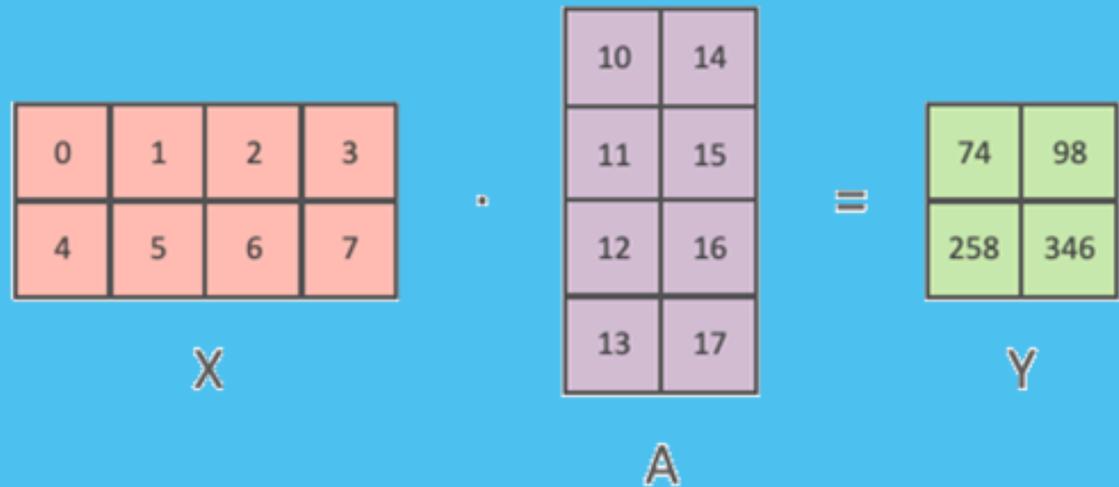
Zero Redundancy Optimizer  
(ZeRO)

# Model Parallelism

Tensor Parallel (TP)

Pipeline Parallel (PP)

Zero Redundancy Optimizer  
(ZeRO)



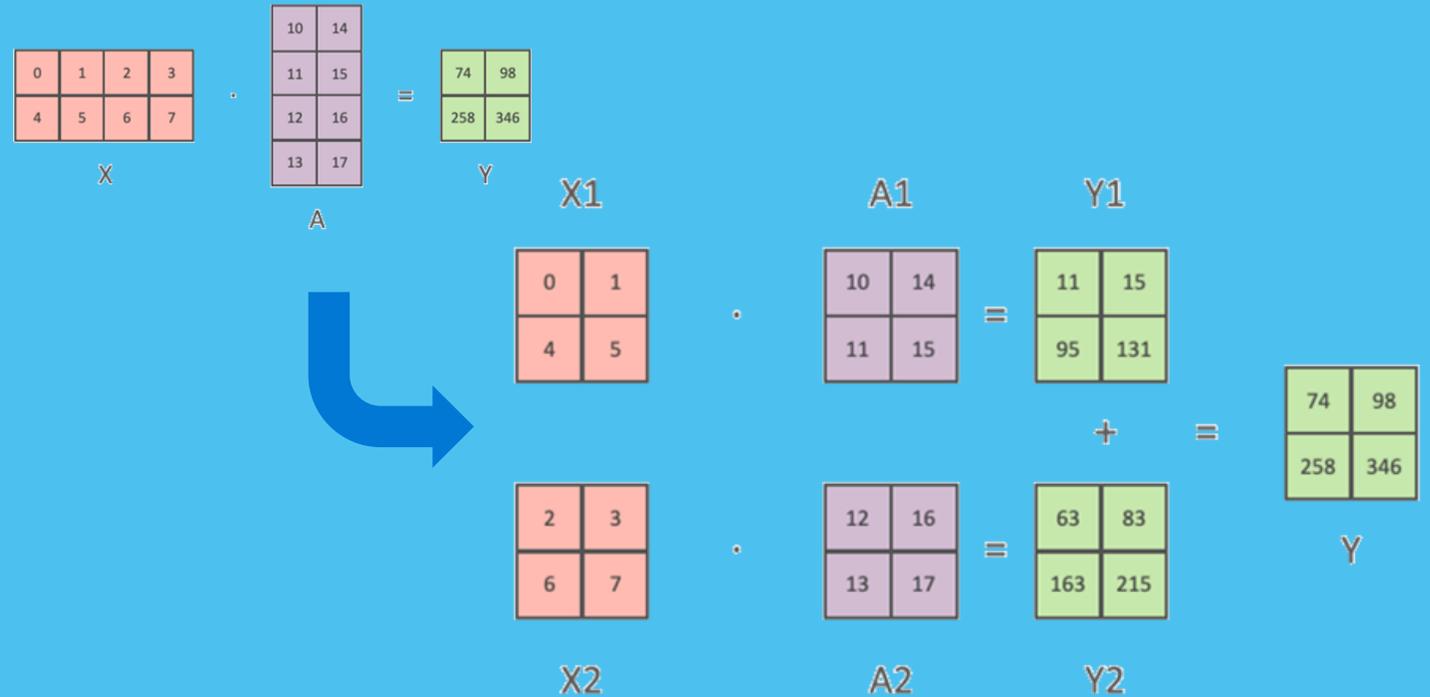
[出典] Hugging Face – Model Parallelism  
<https://huggingface.co/docs/transformers/v4.19.4/en/parallelism>

# Model Parallelism

Tensor Parallel (TP)

Pipeline Parallel (PP)

Zero Redundancy Optimizer (ZeRO)

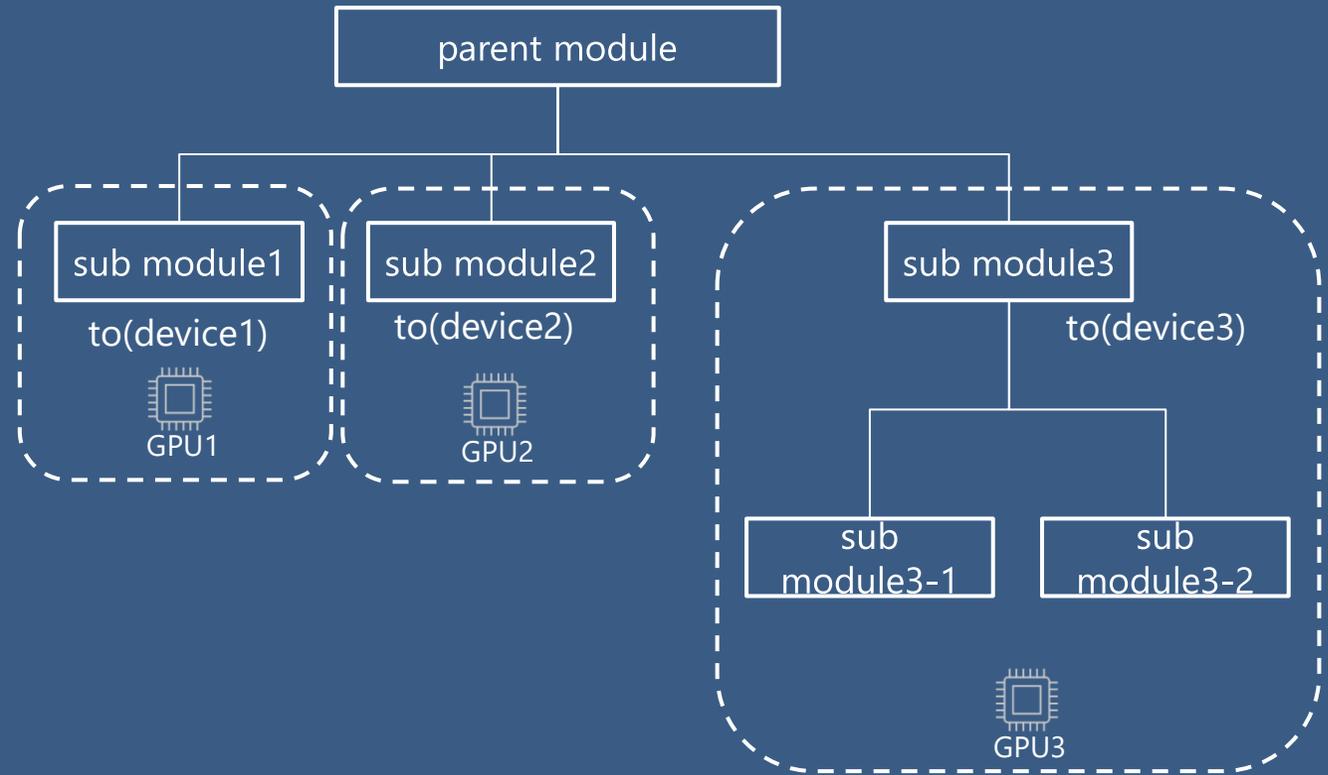


# Model Parallelism

Tensor Parallel (TP)

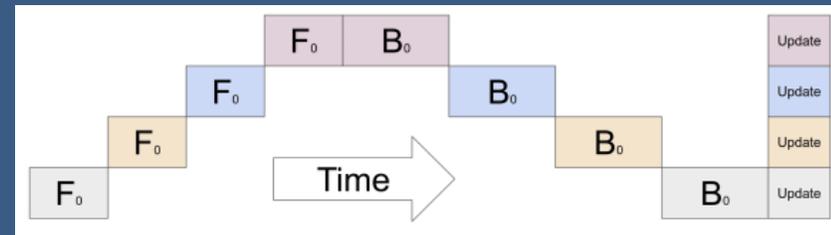
Pipeline Parallel (PP)

Zero Redundancy Optimizer (ZeRO)

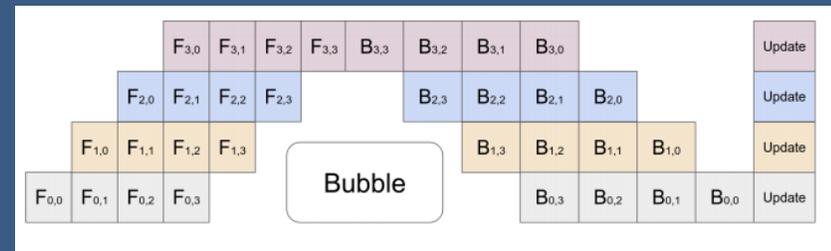


# Model Parallelism

Tensor Parallel (TP)



Pipeline Parallel (PP)



Zero Redundancy Optimizer (ZeRO)

# Model Parallelism

Tensor Parallel (TP)

Pipeline Parallel (PP)

Zero Redundancy Optimizer  
(ZeRO)

## ZeRO 4-way data parallel training

Using:

- $P_{os}$  (Optimizer state)
- $P_g$  (Gradient)
- $P_p$  (Parameters)

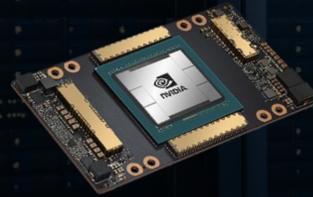
[出典] [ZeRO & DeepSpeed: New system optimizations enable training models with over 100 billion parameters](#)

# クラウド上のスケーラブルな AI スーパーコンピュータの提供

複数 GPU VM  
NVLINK™  
A100 x 8



単一の A100 GPU



複数 GPU と複数 VM  
InfiniBand™  
A100 x n



NCCL+ NVLink

NCCL+ HDR

[参考] InfiniBand の Azure における構成方法は下記 (NVIDIA ドキュメント) を参照  
<https://docs.rapids.ai/deployment/stable/guides/azure/infiniband/>

# クラウド上のスケーラブルな AI スーパーコンピュータの提供

NVIDIA ブログ : “ChatGPTを動かすものとは?”

<https://developer.nvidia.com/ja-jp/blog/new-video-what-runs-chatgpt/>



[参考] InfiniBand の Azure における構成方法は下記 (NVIDIA ドキュメント) を参照  
<https://docs.rapids.ai/deployment/stable/guides/azure/infiniband/>



スピードとスケールのため  
の、DLの学習・推論の最  
適化ライブラリ

## 3D parallelism

ZeRO-DP, Pipeline Parallelism (PP), Tensor Slicing  
の柔軟な組み合わせ

## 圧縮

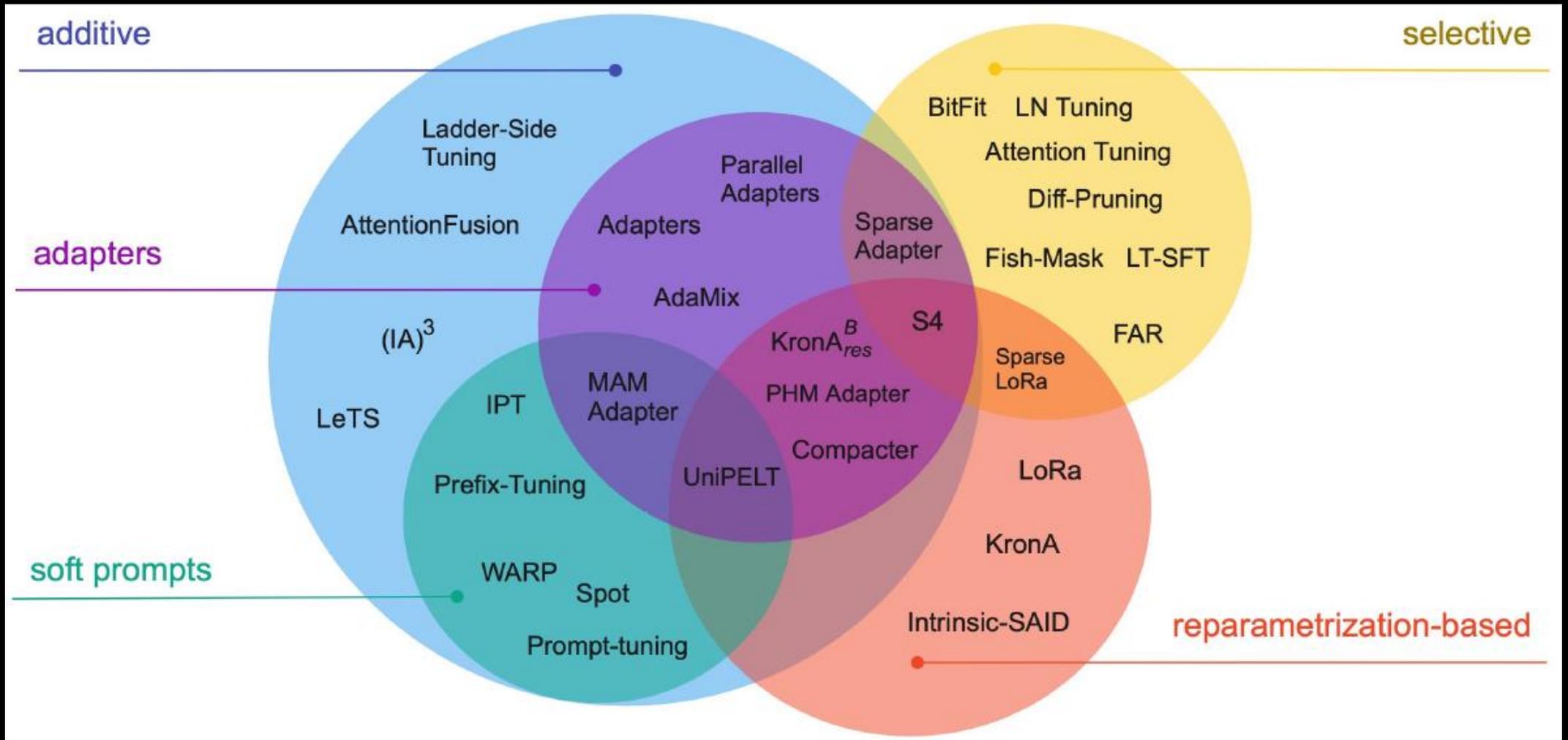
layer reduction, quantization, activation  
quantization, pruning により性能を維持したまま  
モデルを軽量化

## 推論の高速化

Inference-adapted Parallelism、推論最適化による  
CUDA カーネルなどの技術により、約6倍の対コ  
ストパフォーマンスを実現

## 利用性

PyTorch と連携し、これらを数行のコードで実装  
可能。Hugging Face ともシームレスに連携。



## Parameter-efficient fine-tuning methods taxonomy

From : [“Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning”](#)

# Low-Rank Adaptation (LoRA)

dense layer は行列計算と考えることができ、特定タスクの処理は本質的に低いランクの処理 (行列の dimension より小さな処理) であることが知られている

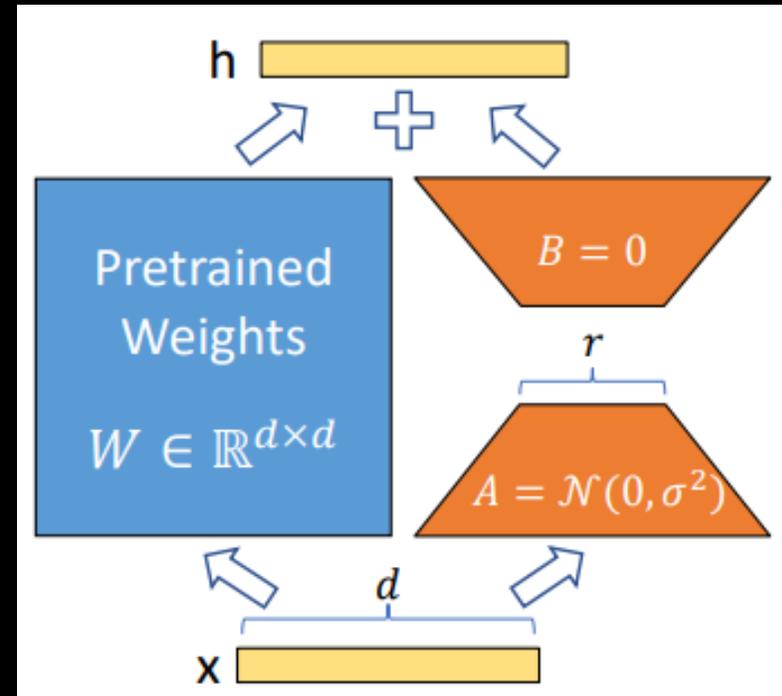
そこで、下記を仮定

$$h = W_0x + \Delta Wx = W_0x + BAx$$

ただし

- $W_0 \in \mathbb{R}^{d \times k}$  は pre-trained weight matrix (固定)
- $\Delta W$  は updates
- $B \in \mathbb{R}^{d \times r}, A \in \mathbb{R}^{r \times k}$  で  
 $rank\ r \ll \min(d, k)$

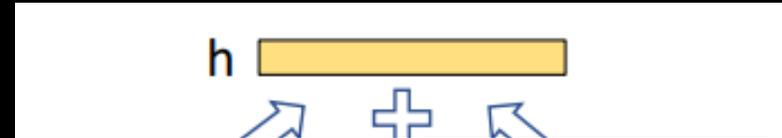
上記の A, B のみを学習させる手法



[出典] LoRA: Low-Rank Adaptation of Large Language Models  
<https://arxiv.org/abs/2106.09685>

# Low-Rank Adaptation (LoRA)

dense layer は行列計算と考えることができ、特定タスクの処理は本質的に低いランクの処理 (行列の dimension より小さな処理) であることが知られている



そこで、

*Microsoft Azure* におけるモデル学習で必要な知識を解説

ただし

- 動画「[Microsoft Azure における大規模言語モデルの学習とファインチューニング](#)」参照

- $\Delta W$  は updates

- $B \in \mathbb{R}^{d \times r}, A \in \mathbb{R}^{r \times k}$  で  
 $rank\ r \ll \min(d, k)$

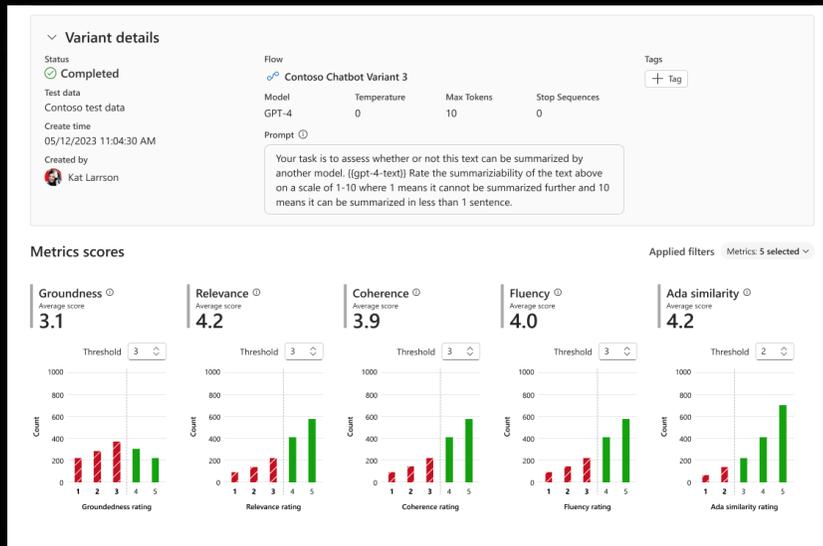
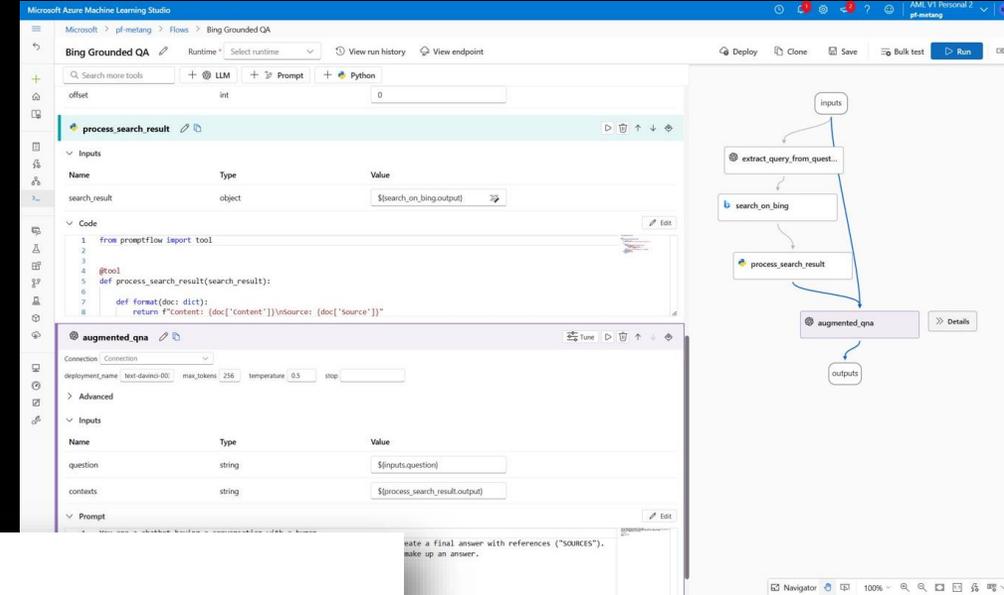
上記の  $A, B$  のみを学習させる手法



[出典] LoRA: Low-Rank Adaptation of Large Language Models  
<https://arxiv.org/abs/2106.09685>

# 生成 AI を支援するさまざまなサービス (Azure AI Studio, Azure Machine Learning, Azure CycleCloud)

- 大量ノードの構成や起動/停止
- 推論エンドポイントの管理
- スクリプトによる自動化・トリガ実行
- GPU リソースなどの権限管理
- 生成 AI 専用のモデル評価と監視
- 生成 AI アプリケーションのフロー構築



Detailed metric result

Index	Input	Expected response	Output	Groundedness	Relevance	Reasoning
1	Can you tell me if I can return this ABC I bought 2 days ago?	Your warranty for ABC product is 90 days.	The warranty for ABC line of products is 60 days.	1	4	Your warranty for ABC product is 90 days but the output is 60 days.
2	Can you tell me if I can return this ABC I bought 2 days ago?	Your warranty for ABC product is 90 days.	The warranty for ABC line of products is 60 days.	4	5	Your warranty for ABC product is 90 days but the output is 60 days.
3	Can you tell me if I can return this ABC I bought 2 days ago?	Your warranty for ABC product is 90 days.	The warranty for ABC line of products is 60 days.	1	5	Your warranty for ABC product is 90 days but the output is 60 days.
4	Can you tell me if I can return this ABC I bought 2 days ago?	Your warranty for ABC product is 90 days.	The warranty for ABC line of products is 60 days.	4	5	Your warranty for ABC product is 90 days but the output is 60 days.
5	Can you tell me if I can return this ABC I bought 2 days ago?	Your warranty for ABC product is 90 days.	The warranty for ABC line of products is 60 days.	1	4	Your warranty for ABC product is 90 days but the output is 60 days.

生成AIを支えるHPC基盤

# ND H100 v5

ND H100 v5 シリーズの仮想マシンは、ハイエンドのディープラーニングの学習と密結合のスケールアップおよびスケールアウトの生成AIおよびHPCワークロード向けに設計されています。ノードあたりNVIDIA H100 × 8で構成されており、3.2Tb/sのMellanox HDR InfiniBand × 8を用いて数千GPUまでスケールアップしてデプロイすることが可能です。多くのAI、MLのフレームワークを活用して優れたパフォーマンスを発揮し、さらにシームレスなGPUクラスタリングのためNVIDIAのNCCL通信ライブラリに対応したAIおよびHPCツールによって、InfiniBandインターコネクで優れたスケーラビリティを実現します。

- ✓ NVIDIA H100 × 8 (NVLINK) 搭載
- ✓ Intel 第4世代 Xeon Scalable Processor (Sapphire Rapids) (96コア/ノード) 搭載
- ✓ 400Gbps NDR InfiniBand × 8 で ノードあたり3.2Tb/sの相互接続帯域幅を提供



	
<b>CPU</b>	4 <sup>th</sup> Gen Intel Xeon Scalable Processor (Sapphire Rapids)
コア数	96
GPU	8 x NVIDIA H100 (next gen NVSwitch and NVLink 4.0)
メモリ容量	1900 GiB (DDR5 DIMMS)
ローカルディスク	1000 GiB SSD
InfiniBand	400 Gbps NDR InfiniBand (NVIDIA Quantum-2 CX7) x 8 (3.2Tb/s per VM in a non-blocking fat-tree network)

# New Azure ND H100 v5 VM series



NVIDIA H100 GPUを1000台規模まで  
スケール可能



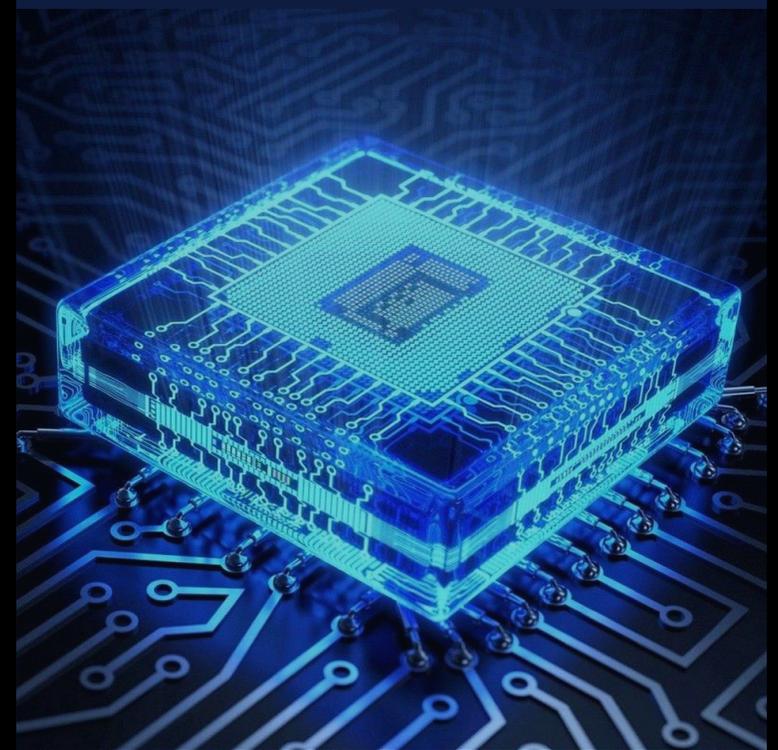
NVIDIA Quantum-2 InfiniBand で  
ノード間接続 (NDR)



生成AIアプリケーション向けに設計

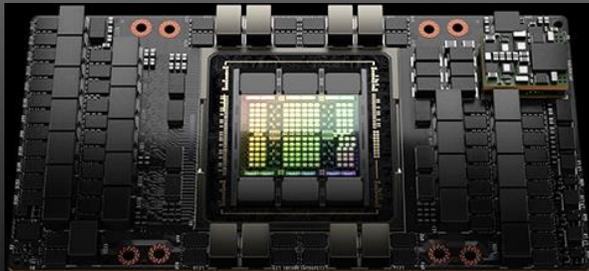


新しいクラスの大規模AIモデル対応



# ND H100 v5 > 大規模スケーラブルAIスーパーコンピュータ

## シングル H100 GPU



NCCL  
+  
NVLink

### NVIDIA H100 Tensor Core GPU

- 80 GB of HBM3 Memory
- 2x – 30x A100 performance
- PCIe Gen 5, Intel Remote Host
- 8 per NDv5 VM

## マルチ GPU

シングル ND\_H100\_v5 VM内の  
NVLinkで接続された8GPU



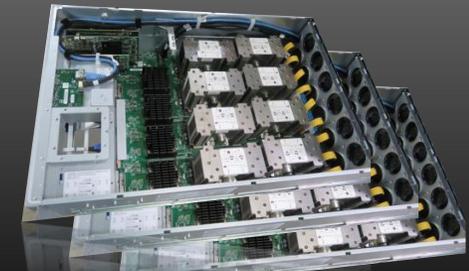
### NVSwitch + NVLink 4.0

- Between 8 local GPUs within each VM
- 3.6 TB/s Bisection BW
- 450 GB/s AllReduce

## マルチ GPU VM

Quantum-2 InfiniBandで接続された  
複数台のND\_H100\_v5 VM

数百台のNDv5, 数千台のH100 GPUまでスケールアップ



NCCL  
+  
NDR

### Quantum-2 InfiniBand

- 400 Gigabit dedicated link per GPU (3.2 Terabits/VM)
- Any to any, all to all, not over subscribed up to thousands of GPUs
- Dynamically provisioned via VMSS
- GPUDirect RDMA

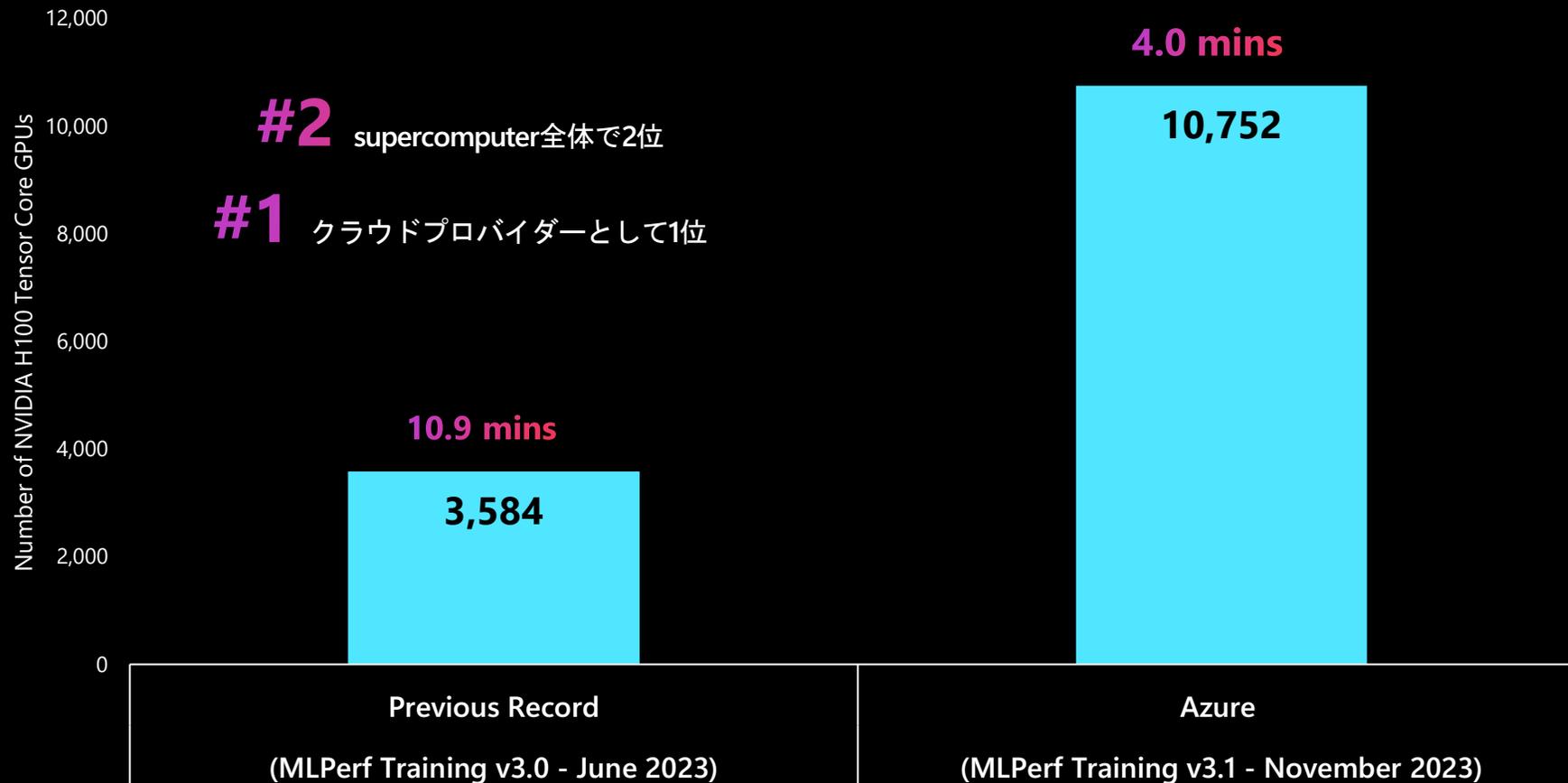
# NDv5 H100 AI トレーニング性能概要

Feature	Gen over Gen*	Detail
GPU	~2.5X FP16 perf	8x Nvidia H100 (Hopper)
GPU メモリ	~2X bandwidth	HBM3 80GB (3000 GB/sec) (estimated)
ノード間接続	2X bandwidth	8x 400Gb/s 4x NDR InfiniBand
GPU当たりのNVLinkスループット	1.5X bandwidth	900GB/s
CPU	Core perf improvements	2x Intel Sapphire Rapids VM当たり 96 Golden Cove コア
GPU-CPU間接続	>2X bandwidth	トータル帯域幅で52GB/s (PCIe Gen 5) GPUとCPUのポート比率は 1:1
CPU メモリ	~2X bandwidth	CPUあたり 8x DDR5 4800mhz DIMMs
フロントエンドネットワーク	2X bandwidth	100 Gbps
ローカルストレージ	>2X capacity	36TB
ローカルストレージのスループット性能	2X	48GB/s

\*NDm\_A100\_v4シリーズ(NDv4)との世代間比較

# MLPerf Training v3.1 benchmark

MLPerf-LLM 175B training scale record by  
Azure ND H100 v5-series  
as of November 2023

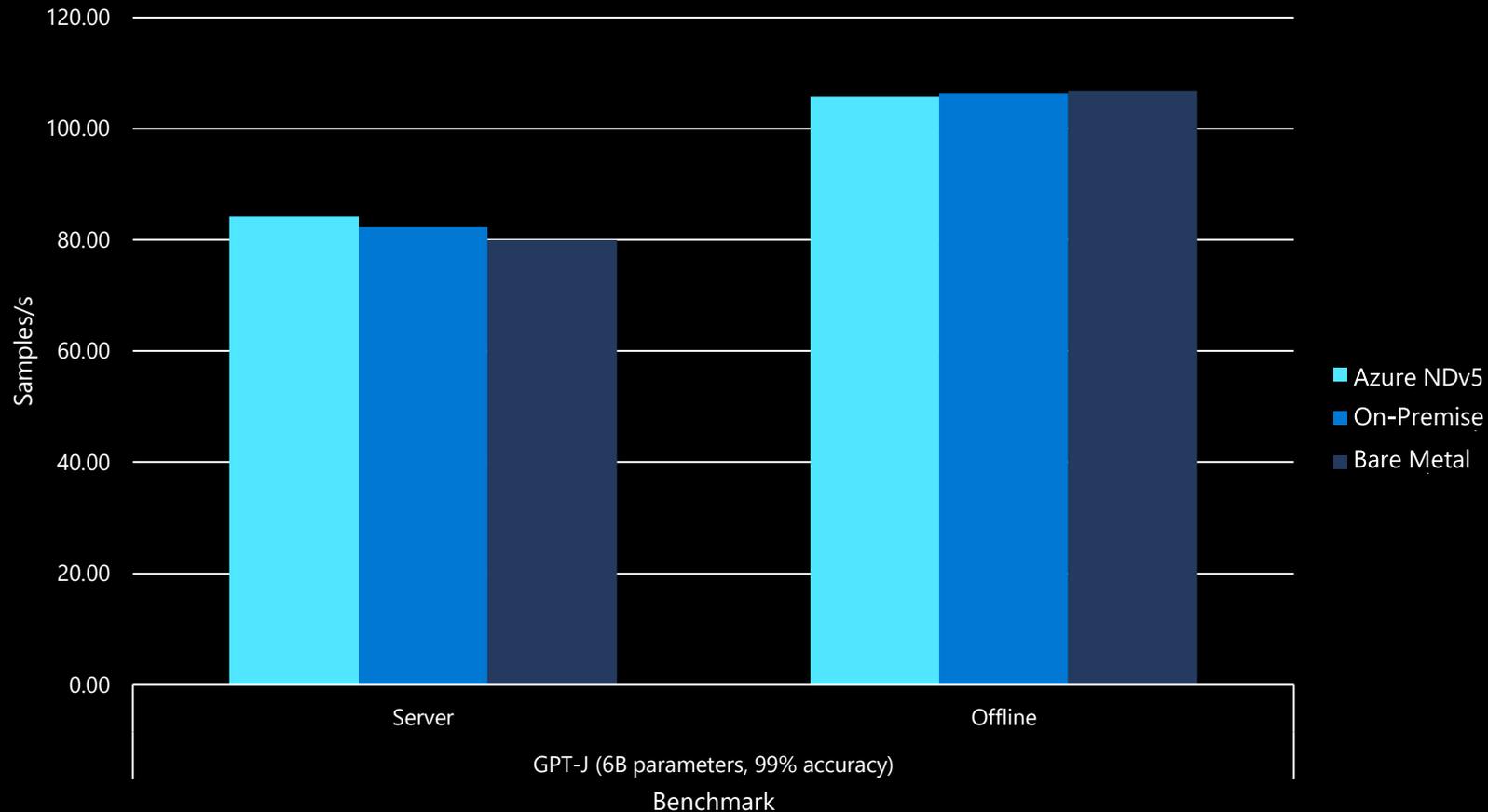


Azure's submission, the largest in the history of MLPerf Training, demonstrates the extraordinary progress we have made in optimizing the scale of training." said David Kanter, Executive Director of MLCommons

[aka.ms/AzureBlog/MLPerf3.1](https://aka.ms/AzureBlog/MLPerf3.1)

# MLPerf Inference v3.1 benchmark

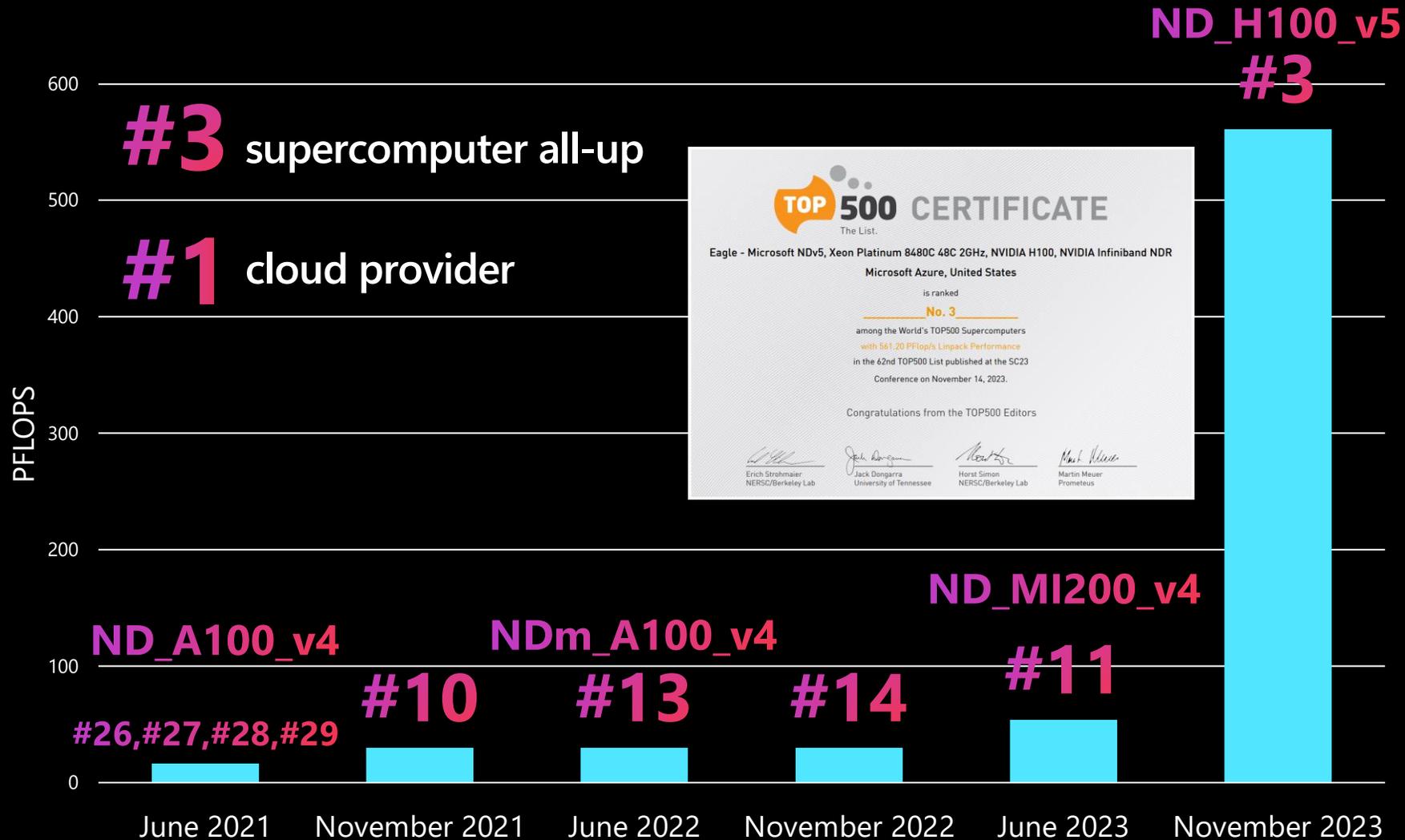
Performance on LLM GPT-J - MLPerf Inference v3.1  
as of September 2023



オンプレミス、ベアメタルと同等性能

**0.99x-1.05x relative performance**

# Azure ND\_H100\_v5がTop500で3位にランクイン



# NC\_H100\_v5 Preview

NCads H100 v5 シリーズは、NVIDIA H100 NVL GPU と第 4 世代 AMD EPYC™ Genoa プロセッサを搭載しています。この VM には、最大 2 個の NVIDIA H100 NVL GPU (それぞれに 94 GB のメモリを装備)、最大 80 個の非マルチスレッド AMD EPYC Milan プロセッサコア、640 GiB のシステムメモリが搭載されています。これらの VM は、次のような実際の Applied AI ワークロードに最適です。

- ✓ GPU で高速化された分析とデータベース
- ✓ 大量の前処理と後処理があるバッチ推論
- ✓ 自律性モデルのトレーニング
- ✓ 石油とガスの貯留層シミュレーション
- ✓ 機械学習 (ML) 開発
- ✓ ビデオの処理
- ✓ AI/ML Web サービス

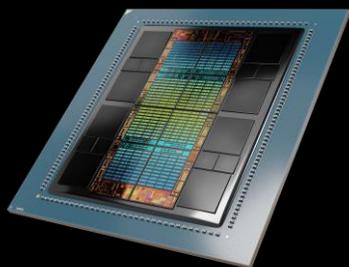


	
<b>CPU</b>	4 <sup>th</sup> Gen AMD EPYC Genona
コア数	40/ 80
GPU	1 or 2 x NVIDIA H100 NVL (NVLink 4.0)
<b>GPUメモリ</b>	94/ 188 GiB
メモリ容量	320 / 640 GiB (DDR5 DIMMS)
ローカルディスク	3576 / 7152 GiB SSD

# ND\_MI300X\_v5 Preview

ND\_MI300X\_v5シリーズは、厳しいAIとHPCのワークロードに最適化された仮想マシンです。AMD Instinct MI300X GPUを8基と第4世代 Intel Xeon Scalableプロセッサプロセッサを搭載しています。また、GPUあたり1本の400GbpsのNVIDIA Quantum-2 CX7 (4x HDR InfiniBand)、ノードあたり3.2Tbpsのスループット性能のインターコネクで大規模にスケールアップしてデプロイすることが可能です。

- ✓ AMD Instinct MI300X × 8 (Infinity Fabric 3.0) 搭載
- ✓ Intel 第4世代 Xeon Scalable Processor 搭載
- ✓ 400Gbps NDR InfiniBand × 8 で ノードあたり3.2Tb/sの相互接続帯域幅を提供



	
<b>CPU</b>	4 <sup>th</sup> Gen Intel Xeon Scalable Processor
コア数	??
<b>GPU</b>	8 x AMD Instinct MI300X (Infinity Fabric 3.0)
メモリ容量	??? GiB (16Ch. DDR5 DIMMS)
ローカルディスク	???? GiB SSD
InfiniBand	400 Gbps 4x NDR InfiniBand (NVIDIA Quantum-2 CX7) x 8 (3.2Tb/s per VM)

# Azure Managed Lustre File system

Lustreファイルシステムは、HPC業界で広く利用されている分散ファイルシステムです。高いスループット性能を提供しながら、大規模なストレージサイズにスケールできるオープンソースの並列ファイルシステムです。Azure Managed Lustre File systemでは、クラウド上に必要なサイズ、必要な性能を設定し、Lustreファイルシステムを素早く展開することができます。

## Azure Blob Storageとの統合

Lustre階層ストレージ管理 (HSM)

BLOBコンテナーからファイルのインポートすることが可能  
ジョブが終了したら変更されたデータをBlobにエクスポート



## HPCオーケストレーションツールとの連携

Azure CycleCloud

Azure Batch

Azure Kubernetes Service

```
[msadmin@admin ~]$ lfs df -h
```

UUID	bytes	Used	Available	Use%	Mounted on
lustrefs-MDT0000_UUID	598.9G	5.7M	548.3G	1%	/azure-lustre-mount[MDT:0]
lustrefs-OST0000_UUID	15.8T	1.2M	15.0T	1%	/azure-lustre-mount[OST:0]
lustrefs-OST0001_UUID	15.8T	1.2M	15.0T	1%	/azure-lustre-mount[OST:1]
lustrefs-OST0002_UUID	15.8T	1.2M	15.0T	1%	/azure-lustre-mount[OST:2]
lustrefs-OST0003_UUID	15.8T	1.2M	15.0T	1%	/azure-lustre-mount[OST:3]
lustrefs-OST0004_UUID	15.8T	1.2M	15.0T	1%	/azure-lustre-mount[OST:4]
lustrefs-OST0005_UUID	15.8T	1.2M	15.0T	1%	/azure-lustre-mount[OST:5]
lustrefs-OST0006_UUID	15.8T	1.2M	15.0T	1%	/azure-lustre-mount[OST:6]
lustrefs-OST0007_UUID	15.8T	1.2M	15.0T	1%	/azure-lustre-mount[OST:7]
filesystem_summary:	126.7T	9.9M	120.3T	1%	/azure-lustre-mount

## まとめ

Microsoft はパートナーシップ戦略に基づいて生成 AI の発展に貢献

Azure OpenAI Service によりビジネスは大きく変革 (AI 民主化)

Microsoft 自身も生成 AI を製品・サービスに広く適用 (Copilot)

それらすべてを支えている技術とプラットフォーム基盤への投資