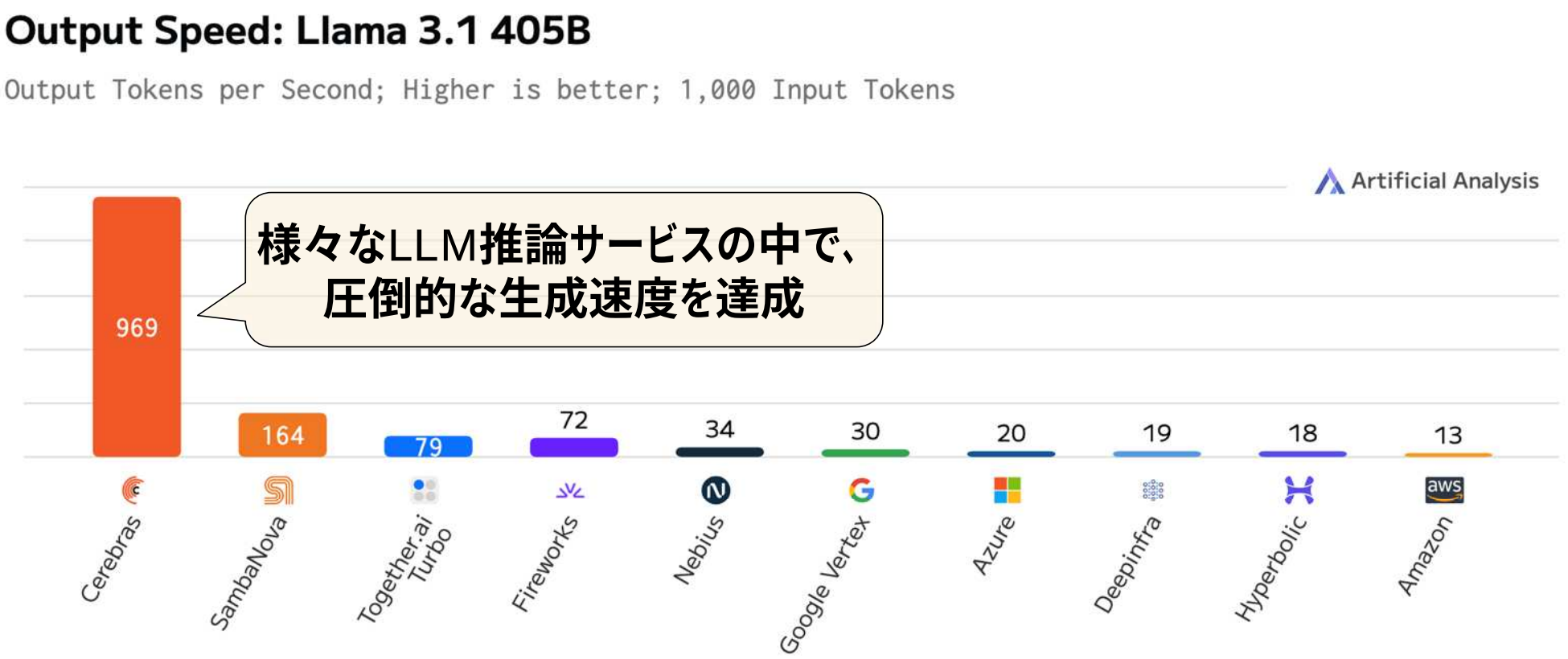
Cerebras CS-3で実現する LLM高速推論

東京エレクトロン デバイス株式会社

CNBU
CN技術本部
システムエンジニアリング部
中田 友康

Cerebrasは圧倒的に高速な推論を実現



Cerebras Inference measured on a private endpoint optimized for 16K context.

https://cerebras.ai/blog/cerebras-inference-3x-faster

本日お話しすること

■そもそもCerebras CS-3とは何か

■何故、推論を高速化できるのか

■高速推論のデモ

■もちろん学習に使うこともできます

自己紹介：中田 友康

Connect Beyond

当社が販売しているCerebrasとNVIDIAのAIエンジニアです。

cerebras

NVIDIA

製品技術に関するブログも書いています。ご興味がある方は以下のURLから

https://cn.teldevice.co.jp/blog/

東京エレクトロンデバイス　ブログ　検索

東京エレクトロン デバイス

Cerebras CS-3のご紹介

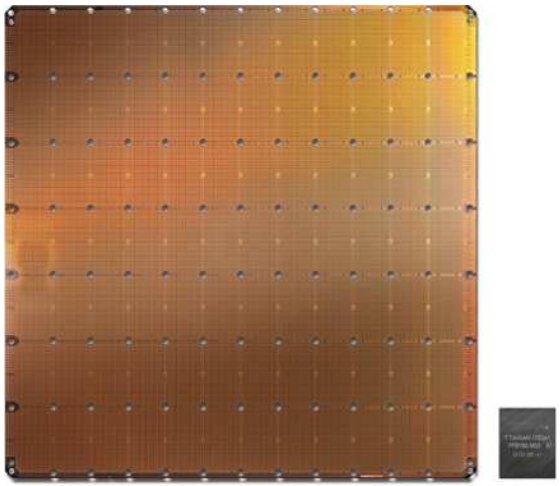
- **巨大AIチップを搭載したCerebras CS-3についてご紹介します**

Cerebras Systems 企業概要

創業:	**2016年4月**
米国本社:	**サニベール カリフォルニア USA**
エンジニアリングオフィス:	**サン･ディエゴ カリフォルニア USA**
カナダテクノロジーセンター:	**オンタリオ　カナダ**
日本法人:	**セレブラス･システムズ合同会社**
	東京都港区西麻布4-14-14
総調達資金:	**7.2億ドル超 (約920億円)**
従業員数:	**500＋ (350＋がエンジニア)**
主要製品:	**CS-3 システム**
テクノロジー:	**Wafer Scale Engine**
取得済･出願中特許:	**80**

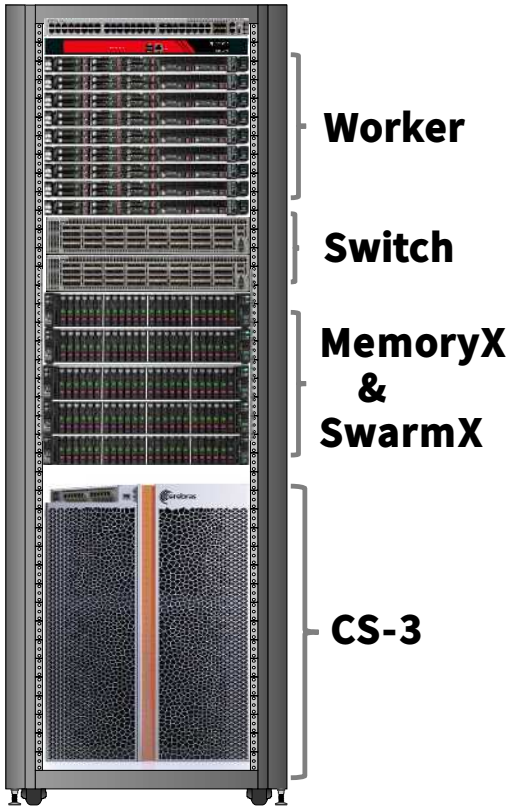
販売･サポートパートナー:	**東京エレクトロン デバイス株式会社**
アライアンスパートナー:	**日本ヒューレット･パッカード合同会社**
事例顧客:	**アルゴンヌ国立研究所、国立科学財団 ピッツバーグ･スーパー･コンピューティングセンター**
	ローレンスリバモア国立研究所、エジンバラ大学パラレルコンピューティングセンター
	国立エネルギー研究所、グラクソ･スミス･クライン、アストラゼネカ、トタルエナジーズ

世界最大チップ Wafer Scale Engine 3



	WSE-3	最新GPU	Cerebrasの優位性
チップサイズ	46,225 mm²(21.5cm角)	826 mm²	57 X
コア数	900,000	16,896 FP32 + 528 Tensor	52 X
オンチップ・メモリ	44 Gigabytes	0.05 Gigabytes	880 X
メモリ帯域幅	21 Petabytes/sec	0.003 Petabytes/sec	7,000 X
ファブリック帯域幅	214 Petabits/sec	0.0576 Petabits/sec	3,715 X

Cerebras CS-3 / Wafer Scale Cluster



Wafer Scale Cluster

世界で最もパワフルなAIコンピューター
一つの筐体システムにすべてのソリューションを満載

■ CS-3

- 世界最大チップ　Wafer Scale Engine (WSE) を搭載したシステム
- 通常の19インチサーバーラックに容易に設置
- PytorchでDeep Learningの学習/推論が可能
- スパース処理をハードウェアで実装
- SDKの利用でユーザープログラミングによるHPC用途にも適用可能

■ Wafer Scale Cluster

- CS-3で巨大AIモデルを処理するために必要な機器を1ラックにオールインワン
- 1ラックに全て搭載されているため、拡張はラック毎に追加するシンプルな構成

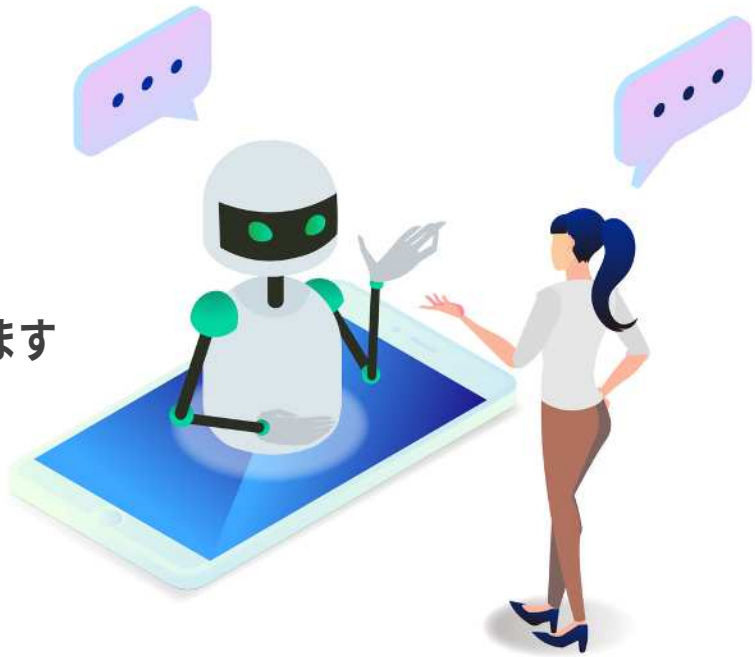
Cerebras CS-3で実現する高速推論

- **今後重要ファクターであるLLM推論高速化についてご紹介します**

LLMの高速推論のニーズ

➢ リアルタイム応答の重要性

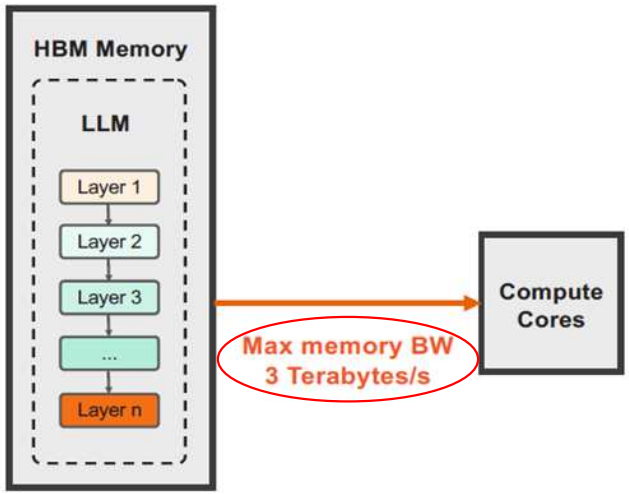
- 人間のリアルタイム会話では、数百ミリ秒の応答遅延があると、会話が途切れたようになり、不自然に感じられます
- Chatbotや動画・音声生成など、LLMを活用したアプリケーション開発では、リアルタイム性を実現する高速推論が今後重要となります
- アプリケーション要件によっては、生成するスピードだけでなく、最初のレスポンス時間も重要になる場合があります



LLM推論パフォーマンスの壁

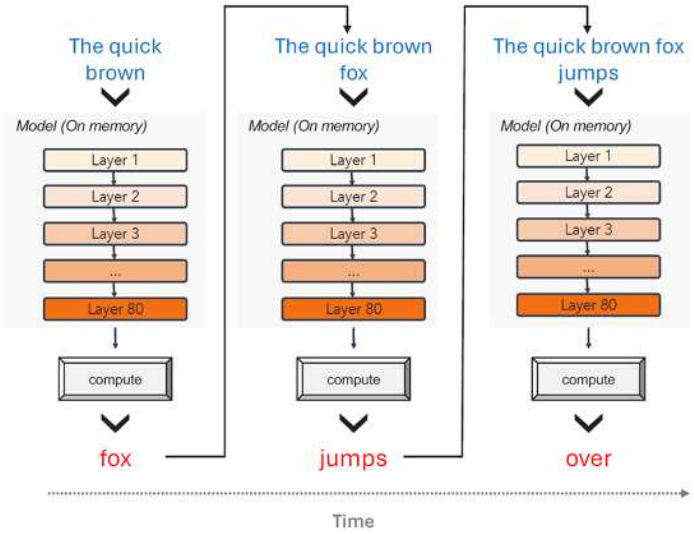
➢ メモリとコンピュートコア間のメモリ帯域の壁

- GPU上ではLLMのパラメータは一旦メモリにロードされ、計算時にコンピュートコアへ転送される
- この際にメモリ帯域幅によるデータ読み出し量の制約を受けるため、処理遅延に繋がる



GPUメモリとコンピュートコア間接続イメージ

- LLMでは1単語を生成する度にモデルの全パラメータを使用するため、生成する単語数が多い場合、コンピュートコアとメモリ間の転送速度がハードウェアの限界に達する場合がある

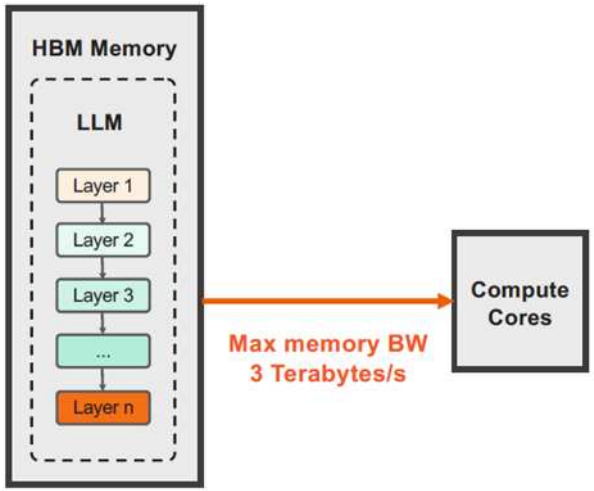


LLMモデルのトークン生成フローのイメージ

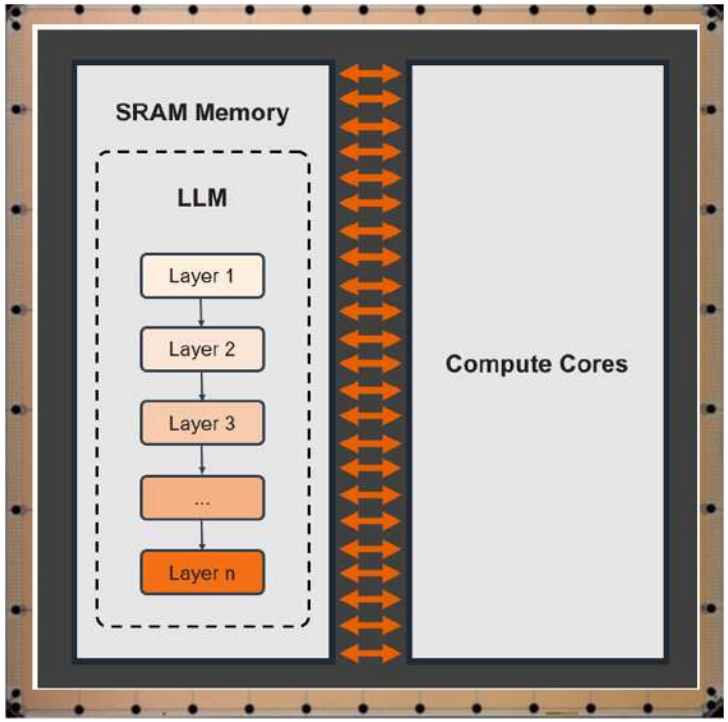
CerebrasによるLLM高速推論のアプローチ

onnect Beyond

- **巨大チップにメモリとコンピュートコアを実装**
  - Cerebras Wafer Scale Engineは、シリコン上にコンピュートコアとメモリを直接統合
  - メモリ帯域幅が通常のGPUより7000倍、速度が10倍以上向上



VS



Cerebras Wafer Scaleチップの最大メモリ帯域は「21 PB / s」

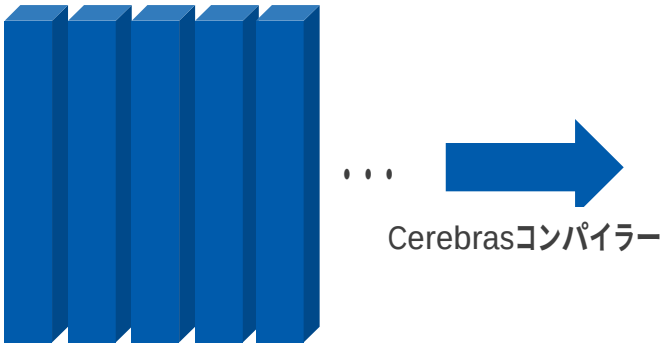
東京エレクトロン デバイス

生成AI推論モデルの展開イメージ

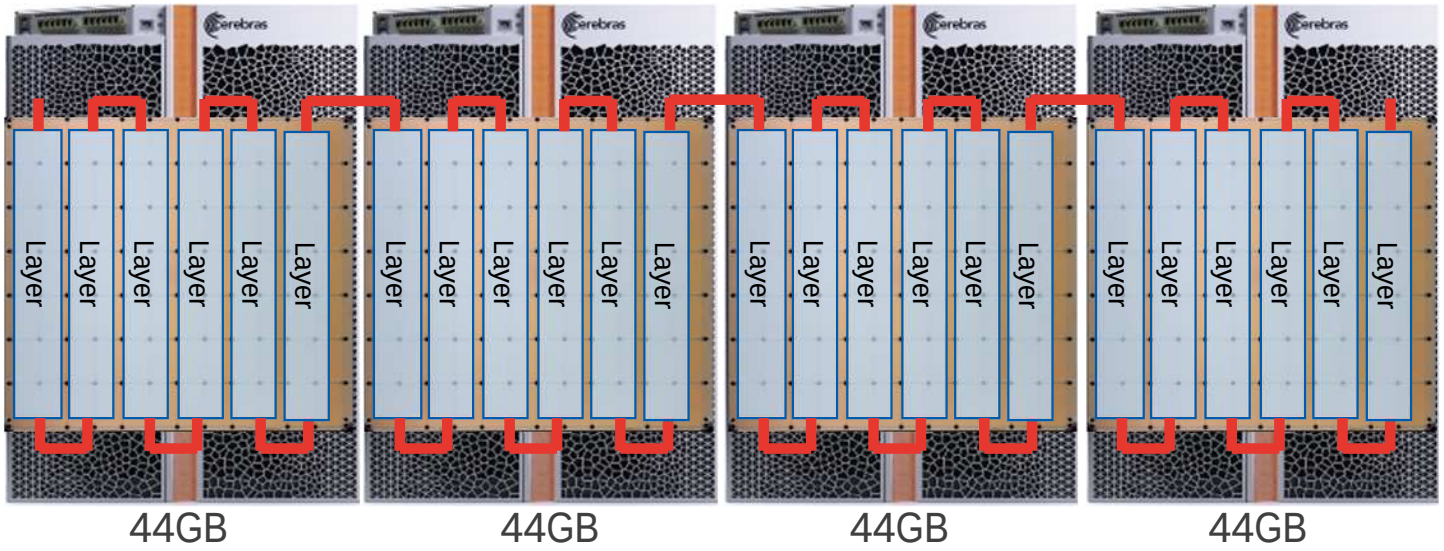
➢ **Llama3.1-70BをCerebras CS-3に展開した場合**

- **Llama 3.1-70b推論に必要なメモリサイズは140GB(FP16)**
- **モデルのレイヤ数は80レイヤ**

Llama 3.1-70bモデル



Cerebrasコンパイラー



Cerebras CS-3 4**台のトータルメモリ容量** ： 176GB

生成AI高速推論デモ

- Llama 70Bモデルを使った高速推論のデモをお見せします

デモを動画でお見せします

- **Llama 70B Speed Test**
  - **比較**

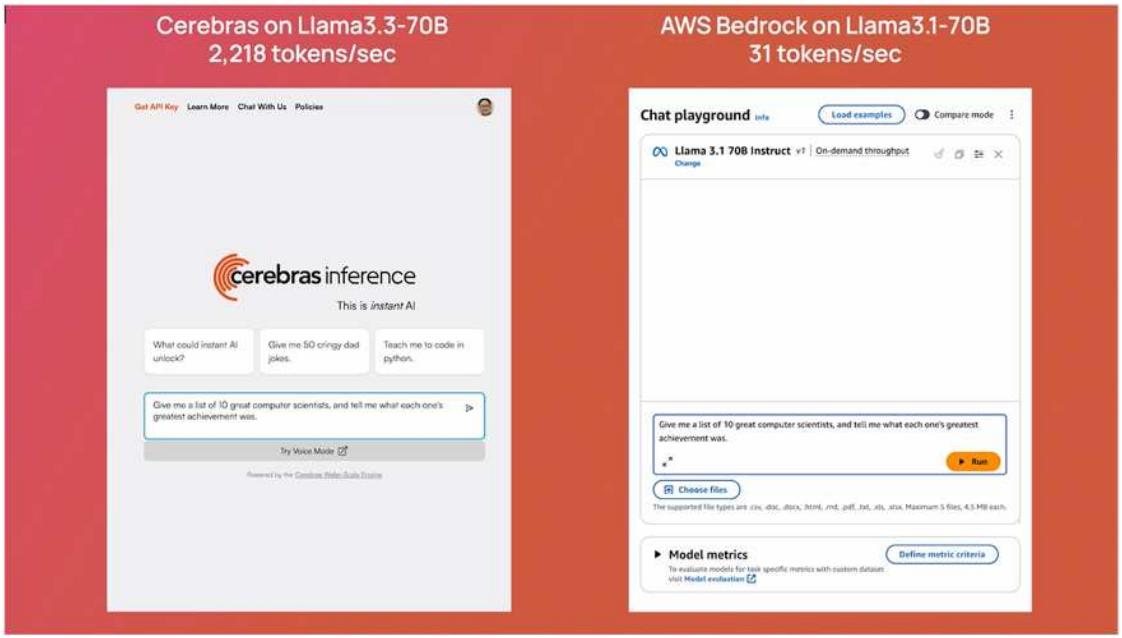
    Cerebras CS-3 **推論サービス**

    AWS Bedrock **推論サービス**

  - **質問内容**

    Give me a list of 10 great computer scientists, and tell me what each one's greatest achievement was.

    (偉大なコンピューター科学者を10人挙げて、それぞれの偉大な業績を教えてください)



Cerebras CS-3 **推論環境**　　AWS Bedrock **推論環境**

当社独自LLMの開発

- 当社でLlamaモデルをベースとして開発した独自LLMについてご紹介します

Llama3-8Bベースの独自大規模言語モデル発表

- 日本語コーパス及び、弊社独自データを使った大規模自然言語モデル
- 学習にはCerebras CS-3を利用

東京エレクトロン デバイス株式会社　会社情報　商品情報　投資家の皆様へ

TOP > プロダクトインフォメーション > ニュースリリース

ニュースリリース

報道関係各位

2024年10月24日
東京エレクトロン デバイス株式会社

企業の膨大な内部データを学習可能とした独自大規模言語モデル（LLM）の開発に成功

企業の生成AI活用に新たな選択肢の提供を可能に
Cerebras CS-3で社内データを活用し、1,730億トークン以上の学習を実施

東京エレクトロン デバイス株式会社（本社：東京都渋谷区、代表取締役社長：徳重 敦之、以下TED）は、Cerebras Systems（以下、Cerebras）と共同で企業の膨大な内部データを学習可能とした独自の日本語大規模言語モデル「Llama3-tedllm-8B-v1）を開発しました。このモデルはmetallama/Meta-Llama-3-8Bを基盤モデルとし、日本語の一般コーパスと社内の豊富なデータを活用した1,730億トークンのデータセットを用いて継続事前学習を行ったものです。これにより英語能力を有する基盤モデルに日本語能力を追加し、さらに社内データの反映も実現しています。

出典:https://www.teldevice.co.jp/pro_info/2024/press_241023.php

Cerebras Systems
41,477人のフォロワー
3週間前 • 編集済み •

+ フォロー

Cerebras and 東京エレクトロン デバイス株式会社 (TED) have trained Llama3-tedllm-8B-v1, a proprietary Japanese large-scale language model based on Meta's Llama3-8B and trained on 173 billion tokens using Cerebras CS-3.

This model offers enhanced Japanese language precision with industry-specific adaptation, efficient training powered by Cerebras CS-3, and effective document generation and decision support.

Learn more about TED's advancements in corporate AI in Japan: https://hubs.li/Q02WT2Zw0

Check out Llama3-tedllm-8B-v0 on Hugging Face: https://hubs.li/Q02WSTFG0

翻訳を表示

119　1件のコメント・9件の再投稿

出典:https://www.linkedin.com/posts/cerebras-systems_cerebras-and-tokyo-electron-device-ltd-activity-7259353755299532801-6EM-

TEDLLM(Llama3-tedllm-8b) 仕様

- Meta Llama3-8bモデルに継続事前学習を実施
- 正式名称は「Llama3-tedllm-8b」
- 日本語精度を上げるために、独自トークナイザーを作成
- LLMスケール則を考慮し、1730億トークンのデータセットを利用

TED版LLM仕様		
LLMベースモデル	Meta Llama3-8B	
言語	日本語･英語	日本語、英語を再学習
メッセージコンテキスト	8192	Llama3仕様と同等
トークナイザー	独自トークナイザー	Llama3-8Bをベースに利用データで再学習
学習手法	Continual Pre-training(継続的事前学習)	Full fine-tuning
利用タスク	Q&A	PEFTチューニング(Lora)で微調整
利用データセット	日本語の一般コーパスとTED独自データ	データサイズ 100GB トータルトークン数 1730億トークン

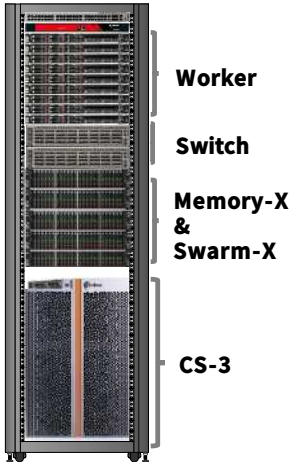
TEDLLM(Llama3-tedllm-8b)開発システム紹介

➢ 学習規模に合わせて、2つのシステムで開発

Cerebras CS-3 in Cerebras Datacenter
- ✓ 計算リソースが必要なフルファインチューニングで利用
- ✓ Cerebraクラウドサービスでご利用できます

DGX A100 in TED AI Lab
- ✓ PEFTチューニング(Lora)及び推論実行で利用
- ✓ TEDエンジニアリングサービスでご利用できます

CS-3 Wafer Scale Cluster



※実際に利用したシステムではなくイメージ写真です

TED AI Lab

NVIDIA DGX A100

東京エレクトロン デバイス