エンタープライズ AI

エンタープライズ向けの生成 AI とドメイン固有のモデル

専用のインテル® AI ハードウェアとソフトウェアで
トレーニングと導入を最適化し、ビジネの変革を支援

intel®

目次

インテルのパートナーになる理由

インテルには、地球上のあらゆる人の生活とあらゆる企業の成果を向上させる機会がありますが

それは一社のみでは実現しません。

インテルはパートナーと共に、**AI**をあらゆる場所に導入し、導入のリスクを
最小限に抑えることで、顧客に真の価値を創造します

インテルのパートナーになることで、**AI**エコシステム全体とパートナーに

インテルの AI 実現テクノロジーの幅広いポートフォリオと、ハードウェア、ソフトウェア、システム・インテグレーターが緊密に連携するパートナーシップにより、業界、企業、コミュニティーに差別化されたビジネスの成果をもたらす、実環境のソリューションを創造します。

そしてお客様のビジネスの成長も支援します。

AI をあらゆる場所に導入する取り組みにぜひ参加してください

インテル® AI ソリューションで顧客に価値を創出

インテルのアプローチにより、AI 関連企業の幅広いオープンなエコシステムは、エンタープライズ固有の GenAI ニーズを満たすソリューションを提供することが可能になります。

NAVER

クラウドからオンデバイスまで、高度な AI サービスをグローバルに展開する強力な大規模言語モデル（LLM）を
開発するために、NAVER は、大規模言語モデル（Transformer モデル）の演算処理を卓越したワット当たりパフォーマンスで実行するインテル® Gaudi® AI アクセラレーターの基本性能を確認しました。

BOSCH

製造異常に関する合成データセットを生成する基礎モデルなど、スマート製造のさらなる機会を模索し、強力かつ均等に分散されたトレーニング・セット（自動光学検査など）を提供しています。

IBM

自社の watsonx.data™ データストアに第 5 世代インテル® Xeon® プロセッサーを採用し、インテルと
密接に協力することで watsonx™ プラットフォームを
インテル® Gaudi® アクセラレーターで検証しています。

OLA

10 言語に対応する生成機能を備えたインド初の基本モデルの事前トレーニングと微調整を実施し、市場にあるソリューションに比較して業界トップクラスのコスト・
パフォーマンスを実現します。Krutrim は、インテル® Gaudi® 2 クラスター上で、より大規模な基本モデルの事前トレーニングを行っています。

seekr

信頼できる AI のリーダー企業である同社は、LLM の開発と本番環境への導入サポートで
インテル® Tiber™ AI クラウドを利用し、インテル® Gaudi® 2、インテル® データセンター GPU マックス・シリーズ、インテル® Xeon® プロセッサーで本番環境ワークロードを実行しています。

IFF

食品、飲料、香水、バイオサイエンスのグローバルリーダー企業 IFF は、高度な酵素設計と発酵プロセス最適化の
ための統合型のデジタル生物学的ワークフローを確立するため、GenAI とデジタルツイン・テクノロジーを活用
しています。

airtel

最先端テクノロジーのパワーを取り入れ、Airtel は豊富な通信データを活用して AI 機能を強化し、顧客体験の飛躍的な向上を図ります。この導入は、テクノロジー・イノベーションの最前線に立って、急速に
進化するデジタル環境の中で新たな収益源を促進するという Airtel のコミットメントに沿ったものです。

Infosys

この次世代デジタルサービスとコンサルティングにおけるグローバルリーダー企業は、第 4 世代および第 5 世代
インテル® Xeon® プロセッサー、インテル® Gaudi® 2 AI
アクセラレーター、インテル® Core™ Ultra プロセッサーを含むインテルのテクノロジーを、Infosys Topaz（生成 AI テクノロジーを使用してビジネス価値を加速させる AI ファーストのサービス、ソリューション、プラットフォームのセット）に導入するための戦略的コラボレーションを発表
しました。

エンタープライズ AI のオープン・プラットフォーム開発を推進するために集結したエコシステム

エンタープライズ AI の価値提案

エンタープライズ AI でビジネスを変革する

今日の競争の激しい環境において、**AI を採用する企業が優位に立ちつつあります。**

AI がワークフローをどう補強し、自動化できるかを理解
するために、あらゆる業界の企業が運用のすべての側面を再検討しています。

**インテルでは、AI を企業の構造に組み込むことを独自の
専門技術としています。**

生産性を変革する AI PC から、最も価値を生むユースケースを把握する長年の専門知識に至るまで、インテルは、AI をあらゆる場所に安全かつ責任を持って導入する信頼できるパートナーです。

生成 AI (GenAI) イノベーションは、インターネット時代、モバイル時代、クラウド時代を上回るスピードで、あらゆる規模の企業で採用されることが期待されています。

AI プラットフォームの次の波は、このような画期的な未来を、低コストかつ柔軟な方法で実現することになるでしょう。

**エンタープライズ AI について、今までとは異なる
考え方をする時期が来ています。**

この支援パッケージは、市場のあらゆる企業が長期的な成功に向けて生成 AI、特にドメイン固有モデルから、どのように大きな価値を得られるかを理解するのに役立ちます。

生成 AI と大規模言語モデルとは

生成 AI (GenAI) は、新しいオリジナルのコンテンツ
作成に焦点を当てた AI のサブセットです。

トレーニング用データセットの例に類似した画像、テキスト、オーディオなどのデータを生成するために、
AI モデルのトレーニングと導入を行います。

GenAI アルゴリズムは、ディープラーニングやニューラル・ネットワークなどの高度な手法を使用し、画像合成、テキスト生成、さらにはクリエイティブなアートワークなどのアプリケーションを可能にする、
現実的で一貫性のある出力を生成します。

大規模言語モデル (LLM)は、ディープ・ニューラル・
ネットワークを使用してテキストを処理・生成する、特定のタイプの自然言語処理モデルです。LLM は大量のテキストデータを使用してトレーニングを実施し、一貫性があり意味のある出力を生成するように設計されています。

詳細を見る

詳しくはこちら

生成 AI のパワーを活用する

企業が GenAI を使用する方法

消費財と小売

- バーチャル試着室
- 配送と設置
- 店舗での製品検索のサポート
- 需要予測と在庫計画
- 新しい製品の設計

ヘルスケアと医療

- 多忙なフロントスタッフの支援
- カルテの書き起こしと要約
- 医療に関する質問に答えるチャットボット
- 診断と治療を伝えるための予測分析

製造

- 技術者向けのエキスパート・コパイロット
- 機械との対話型インタラクション
- 処方的および予防的な現場保守
- 自然言語トラブルシューティング
- 保証のステータスとドキュメント
- プロセスのボトルネックを把握し、リカバリー戦略を考案

メディアとエンターテインメント

- インテリジェント検索、カスタマイズされたコンテンツ発見
- ヘッドラインとコピーの作成
- コンテンツの品質に関するリアルタイムのフィードバック
- パーソナライズされたプレイリスト、ニュースダイジェスト、おすすめ
- 視聴者の選択によるインタラクティブなストーリーテリング
- 対象を絞ったオファー、サブスクリプション・プラン

金融サービス

- 売買シグナルを発見し、トレーダーに脆弱なポジションを警告
- 引受判断の迅速化
- レガシーシステムの最適化と再構築
- 銀行業務と保険モデルのリバース・エンジニアリング
- 潜在的な金融犯罪と不正行為の監視
- 規制遵守のためのデータ収集を自動化
- 企業の情報開示からインサイトを抽出

出典: MIT Technology Review Insights による「Retail in the Age of Generative AI」9、「The Great Unlock: Large Language Models in Manufacturing」10、「Generative AI Is Everything Everywhere, All at Once」および「Large Language Models in Media & Entertainment」12 からのデータのまとめ、Databricks (2023年4月～6月)

生成 AI と大規模言語モデルのユースケース

チャットボットと
仮想アシスタント

カスタマーサポート

コード生成と
デバッグの LLM

会社のドキュメントを
使用したトレーニング

感情分析

顧客満足度を評価

テキストの分類と
クラスタリング

大量のデータを
分類してトレンドを特定

言語翻訳

会社のウェブページ
を他の言語に翻訳

要約と言い換え

議事録のまとめ

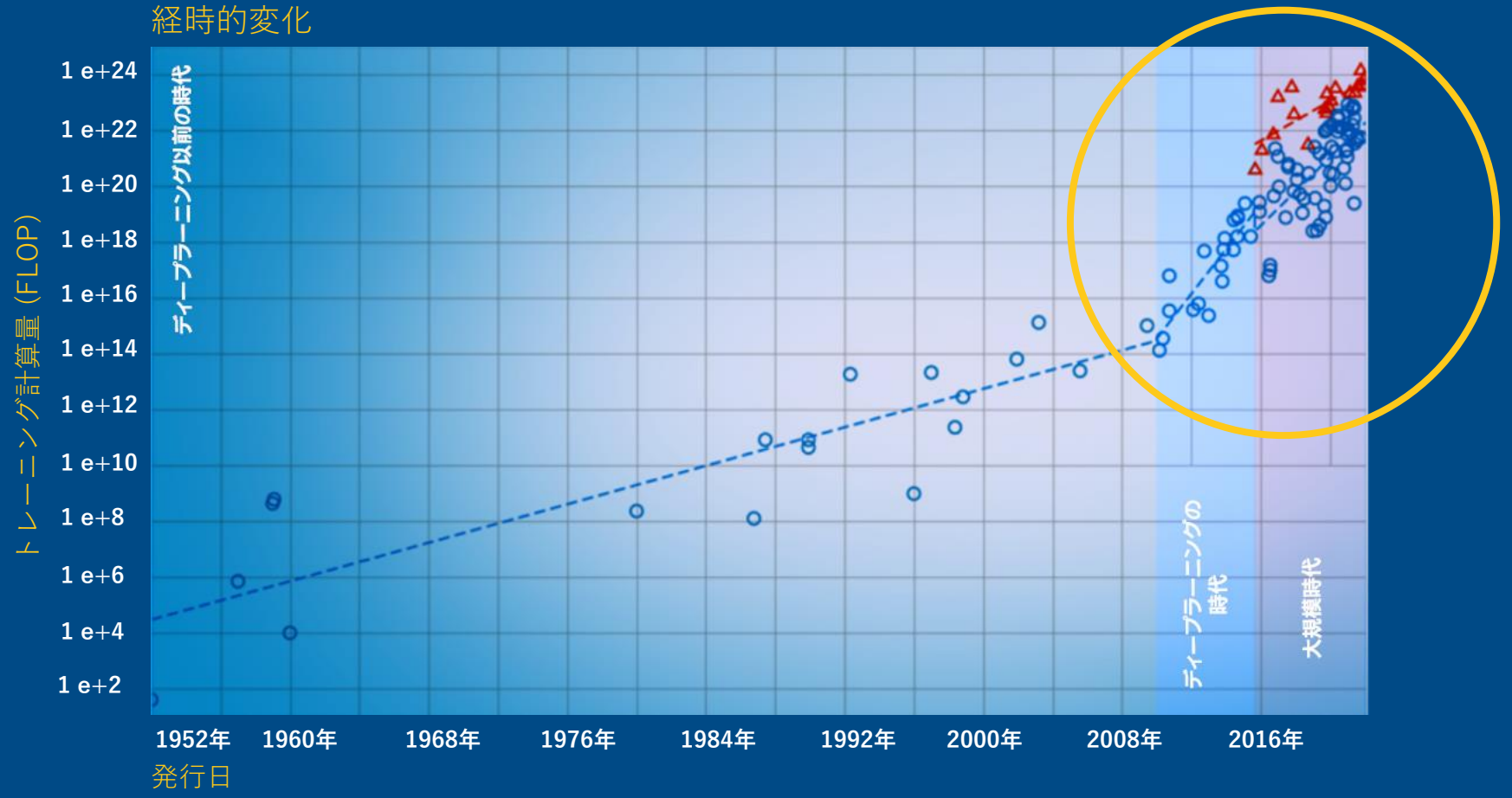
コンテンツ、画像、
ビデオの生成

電子メールの最初の
下書き、アイデア生成、
マーケティング用ビジュア
ル、ショートビデオ

出典:「A Compact Guide to Large Language Models」 Databricks (2023年)

モデルの肥大化に伴い、計算量も増加

milestone マシンラーニング・システムのトレーニング計算量 (FLOP) の経時的変化



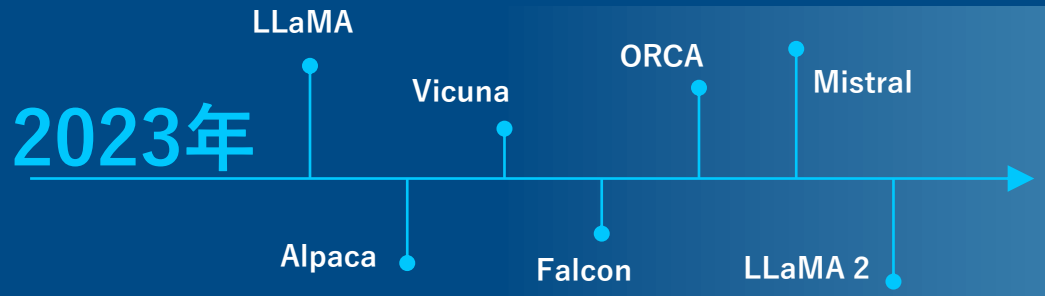
Epoch、アバディーン大学、Center for the Governance of AI、セント・アンドリューズ大学、MIT、Eberhard Karls Universitat Tubingen、Universidad Complutense による調査

大規模モデルだけではない AI モデル

	大規模モデル (サードパーティー	VS.	**小規模で高速** (10 ～ 100 倍)
説明可能性	独自のモデル	VS.	オープンソース・ベースのモデル
精度	オールインワンの汎用モデル	VS.	ターゲット設定、ドメイン固有、カスタマイズ
場所	クラウドベース (as-a-service)	VS.	ローカルで推論を実行、エッジ、クライアント、オンプレミス
コスト	永続的にスケーリングするコスト	VS.	コスト管理
市場投入までのスピード	高速セットアップ (数秒)	VS.	ビルドに要する時間 (時間/日)

多くの小規模モデルの発展

6カ月で、1,000 億 ～ 200 億のパラメーター



databricks

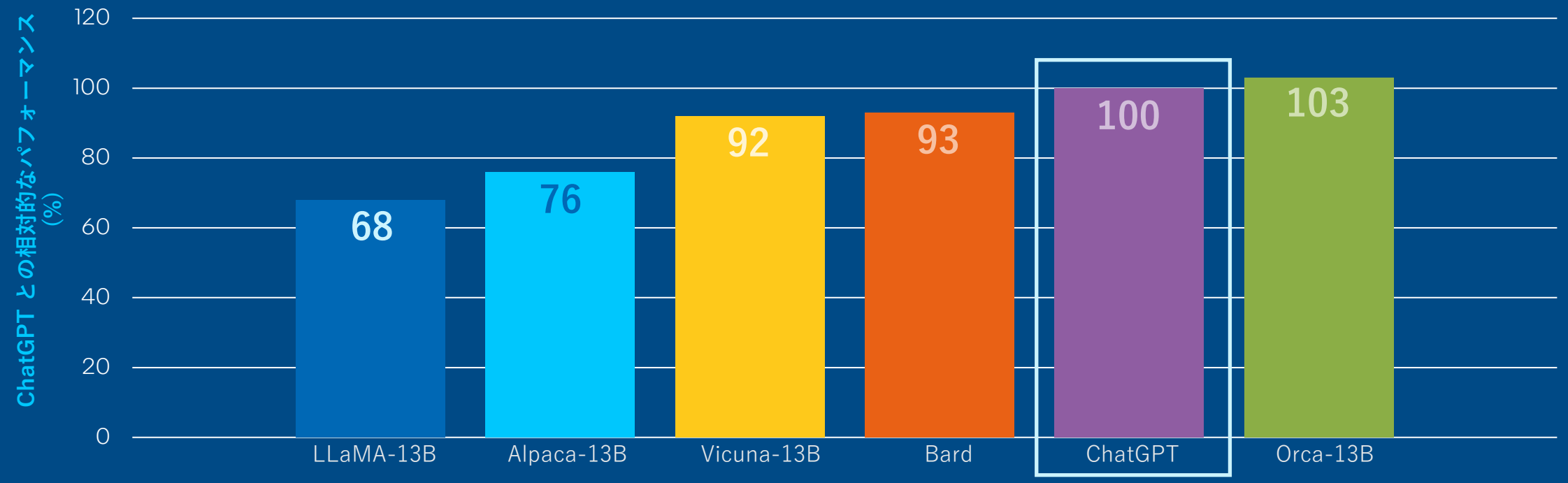
- 毎週、数十個の小規模モデルが登場
- 商用およびオープンソースのライセンス
- 慎重に入手したデータで学習させれば、小規模な
  モデルでも、大規模なモデルの精度を再現できると示されている

- 何千ものドメイン固有の商用モデルと AI プラットフォームが実証されている
- モデルは、少数のプロセッサーを使用してドメイン固有のデータで微調整が可能

ChatGPT と比較して良好なパフォーマンスを発揮する小規模なモデル

ChatGPT のような大規模モデルと比較して、小規模なモデルでも十分実用的で良好なパフォーマンスを発揮することが証明されている

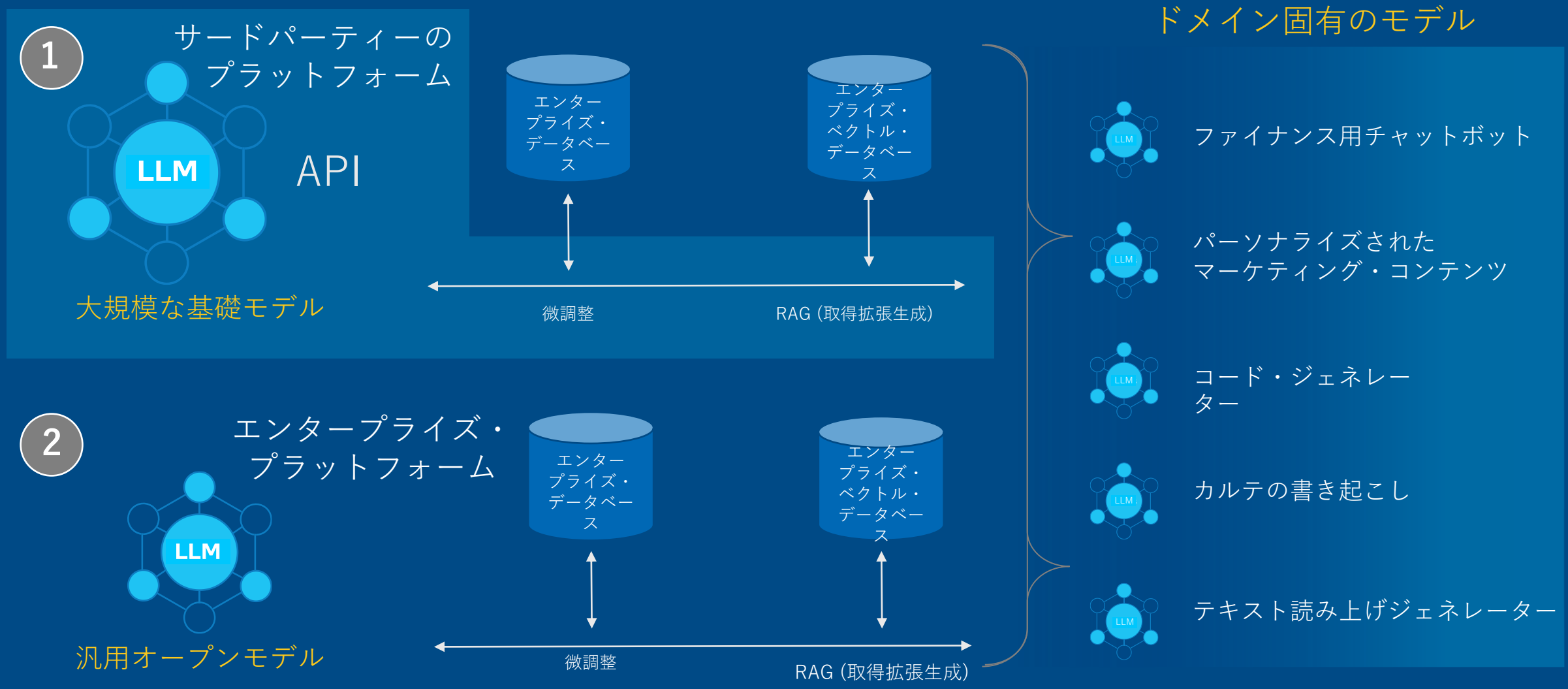
GPT-4 と比較した評価



Orca は、Vicuna 評価セットの GPT-4 による評価の通り、OpenAI ChatGPT を含む幅広い基礎モデルのパフォーマンスを上回っています。

出典: Microsoft Research (2023)。Orca: GPT-4 の複雑な説明トレースからのプログレッシブ・ラーニング

ドメイン固有のモデルを構築



ドメイン固有のモデルは、エンタープライズに多くのメリットをもたらす

小規模でターゲットを絞ったモデルは、同等または優れたパフォーマンスを提供し、時間とコストの投資を削減するため、ROI を向上させることができる

より正確な出力

エンタープライズ独自のデータを使用し、ドメイン固有の精度を向上

コストの削減

事前学習済みモデルの微調整および / または RAG を使用して、より小規模なモデルを推論

好きなプラットフォームでどこでも展開可能

ローカルで推論を実行、エッジ、クライアント、オンプレミス

セキュアで非公開

データのセキュリティーと規制要件の遵守を両立

責任ある AI

微調整と RAG により、モデルにデータのソースを引用する機能を付与

未来

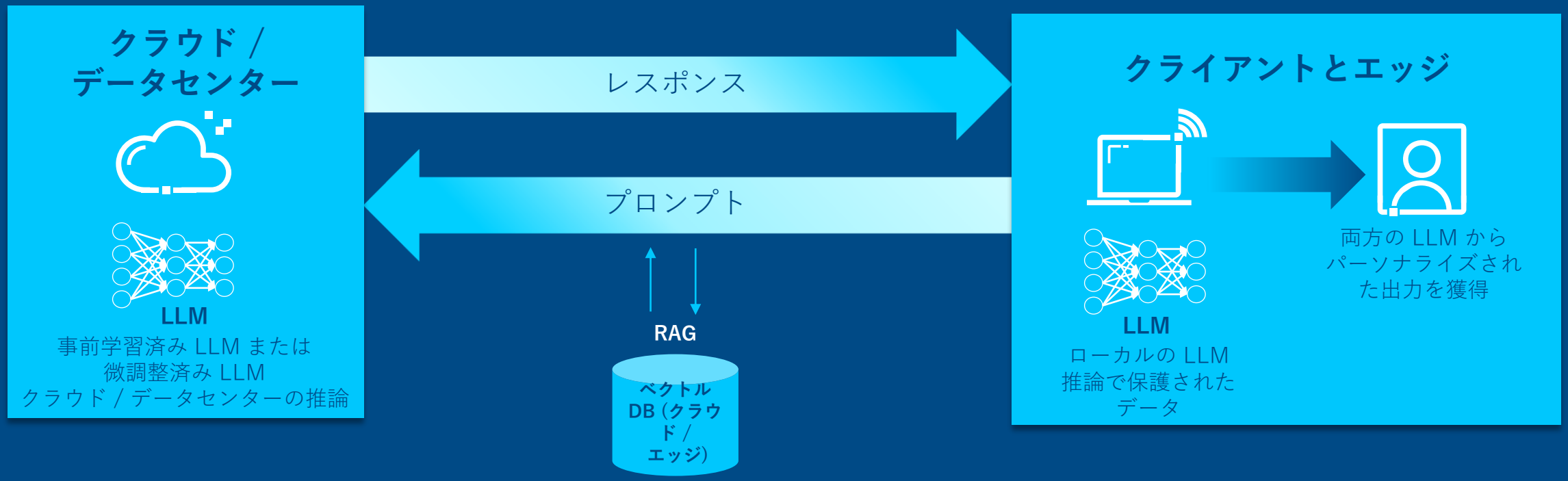
少数の大規模 AI モデルと、膨大な数の小規模かつ軽快な AI モデルが、無数のアプリケーションに組み込まれる[1]

[1]出典:「Survival of the Fittest: Compact Generative AI Models Are the Future for Cost-Effective AI at Scale」

シームレスなクラウドからエッジの AI プラットフォーム

インテル® AI ポートフォリオ

クラウドで学習と推論。RAG を使用してドメインの精度を向上。



生成 AI - 本番環境での 1 年

ドメイン固有だが高度にインテリジェントなモデルの使用が増加している

2022年	2023年	2024年
実験	パイロット	実稼動

巨大なモデルが道を切り開いた

- 汎用的な用途で非常に効果的
- 学習と導入のコストが高価
- 大規模なパブリック・データセット上に構築
- 簡単に使える

小規模のドメイン固有のモデル

- 組織内の非公開データを使用して、ビジネス固有の結果を得る
- 所有しているハードウェアに導入
- 効率、精度、セキュリティー、トレーサビリティーの向上
- ビルドに時間が必要

ブログを読む

Survival of the Fittest: Compact Generative AI Models Are the Future for Cost-Effective AI at Scale

ドメイン固有のモデルに対するインテルのアプローチ

ドメイン固有のモデル

特長

+ 精度を維持 / 向上させる一方で 10～100 倍小型化したモデル
+ 汎用コンピューティング上で経済的
+ 正確性、ソースの帰属、説明可能性
+ 非公開 / エンタープライズ・データの活用
+ 継続的に更新される情報

課題

- タスクの範囲が縮小
- 数回にわたる微調整とインデックス作成が必要

インテルの目標

業界のフレームワーク、事前学習済みモデル、インテルの AI ソフトウェアとツールを使用して、インテルのハードウェア上で数万のモデルを微調整して導入する、最も
コスト効率が高くユビキタスなアプローチを実現する

詳しくはこちら

指先で生成 AI

電子書籍 ・
インフォグラフィック

エンタープライズ AI: 参入障壁の克服を支援

要件	インテルとのパートナーシップによる支援
市場投入までのスピード	インテルと Hugging Face のデベロッパー向けリソース、Gaudi Developer Hub、5 つのリファレンス・キットを使用して、生成 AI のスタートラインに立つことができます。
ユーザー体験 (精度 / レイテンシー)	インテル® Gaudi® アクセラレーターでは 100 億パラメーターを超えるモデルの推論、インテル® AMX 搭載のインテル® Xeon® プロセッサーでは 200 億未満のパラメーターの小規模モデルによる推論で、ユーザーにリアルタイムの体験を提供します[1]。
コンピューティングの入手可能性	インテル® Xeon® CPU + アクセラレーターは、グローバルな GPU 不足に、コスト効率の高い代替手段を提供。インテル® Gaudi® 2 は現在 SuperMicro から入手可能で、インテル® Gaudi® 3 はさらに入手しやすくなります。
使い慣れたテクノロジー	小規模なモデルの推論は、すでにコンピューティング設定の一部になっているかもしれないユビキタス・ソリューションなど、 実質的にあらゆるハードウェア上で実行可能です[2]。
規模に応じた運用	インテル® Gaudi® 2 は、各アクセラレーターに統合された 24 の 100GbE ポートにより、ほぼ線形のスケーラビリティーを実現。 インテル® Xeon® プロセッサーは、すでにデータセンターや現場 (クラウドからエッジまで) で利用できます。データセンターの推論の 65% が、インテル® Xeon® プロセッサーで実行されています[3]。
コスト効率性	実際の業務アプリケーションにおいて、インテルは、パフォーマンスの向上、低価格、AI 推論向けのバランスに優れたプラットフォームを提供することで、業界に破壊的革新をもたらし、AI を民主化しています。NVIDIA、インテル® Gaudi® 2 が H100 の 4 倍優れたコスト・パフォーマンスを証明をご覧ください。

[1]出典:「Four Roadblocks to Implementing Generative AI」

[2]出典:「Survival of the Fittest: Compact Generative AI Models Are the Future for Cost-Effective AI at Scale」

[3]2022年12月時点での AI 推論ワークロードを実行するデータセンター・サーバーの、インテルによる世界のインストール・ベース市場モデリングに基づく。

生成 AI トレーニングと導入を簡素化するソフトウェア・リソース

オープンソースモデル

- BLOOM（a BigScience initiative）— 176B params, 59 languages, Open-access — 176B
- BioGPT — ドメイン 1.5B
- Stable Diffusion — 画像
- Llama2 / GPT-J / MPT / Falcon — 7-65B LLM
- Stanford Alpaca — 微調整可能な 7B LLM
- ナレッジベース

オープン・ソフトウェア

- OpenVINO™
- PyTorch 向けインテル® エクステンション (IPEX)
- Transformers 向けインテル® エクステンション (ITREX)
- DeepSpeed 向けインテル® エクステンション (IDEX)
- haystack
- fastRAG

GenAI プラットフォーム

- intel. GAUDI
- intel. DATA CENTER GPU MAX SERIES
- intel. XEON
- intel. CORE ULTRA

詳しくはこちら ユビキタス・ハードウェアとオープン・ソフトウェアで生成 AI を解き放つ

価値を最大化

インテルのオープン AI アプローチが
AI ビジネスのニーズに適している理由

ベンダーロックインを回避

オープンソースの標準ベースのソフトウェア

MODIN
SciPy
dmlc XGBoost
PyTorch
pandas
NumPy
scikit learn
TensorFlow
Spark
Numba
SigOpt AutoML
oneAPI

インテルのハードウェア・ポートフォリオを活用

AI のユースケース向けに最適化

intel XEON
intel CORE ULTRA
intel ARC GRAPHICS
intel DATA CENTER GPU
intel GAUDI
intel ETHERNET

明日の AI のためにソフトウェアとオープンな標準により最適化されたハードウェアで、
クライアントからエッジ、データセンター、クラウドに至るまで、新たな機会を創出

インテル® AI ソフトウェア・ポートフォリオ

エンジニアのデータ → モデルの作成 → 最適化と導入

MODIN　SciPy　dmlc XGBoost　PyTorch　ONNX RUNTIME　OpenVINO
pandas　NumPy　scikit learn　TensorFlow　DirectML　Write Once Deploy Anywhere
Spark　Numba　SigOpt AutoML　intel Neural Compressor

大規模なデータ分析†

マシンラーニング＆ディープラーニングのフレームワーク、最適化と導入ツール†

oneAPI

Intel® oneAPI Deep Neural Network Library
Intel® oneAPI Collective Communications Library
Intel® oneAPI Math Kernel Library
Intel® oneAPI Data Analytics Library

CPU、GPU、その他のアクセラレーター向けのオープンなクロスアーキテクチャー・プログラミング・モデル

クラウド＆エンタープライズ
クライアント＆ワークステーション
エッジ

intel AI ANALYTICS TOOLKIT
エンドツーエンドのデータサイエンスとAI を加速

インテル® Tiber™ AI クラウドとインテル® Developer Catalog
最新のインテルのツールとハードウェアを試して、最適化された AI モデルにアクセス

cnvrg.io
フルスタックのML オペレーティング・システム

インテル® Geti™
アノテーション/トレーニング/最適化プラットフォーム

Hugging Face
インテルの最適化と微調整レシピ、最適化された推論モデル、モデルの提供

注: スタックの各レイヤーのコンポーネントは、予想される AI の使用モデルに基づいて、その他のレイヤーの対象となるコンポーネントに最適化されています。右側の列にあるソリューションでは、すべてのコンポーネントが活用されるわけではありません。
† このリストには、インテルのハードウェアに最適化された、一般的なオープンソースのフレームワークが含まれています。

エンタープライズ生成AI の導入を簡素化し、要塞化された信頼されるソリューションの製品化までの時間を短縮

OPEA:

エンタープライズ生成AI の導入を簡素化し、
要塞化された信頼されるソリューションの
製品化までの時間を短縮

Open Platform for Enterprise AI

OPEA パートナー

anyscale
CLOUDERA
DATASTAX
kx
DOMINO
Hugging Face
Red Hat
drant
MariaDB
MINIO
sas
Yellowbrick
vmware

OPEA の価値

- 生成 AI (LLM、RAG) を使用して、企業がより迅速かつ簡単にデータから価値を引き出せるよう支援
- 断片化されたエコシステムの複雑さを軽減し、ソリューションの生産規模拡大を支援
- Linux Foundation と提携する業界のリーダーたちのコラボレーションと貢献を推進

効率的
既存のインフラストラクチャー、AI アクセラレーター、または選択するその他のハードウェアを活用できます。

シームレス
システムおよびネットワーク全体で異種サポートと安定性を備えたエンタープライズ・ソフトウェアと統合します。

オープン
ベスト・オブ・ブリードのイノベーションを選んで集めることで、ベンダーロックインを排除します。

ユビキタス
クラウド、データセンター、エッジ、PC 向けに構築された柔軟なアーキテクチャーにより、あらゆる場所で稼働します。

信頼性
安全なエンタープライズ対応のパイプラインと、責任、透明性、トレーサビリティーを実現するツールを搭載しています。

拡張性
ソリューションの構築と拡張を支援するため、パートナーの活気あるエコシステムへのアクセスを提供します。

生成 AI についての Hugging Face とのパートナーシップ

Hugging Face

生成 AI と言語 AI のトレーニングとイノベーションを円滑にするため、インテルは AI モデルとデータセットを共有する人気のプラットフォームである Hugging Face と提携しました。最も注目すべきは、Hugging Face が NLP 向けに構築されたトランスフォーマー・ライブラリーで知られていることです。

intel XEON

インテルは Hugging Face と協力し、トランスフォーマー・モデルによるトレーニング、微調整、予測のための最先端のハードウェアおよびソフトウェア・アクセラレーションを構築しました。

ハードウェア・アクセラレーションは、インテル® Xeon® スケーラブル・プロセッサーにより駆動され、ソフトウェア・アクセラレーションは、最適化された AI ソフトウェア・ツール、フレームワーク、ライブラリーのポートフォリオにより有効化されます。

intel GAUDI

インテル® Gaudi® アクセラレーターのディープラーニング製品もまた、Optimum Habana ライブラリーを通じて Hugging Face のオープンソース・ソフトウェアと共同しており、Hugging Face コミュニティーによって最適化された何千ものモデルをデベロッパーが使いやすくします。

Hugging Face はまた、Stable Diffusion、T5-3B、BLOOMZ 176B および 7B、および新しい BridgeTower モデルといった
生成 AI モデルに対するインテル® Gaudi® 2 のパフォーマンスの評価も公開しています。

インテル、Articul8、BCG が協力して エンタープライズ・グレードの安全な生成 AI を提供

インテル AI スーパーコンピューターを搭載した先駆的なソリューションは、高度なセキュリティーとデータのプライバシーを維持しながら、カスタム・データセットを
活用してビジネスの価値を大きく向上させます。

Articul8* は、インテルのハードウェアに最適化されたターンキーの GenAI ソフトウェア・プラットフォームを提供し、
大企業のお客様における AI の運用と拡張を支援します。このプラットフォームは、インテル® Xeon® スケーラブル・プロセッサーとインテル® Gaudi® アクセラレーターを含むインテルのハードウェア・アーキテクチャー上で立ち上げられ最適化されていますが、
さまざまなハイブリッド・インフラストラクチャーの選択肢をサポートする予定です。

このテクノロジーの Boston Consulting Group (BCG) における初期の導入に続いて、同チームは金融サービス、航空宇宙、半導体、通信など、高度なセキュリティーと専門的なドメインの知識を
必要とする業界セグメントのエンタープライズ顧客向けにプラットフォームを拡張してきました。

詳しくはこちら

Articul8 の発表

Articul8 ウェブサイト

責任あるエンタープライズ AI

課題:

生成 AI モデルは、インターネット上で利用可能な膨大なデータから学習しますが、そのデータには
社会に存在する偏見が含まれている可能性があるため、こういった偏見を誤って適用してしまう
可能性があります。LLM は、誤情報、フィッシング・メール、ソーシャル・エンジニアリング攻撃の生成や拡散のために操作される可能性があります。

LLM に「幻覚」が生じ、不正確な情報を生成することがよくあります。これは、モデルが診断や治療の意思決定に影響を与え、患者に危害を及ぼす
可能性がある、医療などの業界では特に問題に
なる可能性があります。

詳細を見る

生成 AI のリスクを最小限に抑える

ソリューション

AI テクノロジーに取り組む企業や個人は、そのソフトウェアが倫理的な AI の原則に従って開発され、導入されていることを確認する必要があります。

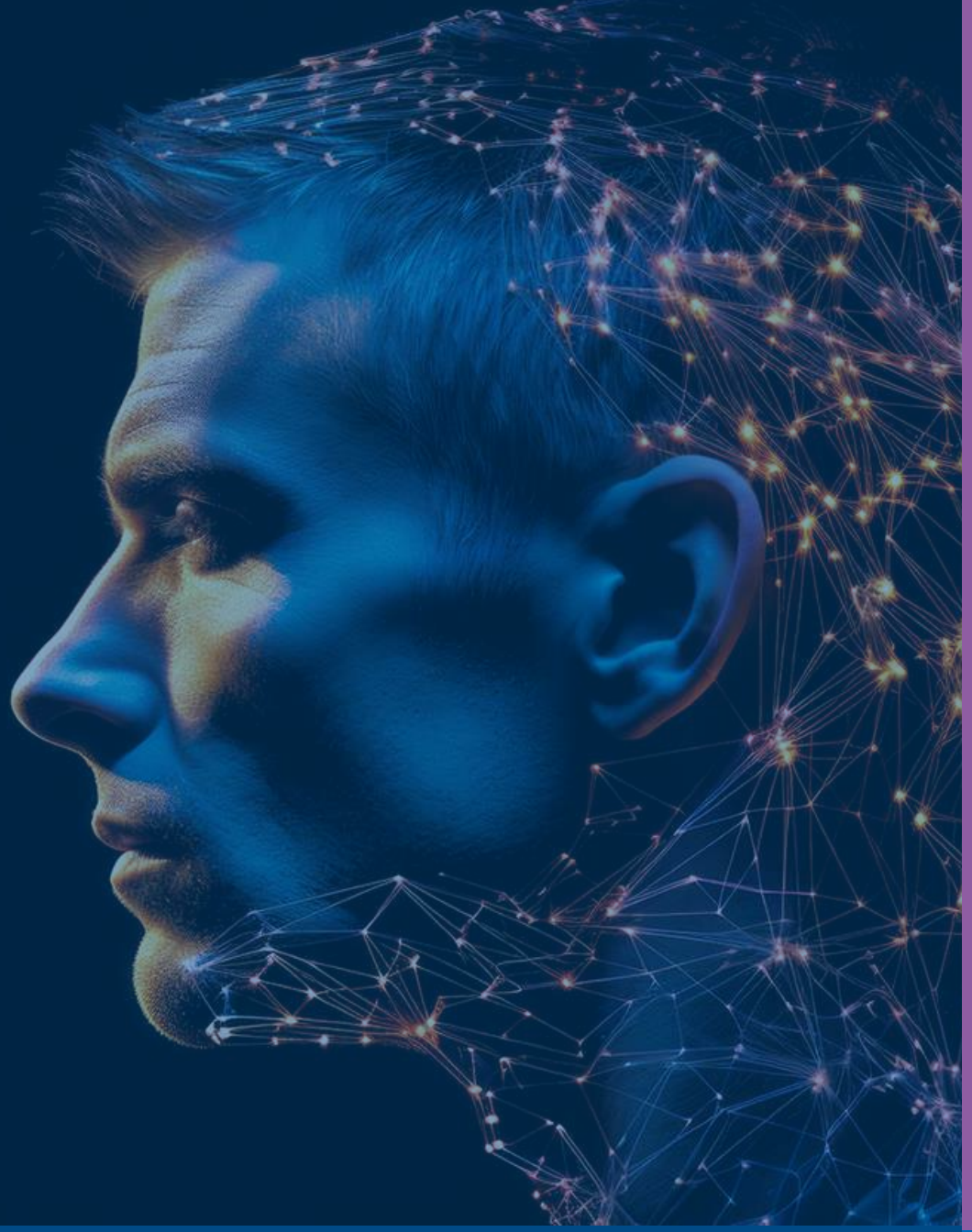
オープンソースのインテル® Explainable AI Tools により、ユーザーはポストホック・モデルの抽出と可視化を
実行し、TensorFlow* と PyTorch* モデルの両方の予測動作を調べることができます。

LLM は通常、大規模な公開データセットで学習し、潜在的にセンシティブなデータ（金融や医療など）を使って微調整されます。

インテルの OpenFL（オープン・フェデレーテッド・
ラーニング）などのテクノロジーは、コンフィデンシャル・コンピューティングを組み込んでいます。そのため、LLM は機密データで安全に微調整を実行し、これにより、幻覚や偏見を減らし、モデルの一般化可能性が向上します。

生成 AI 向けインテル製品

AI をあらゆる
場所に導入



AI 向けのスケーラブルなシステム

トレーニンと微調整

トレーニング

ピーク推論

メインストリーム推論 / 微調整

ベースライン推論

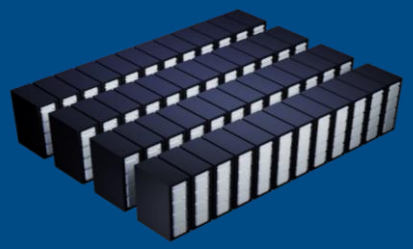
エンドポイント推論

推論と導入

クラウドデータセンター

エッジ

クライアント

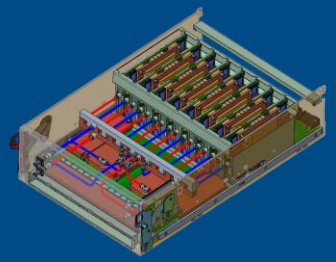


クラスターおよびデータセンター規模

intel. GAUDI

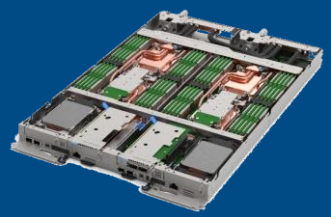
ラックごとのマルチノードの導入

intel. GAUDI　intel. XEON MAX SERIES　intel. XEON



マルチ GPU またはマルチソケット CPU

intel. XEON　intel. XEON MAX SERIES　intel. DATA CENTER GPU FLEX SERIES



シングルソケット CPU または単一 GPU

intel. XEON　intel. DATA CENTER GPU FLEX SERIES　intel CORE ULTRA

クライアント CPU

intel CORE ULTRA　intel. ARC GRAPHICS

intel. ETHERNET

NLP / LLM 向けインテル製品

トレーニングと推論

GAUDI®2

インテル® Gaudi® 2 AI アクセラレーターは、LLM や NLP など、大規模
モデルのトレーニングと推論を高速化するように設計されています。

インテル® Gaudi® 2 で生成 AI と大規模言語モデルを高速化

インテル® Gaudi® 2 は、AI トレーニングで最先端のパフォーマンスと最適なコスト削減を実現

プレスリリース

視聴する

生成 AI と大規模言語モデル (LLM) の可能性を引き出すインテル® Gaudi® 2 AI プロセッサーの最先端機能について紹介するインテルのウェビナーの録画

インテル® Gaudi® 2 ディープラーニング・
アクセラレーターは、ディープラーニングのトレーニングと推論で競争力を発揮し、**NVIDIA A100 よりも最大 2.4 倍高速な**
パフォーマンスを実現[1]

ニュース

技術記事

インテル® Gaudi® 2 は、GenAI のパフォーマンスにおいて、ベンチマークされた唯一の NV H100 の代替製品

[1]性能は、使用状況、構成、その他の要因によって異なります。ワークロードと構成の詳細は、intel.com/performanceindex (英語) を参照してください。結果は状況によって変わります。

Gaudi2: 基礎モデルの効率的なトレーニングと推論に最適

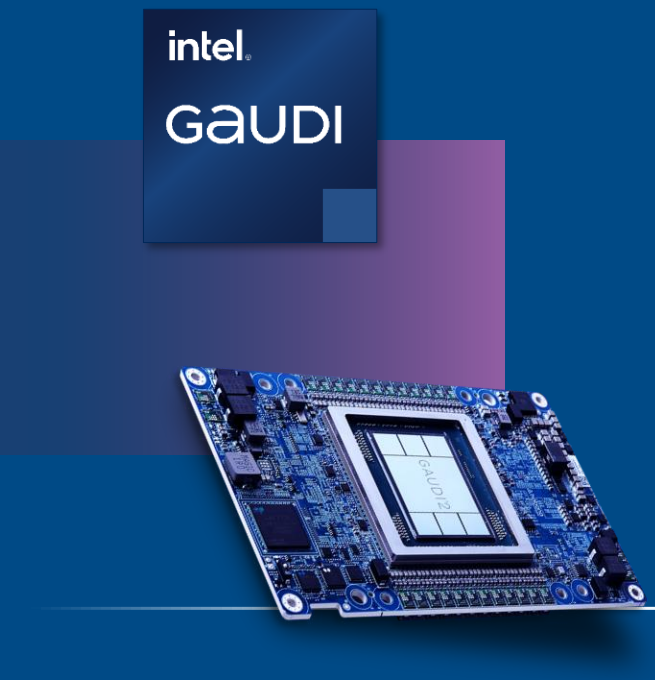
Gaudi2 は、LLM (GPT) や GAI (Stable Diffusion) といった大規模な基礎モデルの要求を満たすディープラーニングのパフォーマンス、効率性、スケーラビリティーを実現するために設計されています。

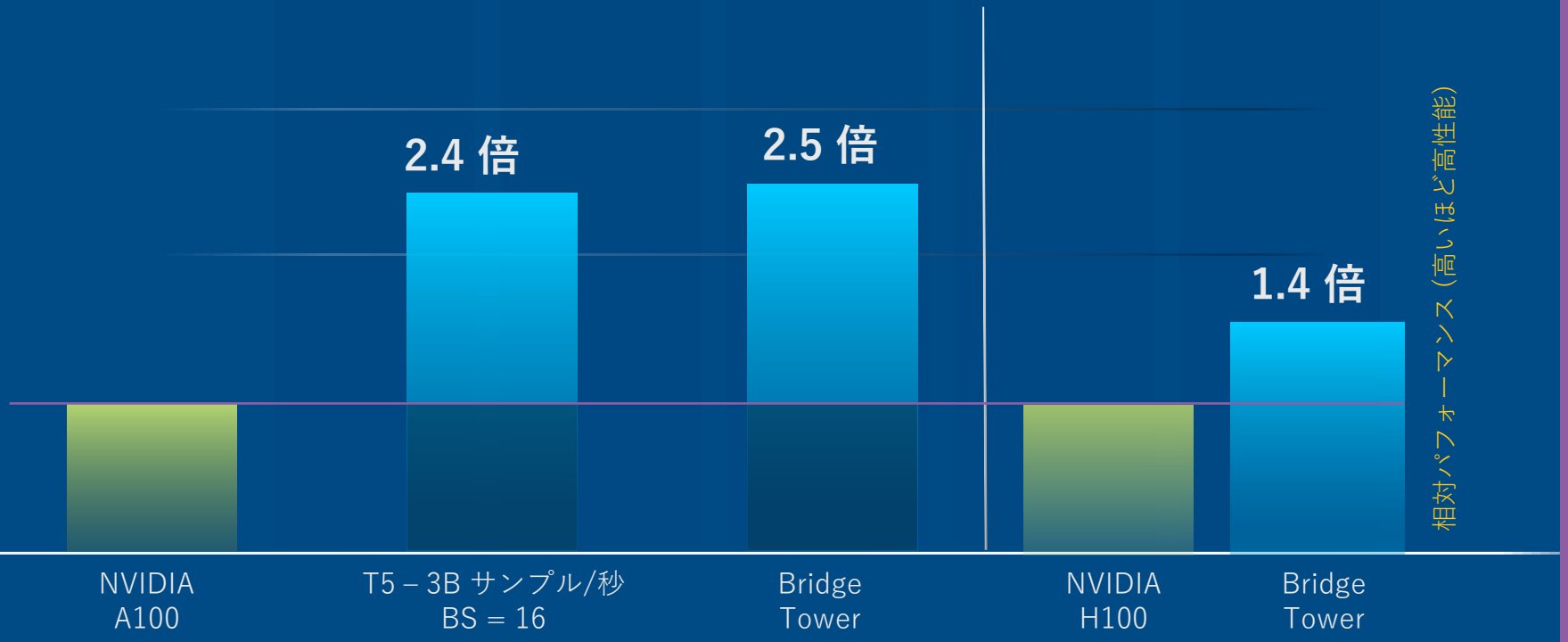
要件	Gaudi2
速度	トレーニングと推論の両方で A100 よりも **1.5～2 倍高速**
メモリー	各 Gaudi2 デバイスは、**96GB のオンチップの高帯域幅メモリー**を搭載しているため、大規模な基礎モデルをメモリーに簡単に取り込み、トレーニングと導入を実行可能
スケーラビリティー	**オンチップに統合された 24x100GbE ポート**、サーバーにある 8 枚のカード間のオールツーオールの直接接続、サーバー内およびサーバー間のオープンな ROCEv2 ベースの通信により、効率の高いスケーリングを実現
使いやすさ	SynapseAI、PyTorch、DeepSpeed を使用し、**最小限のコード変更**でモデルを移行または構築
電力効率	**A100 と比較して、ワット当たりスループットが最大 1.8 倍向上**
コスト効率	専用の第 1 世代の Gaudi アーキテクチャーに基づいて、Amazon クラウドで A100 より最大 **40% 優れたコスト・パフォーマンス**を実現

出典: https://habana.ai/habana-claims-validation/

多数の LLM での微調整

Hugging Face の評価では、NVIDIA A100 および H100 と比較した
インテル® Gaudi® 2 アクセラレーターの LLM のパフォーマンスが実証されました





ワークロードと構成については、https://habana.ai/habana-claims-validation を参照してください。 結果は状況によって変わります。

https://huggingface.co/blog/habana-gaudi-2-benchmark
https://huggingface.co/blog/bridgetower

GPT-J: インテル® Gaudi® 2 の結果

GPT-J のインテル® Gaudi® 2 推論パフォーマンスの結果によって、その競争力あるパフォーマンスが実証されました

- インテル® Gaudi® 2 推論パフォーマンス (GPT-J-99 と GPT-J-99.9) は、サーバークエリーで **78.58/秒、オフライン・サンプルで 84.08/秒**[1]
- **インテル® Gaudi® 2 は、NVIDIA の H100 と比較して説得力のあるパフォーマンスを実現。**H100 は、Gaudi 2 と比較して、1.09 倍
(サーバー) と 1.28 倍 (オフライン) というわずかな優位性を示した[1]
- **インテル® Gaudi® 2 は、NVIDIA の A100 を 2.4 倍 (サーバー) および 2 倍 (オフライン) 上回る性能を発揮**[1]
- インテル® Gaudi® 2 は FP8 を採用し、この新しいデータタイプで **99.9% の精度**を達成[1]

[1]性能は、使用状況、構成、その他の要因によって異なります。ワークロードと構成の詳細は、https://mlcommons.org/2023/09/mlperf-results-highlight-growing-importance-of-generative-ai-and-storage/ (英語) を参照してください。結果は状況によって変わります。

詳しくはこちら

**インテル® Gaudi® 2 ソフトウェアのアップデートは 6～8 週間ごとにリリースされるため、インテルは、MLPerf ベンチマークにおけるパフォーマンスの
向上とモデルカバレッジの拡大を継続的に
実現できる見込みです。**

ニュースルームの記事

MLCommons の発表

インテル® Gaudi® 2: ベンチマークの結果

Supermicro が提供するベンチマーク結果: 業界初のインテル® Gaudi® 2 の OEM

Gaudi に関する主張の検証

インテル® Gaudi® 2 AI アクセラレーターによる LLM のトレーニングと推論

ベンチマーク

より高速なトレーニングと推論: インテル® Gaudi® 2 と NVIDIA A100 80GB

ベンチマーク

結果は状況によって変わります。

インテル® Gaudi® 2: 基礎モデルのトレーニングと推論

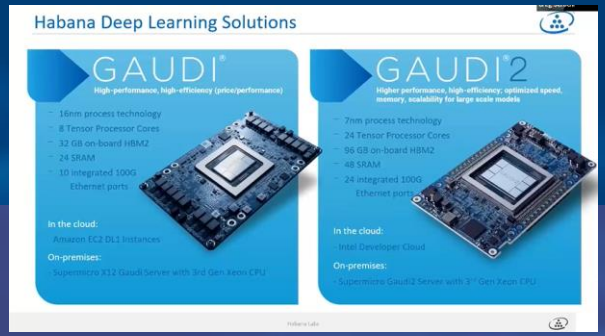
利用可能な Gaudi 対応モデルはこちらからアクセス可能

デベロッパー・カタログ

GAUDI®2

インテル® Gaudi® 製品デベロッパー・トレーニング

入門ガイド: Gaudi におけるディープラーニングと推論



インテル® Gaudi® 2 のパワーを最大限に活用: 生成 AI と大規模言語モデルの高速化

インテル® Gaudi® プロセッサーでモデルのパフォーマンスを最大限に高める: 最適な結果を得るための高度なツールと戦略

インテル® Gaudi® ソフトウェア（SynapseAI® ソフトウェア・スイート）

開発の簡素化: 思い通りの方法で開発が可能に

目標: 既存のソフトウェアのインテル® Gaudi® AI アクセラレーターへの**移行を容易**にすることで、ソフトウェアへの投資を維持し、ディープラーニング、生成 AI、大規模言語モデルを定義する多数の成長
モデルのトレーニングと導入の両方に対応する、新しいモデルの構築を容易にします。 データ・サイエンティスト、デベロッパー、IT およびシステムの管理者向けの広範なサポート:

- デベロッパー・サイト
- GitHub

インテル® Gaudi® AI アクセラレーター

ディープラーニングのソフトウェア・エコシステムでは、主要なソフトウェア・プロバイダー、ツール、コードを集めて、PyTorch、TensorFlow、PyTorch Lightning、DeepSpeed フレームワークに基づく最先端のディープラーニング・モデルの開発を高速化します。

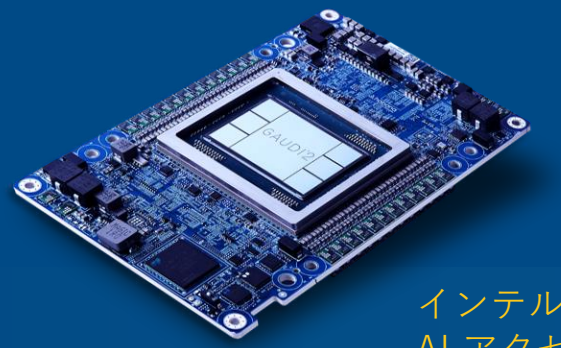
cnvrg.io

PyTorch Lightning

インテル® Gaudi® ソフトウェアを使用しましょう

インテル® Gaudi® 2 AI アクセラレーター Denvr Cloud 限定で現在入手可能

インテル® Gaudi® 2
ソフトウェア・エコシステム



インテル® Gaudi® 2
AI アクセラレーター(7nm)

インテル® Gaudi® 2 – 生成 AI の要求に最適

- 現在入手可能 - Denvr Cloud 上の Gaudi 2 クラスター
- 最大 8 つの Gaudi 2 ノードのテストドライブ
- インテルのお客様向け優先 VIP 価格
- Denvr Dataworks High-Touch 商用サービスとサポート
- Denvr Cloud 上の Gaudi 2 クラスターへのシームレスな移行
- Denvr Cloud 上の Gaudi 3 クラスターへの独占的な優先配置 – 近日公開

今すぐ始める

近日公開

intel GAUDI

トレーニングと推論

インテル® Gaudi® 3

より多くの顧客に、そのパフォーマンス、スケーラビリティー、効率を
もたらすインテル® Gaudi® 3 アクセラレーターは、企業のインサイト、イノベーション、収益の向上を支援します。

近日公開 - インテル® Gaudi® 3 AI アクセラレーター

パフォーマンス、スケーラビリティー、効率性を備えた GenAI の新たな選択肢

インテル® Gaudi® 3 は、GenAI の大規模な展開を目指すグローバル企業において、AI のトレーニングと推論の飛躍的進歩を実現します。

プレスリリース

NVIDIA H100 と比較した インテル® Gaudi® 3 アクセラレーターのパフォーマンス

インテル® Gaudi® 3 は、7B および 13B のパラメーターを有する Llama2 モデルと、175B のパラメーターを有する GPT-3 モデルで、**平均 50% の学習時間の短縮**[3]を実現すると予測されています。

詳しくはこちら

インテル® Gaudi® 3 は、以下のように H100 を上回るパフォーマンスを発揮すると予測されています。

50% アクセラレーター推論のスループット向上[1]

40% 推論の電力効率の向上[2]

Llama 7B/70B パラメーター・モデルおよび Falcon 180B パラメーター・モデルで実現

ホワイトペーパー

インテル® Gaudi® 3 は、2024年第 2 四半期から以下を含む OEM 向けに提供予定:

Lenovo DELL Hewlett Packard Enterprise SUPERMICRO

[1]NV H100 の比較は、https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8 に基づく。報告された数値は GPU 当たりの数値です。比較対象は、LLAMA2-7B、LLAMA2-70B & Falcon 180B に対するインテル® Gaudi® 3 の予測値です。結果は状況によって変わります。
[2]NV H100 の比較は、https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8 に基づく。報告された数値は GPU 当たりの数値です。比較対象は、LLAMA2-7B、LLAMA2-70B & Falcon 180B に対するインテル® Gaudi® 3 の予測値です。社内の推定に基づく、NVIDIA と Gaudi 3 の電力効率。結果は状況によって変わります。
[3]NV H100 の比較は、2024年3月28日時点における https://developer.nvidia.com/deep-learning-performance-training-inference/training「Large Language Model」タブと、LLAMA2-7B、LLAMA2-13B、GPT3-175B のインテル® Gaudi® 3 予測値との比較に基づく。結果は状況によって変わります。

NLP / LLM 向けインテル製品

推論

第 4 世代および第 5 世代インテル® Xeon® スケーラブル・プロセッサーは、
インテル® DL ブースト、インテル® AMX、インテル® AVX-512 により NPL を高速化しています。ハイパフォーマンス・コンピューティング向けに設計
されており、NLP ワークロードの高速化に使用できます。多数のスレッド、
大容量のメモリー容量、高メモリー帯域幅に対応しているため、言語翻訳、
テキストの要約、テキスト読み上げなどの NLP ワークロードに適しています。

第 5 世代インテル® Xeon® プロセッサー: AI 向けに設計されたプロセッサー

すべてのコアに AI アクセラレーションを搭載した第 5 世代インテル® Xeon® プロセッサーにより、お客様がディスクリート・アクセラレーターを追加する前に、要求の高いエンドツーエンドの AI ワークロードに対応

AI 推論のパフォーマンス向上

最大 42%

前世代との比較[1]

一般的なコンピューティング・パフォーマンスの向上

平均 21%

前世代との比較[1]

自然言語処理の高速化

最大 23%

前世代との比較[1]

インテル コーポレーション主席副社長 データセンター & AI 事業本部 本部長
サンドラ・リベラ

「AI 向けに設計された第5 世代インテル® Xeon® プロセッサー・ファミリーは、クラウド、ネットワーク、エッジの各ユースケースで AI 機能を導入するお客様に、より高いパフォーマンスを提供します。顧客、パートナー、開発者エコシステムとの長年の協力により、第5 世代インテル® Xeon® プロセッサーは、TCO を削減した迅速な導入と拡張を可能にする、実証済みの基盤の上に、リリースされています」

詳細情報

ウェブサイト

製品概要

ワークロードと構成については、intel.com/processorclaims の「第 5 世代インテル® Xeon® スケーラブル・プロセッサー・ファミリー」をご覧ください。結果は状況によって変わります。

インテル® Xeon® プロセッサー: 実世界の AI アプリケーションにおける CPU パフォーマンスのリーダーシップ

実際の業務アプリケーションにおいて、インテルは、以下の方法で、優れたパフォーマンス、低価格、バランスの取れたプラットフォームを提供することで、業界に破壊的変革をもたらし、**AI** の民主化を実現します。

- データ・ローカリティーに役立つキャッシュ容量の増加と、より大きな問題を解決できる大容量のメモリー
- コア周波数の向上、複数のスカラーポート、アウトオブオーダー実行により、シングルスレッドまたはマルチスレッドでありながらスカラーである計算処理の高速化を支援
- 非 DL ベクトル演算を支援するインテル® アドバンスト・ベクトル・エクステンション 512 (インテル® AVX-512)
- AI アクセラレーションを内蔵ハードウェアでサポートするインテル® アドバンスト・マトリクス・エクステンション (インテル® AMX)

インフォグラフィック

技術記事の全文

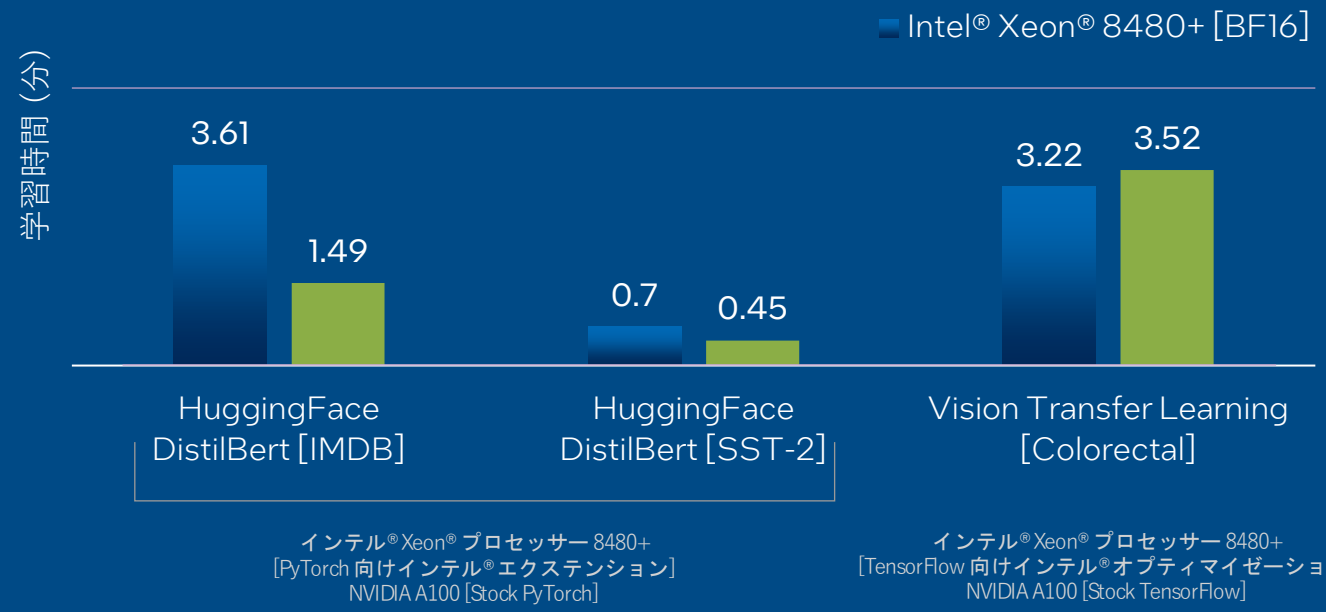
GPU 神話の虚像: 内蔵アクセラレーター搭載 CPU が AI に革命を起こす方法

インテル® Xeon® スケーラブル・プロセッサーにより、4 分未満でモデルの微調整が可能[1]

Hugging Face

インテル® Xeon® Platinum 8480+ プロセッサーと NVIDIA A100 GPU の比較による微調整の学習時間パフォーマンス

値が小さいほど高性能



インテル® Xeon® プロセッサー 8480+
[PyTorch 向けインテル® エクステンション]
NVIDIA A100 [Stock PyTorch]

インテル® Xeon® プロセッサー 8480+
[TensorFlow 向けインテル® オプティマイゼーション]
NVIDIA A100 [Stock TensorFlow]

参照：

パフォーマンスの向上:
インテル製 CPU の Numenta と NVIDIA GPU の比較

[1]第 4 世代インテル® Xeon® スケーラブル・プロセッサーのパフォーマンス・インデックス [A221] を参照してください。結果は状況によって変わります。

第 4 世代インテル® Xeon® プロセッサー上の LLMS

人工知能 (AI) チャットボット・テクノロジーは、顧客と
対話して顧客サービスを向上させる方法として、企業や
組織の間でますます人気が高まっていますが、**特定の
ユースケース向けのチャットボットの構築、最適化、メンテナンスは、多くの組織にとってコストが高く、経済的に困難な場合があります。**

詳細情報

第 4 世代インテル® Xeon® スケーラブル・プロセッサー搭載 AI 向けチューニング・ガイド

ガイドへのリンク ›

第 4 世代インテル® Xeon® プロセッサーは、**インテル® アドバンスト・マトリクス・エクステンション (インテル® AMX)** を通じて、強化されたデータ管理と効率的な計算処理を提供します。また、PyTorch 向けインテル® エクステンションで利用可能な **Auto Mixed Precision (AMP)** 機能と組み合わせることで、このテクノロジー・スタックは、転移学習や小 / 中規模のモデルをゼロからトレーニングするようなワークロードで、高い競争力を発揮します。

方法に関する技術記事

生成 AI 向け第 5 世代 / 第 4 世代インテル® Xeon® プロセッサー搭載 Cisco UCS

小さいほうが良い: Q8-Chat LLM は、インテル® Xeon® プロセッサー・ファミリーで利用できる効率的な生成 AI エクスペリエンス

LLM は、検索や対話アプリケーションなど、低レイテンシーのユースケースで十分な高速予測を行うために、一般的に
ハイエンド GPU に搭載されるような膨大な処理能力を必要とします。残念なことに、多くの組織にとって**関連コストは法外なものとなるため、アプリケーションで最新の LLM を
使用することは困難です。**

Hugging Face

「多くの企業は、トレーニングと実行のコストが低い、小規模な固有モデルに注力したほうが良好な結果が得られます。」

LLM のサイズと推論レイテンシーを削減し、インテルの CPU 上で効率的に実行できるよう支援する最適化の手法についてご覧ください。

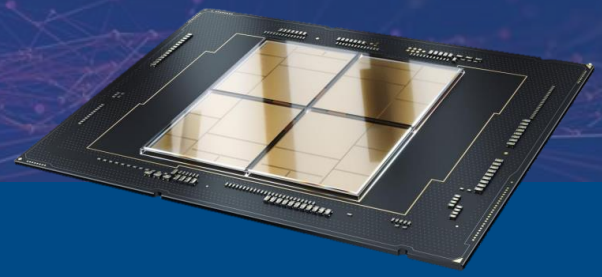
方法に関する技術記事 ›

第 4 世代インテル® Xeon® プロセッサーで Hugging Face を始める

LLM 向けインテル® Xeon® プロセッサー

まとめ

- 専門ドメインの LLM の推論に最適
- 転移学習のユースケースに対応
- オープンソースのソフトウェアでインテル® Xeon® プロセッサーに LLM を導入することで最適なパフォーマンスを容易に実現可能



LLM 向けインテル® Xeon® スケーラブル・プロセッサー

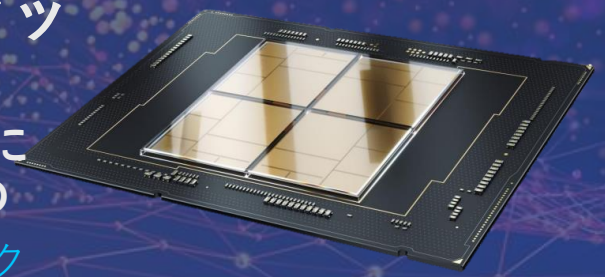
最も人気の高い AI フレームワークとライブラリーを使用した、汎用 AI ワークロードの構築と導入に最適

- ドメイン固有の LLM の推論に既存のインフラストラクチャーを活用
- 転移学習のユースケースに対応
- オープンソースのソフトウェアでインテル® Xeon® プロセッサーに LLM を導入することで最適なパフォーマンスを容易に実現可能

インテル® Xeon® プロセッサー

実環境の AI アプリケーションにおける CPU パフォーマンスのリーダーシップ

技術記事 • インフォグラフィック



GPT-J

第 4 世代インテル® Xeon® プロセッサーの結果

2 **段落 / 秒をオフライン・モードで実行[1]**

1 **段落 / 秒をリアルタイム・サーバー・モードで実行[1]**

ニュースルームの記事 • MLCommons の発表

GPU 神話の虚像: 内蔵アクセラレーター搭載 CPU が AI に革命を起こす方法

インテル® AMX 搭載第 4 世代インテル® Xeon® プロセッサーにおける Alibaba NLP のケーススタディー

詳しくはこちら

[1]性能は、使用状況、構成、その他の要因によって異なります。ワークロードと構成の詳細は、https://mlcommons.org/2023/09/mlperf-results-highlight-growing-importance-of-generative-ai-and-storage/ (英語) を参照してください。結果は状況によって変わります。

NLP / LLM 向けインテル製品

クライアントにおける小規模の推論

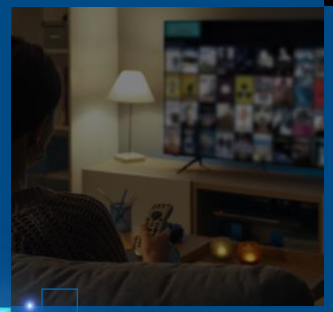
AI PC 時代の到来を告げるインテル® Core™ Ultra プロセッサー

インテル® Core™ Ultra プロセッサーは、プレミアムな薄型でパワフルなノートブック PC 向けに最適化され、3D パフォーマンス・ハイブリッド・アーキテクチャー、先進の AI 機能を搭載し、インテル® Arc™ GPU を内蔵しています。新しい Intel 4 プロセスを採用したインテル® Core™ Ultra
プロセッサーは、ゲーム、コンテンツ作成、外出先での生産性にパフォーマンスと電力効率の最適なバランスを提供します。

ユースケース: PC への AI 導入

クリエーター:写真と動画の検索と編集

自動化され、迅速化された検索で、より高速で自然なフィルター、より高品質なプレビュー、エクスポート時間の短縮を実現。



コラボレーション/ストリーミング

次世代のビデオ会議、ストリーミング、コラボレーション、バッテリー持続時間の維持を実現する新しい AI 機能。

メインストリームのゲーム

ゲーム内の臨場感を高める 3D アニメーション、文字起こし、チャット翻訳などの新しい AI 機能。

PC への AI 導入

生産性

執筆、創作、コーディング、テキストと文法予測などのオフライン機能の AI アシスタント。

クリエーター:テキストから画像への変換

マーケティング、広告、デザインなど、わずかな言葉で画像を生成できる新しい AI エフェクトと機能。

アクセシビリティー

多様なユーザーのニーズに対応する AI 支援のオーディオビジュアル機能により、PC での制作が容易になり生産性が向上。

「AI が可能性を解き放つ」

*インテルの 2023年12月14日「AI Everywhere」イベントのニュースリリース

生成 AI 向けインテル® Core™ Ultra プロセッサー

AI PC 時代の到来を告げるインテルの最も電力効率に優れたクライアント・プロセッサー

効率とパフォーマンスの大幅な向上

AI の効率
最大 **70%**
より高速な生成 AI パフォーマンス[2]

省電力性能
最大 **25%**
消費電力の削減[3]

詳しくはこちら

お知らせ ・ 製品概要 ・ ウェブサイト

インテル® Core™ Ultra プロセッサーは、インテル初のクライアント・オンチップ AI アクセラレーター (ニューラル・プロセシング・ユニット、NPU) を搭載し、前世代と比較して 2.5 倍の
電力効率を実現し、新次元の電力効率に優れた AI アクセラレーションを実現します。[1]

インテル® Core™ Ultra プロセッサー (H/U シリーズ) のチップには、低負荷ワークロード向けに 2 つの新しい低消費電力 (LP-E) コアが搭載されており、インテル® AI NPU 内の 2 つのニューラル・コンピューティング・エンジンは生成 AI 推論に対応するように設計されています。

[1]インテル® Core™ Ultra プロセッサー 165H の NPU と、インテル® Core™ i7-1370P プロセッサーの GPU を比較し、int8 モデルの実行中に UL Procyon AI ベンチマークのワット当たり性能を測定。
[1,2,3]ワークロードと構成については、www.intel.com/PerformanceIndex (英語) を参照してください。結果は状況によって変わります。

AI イノベーションの加速

インテルは、業界の主要な ISV と連携して、AI を用いた体験の最適化に努めています。

AI PC アクセラレーション・プログラムは、独立系ハードウェア・ベンダー (IHV)
および独立系ソフトウェア・ベンダー (ISV) と、人工知能 (AI) ツールチェーン、
トレーニング、共同エンジニアリング、ソフトウェアの最適化、ハードウェア、設計リソース、技術的専門知識、共同マーケティング、販売機会などの
インテルのリソースを結びつけることを
目的としています。

詳細を見る

インテル® Tiber™ AI クラウド で エンタープライズ AI 開発を加速

インテル最新のハードウェアとソフトウェアのクラスターでアプリケーションや
ワークロードを学習、プロトタイプ、テスト、実行しましょう

この開発環境では、最新のハードウェアとソフトウェアのイノベーションを活用して
AI を加速・拡張することができます。
ソフトウェアと生成 AI を微調整するために、計算能力を増強して選択肢を増やしましょう。

インテルと始める

最新のインテル製品をハンズオンで体験。インテルで AI スキルを強化しましょう。

テクノロジーへの早期アクセス

インテルのプラットフォームとインテルに最適化された関連ソフトウェア・スタックのプレリリース版を評価できます。

AI を大規模に導入

インテルによる最新のマシンラーニング・ツールキットと、インテル® Tiber™ AI クラウドでホストされているライブラリーにより、AI 導入を加速できます。

技術記事を読む ›

始める ›

実施すること

教育

インテルのテクノロジーが、
生成 AI とドメイン固有のモデル
にどのように使用できるか、そして
インテル® Xeon® プロセッサー
およびインテル® Gaudi® 製品
ラインが、ビジネスの拡大にどの
ように役立つかを説明します。
始める

エンゲージメント

始めましょう:

インテル® Tiber™ AI クラウド

この開発環境では、最新のハードウェアと
ソフトウェアのイノベーションを活用して
AI を加速・拡張することができます。

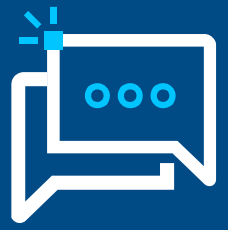
&

AI リファレンス・キットを使用する

お問い合わせ

詳細については、
インテルの担当者に
お問い合わせください。

インテル® パートナー・アライアンスのカスタマーサポートにアクセスする方法



Intel® Virtual Assistant

このチャットボットは、パートナー・アライアンスの各ウェブページの右下に設置されており、ほとんどの質問に対するセルフヘルプ、またはライブサポート・エージェントへのクイックリンクを提供します。

「サポート」ブレード

オンライン・サポートのリクエストを送信します。

このリンクは、パートナー・アライアンスのウェブサイトでは、多くのページのフッターに表示されています。

パートナー・アライアンスの「サポート」ページ

サポートページでは、パートナー・アライアンスのメンバーが利用できるほとんどのツールや特典に関する詳細なセルフヘルプ・ガイドを提供しています。

AI アクティベーション・ゾーン

重要なリソース、ツール、利点などを整理するデジタルファーストの AI ワークスペース - インテルのテクノロジーに基づくソリューションの構築、マーケティング、販売パートナーのアクティベーションを支援します。

技術支援

セールスとマーケティング支援

技術支援

セールスとマーケティング支援

技術支援

セールスとマーケティング支援

AI リファレンス・キット

これらのリファレンス・キットを活用することで、企業は解決までの**時間を大幅に短縮**できるほか**パフォーマンスの大幅な向上**を体験できる

金融と保険

不正検出

GitHub・ブログ・ブループリント

医療とライフサイエンス

疾病の予防

GitHub・ブログ

製造と公益事業

異常検知

GitHub・ブログ

フリート管理

予測メンテナンス

GitHub

プロセスの自動化

ドキュメント・オートメーション

GitHub・ブログ・ブループリント

ワークフロー

- DL 転移学習
- HF 微調整と推論の最適化
- DL 分散圧縮
- 分散型の古典的 ML ワークフロー
- インテル® アクセラレーターによる DL 事前トレーニング
- DGL & PyG によるグラフ分析と GNN
- Big-DL における分散トレーニング / 推論
- Ray での LLM 事前トレーニングと微調整

ツール

- インテル® ディストリビューションの Python
- インテル® Optimized Modin
- インテル® Optimized XGBoost
- scikit-learn 向けインテル® エクステンション
- インテル® Optimized Tensorflow (ITEX)
- インテル® Optimized PyTorch (stock & IPEX)
- インテル® ニューラル・コンプレッサー
- SigOpt Python SDK & CLI
- CNVRG Python SDK & CLI
- インテルに最適化された Horovod
- DeepSpeed

ドメインキット

- 時系列
- PPML
- 転移学習
- Transformer/NLP

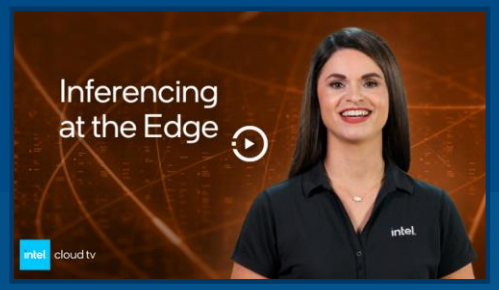
リファレンス・キットは**コンテナ**として提供され、**オンプレミスだけでなく、主要クラウド**でも使用できます。**リファレンス・キット**は、**ワークフローと**

ドメイン・ツールキット上に階層化されています。それぞれ独立して活用することで、**複数の業界の幅広い多様なユースケースをサポートできます。**

クラウド TV

インテル® クラウド TV は、お客様を成功に導くため、クラウド・コンピューティングのニュース、トレンド、戦略を探ります

インテル® Gaudi® AI アクセラレーターによる GenAI の機会



エッジでのデータ推論を使用してインサイトを獲得

クラウドでの AI による競争優位性の創出

クラウド・テクノロジーを用いた AI 推論

クラウドにおける AI

高速パスで、 場所を問わずに AI を拡張

トレーニング

AI をあらゆる場所に導入
生成 AI エンタープライズのユースケース

生成 AI は、単なるインターネット・チャットボット用では
ありません。無数の企業が、日々の業務を支援するために、
生成 AI と大規模言語モデルのパワーを活用する方法を
模索しています。このセッションでは、エンタープライズに
おける生成 AI のユースケースを探索し、お客様の組織が
日常的な業務に対し、どのように AI を適用できるかを
考察します。

登録 ›

データ生成と大規模言語モデル向けの
AI の効率化

登録 ›

AI を組織のワークロードに組み込む、または既存のインフラストラクチャーを拡張するには、高度なスキルと多大な計算処理が不可欠で
あり、膨大なデータセットで学習した強力なモデルの開発と、それらを
適切に実行する強力な GPU を必要とします。すべての組織が、このようなタスクの達成に必要なリソースを持っているわけではありません。

このセッションでは、Accenture* とインテルが開発したソリューション、つまり、組織が利用しやすい AI を実現し、学習と推論の時間を短縮で
きるよう最適化されたオープンソースの AI リファレンス・キットの
コレクションに焦点を当てます。

その他のトレーニング

テクニカル

アセットタイプ	タイトルとリンク
コンピテンシー	「クラウドにおける AI」コンピテンシー
ウェビナー	Hugging Face でインテル® ハードウェア向けに AI を最適化
ウェビナー	LLM を微調整するクラウドベースの分散トレーニングを設定する方法
トレーニング・コース	迅速なコスト削減とインコンテキスト・ラーニングによる LLM の改善
トレーニング・コース	データ生成と大規模言語モデル向けの AI の効率化
トレーニング・コース	自然言語処理
トレーニング・コース	TensorFlow* による応用ディープラーニング
トレーニング・コース	小規模で迅速なエンタープライズ GenAI への高速パス
トレーニング・コース	GenAI の次の波 - ドメイン固有の LLM
ガイド	生成 AI の導入のための開発者ガイド: ユースケース別のアプローチ
トレーニング・コース	インテル® Xeon® プロセッサーで AI をソリューション分野へ導入

その他のトレーニング

非技術的

アセットタイプ	タイトルとリンク
ビデオシリーズ	生成 AI の採用
トレーニング・コース	小規模で迅速なエンタープライズ GenAI への高速パス
トレーニング・コース	GenAI の次の波 - ドメイン固有の LLM
トレーニング・コース	「あらゆる場所での AI 活用の原則」コンピテンシー
トレーニング・コース	「AI ソフトウェアとエコシステムの原則」コンピテンシー
トレーニング・コース	AI エコシステムの活用: ソフトウェアで勝ち、SI で拡張し、ソリューションを販売
トレーニング・コース	実用世界向けの生成 AI と大規模言語モデル

関連情報

アセットタイプ	タイトルとリンク
ウェビナー	生成 AI ウェビナーシリーズ
ウェビナー	あらゆる場所に GenAI を導入する
ポッドキャスト	Copilot、ChatGPT、Stable Diffusion、生成 AI によって開発、仕事、生活はどう変わるのか
ビジネス概要	AI をあらゆる場所に導入
ブログシリーズ	第 4 世代インテル® Xeon® プロセッサーによる生成 AI の調整と推論
ソリューション概要	Lenovo ThinkSystem SR650 V3 / 第 4 世代インテル® Xeon® プロセッサーによる生成 AI 推論の導入と拡張 新しいインテルと VMware のテクノロジーが Lenovo ThinkAgile VX V3 Systems を高速化
技術記事	インテル® AI ハードウェアとソフトウェアの最適化による Llama 2 の高速化
リサーチ PR	調査対象の組織のうち 10% が、2023年に GenAI ソリューションを本番環境で稼働
Fireside Chat のビデオ	生成 AI に立ちはだかるコンピューティングとサステナビリティーの課題に取り組む
ポッドキャスト	Hugging Face とインテル - 実用的かつ高速、民主化された倫理的な AI ソリューションの実現に向けて
Twitter / X のスレッド	民主化された大規模言語モデルが AI 開発を促進する方法
Supermicro ベンチマーク	Habana に関する主張の検証
Hugging Face ベンチマーク	ベンチマーク
トレーニング / ウェビナー	クラウド・ソリューション・アーキテクト (CSA) Tech Talk: Habana による AI
ホワイトペーパー	エンタープライズ AI は、開発者向けホワイトペーパーのすべて
インフォグラフィック	CPU はエンタープライズ AI の鍵

関連情報

アセットタイプ	タイトルとリンク
ソリューション概要	インテル® エンタープライズ AI と Red Hat® OpenShift® AI を使用して AI の採用および導入を合理化
ガイド	AI ガイド
リファレンス・キット	AI 非構造化テキストデータ生成
ホワイトペーパー	Zoho、ビデオ AI ワークロードの最適化と高速化を実現
ホワイトペーパー	Seekr、信頼できる AI スクリーニング・システムを開発
ソリューション概要	教育におけるセキュリティー: AI とコンフィデンシャル・コンピューティングが安全なリモート受験を実現
ケーススタディーとビデオ	Nature Fresh Farms、種から店舗まで AI を活用
ケーススタディー	QMed Asia、早期がん発見率を促進向上
ケーススタディーとビデオ	MetaApp、AI ベースのレコメンデーション・システムを刷新
ソリューション概要	自動光学検査（AOI）のための AI モデル・トレーニングと改良の最適化
ブログ	LLM のためのプロンプト主導の効率性

法務情報および免責事項

通知と免責事項


intel