CTC Engineer's Insight＃01

【session3】「AI専門外のエンジニアが挑戦する、安全な「生成AI」の作り方（暗中模索）」

伊藤忠テクノソリューションズ株式会社

金融システム技術第４部インフラアーキテクト課

荻野　晃浩

自己紹介

- 氏名：荻野　晃浩（おぎの　あきひろ）
- 出身：石川県
  千里浜なぎさドライブウェイの近く
- 趣味：畑いじり **8年**（市民農園レベル）
- 経歴：新卒CTC入社 金融業界一筋 **18年**
  1. メガバンク系インフラSE
  2. 金融系インフラPM
  3. 金融系アーキテクト
- 主な業務：インフラ構想策定、ロードマップ策定

このセッションの目的

- 生成AIで何やるの？どうやって作るの？
  要件も方式も定まらない中で、何もわからない**AI専門外**の
  リードアーキテクトがチャレンジしてみました

- セキュリティ要件が厳しい金融業界でも利用可能な、
  安全な生成AI開発の**第一歩**を紹介します

このセッションのアジェンダ

無限の未来と、幾千のテクノロジーをつなぐ。

CTC Financial Services Group

生成AIと金融機関

無限の未来と、幾千のテクノロジーをつなぐ。

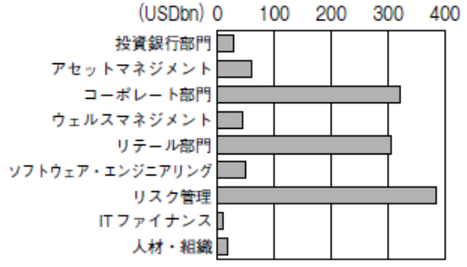
CTC Financial Services Group

金融機関でのユースケース

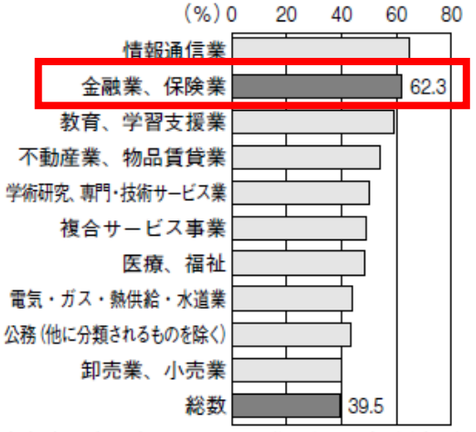
国内金融機関における生成 AI 動向

・日本の労働市場では、総就業者数の約80%が生成AIの影響を受ける可能性があり、約40%の就業者が仕事の半分以上を自動化できるとの推計もある。生成AIはデータ分析・プログラムコードの生成・文章の要約等を扱うオフィスワーク中心の産業で影響が大きく、金融業の生産性向上に大きく寄与すると考えられている（図表4）。
・国内金融機関は、業務効率化・新たなサービスの開発・与信判断能力の向上・投資判断の向上・リスク管理とコンプライアンス対応の高度化に生成AIを活用していくことが見込まれ、実際に各社生成AIの活用が進んでいる（図表5）。生成AI活用によるバックオフィスのスリム化、人的リソースの配置の見直しが図られ、新たな成長機会を生み出していくことが期待される。
・国内金融機関の生成AI関連の投資額は、2023年の114億円から2028年に1,041億円まで拡大する見通しである（図表6）。

（図表3）銀行業で期待されるAIの付加価値



（図表4）生成AI導入によるタスクの自動化対象率（産業ベース）の一部抜粋

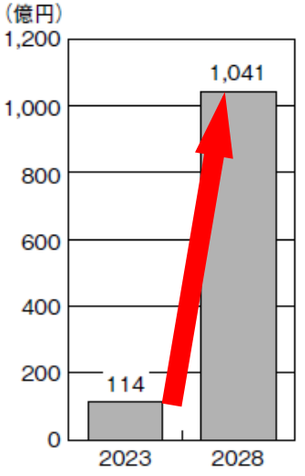


（注）職業別（小分類）の就業者数で加重平均した値。宗教家や音楽家など自動化対象率のデータが欠落している職業は除外した。

（図表5）国内金融機関の生成AIの活用事例

三菱UFJ銀行	生成人工知能（AI）を行内の事務の手続き照会や通達の添削など110を超える業務で導入する。年内に全行員に利用を開放して法律相談やメールの案文作成、リポートの要約などを含め順次対応する業務を増やす。回答の精度を向上させるため、生成AIが参照する行内情報の基盤を2024年度にも整備する。
三井住友銀行	SMBC全従業員に対し「生成AIアシスタント」を公開。生成AIを活用した「行内規定検索サービス」の実証実験を開始。
みずほ銀行	国内全社員が使える社内向けテキスト生成AIの導入。AIチャット導入による事務手続照会の時間の短縮をめざし概念実証を進行中。ワンクリックで稟議資料ドラフトを自動作成できる「稟議作成サポートAI」開発も進めている。
損保ジャパン	テストコードやテストパターンの生成を補助。ChatGPTのように大規模言語モデルをベースとして、日本語に特化した生成AIエンジンによるコールセンター支援システムを開発中。
SBI生命保険	ChatGPTのような大規模言語モデルを活用し、コールセンターへの適用などを想定。
住友生命保険	生成AI（人工知能）技術を生かし、主力保険商品の強化に取り組んでいる。ChatGPTの技術を基に開発したチャットシステム「Sumisei AI Chat Assistant」を使い、新たな企画づくりや日常業務の生産性向上に取り組み保険商品自体を強化。
クレジットエンジン	生成AIを活用した、延滞債権／支払に特化した督促交渉AI機能。返済や支払いに関する質問や相談、各種ヒアリングの実施し、連絡約束や入金約束を取得が可能。

（図表6）国内金融機関における生成AI関連の投資額の予測



出展：財務省広報誌「ファイナンス」2024 Apr. -金融機関における生成AIの成長性について

金融機関でのAI活用の悩み

AI推進部署

- 生成AIを業務に活用せよと言

生成AI活用の機運は高まっているから、システム部門に導入してもらおう

用すればペイできる？

IT部署

- 生成AIサービスってセキュア

どうすればいいか分からないから、いつものSIerに宜しくやってもらおう

- AI開発者向け社内クラウド基盤をIT部門主導で作りたいけど、どうしたらいいの？

SIer

- 社内で検討します！

(AIチームは忙しそうだな…)

~~(言っていることが宇宙人…)~~

(ChatGPTじゃ…だめか)

(Azure OpenAIとかかな…)

暗中模索

昨今のムーブメント

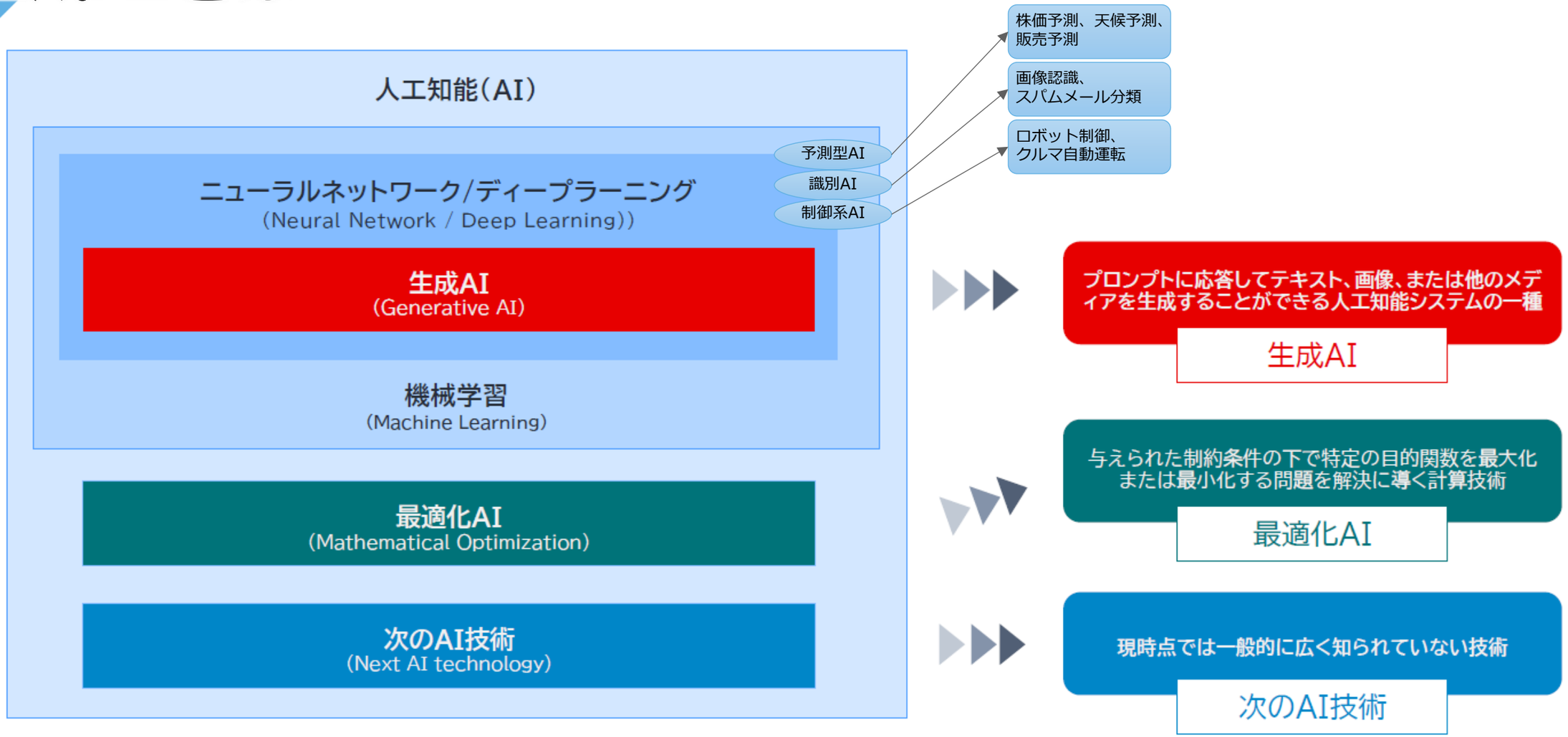
- 2022年12月 OpenAI社 ChatGPT(GPT3.5ベース)発表
  - ⇒ 第一の波、ChatGPTを業務活用しよう
- 2023年1月 Microsoft社 Azure OpenAI Service一般提供開始
- 2023年4月 AWS Amazon Bedrock発表
- 2023年7月 Google Vertex AIで生成AI対応を発表
  - ⇒ 第二の波、生成AIサービスを構築してみよう
  - ⇒ クラウドサービスでの構築の限界、オンプレ生成AI基盤へ
- Copilot+PC、Google Gemini、Apple Intelligence
  - ⇒ 第三の波、エンドポイントAI

生成AIと私

無限の未来と、幾千のテクノロジーをつなぐ。

CTC Financial Services Group

生成AIとは



- 出展：「CTCが提供する生成AI サービスのご紹介」 - ターゲット領域：「生成AI 」「最適化AI 」「次のAI 技術」

無限の未来と、幾千のテクノロジーをつなぐ。

CTC Financial Services Group

LLM(Large Language Model)とは

- Large Language Model：「大規模」＋「言語モデル」

- 言語モデル
  - **言語データを学習し**、ある単語の後に続くワードがどのくらいの確率で**出現するかを予測する**計算モデル

- 大規模
  - 機械学習に利用される**データ量**が大規模
  - 学習、推論に利用される**コンピュータの計算リソース**が大規模

参考）SSM/SLMも存在する

伊藤忠テクノソリューションズ株式会社

ENGLISH　アクセス　お問い合わせ

ニュース　会社情報　決算関連情報　サステナビリティ

生成AIを活用した投資のアドバイザリーソリューションを開発

高い専門知識を必要とする業務の効率化を目指す

PDF（228KB）　印刷する

2023年12月12日
伊藤忠テクノソリューションズ株式会社

伊藤忠テクノソリューションズ株式会社（代表取締役社長：柘植 一郎、本社：東京都港区、略称：CTC）は、金融業や製造業などの特定分野向けの生成AIを手掛ける米Aitomatic, Inc.（本社：米国カリフォルニア州サンフランシスコ、CEO：Christopher Cuong T. Nguyen、以下：AITOMATIC社）の製品を活用して、投資ポートフォリオの提案を行うAI投資アドバイザリーソリューションを開発しました。投資に関する質問や相談を対話形式で対応するソリューションで、今期から提供を目指します。

各企業では業務効率化、省人化、顧客満足度の向上などの課題解決のために生成AIの導入に取り組んでいます。しかし、効果的に活用していくためには、業務に適合するよう、一般的な知識に加え、特定の業界や領域に特化した専門知識を有する学習データの準備や継続的な改善及びチューニングなどが必要となります。

AI投資アドバイザリーソリューションは、AITOMATIC社の目的特化型の生成AIを活用して、投資を希

関連情報

金融機関のDXならCTC金融事業グループ

【Pick Up】事例・コラム

伊藤忠テクノソリューションズ（自社）

再エネ主力電源化を目指すアグリゲーションビジネスに取り組むCTC 各種機能・サービスを提供するデジタル基盤にAWSクラウドを採用

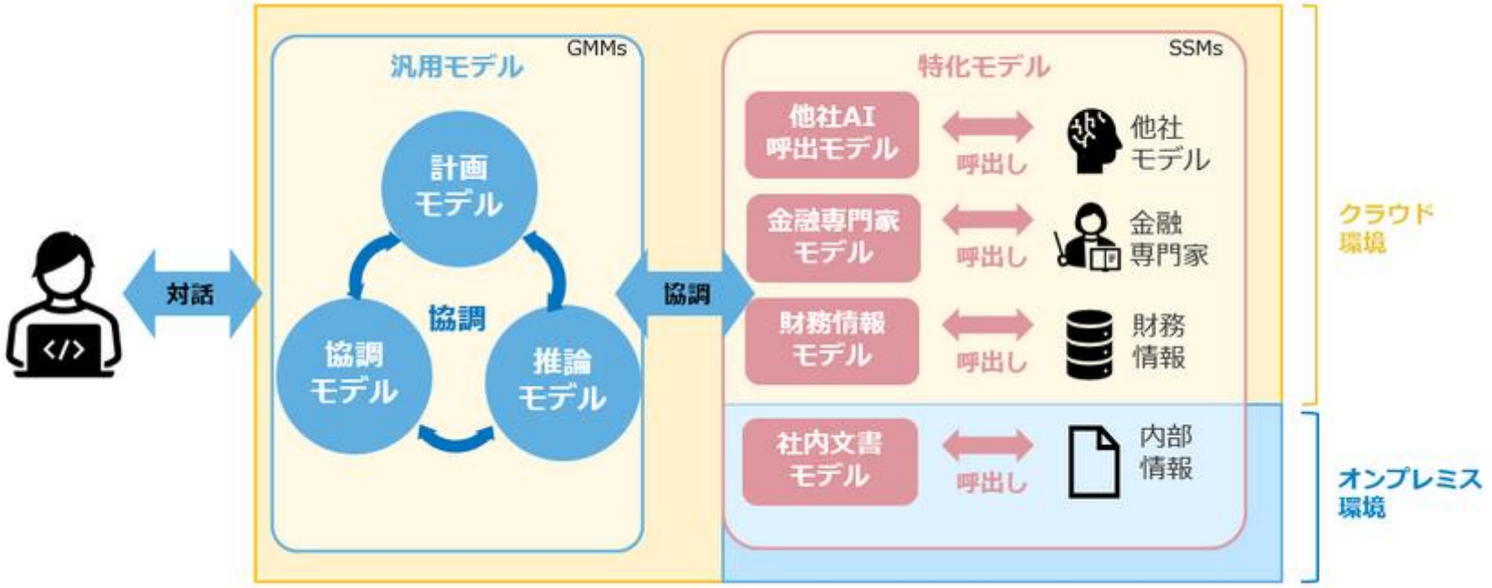
DX　クラウド　AI

チャットボットの導入事例

チャットボットを"社内の問合せ対応"に活用するメリットと導入事例

クラウド　AI

参照：https://www.ctc-g.co.jp/company/release/20231212-01664.html



- SSM：Small Specialist Model　：「目的特化型」＋「モデル」
- SLM：Small Language Model　：　「小規模」＋「言語モデル」

無限の未来と、幾千のテクノロジーをつなぐ。

CTC Financial Services Group

LLMの種別

ここまでのまとめ: オープンなLLM/クローズドなLLM

求められる性能や使いたい環境にあわせて
選択する必要がある

LLM種別	具体例	パラメータ	利用環境	性能
オープン	Llama 2 Falcon Qwen …	公開	場面に合わせて 柔軟に 利用可能	全体的にはやや ビハインドだが 一部には匹敵
クローズド	GPT-4 Bard（Gemini） Claude …	非公開 （サービス/API からのみ利用可）	インターネット 接続は必須	オープンLLMを 大きく上回るもの がいくつか存在

オープンなLLMとは

クローズドなLLMとは異なり、その中身（パラメータやソースコード）が
公開されているLLM

オープンなLLMの特徴: パラメータの公開

Hugging Face Hubのようなサイト上でダウンロードすることが可能
このため、計算機環境さえあれば直接使える

オープンなLLMの特徴: 透明性

パラメータの公開のほか、モデルによっては学習データ等の詳細が
明らかにされていることも多い

Llama 2については論文で詳細に説明されている

また、学習中の各種ログが公開されているケースも

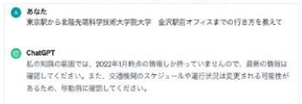
オープンなLLMの特徴: 場合によってはクローズドLLMより安価

解きたい課題の種類・難易度などに依存するが、
大幅にコストダウンできる場合もある

クローズドなLLMとは

OpenAIのChatGPTやGoogleのBardなど、
当該サービス/APIでの利用が前提のLLM

ChatGPT（OpenAI）



Bard（Google）

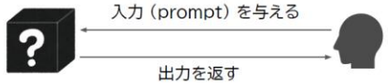
クローズドなLLMの特徴: 性能の高さ

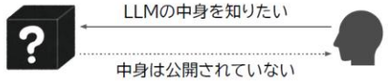
ChatGPTは2022年11月にリリースされたが、バージョンアップを
重ねつつ、現時点（2024年1月）でも最高性能を維持している

クローズドなLLMの特徴: ブラックボックス

当該サービス/APIでの提供となるため、
その中身を利用者が知ることは不可能

またその仕様上、インターネット接続が事実上必須



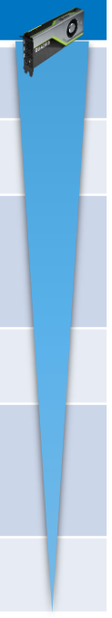


※ ちなみにクローズドなLLMは
技術詳細についても完全には論文化
されていないことが多い

- 出展：JAIST産学官共創フォーラム・令和５年度 第３回定期講演会 ELYZA オープンなLLM（大規模言語モデル）の発展と、それを活かした日本語LLMの開発について

LLM開発フェーズ

開発フェーズ		概要	GPU
Pre-Training	事前学習	LLMフルスクラッチ開発(基盤モデル)	数百～
	継続事前学習	基盤モデルに大量データで追加学習	
Post-Training	指示学習	教師・指示データを元にユーザが 求める出力になるよう調整	数十～
	微調整/Fine-Tuning	個別タスクに合わせて追加学習	
Inference	情報参照/RAG	情報参照を利用してより正確な 応答を目指す	
	推論	LLMの利用	数枚～



Ex.

基盤モデル	継続事前学習モデル
Llama 2 70B（Meta）	ELYZA-japansese-Llama-2 70B（ELYZA） Swallow 70B（東工大） KARAKURI LM（カラクリ）
Mistral-7B（Mistral）	Japanese Stable LM Gamma 7B（Stability AI） Karasu（Lightblue）

基盤モデル(独自開発)
LLM-jp-13B（LLM-jp）
PLaMo-13B（Preferred Networks）
CALM2（サイバーエージェント）

無限の未来と、幾千のテクノロジーをつなぐ。

CTC Financial Services Group

RAG(Retrieval Augmented Generation)とは

- RAG：検索拡張生成

- LLMに外部の情報源から**検索したデータ**を取得させ、その情報を元に**回答の文章を要約**させる技法
  - 学習済みのデータに含まれないが、ユースケースに必要となる情報を都度検索させることで補完
    ex) 今日の株価、行内規約、社内用語
- モデルに情報を追加で獲得させるという目的では微調整/Fine-Tuningもあるが、、、
  - 追加学習には時間、コストが大量に必要
  - 「最新の情報を踏まえて回答」の用途には適さない

  ⇒ 様々なRAG手法が登場しており、Fine-Tuningより**RAGで回答精度を高めるのがトレンド(らしい)**

参考）RAGを活用した推論の実装例

伊藤忠テクノソリューションズ株式会社

ENGLISH　アクセス

ニュース　会社情報　決算関連情報　サステナビリティ

生成AIの利用環境を短期間で構築

Azure OpenAI Serviceの環境が最短2週間で導入可能

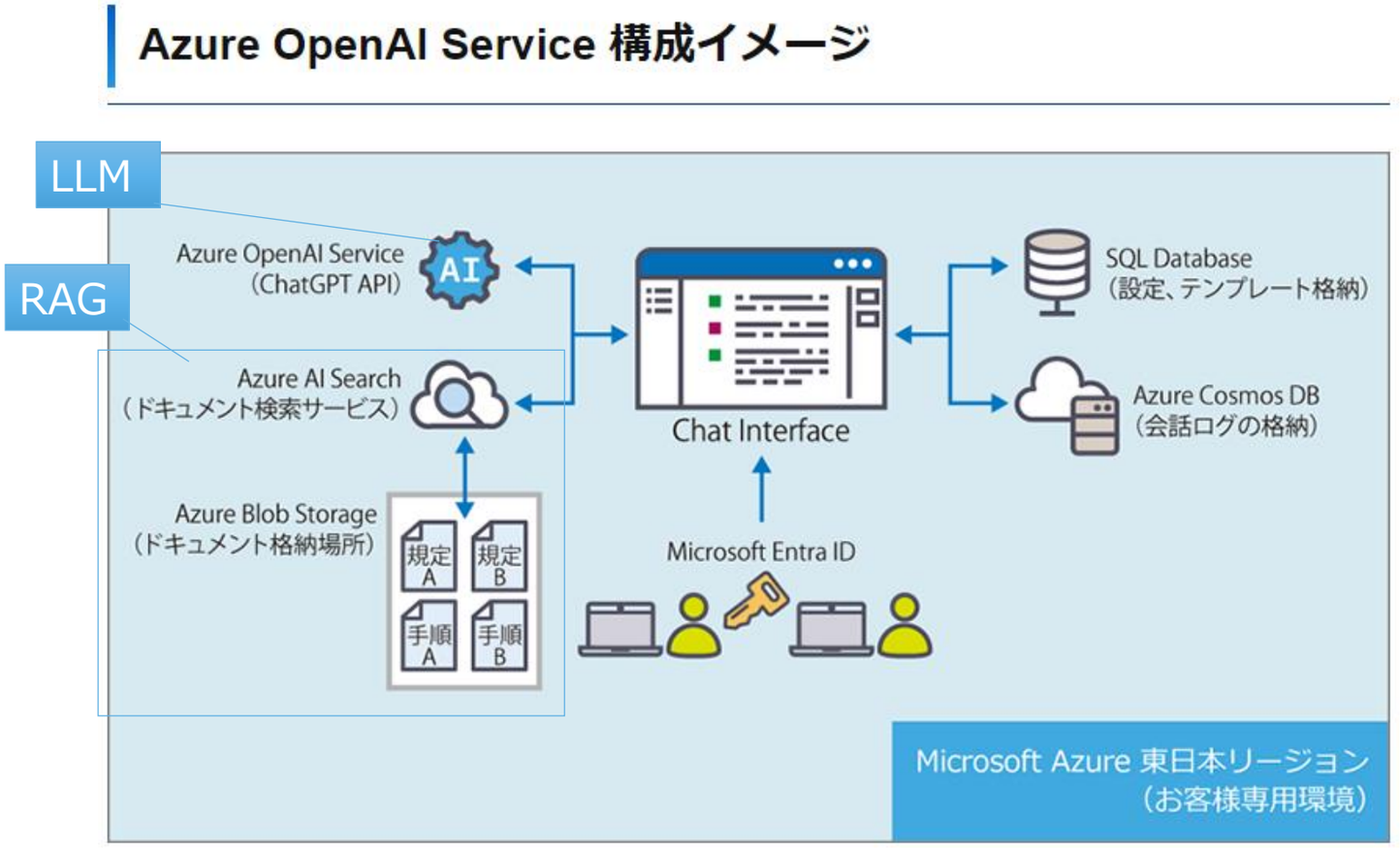
PDF（268KB）　印刷する

2024年10月15日
伊藤忠テクノソリューションズ株式会社

伊藤忠テクノソリューションズ株式会社（代表取締役社長：新宮 達史、本社：東京都港区、略称：CTC）は、マイクロソフトの生成AIクラウドサービス「Azure OpenAI Service」の利用環境を短期間で構築する「Azure OpenAI Serviceクイック導入パッケージ」を本日から提供します。生成AIの業務利用を検討するお客様に向け、最短2週間で環境を構築するサービスです。価格は50万円（税抜）で、3年間で60社の導入を目指します。

CTCは2023年からAzure OpenAI Serviceの導入から実装までをカバーするコンサルティングサービス「生成AIアドバイザリサービス」や、質問や回答の会話、利用した社内文書の記録（ログ）から回答の精度向上につなげる「生成AIデータ分析サービス」を提供しています。検証やトライアル目的での簡易な利用や短期間かつ安価な環境の構築などの要望を受け、両サービスに加えて新たにAzure OpenAI Serviceクイック導入パッケージを開始します。

Azure OpenAI Serviceクイック導入パッケージは、申し込みから利用開始まで最短2週間で実現する、



参照：https://www.ctc-g.co.jp/company/release/20241015-01807.html

無限の未来と、幾千のテクノロジーをつなぐ。

CTC Financial Services Group

オンプレ生成AI基盤

無限の未来と、幾千のテクノロジーをつなぐ。

CTC Financial Services Group

生成AI基盤を構築するためのポイント

自社の**付加価値**、**差別化要素**を組み込むならオンプレ

クラウド上で**個人情報**や**機密情報**を扱うのに障害があるならオンプレ

	クラウド	オンプレ
求めるサービス・基盤	GPT-4やPaLM2など既に存在するクラウドサービス	自社オリジナル 自分たちでサービスを構築
カスタマイズの柔軟性	提供されているサービスの範囲で	技術的に可能な範囲で自由
費用の考え方	トークン毎、文字数単位で 従量費用が発生する	導入時に初期費用が発生する
データ活用のための セキュリティ	サービス提供者のポリシーに準拠	自社のポリシーに準拠 安心

一般的な情報を元に
他社・他者に追いつくなら
クラウドサービス
ex) 翻訳、コード生成

無限の未来と、幾千のテクノロジーをつなぐ。

CTC Financial Services Group

再掲

LLM開発フェーズ

開発フェーズ		概要	GPU
Pre-Training	事前学習	LLMフルスクラッチ開発(基盤モデル)	数百～
	継続事前学習	基盤モデルに大量データで追加学習	
Post-Training	指示学習	教師・指示データを元にユーザが 求める出力になるよう調整	数十～
	微調整/Fine-Tuning	個別タスクに合わせて追加学習	
Inference	情報参照/RAG	情報参照を利用してより正確な 応答を目指す	数枚～
	推論	LLMの利用	

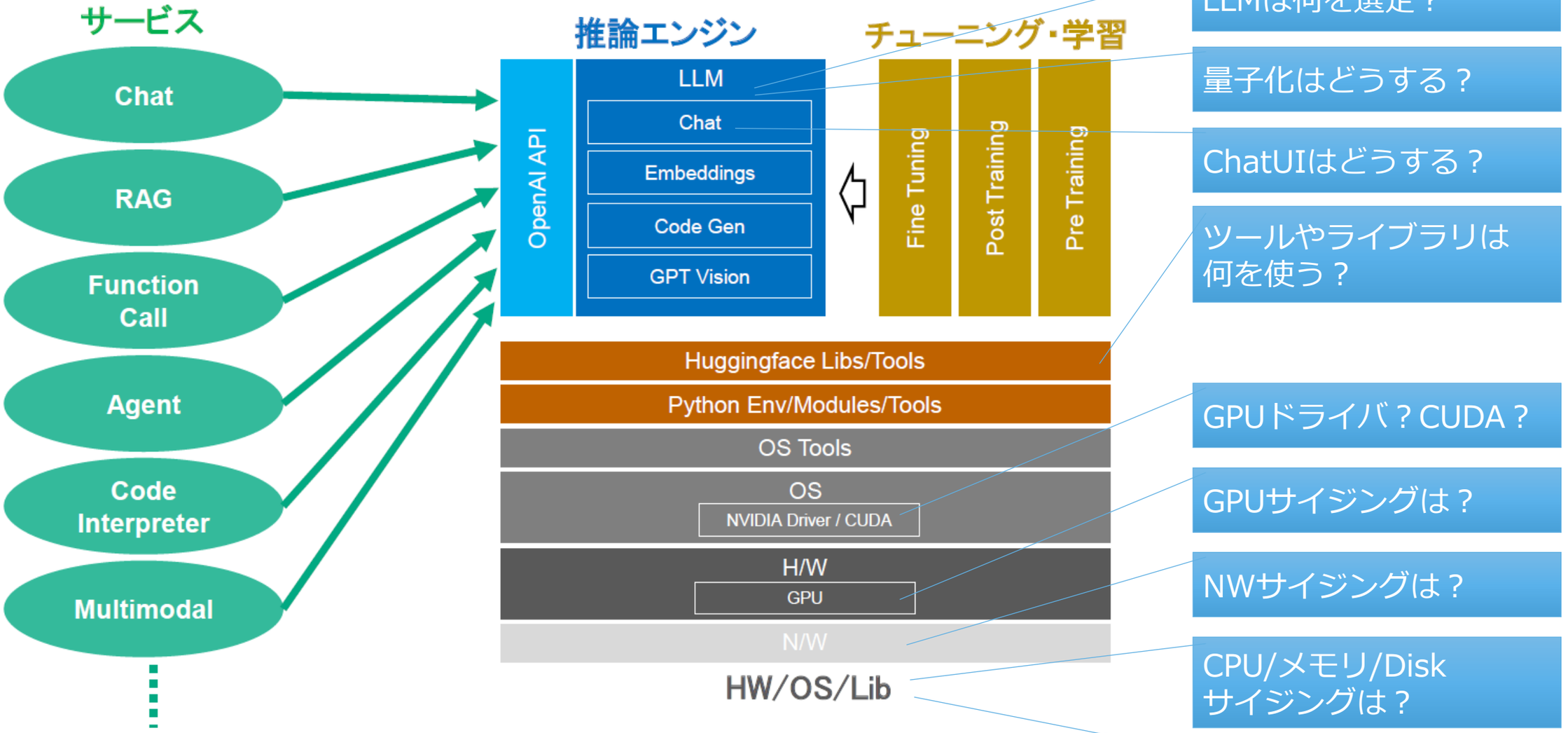
生成AI基盤で**何をする**？
・RAG？Fine-Tuningも？
・事前学習から？

Ex.

基盤モデル	継続事前学習モデル
Llama 2 70B（Meta）	ELYZA-japansese-Llama-2 70B（ELYZA） Swallow 70B（東工大） KARAKURI LM（カラクリ）
Mistral-7B（Mistral）	Japanese Stable LM Gamma 7B（Stability AI） Karasu（Lightblue）

基盤モデル(独自開発)
LLM-jp-13B（LLM-jp）
PLaMo-13B（Preferred Networks）
CALM2（サイバーエージェント）

オンプレLLM – 基本構成



無限の未来と、幾千のテクノロジーをつなぐ。

CTC Financial Services Group

LLM初期導入役割分担

タスク		AI推進部	開発部	基盤部	運用部
AI基盤	サーバ構築	-	-	○	-
	コンテナ/k8sインストール	-	-	△	-
	ミドルウェア/フレームワークインストール	-	-	○	-
LLM	LLMインストール	○	-	△	-
	フロントエンド(チャットアプリなど)構築	-	○	△	-
RAG	ベクトルDB構築(ドキュメントのDB化)	○	-	-	-
	エンタープライズ検索エンジン構築	-	-	○	-
	RAG設定(情報のリランキング、キーワード抽出、質問の言い換え、インデックス化、など)	○	△	-	-
Fine-Tuning	教師データの作成	△	○	-	-
	LLMモデルのチューニング	○	△	-	-
Pre-Training	○○○	○	△	-	-

キーパーソンを洗い出し、役割分担を整理するのは大事

運用工程も同様

Fine-Tuningなど追加学習を継続的に実施する場合は更に複雑

無限の未来と、幾千のテクノロジーをつなぐ。

CTC Financial Services Group

さいごに

無限の未来と、幾千のテクノロジーをつなぐ。

CTC Financial Services Group

さいごに

- 生成AIのクラウドサービスやそれを活用する支援サービスは無数存在し、SIerに加えAIコンサルやAIベンチャーなど競合も多数です
- 用途に応じてそれぞれ使い分けをすることが大事になります

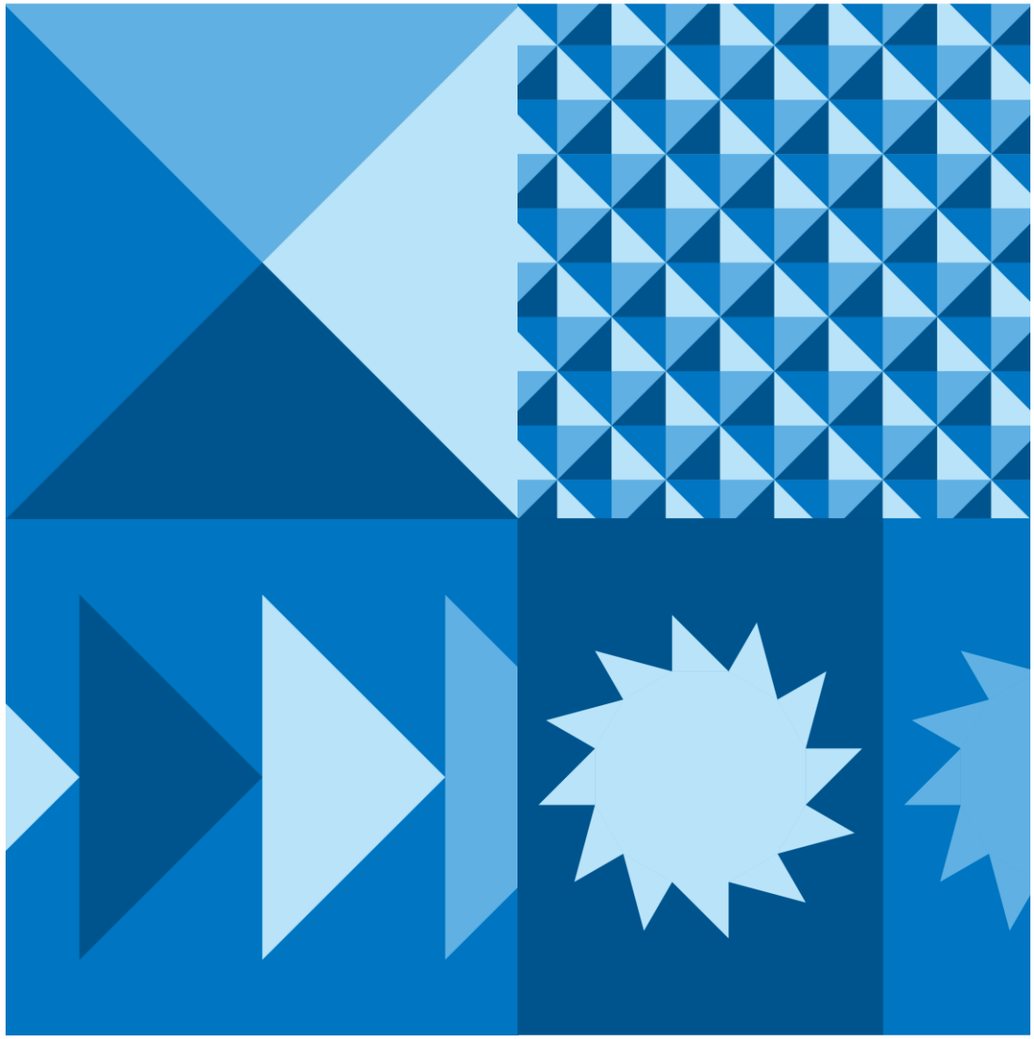
- 「オンプレ生成AIを構築する」というニーズは一定存在する一方、それを金融機関向けのクオリティで実現できるのは…極少数

生成AI界隈で迷える人の一助になれば幸いです

無限の未来と、
幾千のテクノロジーをつなぐ。

CTC Financial Services Group

CTC
Challenging Tomorrow's Changes



Appendix

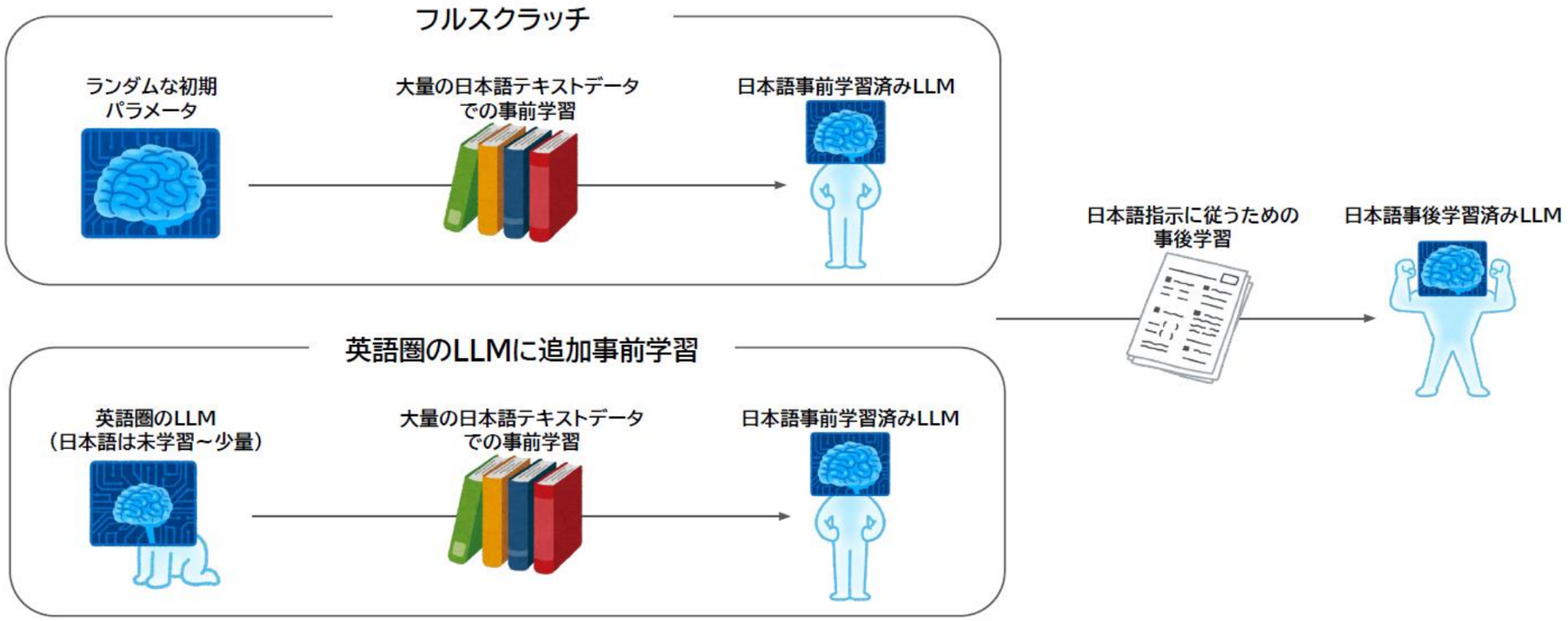
無限の未来と、幾千のテクノロジーをつなぐ。

CTC Financial Services Group

日本語LLM

日本語LLMを作るにあたっての方向性

特に事前学習（基本的な言語能力を得るための学習）
について、大まかに2つの方向性が存在



無限の未来と、幾千のテクノロジーをつなぐ。

CTC Financial Services Group

- 出展：JAIST産学官共創フォーラム・令和５年度 第３回定期講演会 ELYZA オープンなLLM (大規模言語モデル) の発展と、それを活かした日本語LLMの開発について

日本語LLM：ELYZA

ELYZA-japanese-Llama-2-13b

Llama 2（13B）をもとに日本語を追加学習したLLM

評価用データセットELYZA-tasks-100（後述）では
GPT-3.5（text-davinci-003）を上回る性能

モデル	スコア	開発元	パラメータ数	Open / Closed	商用利用
Claude 2.1	3.84	Anthropic	不明	Closed	-
Gemini Pro	3.72	Google	不明	Closed	-
ELYZA-japanese-Llama-2-13b-instruct	3.01	ELYZA	13B	Open	○
Qwen-14B-Chat	2.78	Alibaba	14B	Open	○
GPT-3.5 (text-davinci-003)	2.77	OpenAI	175B	Closed	-
ELYZA-japanese-Llama-2-13b-fast-instruct	2.73	ELYZA	13B	Open	○
calm2-7b-chat	2.63	CyberAgent	7B	Open	○
japanese-stablelm-instruct-beta-70b	2.62	Stability AI	70B	Open	○
Swallow-70b-instruct	2.50	TokyoTech-LLM	70B	Open	○
nekomata-14b-instruction	2.50	rinna	14B	Open	○
Swallow-13b-instruct	2.34	TokyoTech-LLM	13B	Open	○
nekomata-7b-instruction	2.23	rinna	7B	Open	○
Swallow-7b-instruct	2.22	TokyoTech-LLM	7B	Open	○
youri-7b-instruction	2.00	rinna	7B	Open	○
japanese-stablelm-instruct-gamma-7b	1.87	Stability AI	7B	Open	○
plamo-13b-instruct	1.49	Preferred Networks	13B	Open	○
japanese-stablelm-instruct-beta-7b	1.43	Stability AI	7B	Open	○
llm-jp-13b-instruct-full-jaster-v1.0	1.31	LLM-jp	13B	Open	○

ELYZA-japanese-CodeLlama-7b

Code Llama（7B）をもとに日本語を追加学習したLLM

コード生成およびコード補完に関する能力に特化

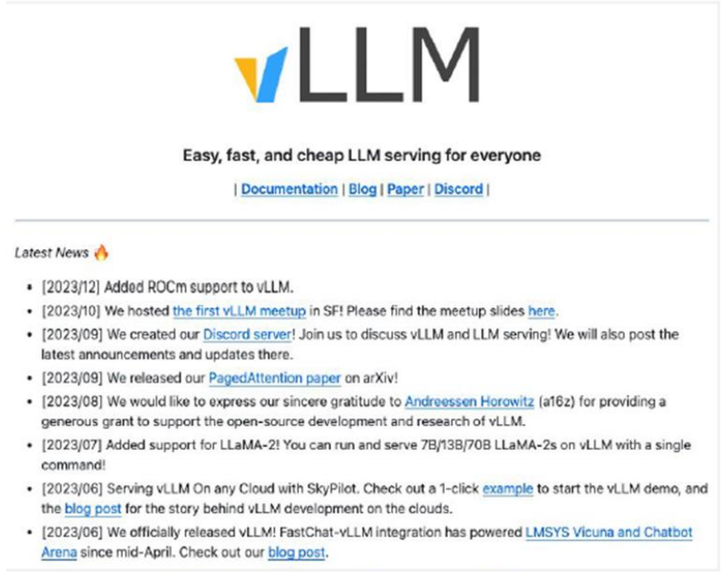
無限の未来と、幾千のテクノロジーをつなぐ。

CTC Financial Services Group

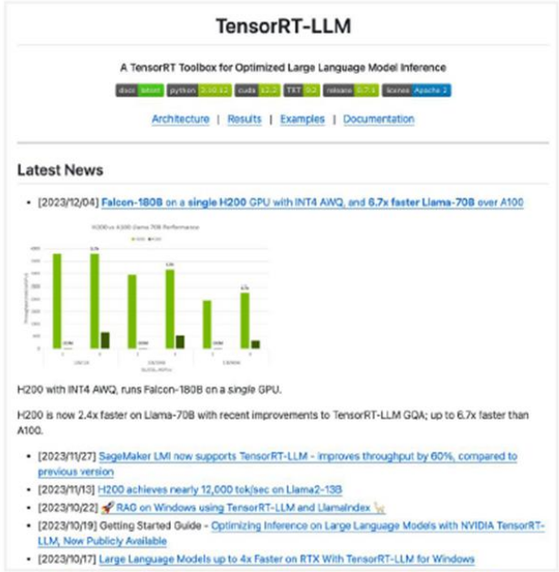
- 出展：JAIST産学官共創フォーラム・令和５年度 第３回定期講演会 ELYZA オープンなLLM (大規模言語モデル) の発展と、それを活かした日本語LLMの開発について

推論の効率化技術

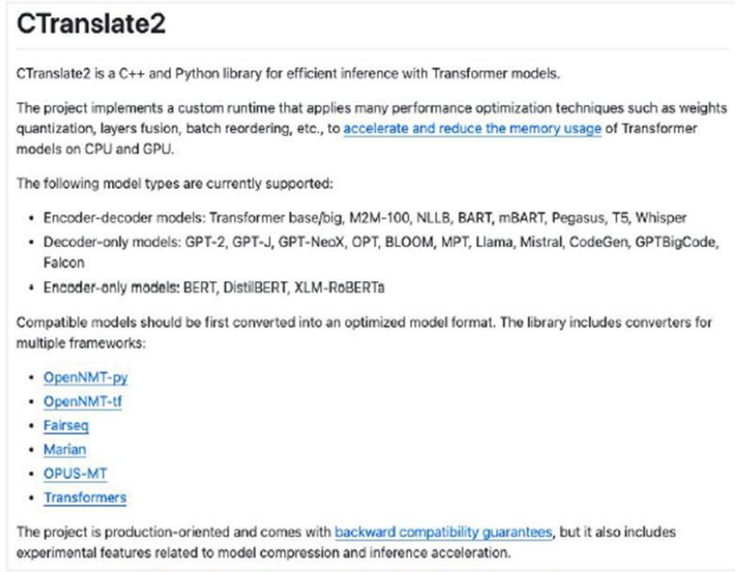
推論を高速化したりスループットを高めたりといった、
効率化のためのライブラリも発展しつつある



https://github.com/vllm-project/vllm



https://github.com/NVIDIA/TensorRT-LLM



https://github.com/OpenNMT/CTranslate2

弊社が先日公開した以下デモでも活用しています
https://huggingface.co/spaces/elyza/ELYZA-japanese-Llama-2-13b-instruct-demo

無限の未来と、幾千のテクノロジーをつなぐ。

CTC Financial Services Group

- 出展：JAIST産学官共創フォーラム・令和５年度 第３回定期講演会 ELYZA オープンなLLM (大規模言語モデル) の発展と、それを活かした日本語LLMの開発について

量子化

巨大なモデルをより少ない計算リソースで学習するための手法
LoRA/QLoRAが特に流行中

LoRAはもとの巨大なパラメータを直接更新せずに
別途少量のパラメータを学習することで効率化

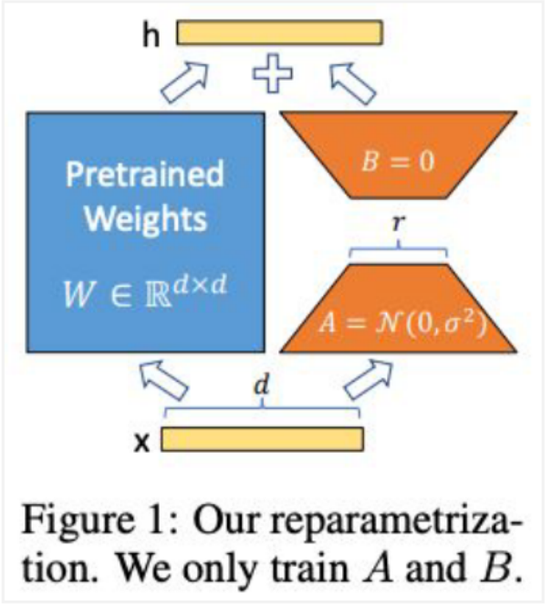


Figure 1: Our reparametrization. We only train A and B.

https://arxiv.org/abs/2106.09685

QLoRAはそれに加えて、パラメータを量子化
するなどの工夫により更に効率化を実現

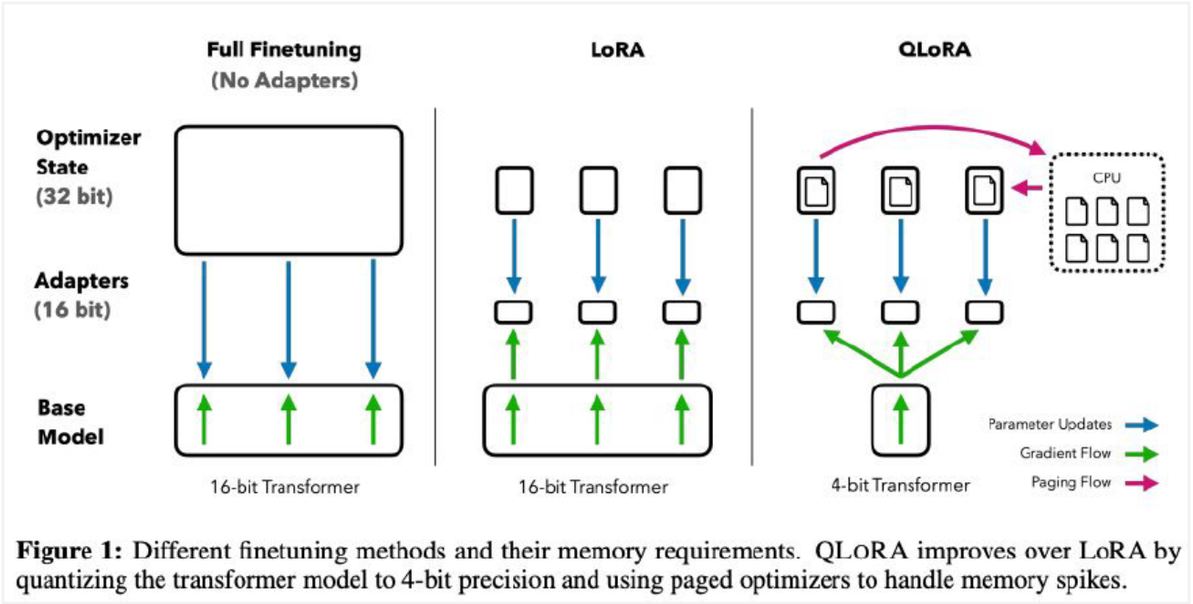


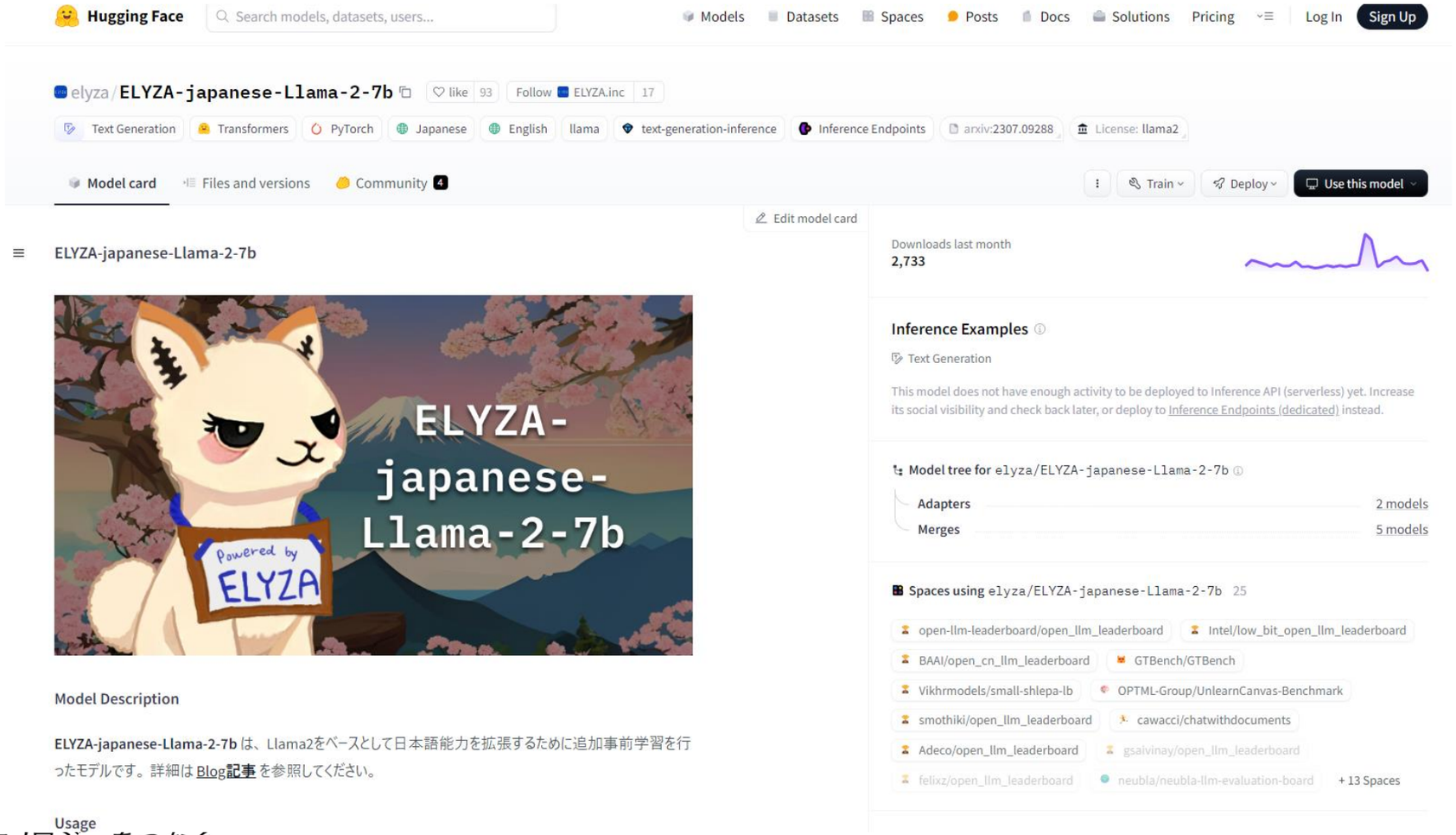
Figure 1: Different finetuning methods and their memory requirements. QLoRA improves over LoRA by quantizing the transformer model to 4-bit precision and using paged optimizers to handle memory spikes.

https://arxiv.org/abs/2305.14314

- 出展：JAIST産学官共創フォーラム・令和５年度 第３回定期講演会 ELYZA オープンなLLM (大規模言語モデル) の発展と、それを活かした日本語LLMの開発について

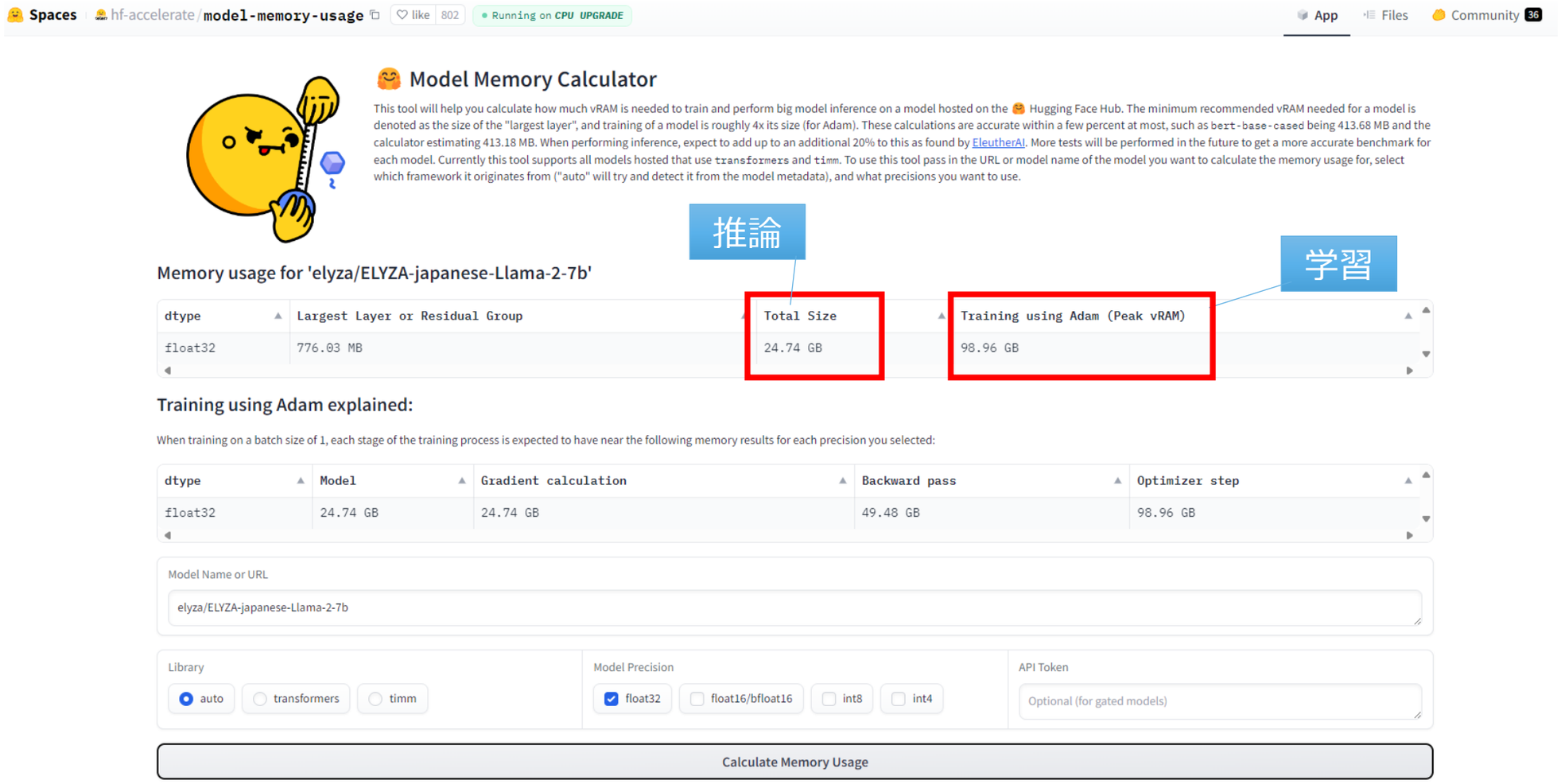
Hugging Face

Transformersをベースに機械学習モデルの開発と共有、公開をするためのプラットフォーム



参照：https://huggingface.co/elyza/ELYZA-japanese-Llama-2-7b

Hugging Face – Model Memory Calculator



無限の未来と、幾千のテクノロジーをつなぐ。

CTC Financial Services Group

参照：https://huggingface.co/spaces/hf-accelerate/model-memory-usage

NVIDIA H100 Tensor コア GPU

製品仕様

フォーム ファクター	H100 SXM	H100 PCIe	H100 NVL[1]
FP64	34 teraFLOPS	26 teraFLOPS	68 teraFLOPs
FP64 Tensor コア	67 teraFLOPS	51 teraFLOPS	134 teraFLOPs
FP32	67 teraFLOPS	51 teraFLOPS	134 teraFLOPs
TF32 Tensor コア	989 teraFLOPS[2]	756 teraFLOPS[2]	1,979 teraFLOPs[2]
BFLOAT16 Tensor コア	1,979 teraFLOPS[2]	1,513 teraFLOPS[2]	3,958 teraFLOPs[2]
FP16 Tensor コア	1,979 teraFLOPS[2]	1,513 teraFLOPS[2]	3,958 teraFLOPs[2]
FP8 Tensor コア	3,958 teraFLOPS[2]	3,026 teraFLOPS[2]	7,916 teraFLOPs[2]
INT8 Tensor コア	3,958 TOPS[2]	3,026 TOPS[2]	7,916 TOPS[2]
GPU メモリ	80GB	80GB	188GB
GPU メモリ帯域幅	3.35TB/秒	2TB/秒	7.8TB/秒[3]
デコーダー	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG	14 NVDEC 14 JPEG
最大熱設計電力 (TDP)	最大 700W (構成可能)	300-350W (構成可能)	2x 350-400W (構成可能)
マルチインスタンス GPU	最大 7 個の MIG @ 10GB		各 12GB の最大 14 のMIG
フォーム ファクター	SXM	PCIe デュアルスロット空冷	2x PCIe デュアルスロット空冷
相互接続	NVLink: 900GB/秒 PCIe Gen5: 128GB/秒	NVLINK: 600GB/秒 PCIe Gen5: 128GB/秒	NVLink: 600GB/秒 PCIe Gen5: 128GB/秒
サーバー オプション	4 または 16 GPU 搭載の NVIDIA HGX™ H100 パートナーおよび NVIDIA-Certified Systems™ 8 GPU 搭載の NVIDIA DGX™ H100	1～8 GPU 搭載のパートナーおよび NVIDIA Certified Systems™	2-4 組のパートナーおよび NVIDIA Certified Systems
NVIDIA AI Enterprise	アドオン	含む	含む

1. 参考仕様。仕様は変更される場合があります。H100 NVL PCIe カード 2 枚と NVLink Bridge を組み合わせた場合の仕様です。
2. 疎性あり。
3. HBM 帯域幅の総計

無限の未来と、幾千のテクノロジーをつなぐ。

CTC Financial Services Group

参照：https://www.nvidia.com/ja-jp/data-center/h100/

NVIDIA-Certified Systems Configuration Guide

2. Configurations

2.1. Inference System Configurations

Inference application performance is greatly accelerated with the use of NVIDIA GPUs and includes workloads such as:

- Large Language Model Inference
- Natural Language Recognition (NLR)
- Omniverse applications
- DeepStream – GPU-accelerated Intelligent Video Analytics (IVA)
- NVIDIA® TensorRT™, Triton – inference software with GPU acceleration

A GPU server designed for executing inference workloads can be deployed at the edge or in the data center. Each server location has its own set of environmental and compliance requirements. For example, an edge server may require NEBS compliance with more stringent thermal and mechanical requirements.

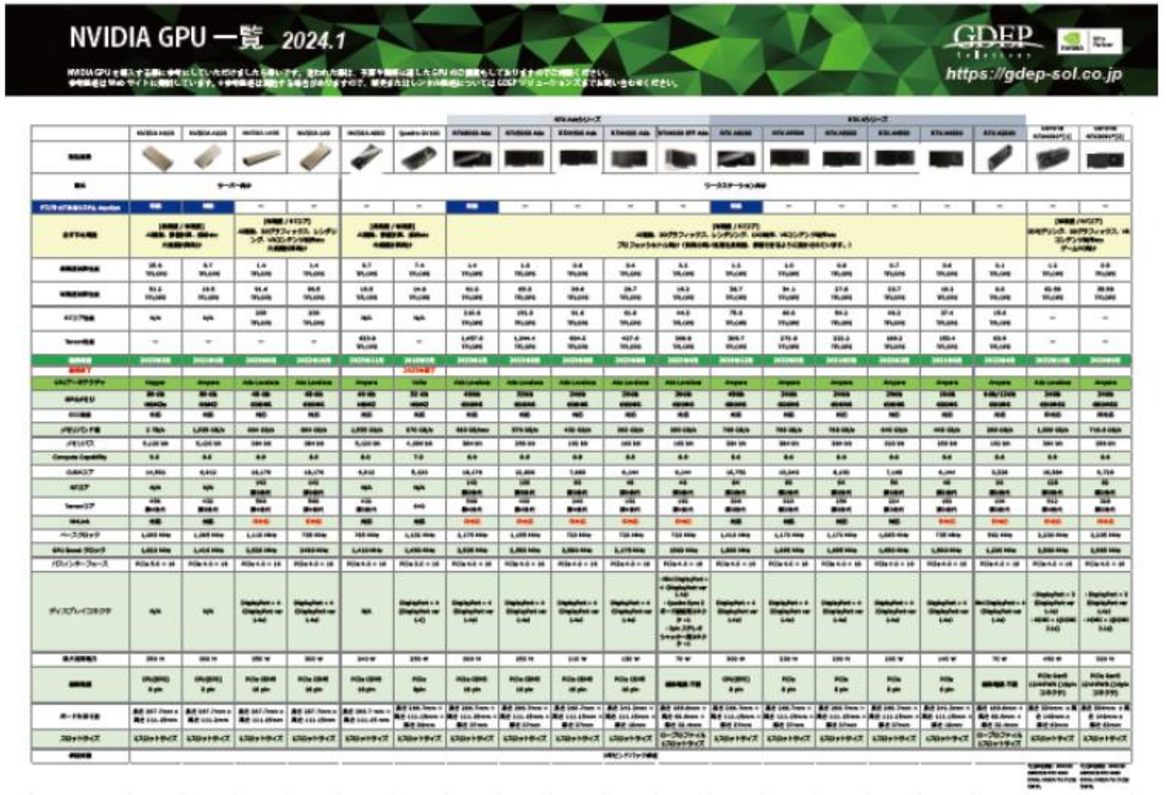
Table 1 provides the system configuration requirements for an inference server using NVIDIA GPUs. Large Language Models should target the higher-end specs. Omniverse and visualization application usage will need L40S/L40.

Table 1. Inference Server System Configuration

Parameter	Inference Server Configuration
GPU	L40S L40 L4 H100 H100 HGX
GPU Configuration	2x / 4x / 8x GPUs per server 4x is recommended to remove the need for a PCIe switch. GPUs should be balanced across CPU sockets and root ports.
CPU	x86 PCIe Gen5 capable CPUs are recommended, such as Intel Xeon scalable processor (Sapphire Rapids) or AMD Genoa.
CPU Sockets	2 CPU sockets minimum
CPU Speed	2.1 GHz minimum base clock
CPU Cores	6x physical CPU cores per GPU
System Memory	Minimum 1.5x of total GPU memory / 2.0x is recommended. Evenly spread across all CPU sockets and memory channels.
DPU	One Bluefield®-3 DPU per server
PCI Express	Minimum of one Gen5 x16 link per Gen5 GPU is recommended. Minimum of one Gen4 x16 link per Gen4 GPU is recommended. Minimum of one Gen5 x16 link per 2x GPUs for PCIe Switch configurations.
PCIe Topology	For balanced PCIe architecture, GPUs should be evenly distributed across CPU sockets and PCIe root ports. NICs and NVMe drives should be placed within the same PCIe switch or root complex as the GPUs. It's important to note that a PCIe switch may be optional for cost-effective inference servers.
PCIe Switches	Direct CPU attach is preferred. ConnectX®-7 Gen5 PCIE Switches as needed.
Network Adapter (NIC)	ConnectX®-7 (up to 400 Gbps) BlueField®-3 DPU in NIC mode (up to 400 Gbps). See Section **Network** for details.
NIC Speed	Up to 400 Gbps per GPU

参照：https://docs.nvidia.com/certification-programs/nvidia-certified-configuration-guide/index.html#configurations

NVIDIA-Certified Systems Configuration Guide



NVIDIA GPU 一覧 ダウンロードフォーム

無限の未来と、幾千のテクノロジーをつなぐ。

CTC Financial Services Group

参照：https://gdep-sol.co.jp/news-list-nvidia-gpus/

弊社実績事例(3)_サービス業様

大規模AI開発基盤検討時の課題、要望

- 最新GPUを利用しても処理能力が足りない
- 常にGPUリソースを使い倒したい
- 処理が終わったのち自動的に次の処理を実施したい
- 大規模GPUインフラの設計・導入・運用のノウハウが無い
- 不具合発生時、システム全体を見て相談できる窓口が無い

大規模AI開発基盤 NVIDIA DGX SuperPOD導入による効果

- 短い期間で大規模AI開発基盤導入が完了　システムの追加もスムーズに
- テイクオフトレーニングやJobスケジューラの導入によりGPUの利用率向上ができた
- TAMの的確なアドバイスや障害発生時の一元窓口対応により、問題に即座に対応することができるように

プラン／デプロイ

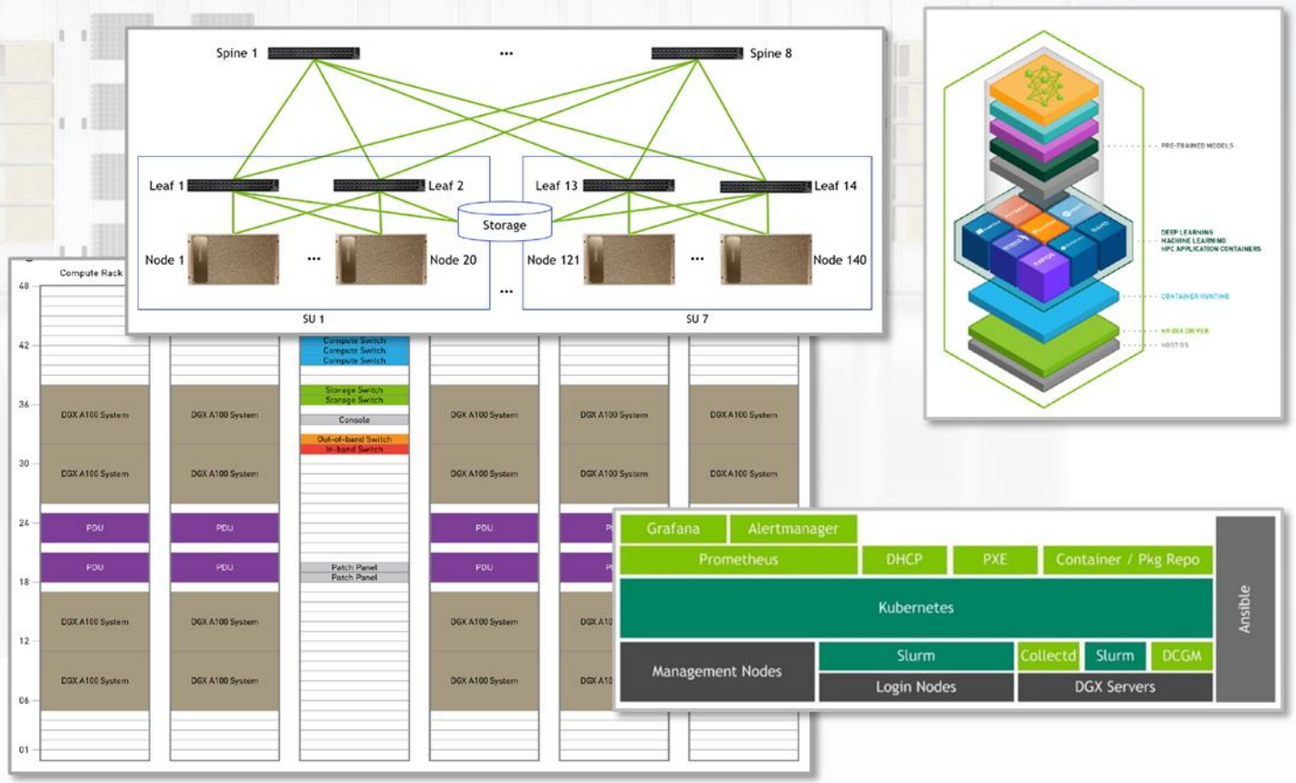
- キャパシティプランニング
- データセンターのデザイン
- パフォーマンスの予測
- サイトの評価／準備
- インストール
- インストール後のテスト
- プロビジョニング／管理

トレーニング／最適化

- アプリケーションのパフォーマンステスト
- サイトのドキュメント一式
- ユーザー／DevOps のトレーニング
- ワークロードベースの DLI
- カスタム システムの手引書
- 受け渡しセッション

ハードウェア／ソフトウェア
- 100 PFLOPS AI システム
- 20 DGX A100 システム
- 1.6 PB DDN ハイパフォーマンス ストレージ
- 200Gbps Mellanox ネットワーキング ファブリック
- CUDA-X／DGX ソフトウェア スタック
- cnvrg.io の MLOps ツール

サービス／サポート
- 資格を取得した優秀な NVIDIA 認定パートナーと NVIDIA アドバンスト サービス チームの共同体制
- NVIDIA エンタープライズ サポートによる継続的なメンテナンス
- オンプレミスまたは DGX 対応のデータセンターにデプロイ



無限の未来と、幾

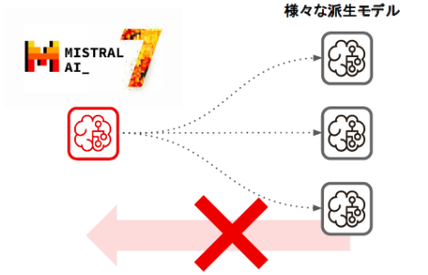
CTC Financial Services Group

生成AI真のオープンソース化 by Red Hat

現在のLLMコミュニティの課題

先進的なスタートアップの取り組みやHugging Faceの流行により、高性能なモデルが商用利用可能なライセンスで公開されていますが、そうしたモデルへの貢献や、コミュニティを通じた開発についてはまだ方法が確立されておらず、一般的なOSS開発とは異なる状況にあります。

アップストリームへの貢献



オープンなモデルをベースに様々な派生モデルが作られているがその成果がアップストリームのモデルに還元されていない。

学習データ作成のノウハウ

Model Series	ライセンス	学習データ
Llama3	META LLAMA 3 COMMUNITY LICENSE	非公開
Mistral	Apache 2.0	非公開

モデル自体はオープンであっても、学習データの作成はノウハウが秘匿され、他のソフトウェアのようなコミュニティによる開発は限定的。

オープンソースが AIのポテンシャルを解放する

初日の基調講演ではCEO Matt HicksからAI領域においてもオープンソースモデルがイノベーションを加速しており、
Red HatもLLM Granite のオープンソース化、InstructLabの公開を通じて、オープンソースエコシステムを強力に支援すると共に、
オープンソースの取り組みをビジネス価値に繋げていくという方針が語られました。

CEOのMatt Hicksは基調講演の中で、現在の LLMコミュニティではモデル自体は公開されているものの、そうしたモデルへのコントリビュートやファインチューニングが一部の人々に限定される点を指摘し、 Red Hatはこういた状況をオープンソースモデルにより誰でも貢献できるように変えると述べた。

1 IBM Researchと共同でGraniteをオープンソース化

対話やコーディング支援を行う LLMであるGranite FamilyをApache 2.0ライセンスにてHugging Face上で公開。利用した学習データについても開示。

2 InstructLabによるオープンなLLM開発モデル

LLMの開発を一部の組織のみがリードするのではなく、他のオープンソースソフトウェア開発と同じように誰でもコントリビュートできるようにするためのプロジェクト。
（=>次ページで紹介）

V0000000 Red Hat

無限の未来と、幾千のテクノロジーをつなぐ。

CTC Financial Services Group

参照：Red Hat Summit 2024 Recap