

Connect Beyond **TEDLLMについて**

東京エレクトロン デバイス株式会社

**CN BU**

システムエンジニアリング部

中田 友康

# 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 | サ

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。

Cerebras Systems | + フォロー ...  
41,477人のフォロワー  
3週間前 • 編集済み • 🌐

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

## Llama 3 TEDLLM

8B Parameters, trained on 173B tokens  
Enhanced Japanese precision  
with industry-specific adaptation  
Trained on Cerebras with Tokyo Electron Device Limited

🌐 119 | 1件のコメント・9件の再投稿

# 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 トータルトークン数	100GB 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

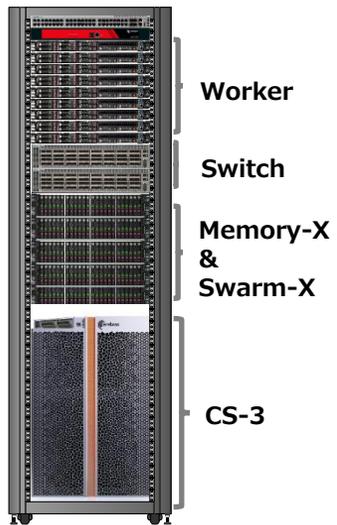
## Cerebras CS-3 in Cerebras Datacenter

- ✓ 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



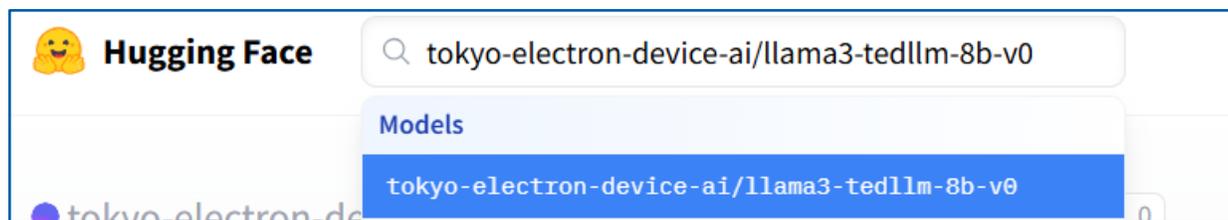
※実際に利用したシステムではなくイメージ写真です



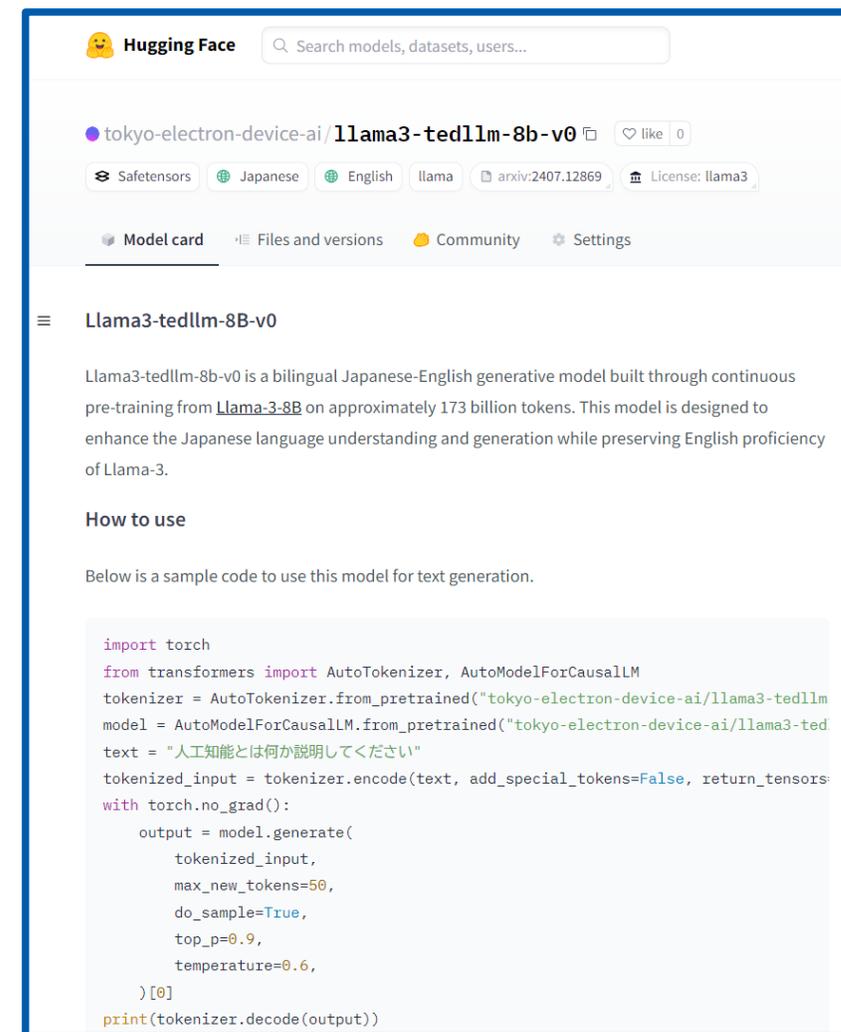
**NVIDIA DGX A100**

# Hugging Faceにてモデルを公開

- 一般コーパスで学習したモデルをHugging Faceにて公開

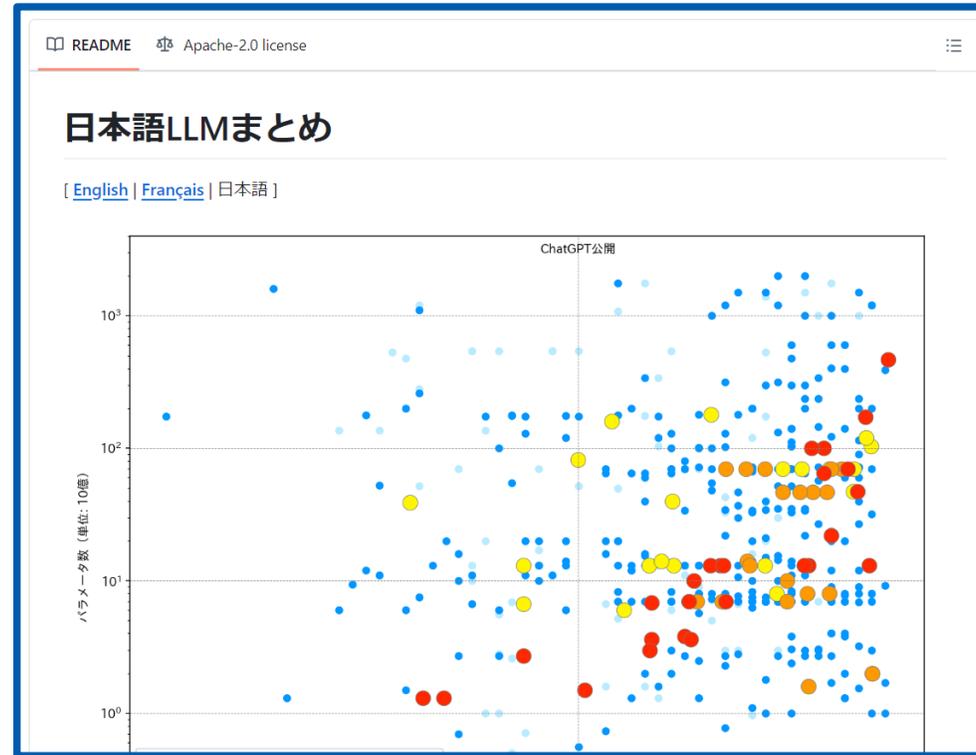


- モデルとトークナイザーを公開
- TED独自データを学習させたv1モデルは非公開
- 8Bモデルのため、GPU 1枚でどなたでも利用可能  
約16GBのGPUメモリが必要です



# LLM-jp 日本語LLMまとめサイトに掲載

➤ 「海外モデルに日本語で継続事前学習を行ったモデル」としてリストに掲載

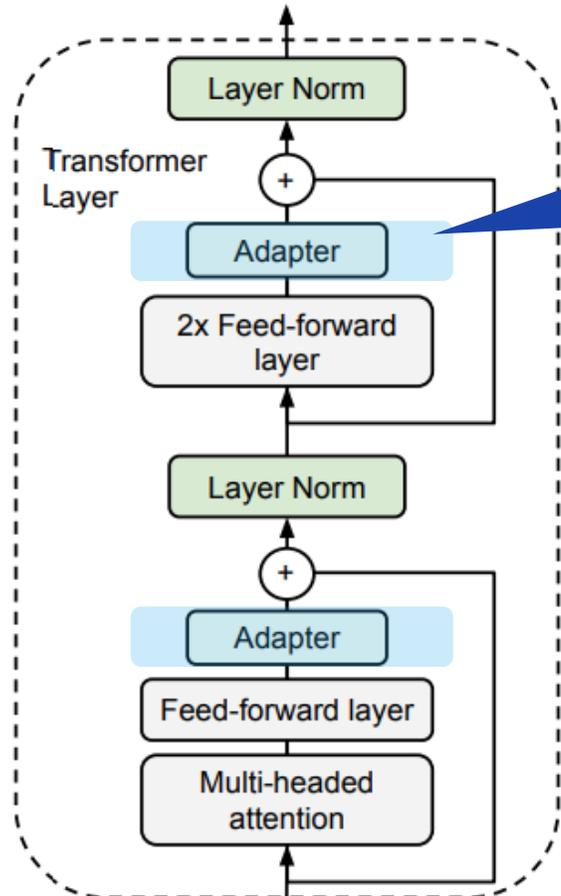


<u>Llama 3 tedllm</u> (v0)	Llama 3 (8b)	事前学習: 日本語の一般コーパス	東京エレクトロンデバイス	Llama 3 Community License
-------------------------------	--------------	------------------	--------------	---------------------------

# 独自LLMの活用に向けて

## ➤ PEFT(Parameter-Efficient Fine Tuning)による様々なタスクへの活用

当社のInfoblox製品の技術QAタスクでの利用を既に始めていて、社内から当社DXをトライ中です！



PEFTとは：学習済みモデルに対して、新しいタスクに効率的に適応させるための手法

- ・モデル全体ではなく、一部のパラメータだけをファインチューニング
- ・目的タスクに合わせたチューニングを「計算コスト・データ量」も抑えつつチューニング

### PEFT(Parameter Effective Fine Tuning)の効果

Setting the style, tone, format, or other qualitative aspects  
(スタイル、トーン、フォーマット、またはその他の質的側面の設定)

Improving reliability at producing a desired output  
(目的のアウトプットを生成する際の信頼性向上)

Correcting failures to follow complex prompts  
(複雑なプロンプトに従えない場合の修正)

Handling many edge cases in specific ways  
(多くの特殊なケースを特定の方法で処理)

Performing a new skill or task that's hard to articulate in a prompt  
(プロンプトで表現するのが難しい新しいスキルやタスクを実行)

出典:OpenAI Finetuning : <https://platform.openai.com/docs/guides/fine-tuning>

出典:Parameter-Efficient Transfer Learning for NLP  
<https://arxiv.org/pdf/1902.00751>

Copyright © Tokyo Electron Device LTD. All Rights Reserved.

# Chatbotでの利用イメージ (Text-generation-webui)

**TED LLM**  
なんでも質問してください。

**You**  
Infoblox製品について教えて

**TED LLM**  
はい、Infoblox製品はネットワーク自動化とセキュリティに特化したソリューションを提供しています。主な製品には、DNS/DHCPサーバー、IPアドレス管理(IPAM)、クラウド・ネットワーキング、脅威防御などがあります。これらの製品は、ネットワークの可視性、制御、および保護を強化し、IT運用効率を向上させることを目的としています。

**You**  
IPアドレス管理(IPAM)とは何ですか？

**TED LLM**  
IPアドレス管理(IP Address Management、略称IPAM)とは、ネットワーク内のIPアドレスを計画、割り当て、監視するプロセスです。IPAMは、ネットワーク管理者がIPアドレスを効率的に使用して、ネットワークのパフォーマンスとセキュリティを向上させるために役立ちます。IPAMシステムは、IPアドレスの在庫を追跡し、重複や誤った割り当てを防ぐことができます。また、IPアドレスの使用状況を追跡し、ネットワークの問題を特定するのにも役立ちます。



東京エレクトロン デバイス

