IA d'entreprise

IA générative et modèles spécialisés pour les entreprises

Optimisez l'entraînement et le déploiement grâce à du matériel et à des logiciels Intel® AI conçus spécifiquement pour vous aider à transformer votre entreprise

intel®

Sommaire

Pourquoi s'associer à Intel ?

Chez Intel, nous avons la possibilité d'améliorer le quotidien et l'avenir de chaque personne et de chaque entreprise dans le monde.

Mais nous ne travaillons pas seuls !

En collaboration avec nos partenaires, nous créons une véritable valeur ajoutée pour nos clients en **intégrant l'IA partout** et en minimisant les risques inhérents à son déploiement

Lorsque vous devenez partenaire d'Intel, vous rejoignez un écosystème d'IA complet

Notre vaste portefeuille de technologies favorisant l'IA et nos partenariats avec des intégrateurs de matériel, de logiciels et de systèmes, qui travaillent tous en étroite collaboration, créent des solutions concrètes qui permettent d'obtenir des résultats commerciaux différenciés pour l'industrie, les entreprises et les communautés.
Nous pouvons ainsi vous aider à développer votre activité.

Rejoignez-nous dans notre mission : intégrer l'IA partout

Créer de la valeur pour les clients à l'aide des solutions Intel® AI

L'approche d'Intel permet à un vaste écosystème ouvert d'acteurs de l'IA d'offrir des solutions qui répondent aux besoins spécifiques des entreprises en matière d'IA générative

NAVER

Développement d'un grand modèle de langage (Large Language Model ou LLM) puissant pour déployer des services d'IA avancés à l'échelle globale, du Cloud aux appareils. NAVER a confirmé la capacité fondamentale d'Intel® Gaudi® à exécuter les calculs des modèles de transformateurs à grande échelle avec des performances par watt exceptionnelles.

BOSCH

Exploration d'autres possibilités de fabrication intelligente, notamment des modèles fondamentaux pour la génération d'ensembles de données synthétiques d'anomalies de fabrication, afin de fournir des ensembles d'entraînement robustes et répartis de manière uniforme (par exemple, pour l'inspection optique automatisée).

IBM

Utilisation des processeurs Intel® Xeon® de 5e génération pour son magasin de données watsonx.data™ et collaboration étroite avec Intel® pour valider la plateforme watsonx™ d'accélérateurs Intel® Gaudi®.

OLA

Pré-entraînement et perfectionnement de son premier modèle fondamental en Inde avec des capacités génératives en 10 langues, afin de produire des solutions dont le rapport prix/performances est le meilleur du marché. Krutrim pré-entraîne actuellement un modèle fondamental plus important sur un cluster Intel® Gaudi® 2.

seekr

Ce leader de l'IA fiable exécute des charges de travail de production sur Intel® Gaudi® 2, Intel® Data Center GPU Max Series et sur les processeurs Intel® Xeon® dans Intel® Tiber™ Developer Cloud pour prendre en charge le développement et le déploiement de production de LLM.

IFF

Ce leader mondial dans les domaines de l'alimentation, des boissons, des parfums et de la biologie compte exploiter l'IA générative et la technologie des jumeaux numériques pour établir un flux de travail de biologie numérique intégré pour la conception avancée d'enzymes et l'optimisation des processus de fermentation.

airtel

Airtel prévoit de profiter de la puissance des technologies de pointe d'Intel pour exploiter ses riches données de télécommunications afin d'améliorer ses capacités d'IA et l'expérience de ses clients. Les déploiements respecteront l'engagement d'Airtel à rester à la pointe de l'innovation technologique et permettront de générer de nouvelles sources de revenus dans un paysage numérique en évolution rapide.

Infosys

Ce leader mondial des services numériques et de conseil de nouvelle génération a annoncé une collaboration stratégique visant à intégrer les technologies Intel® (notamment les processeurs Intel® Xeon® de 4e et 5e génération, les accélérateurs d'IA Intel® Gaudi 2 et les processeurs Intel® Core™ Ultra) à Infosys Topaz, un ensemble de services, de solutions et de plateformes axés sur l'IA qui accélèrent la valeur commerciale à l'aide de technologies d'IA générative.

Proposition de valeur de l'IA d'entreprise

Transformer votre activité avec l'IA d'entreprise

Dans l'environnement hyperconcurrentiel actuel, **les entreprises qui adoptent l'IA prennent de l'avance.**

Les entreprises de tous les secteurs tentent de réimaginer chaque aspect de leurs opérations pour comprendre comment l'IA peut améliorer ou même automatiser les flux de travail.

Chez Intel, notre expertise unique consiste à intégrer l'IA dans le tissu de l'entreprise.

Qu'il s'agisse de bénéficier de PC accélérés par l'IA qui transforment la productivité, ou de tirer parti de nos années d'expérience pour comprendre quels cas d'utilisation génèrent le plus de valeur, Intel® est votre partenaire de confiance pour intégrer l'IA partout, en toute sécurité et de manière responsable.

L'adoption des innovations en matière d'IA générative (GenAI) par les entreprises de toutes tailles sera encore plus rapide que celle de l'Internet, de la téléphonie mobile ou du Cloud.

La prochaine vague de plateformes d'IA adoptera ces réalités passionnantes d'une manière abordable et flexible.

Il est temps d'adopter une approche différente de votre IA d'entreprise.

Ce package d'habilitation vous aidera à comprendre comment les entreprises de tous les secteurs peuvent profiter au maximum de l'IA générative, notamment des modèles spécialisés, pour réussir à long terme.

Qu'est-ce que l'IA générative et que sont les grands modèles de langage ?

L'IA générative (GenAI) est un sous-ensemble d'IA axé sur la création de nouveaux contenus originaux.

Elle implique l'entraînement et le déploiement de modèles d'IA qui permettent de générer des données telles que des images, du texte ou de l'audio, qui ressemblent étroitement aux exemples utilisés dans l'ensemble de données d'entraînement.

Les algorithmes d'IA générative utilisent des techniques avancées telles que le Deep Learning et les réseaux neuronaux pour produire des résultats réalistes et cohérents qui permettent, entre autres, la synthèse d'images, la génération de texte et même la création d'œuvres d'art.

Un grand modèle de langage (LLM) est un type spécifique de modèle de traitement du langage naturel qui utilise des réseaux neuronaux profonds pour traiter et générer du texte. Les LLM sont formés à partir d'énormes quantités de textes et sont conçus pour générer des résultats cohérents et ayant un sens.

En savoir plus

LIRE LA SUITE

Tirez parti de la puissance de l'IA générative

Comment les entreprises utiliseront-elles l'IA générative ?

Biens de consommation et vente au détail

- Salles d'essayage virtuelles
- Livraison et installation
- Assistance à la recherche de produits en magasin
- Prévision de la demande et planification des stocks
- Designs de produits novateurs

Soins de santé et médecine

- Assistance au personnel en contact avec le public
- Transcription et résumé de notes médicales
- Chatbots pour répondre aux questions médicales
- Analytique prédictive pour accompagner le diagnostic et les traitements

Industrie

- Copilote expert pour les techniciens
- Interactions conversationnelles avec les machines
- Service de terrain prescriptif et proactif
- Dépannage en langage naturel
- Statut de la garantie et documentation
- Analyse des goulots d'étranglement, élaboration de stratégies de récupération

Médias et divertissement

- Recherche intelligente, découverte de contenu sur mesure
- Élaboration de titres et de textes
- Commentaires en temps réel sur la qualité du contenu
- Listes de titres personnalisées, bulletins d'information, recommandations
- Récit interactif selon les choix de l'utilisateur
- Offres ciblées, plans d'abonnement

Services financiers

- Découverte de signaux de trading, alertes pour traders en positions vulnérables
- Accélération des décisions de placement
- Optimisation et reconstruction des systèmes existants
- Rétroingénierie des modèles bancaires et d'assurance
- Surveillance des infractions et des fraudes financières potentielles
- Automatisation de la collecte de données pour respecter la réglementation
- Extraction d'informations des renseignements fournis par les entreprises

Source : compilé par MIT Technology Review Insights, d'après les données de « 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 » et « Large Language Models in Media & Entertainment »12, Databricks, avril-juin 2023.

Cas d'utilisation de l'IA générative et des grands modèles de langage

Chatbots et assistants virtuels

Assistance à la clientèle

LLM pour la génération de code et le débogage

Entraînement sur les documents de l'entreprise

Analyse des sentiments

Évaluation de la satisfaction des clients

Classification et regroupement de textes

Catégorisation de grands volumes de données pour identifier les tendances

Traduction linguistique

Traduction des pages Web d'une entreprise dans d'autres langues

Résumer et paraphraser

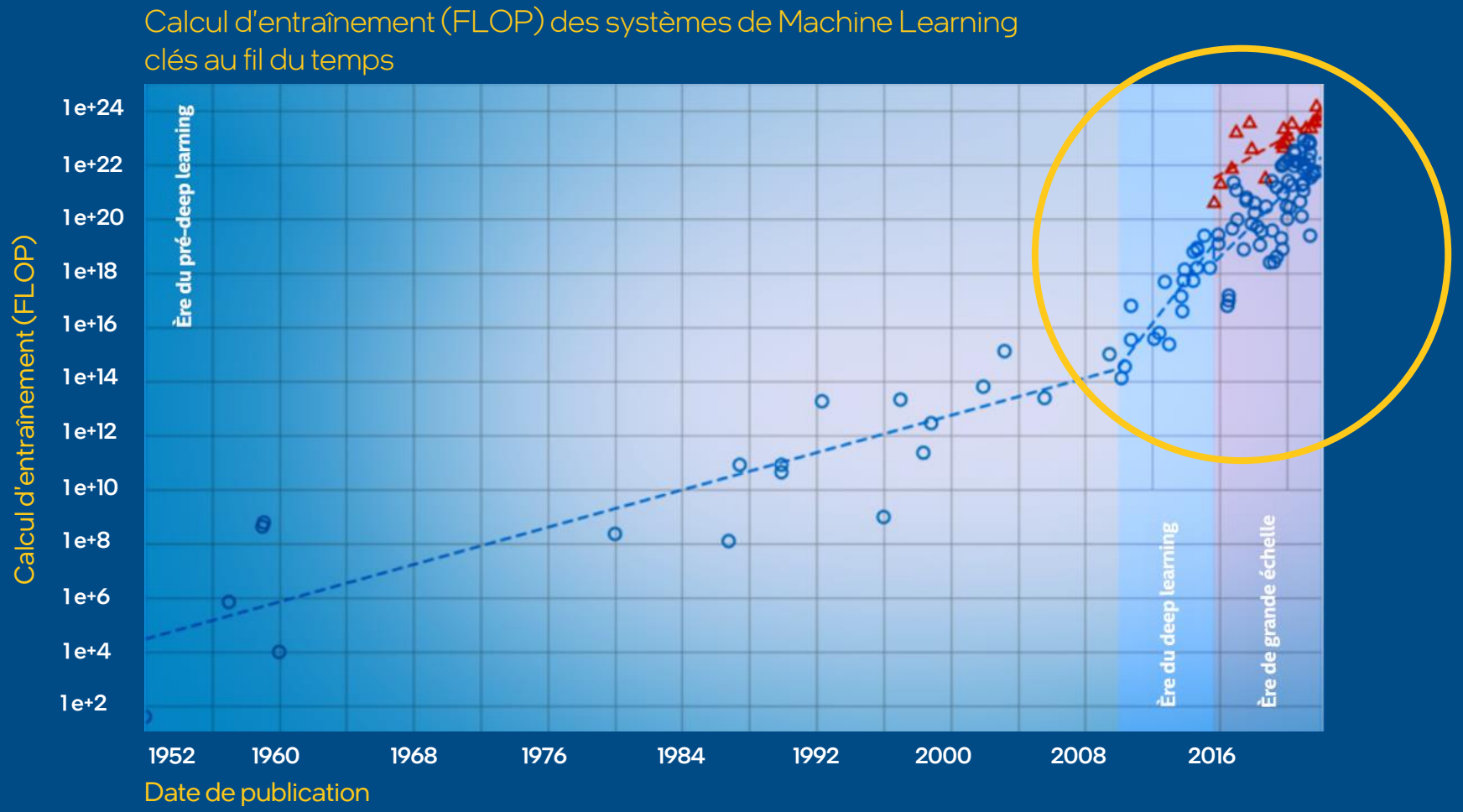
Résumé des notes de réunion

Génération de contenu, d'images et de vidéos

Premiers brouillons d'e-mails, génération d'idées, images de marketing, courtes vidéos

Source : « A Compact Guide to Large Language Models », Databricks, 2023.

L'accroissement de la taille des modèles s'accompagne d'une augmentation des calculs



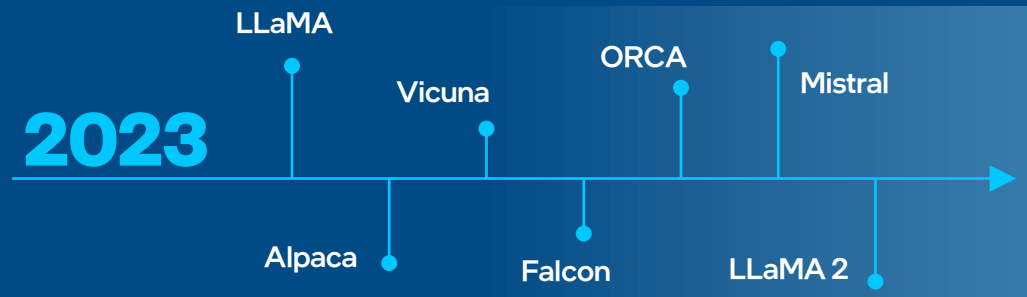
Étude d'Epoch, Université d'Aberdeen, Centre pour la gouvernance de l'IA, Université de St. Andrews, MIT, Eberhard Karls Universitat Tubingen, Universidad Complutense

Au-delà des modèles géants

	Géant (tiers)	VS.	**Petit et agile** (de 10 à 100X)
Explicabilité	Modèle propriétaire	VS.	Modèle Open Source
Précision	Tout-en-un généraliste	VS.	Ciblé, spécialisé, personnalisé
Emplacement	Basé sur le Cloud (en tant que service)	VS.	Inférence exécutée localement ; Edge, client et sur site
Coût	Coût perpétuel de mise à l'échelle	VS.	Gestion des coûts
Délais de commercialisation	Configuration rapide (secondes)	VS.	Délais de création (heures/jours)

La croissance de nombreux petits modèles

De centaines de milliards à <20 milliards de paramètres en 6 mois



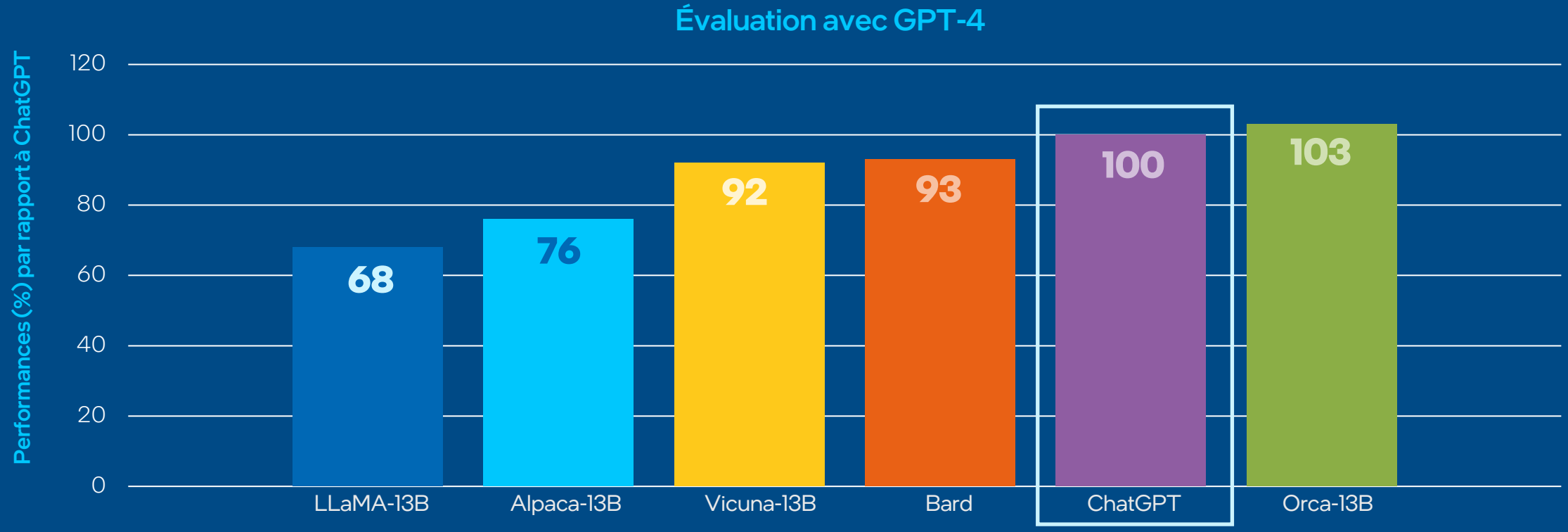
databricks

- Des dizaines de petits modèles émergent chaque semaine
- Licences commerciales et Open Source
- Indication que les petits modèles peuvent offrir la même précision que les grands modèles s'ils sont entraînés sur des données soigneusement obtenues

- Des milliers de modèles commerciaux et de plateformes d'IA spécialisés sont en cours de démonstration
- Les modèles peuvent être affinés sur quelques processeurs avec des données spécifiques à un domaine

Les petits modèles donnent de bons résultats par rapport à ChatGPT

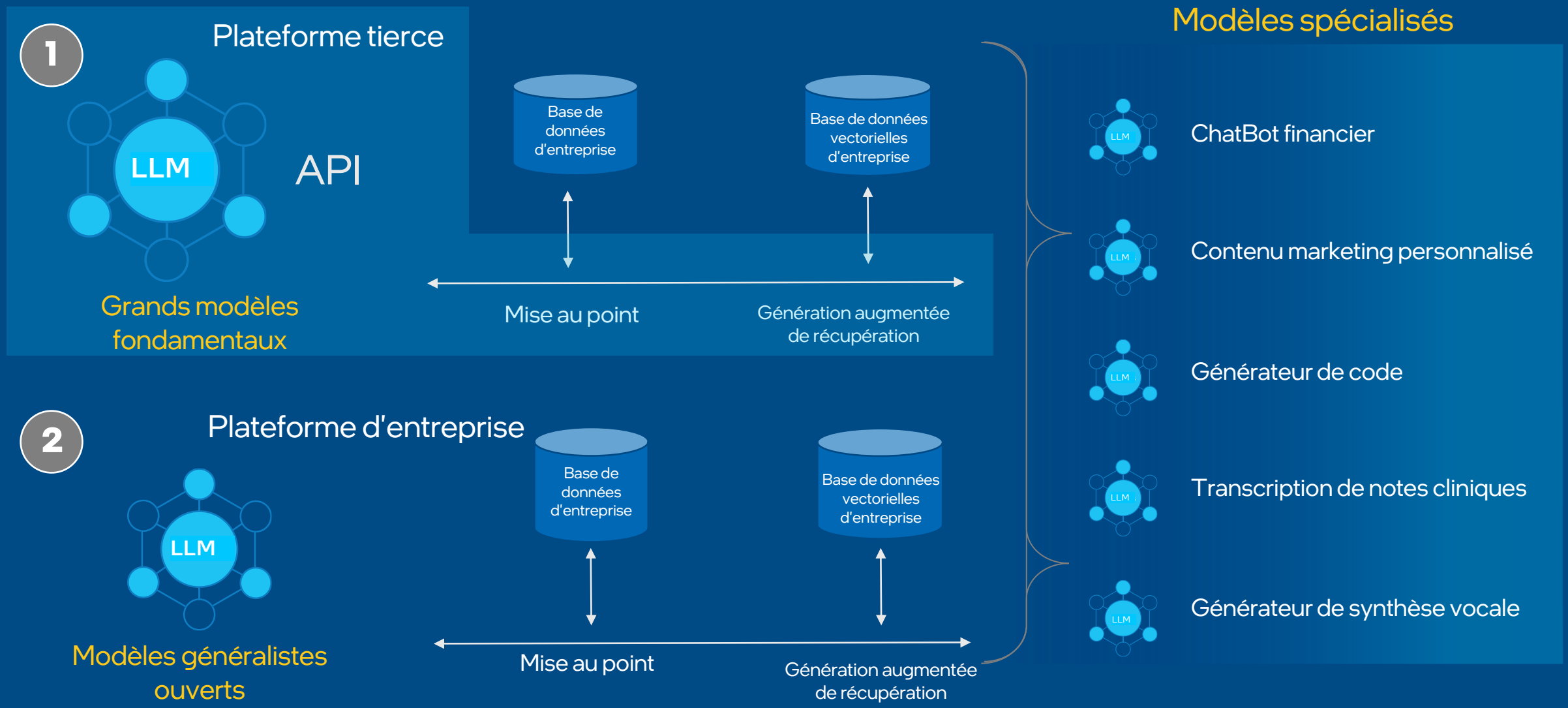
Cela prouve que les petits modèles constituent une option viable et restent performants par rapport aux grands modèles comme ChatGPT



Orca surpasse un large éventail de modèles fondamentaux, notamment OpenAI ChatGPT tel qu'évalué par GPT-4 dans l'ensemble d'évaluation Vicuna

Source : Microsoft Research (2023). Orca : apprentissage progressif à partir de traces d'explication complexes de GPT-4

Création de modèles spécialisés



Les modèles spécialisés présentent de nombreux avantages pour les entreprises

Les petits modèles ciblés peuvent offrir des performances équivalentes ou supérieures, ce qui accroît le ROI en réduisant les investissements en temps et en coûts

Résultats plus précis

Utilisez les données de votre entreprise pour obtenir des résultats plus précis dans un domaine spécifique

Coûts réduits

Mise au point d'un modèle pré-entraîné et/ou utilisation de la RAG, et production d'inférences sur un modèle plus petit

Déploiement n'importe où sur la plateforme choisie

Inférence exécutée localement ; Edge, client et sur site

Sécurité et confidentialité

Respect des exigences règlementaires et relatives à la sécurité des données

IA responsable

Le modèle peut citer la source des données grâce à une mise au point et à la RAG

L'AVENIR

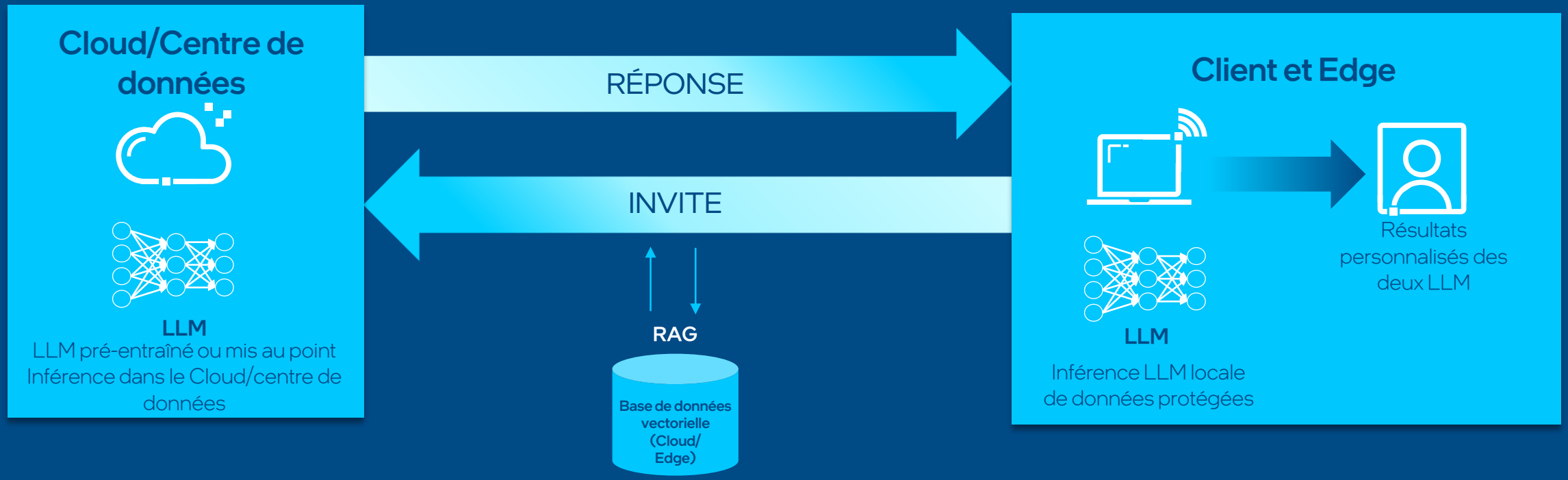
Il y aura un petit nombre de modèles gigantesques et un nombre gigantesque de petits modèles d'IA plus agiles intégrés dans d'innombrables applications[1]

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

Plateforme d'IA **transparente** du Cloud à l'Edge

Portefeuille des solutions d'IA Intel®

Entraînement et production d'inférences dans le Cloud. Utilisation de la RAG pour améliorer la précision dans des domaines spécifiques.



IA générative : une année de production

L'utilisation de modèles spécialisés mais hautement intelligents s'accroît

2022 EXPÉRIMENTATION → **2023** OPÉRATIONS PILOTES → **2024** PRODUCTION

Les modèles énormes ont ouvert la voie

- Très efficaces pour un usage général
- Coûteux à former et à déployer
- Basés sur de grands ensembles de données publics
- Simplicité d'emploi

Petits modèles spécialisés

- Utilisez vos données privées pour obtenir des résultats spécifiques à votre activité
- Déployez-les sur le matériel dont vous disposez
- Efficacité, précision, sécurité et traçabilité accrues
- Délais de création

LIRE LE BLOG

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

Approche d'Intel concernant les modèles spécialisés

MODÈLES SPÉCIALISÉS

Avantages

+ Des modèles 10 à 100 fois plus petits qui conservent ou améliorent la précision
+ Économiques sur les plateformes informatiques généralistes
+ Exactitude ; attribution de la source ; explicabilité
+ Utilisation de données privées/d'entreprise
+ Informations mises à jour en permanence

Difficultés

- Plage de tâches réduite
- Nécessite une mise au point et une indexation en plusieurs étapes

OBJECTIF D'INTEL

Adoption de l'approche la plus rentable et la plus omniprésente pour mettre au point et déployer des dizaines de milliers de modèles sur le matériel Intel en utilisant des frameworks de l'industrie, des modèles pré-entraînés et des outils et logiciels Intel AI

LIRE LA SUITE

L'IA générative au bout des doigts

E-book ▪ Infographie

IA d'entreprise : élimination des obstacles

Exigences	Les avantages d'un partenariat avec Intel®
Délais de commercialisation	Utilisez les ressources pour développeurs d'Intel et de Hugging Face, Gaudi Developer Hub et 5 kits de référence pour vous lancer dans le domaine de l'IA générative
Expérience utilisateur (précision/latence)	Inférence sur des modèles de plus de 10 milliards de paramètres sur l'accélérateur Intel® Gaudi® et sur des petits modèles de moins de 20 milliards de paramètres sur les processeurs Intel® Xeon® avec Intel® AMX, pour offrir aux utilisateurs une expérience en temps réel[1]
Disponibilité des ressources de calcul	Les CPU Intel® Xeon® accompagnés d'accélérateurs offrent une alternative rentable dans le cadre de la pénurie mondiale de GPU. **Intel® Gaudi® 2 est disponible dès maintenant chez SuperMicro, avec une plus grande disponibilité d'Intel® Gaudi® 3.**
Une technologie familière	L'inférence de petits modèles peut se faire pratiquement sur n'importe quel matériel, y compris sur des solutions omniprésentes qui font peut-être déjà partie de votre configuration informatique[2]
Opérationnaliser à grande échelle	Intel® Gaudi® 2 offre une évolutivité quasi linéaire, avec 24 ports 100 GbE intégrés sur chaque accélérateur. Intel® Xeon® est déjà présent dans votre centre de données, sur le terrain, du Cloud à l'Edge. **65 % des inférences de centres de données s'exécutent sur Intel® Xeon®[3]**
Rentabilité	Dans les applications professionnelles réelles, Intel® révolutionne le secteur et démocratise l'IA en offrant de meilleures performances, des tarifs inférieurs et une plateforme plus équilibrée pour l'inférence de l'IA. Consultez l'article « NVIDIA shows Intel® Gaudi 2 is 4x better performance per dollar than its H100 » (NVIDIA montre qu'Intel® Gaudi 2 est quatre fois plus performant par unité de prix que son GPU H100)

[1]Source : Four Roadblocks to Implementing Generative AI

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

[3]D'après la modélisation de marché Intel® de la base installée à l'échelle mondiale de serveurs de centres de données exécutant des charges de travail d'inférence de l'IA en décembre 2022.

Ressources logicielles qui simplifient l'entraînement et le déploiement de l'IA générative

Modèle Open Source

- BLOOM — 176 Md
- BioGPT — Domaine 1,5 Md
- Stable Diffusion — Image
- Llama2 GPT-J MPT Falcon — LLM 7-65 Md
- Stanford Alpaca — LLM 7 Md affiné
- Base de connaissances

Logiciels ouverts

- OpenVINO
- Intel® Extension for PyTorch (IPEX)
- Intel® Extension for Transformers (ITREX)
- Intel® Extension for DeepSpeed (IDEX)
- haystack
- fastRAG

Plateforme d'IA générative

- intel. Gaudi
- intel. Data Center GPU Max Series
- intel. Xeon
- intel. Core Ultra

LIRE LA SUITE

Accéder à l'IA générative grâce à du matériel omniprésent et des logiciels ouverts

Optimiser la valeur

Évitez la dépendance vis-à-vis des fournisseurs

Logiciels open source basés sur des normes

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

Tirez parti du portefeuille de matériel d'Intel

optimisé pour les cas d'utilisation de l'IA

intel Xeon · intel Core Ultra · intel Arc Graphics · intel Data Center GPU · intel Gaudi · intel Ethernet

Créez de nouvelles opportunités, du client et de l'Edge, au centre de données et au Cloud, avec du matériel optimisé par des logiciels et des normes ouvertes pour l'IA de demain

Portefeuille de logiciels d'IA Intel®

Données d'ingénieur → **Création de modèles** → **Optimisation et déploiement**

MODIN · SciPy · pandas · NumPy · Spark · Numba

dmlc XGBoost · scikit learn · SigOpt AutoML · PyTorch · TensorFlow · intel Neural Compressor · ONNX RUNTIME · DirectML · OpenVINO Write Once Deploy Anywhere

Analytique des données à grande échelle †

Cadres de Machine Learning et de Deep Learning, outils d'optimisation et de déploiement†

oneAPI

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

Modèle de programmation ouvert et inter-architecture pour les processeurs, les GPU et d'autres accélérateurs

CLOUD ET ENTREPRISE

CLIENT ET STATION DE TRAVAIL

EDGE

intel AI ANALYTICS TOOLKIT

Accélérez l'IA et la science des données de bout en bout

Intel® Tiber™ AI Cloud et Intel® Developer Catalog

Essayez les derniers outils et matériels Intel et accédez à des modèles d'IA optimisés

Intel® Geti

Plateforme d'annotation/entraînement/optimisation

Hugging Face

Optimisations et recettes d'ajustement, modèles d'inférence optimisés et service de modèles Intel

Remarque : les composants de chaque couche de la pile sont optimisés pour les composants ciblés des autres couches sur la base des modèles d'utilisation de l'IA prévus, et tous les composants ne sont pas utilisés par les solutions figurant dans la colonne de droite.

† Cette liste comprend des cadres open-source populaires qui sont optimisés pour le matériel Intel

Simplifier l'adoption de l'IA générative par les entreprises et réduire le temps de production de solutions renforcées et fiables

OPEA :

Simplifier l'adoption de l'IA générative par les entreprises et réduire le temps de production de solutions renforcées et fiables

Open Platform for Enterprise AI

Partenaires de l'OPEA

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

Valeur de l'OPEA

- Aide les entreprises à libérer plus rapidement et plus facilement la valeur de leurs données en utilisant l'IA générative (LLM, RAG)
- Réduit la complexité d'un écosystème fragmenté et aide les solutions à s'adapter à la production
- Favorise la collaboration et la contribution des leaders du secteur en partenariat avec la Linux Foundation

Efficace
Exploite l'infrastructure existante, l'accélérateur d'IA ou tout autre matériel de votre choix.

Fluide
S'intègre aux logiciels d'entreprise, avec une prise en charge et une stabilité hétérogènes à travers le système et le réseau.

Ouverte
Réunit les meilleures innovations et n'est pas liée à un fournisseur exclusif.

Omniprésente
Fonctionne partout grâce à une architecture flexible conçue pour le Cloud, les centres de données, l'Edge et les PC.

Fiable
Comprend un pipeline sécurisé prêt pour l'entreprise et des outils pour la responsabilité, la transparence et la traçabilité.

Évolutive
Permet d'accéder à un écosystème dynamique de partenaires pour vous aider à créer et à développer votre solution.

Partenariat Hugging Face pour l'IA générative

Hugging Face

Pour faciliter l'entraînement et l'innovation en matière d'IA générative et d'IA linguistique, Intel s'est associé à Hugging Face, une plateforme populaire de partage de modèles d'IA et d'ensembles de données. Hugging Face est notamment connue pour sa bibliothèque de transformateurs conçus pour le traitement du langage naturel (NLP).

intel. XEON

Intel® a collaboré avec Hugging Face pour créer des accélérations matérielles et logicielles de pointe qui permettent d'entraîner et d'affiner des modèles de transformateurs pour générer des prévisions.

L'accélération matérielle est pilotée par les processeurs Intel® Xeon® Scalable, alors que l'accélération logicielle s'appuie sur notre portefeuille de frameworks, de bibliothèques d'IA et d'outils logiciels optimisés.

intel. GAUDI

Les accélérateurs de Deep Learning Intel® Gaudi® sont également couplés à des logiciels Open Source Hugging Face par le biais de la bibliothèque Habana Optimum pour permettre aux développeurs d'utiliser facilement des milliers de modèles optimisés par la communauté Hugging Face.

Hugging Face a également publié plusieurs évaluations des performances d'Intel® Gaudi® 2 sur des modèles d'IA générative : Stable Diffusion, T5-3B, BLOOMZ 176B et 7B et le nouveau modèle BridgeTower.

Intel®, Articul8 et BCG s'allient pour fournir une IA générative sécurisée de qualité professionnelle

Une solution d'avant-garde qui s'appuie sur un superordinateur d'IA Intel® produit de la valeur commerciale grâce à des ensembles de données personnalisés tout en maintenant des niveaux élevés de sécurité et de confidentialité des données

Articul8* propose une plateforme logicielle d'IA générative clé en main qui offre vitesse, sécurité et rentabilité pour aider les grandes entreprises clientes à opérationnaliser et à faire évoluer l'IA. La plateforme a été lancée et optimisée sur des architectures matérielles Intel®, notamment les processeurs Intel® Xeon® Scalable et les accélérateurs Intel® Gaudi®, mais prendra en charge toute une gamme d'infrastructures hybrides alternatives.

intel GAUDI

intel XEON

Suite au déploiement précoce de cette technologie chez Boston Consulting Group (BCG), l'équipe a étendu la plateforme à des entreprises clientes dans des segments qui nécessitent des niveaux de sécurité et de connaissances spécialisées élevés, notamment dans les domaines des services financiers, de l'aérospatiale, des semiconducteurs et des télécommunications.

LIRE LA SUITE

Annonce Articul8

Site Web Articul8

IA responsable pour les entreprises

DÉFI :

Les modèles d'IA générative sont entraînés sur la base de grandes quantités de données disponibles sur Internet qui peuvent être imprégnées de préjugés présents dans la société, avec le risque que les modèles les appliquent par inadvertance. Les LLM peuvent être manipulés pour générer ou propager des fausses informations, des e-mails d'hameçonnage ou des attaques d'ingénierie sociale.

Les LLM ont souvent des « hallucinations » et peuvent produire des informations inexactes, une occurrence particulièrement problématique dans des secteurs comme les soins de santé, où les modèles peuvent influencer les diagnostics et les décisions thérapeutiques et potentiellement nuire aux patients.

En savoir plus

Réduire les risques de l'IA générative

SOLUTIONS :

Les entreprises et les personnes qui travaillent dans le domaine de l'IA doivent veiller à ce que leurs logiciels soient développés et déployés selon les principes éthiques de l'IA

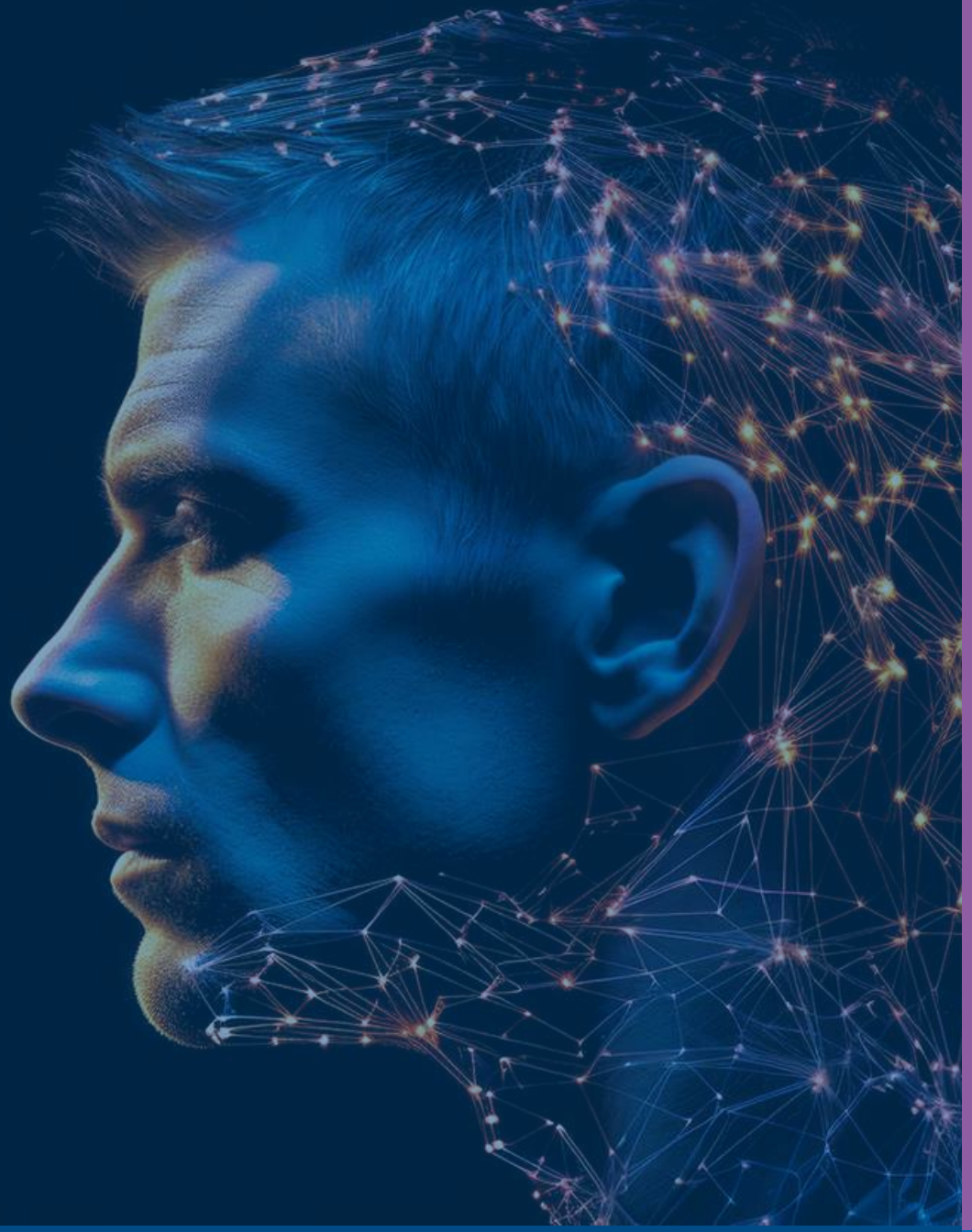
Les outils Open Source Intel® Explainable AI Tools permettent aux utilisateurs d'exécuter ultérieurement une distillation et une visualisation des modèles pour examiner le comportement prédictif des modèles TensorFlow* et PyTorch*

Les LLM sont généralement entraînés sur de grands ensembles de données publics, puis affinés sur des données potentiellement sensibles (par exemple, financières et médicales)

Certaines technologies, comme OpenFL (Open Federated Learning) d'Intel, intègrent l'informatique confidentielle afin que les LLM puissent être mis au point en toute sécurité sur des données sensibles, ce qui améliore la généralisabilité des modèles tout en réduisant les hallucinations et les préjugés

Produits Intel® conçus pour l'IA générative

Intégrer l'IA partout



Systèmes d'IA évolutifs

Entraînement et mise au point

Entraînement | Inférence de pointe | Inférence/mise au point grand public | Inférence de référence | Inférence au point de terminaison

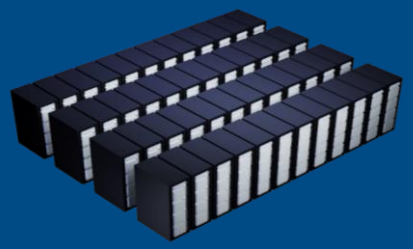
Inférence et déploiement

Cloud

Centre de données

Edge

Client

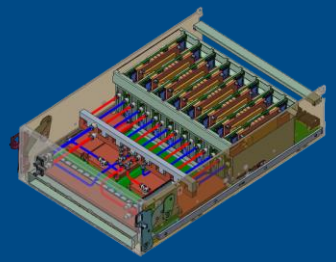


Échelle des clusters et des centres de données

intel. Gaudi

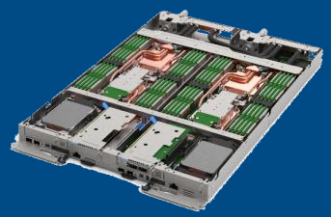
Déploiement multinœud par rack

intel. Gaudi | intel. Xeon | intel. Xeon Max Series



Multi-GPU ou CPU multi-socket

intel. Xeon | intel. Xeon Max Series | intel. Data Center GPU Flex Series



CPU à socket unique ou GPU unique

intel. Xeon | intel. Data Center GPU Flex Series | intel Core Ultra

CPU client

intel Core Ultra | intel. Arc Graphics

intel. Ethernet

Produits Intel® pour les NLP/LLM

Entraînement et inférence

GAUDI®2

Les accélérateurs d'IA Intel® Gaudi® 2 sont spécifiquement conçus pour accélérer l'entraînement et l'inférence de modèles à grande échelle, tels que les LLM et les NLP.

Accélération des modèles d'IA générative et des grands modèles de langage avec Intel® Gaudi® 2

Intel® Gaudi® 2 offre des performances de pointe et des économies de coûts optimales pour l'entraînement de l'IA

Communiqué de presse

Regarder le webinaire

Enregistrement d'un webinaire d'Intel traitant des capacités de pointe du processeur d'IA Intel® Gaudi® 2 pour capturer le potentiel de l'IA générative et des grands modèles de langage (LLM)

L'accélérateur de Deep Learning Intel® Gaudi® 2 offre des performances compétitives en matière d'entraînement Deep Learning et d'inférence, avec des **performances jusqu'à 2,4 fois plus rapides que celles du GPU Nvidia A100**[1]

Espace presse ▪ Article technique

Intel® Gaudi® 2 demeure la seule alternative (reconnue par un banc d'essai) au BPU NV H100 en matière de performances d'IA générative

[1]Les performances varient en fonction de l'utilisation, de la configuration et d'autres facteurs ; les charges de travail et les détails de la configuration sont disponibles à l'adresse : intel.com/performanceindex.
Les résultats effectifs peuvent varier.

Gaudi2 : idéal pour l'entraînement et l'inférence efficaces des modèles de base

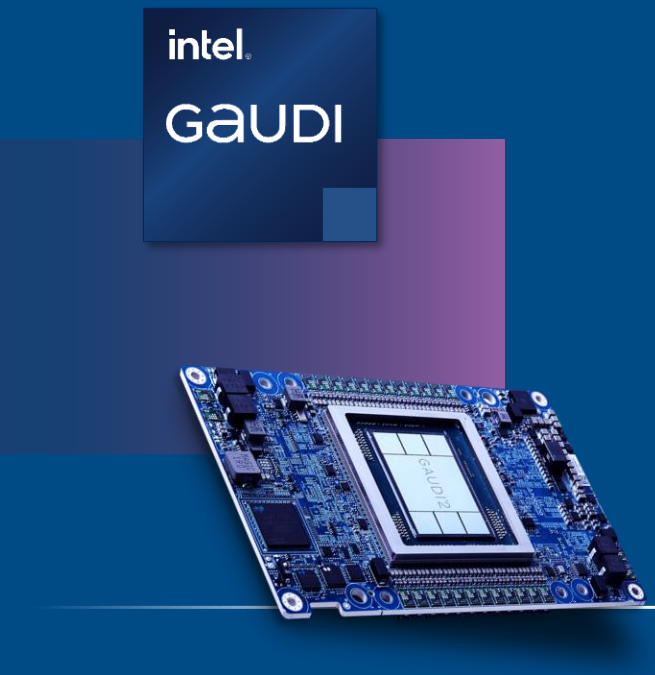
Gaudi2 est conçu pour les performances, l'efficacité et l'évolutivité du Deep Learning afin de répondre aux exigences des modèles de fondation à grande échelle tels que les LLM (GPT) et les GAI (Stable Diffusion)

Exigences	Gaudi2
Vitesse	**1,5 à 2 fois plus rapide** que l'A100 pour l'entraînement et l'inférence
Mémoire	Chaque appareil Gaudi2 est doté **d'une mémoire à large bande passante de 96 Go sur la puce**, ce qui permet d'intégrer plus facilement de grands modèles de base dans la mémoire, de les entraîner et de les déployer à grande échelle
Évolutivité	Mise à l'échelle efficace, avec **24 ports 100GbE intégrés sur la puce**, une connectivité directe entre 8 cartes dans un serveur, et une communication ouverte basée sur ROCEv2 en interne et entre serveurs.
Simplicité d'utilisation	Migrez ou développez des modèles **avec une modification minimale du code**, grâce à SynapseAI, PyTorch et DeepSpeed
Efficacité énergétique	**Débit/watt environ 1,8 fois supérieur par rapport au A100**
Rentabilité	Basé sur une architecture Gaudi de première génération spécialement conçue, qui permet d'obtenir des prix jusqu'à **40 % plus avantageux** que ceux de l'A100 sur le Cloud d'Amazon

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

Mise au point de nombreux LLM

Les évaluations de Hugging Face corroborent les performances de l'accélérateur Intel® Gaudi® 2 par rapport aux GPU NVIDIA A100 et H100 en matière de LLM



Consultez https://habana.ai/habana-claims-validation pour connaître les charges de travail et les configurations. Les résultats effectifs peuvent varier.

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

GPT-J : résultats d'Intel® Gaudi® 2

Les résultats des inférences avec Intel® Gaudi® 2 sur GPT-J confirment clairement ses performances concurrentielles

- Les performances d'inférence d'Intel® Gaudi® 2 sur GPT-J-99 et GPT-J-99.9 pour les requêtes de serveur et les échantillons hors ligne correspondent **respectivement à 78,58 et 84,08 par seconde**[1]
- **Intel® Gaudi® 2 offre des performances convaincantes par rapport au GPU NVIDIA H100**. Celui-ci offre des performances légèrement supérieures (1,09 x sur serveur et 1,28 x hors ligne) par rapport à Gaudi 2[1]
- **Intel® Gaudi® 2 surpasse le GPU NVIDIA A100 (2,4 x sur serveur et 2 x hors ligne)**[1]
- La soumission d'Intel® Gaudi® 2 a utilisé FP8 et a atteint une précision de **99,9 %** sur ce nouveau type de données[1]

LIRE LA SUITE

Avec des mises à jour logicielles d'Intel® Gaudi® 2 publiées toutes les six à huit semaines, Intel® prévoit d'améliorer continuellement les performances et d'étendre la couverture des modèles dans les bancs d'essai MLPerf

Article de la salle de presse

Annonce de MLCommons

[1]Les performances varient en fonction de l'utilisation, de la configuration et d'autres facteurs ; les charges de travail et les détails de la configuration sont disponibles à l'adresse : https://mlcommons.org/2023/09/mlperf-results-highlight-growing-importance-of-generative-ai-and-storage/ Les résultats effectifs peuvent varier.

Intel® Gaudi® 2 : résultats des bancs d'essai

Résultats des bancs d'essai fournis par Supermicro, le premier OEM basé sur Intel® Gaudi® 2

Validation des déclarations sur Gaudi

Entraînement et inférence de LLM avec les accélérateurs d'IA Intel® Gaudi® 2

Bancs d'essai

Entraînement et inférence plus rapides : Intel® Gaudi® 2 par rapport à NVIDIA A100 80 Go

Bancs d'essai

Les résultats effectifs peuvent varier.

Intel® Gaudi® 2 : entraînement et inférence des modèles fondamentaux

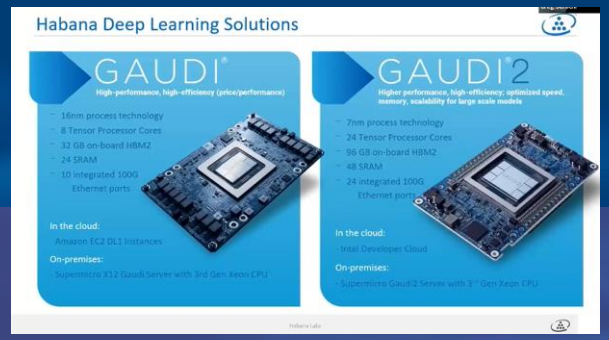
Les modèles disponibles compatibles avec Gaudi sont accessibles dans le

Catalogue développeurs

GAUDI®2

Formation des développeurs : Intel® Gaudi®

[Démarrage : Deep Learning et inférence sur Gaudi]



[Maximiser la puissance d'Intel® Gaudi® 2 : accélérer l'IA générative et les grands modèles de langage]

[Maximiser les performances des modèles avec les processeurs Intel® Gaudi® : outils et stratégies avancés pour des résultats optimaux]

Logiciel Intel® Gaudi® (suite logicielle SynapseAI®)

Développement simplifié : la bonne façon de développer

Objectif : faciliter la migration des logiciels existants vers les accélérateurs d'IA Intel® Gaudi®, préserver les investissements logiciels et faciliter la création de nouveaux modèles, à la fois dans le cadre de l'entraînement et du déploiement de nombreux modèles en pleine croissance qui définissent le Deep Learning, l'IA générative et les grands modèles de langage.

Assistance complète pour les scientifiques des données, développeurs

et administrateurs informatiques et système avec :

- Le site des développeurs
- GitHub

Accélérateur d'IA Intel® Gaudi®

L'écosystème logiciel pour le Deep Learning rassemble les principaux éditeurs de logiciels, outils et codes pour accélérer le développement de modèles de Deep Learning de pointe basés sur les infrastructures PyTorch, TensorFlow, PyTorch Lightning et DeepSpeed

cnvrg.io

PyTorch Lightning

Prêt à utiliser le logiciel Intel® Gaudi® ?

Accélérateurs d'IA Intel® Gaudi® 2 DISPONIBLES exclusivement sur Denvr Cloud

Écosystème logiciel
Intel® Gaudi® 2

Accélérateur d'IA
Intel® Gaudi® 2 (7 nm)

Intel® Gaudi® 2 - Idéal pour les exigences de l'IA générative

- Disponible dès maintenant ! Clusters Gaudi 2 sur Denvr Cloud
- Testez jusqu'à 8 nœuds Gaudi 2
- Tarification VIP prioritaire pour les clients Intel
- Service et assistance commerciaux de haute qualité de Denvr Dataworks
- Migration transparente vers les clusters Gaudi 2 sur Denvr Cloud
- Positionnement prioritaire exclusif pour les clusters Gaudi 3 sur Denvr Cloud - Bientôt disponible !

Se lancer

BIENTÔT DISPONIBLE

intel GAUDI

Entraînement et inférence

Intel® Gaudi® 3

Avec des performances, une évolutivité et une efficacité qui offrent plus de choix à plus de clients, les accélérateurs Intel® Gaudi® 3 aident les entreprises à débloquer des informations, des innovations et des revenus

BIENTÔT DISPONIBLE : accélérateur d'IA Intel® Gaudi® 3

Un plus grand choix dans le domaine de l'IA générative en matière de performances, d'évolutivité et d'efficacité

Intel® Gaudi® 3 constituera un bond en avant important dans l'entraînement et l'inférence de l'IA pour les entreprises mondiales qui cherchent à déployer l'IA générative à grande échelle

Communiqué de presse

Performances de l'accélérateur Intel® Gaudi® 3 par rapport au GPU NVIDIA H100

Intel® Gaudi® 3 devrait offrir des délais d'entraînement **50 % plus rapides en moyenne**[3] sur les modèles Llama2 avec 7 Md et 13 Md de paramètres et le modèle GPT-3 de 175 Md de paramètres

Intel® Gaudi® 3 devrait surpasser le H100 de :

50 % en termes de débit d'inférence de l'accélérateur[1]

40 % en termes d'efficacité énergétique de l'inférence[2]

sur les modèles Llama de 7 Md et 70 Md de paramètres, et les modèles Falcon de 180 Md de paramètres

LIRE LA SUITE

LIVRE BLANC

Intel® Gaudi® 3 sera disponible pour les OEM à partir du 2e trimestre 2024, notamment :

Lenovo DELL Hewlett Packard Enterprise SUPERMICRO

[1]Comparaison du GPU NV H100 basée sur https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8, les chiffres indiqués sont par GPU. Par rapport aux projections Intel® Gaudi® 3 pour les projections LLAMA2-7B, LLAMA2-70B et Falcon 180B. Les résultats effectifs peuvent varier.
[2]Comparaison du GPU NV H100 basée sur https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8, les chiffres indiqués sont par GPU. Par rapport aux projections Intel® Gaudi® 3 sur les modèles LLAMA2-7B, LLAMA2-70B et Falcon 180B. Efficacité énergétique des produits NVIDIA et Gaudi 3 basée sur des estimations internes. Les résultats effectifs peuvent varier.
[3]Comparaison du GPU NV H100 basée sur https://developer.nvidia.com/deep-learning-performance-training-inference/training, onglet « Large Language Model » (Grand modèle de langage) par rapport aux projections Intel® Gaudi® 3 sur les modèles LLAMA2-7B, LLAMA2-13B et GPT3-175B au 28/3/2024. Les résultats effectifs peuvent varier.

Produits Intel® pour les NLP/LLM

Inférence

Les processeurs Intel® Xeon® Scalable de 4e et 5e génération accélèrent le NLP grâce à Intel® DL Boost, Intel® AMX et Intel® AVX-512. Ils sont conçus pour le calcul intensif (HPC) et peuvent être utilisés pour accélérer les charges de travail NLP. Ils peuvent gérer un grand nombre de threads, une grande capacité de mémoire et une large bande passante mémoire, ce qui convient aux charges de travail NLP comme la traduction linguistique, la synthèse de texte et la synthèse vocale.

Processeur Intel® Xeon® de 5e génération : le processeur conçu pour l'IA

Les processeurs Intel® Xeon® de 5e génération, qui intègrent l'accélération de l'IA dans chaque cœur, répondent aux charges de travail exigeantes de l'IA de bout en bout avant que les clients n'aient besoin d'ajouter des accélérateurs dédiés

Performances accrues pour l'inférence de l'IA :

jusqu'à 42 %

par rapport à la génération précédente[1]

Gains de performance en calcul général

moyens de 21 %

par rapport à la génération précédente[1]

Traitement automatique du langage naturel plus rapide

jusqu'à 23 %

par rapport à la génération précédente[1]

Sandra Rivera, vice-présidente exécutive d'Intel et directrice générale de Data Center and AI Group

« Conçus pour l'IA, nos processeurs Intel® Xeon® de 5e génération offrent de meilleures performances aux clients qui déploient des capacités d'IA sur des cas d'utilisation du Cloud, du réseau et de l'Edge. Grâce à notre travail de longue date avec les clients, les partenaires et l'écosystème des développeurs, nous lançons les processeurs Intel® Xeon® de 5e génération en nous appuyant sur un socle solide qui permettra une adoption et une mise à l'échelle rapides pour un TCO réduit. »

Complément d'infos

Site Web

Fiche produit

Pour les charges de travail et les configurations, visitez intel.com/processorclaims : processeurs Intel® Xeon® Scalable de 5e génération. Les résultats effectifs peuvent varier.

Intel® Xeon® : le leader des performances de processeurs pour les applications concrètes de l'IA

Intel révolutionne le secteur et démocratise l'IA en proposant une plateforme plus performante, moins chère et plus équilibrée pour l'inférence de l'IA dans le cadre d'applications professionnelles réelles :

- Un cache plus important qui facilite la localisation des données et une grande capacité de mémoire qui permet de résoudre des problèmes plus importants
- Une fréquence de cœur plus élevée, des ports scalaires multiples et une exécution dans le désordre permettant d'accélérer le calcul scalaire monothread ou multithread
- Intel® Advanced Vector Extensions 512 (Intel® AVX-512) pour faciliter le calcul vectoriel non DL
- Intel® Advanced Matrix Extensions (Intel® AMX) pour prendre en charge l'accélération de l'IA

Article technique complet

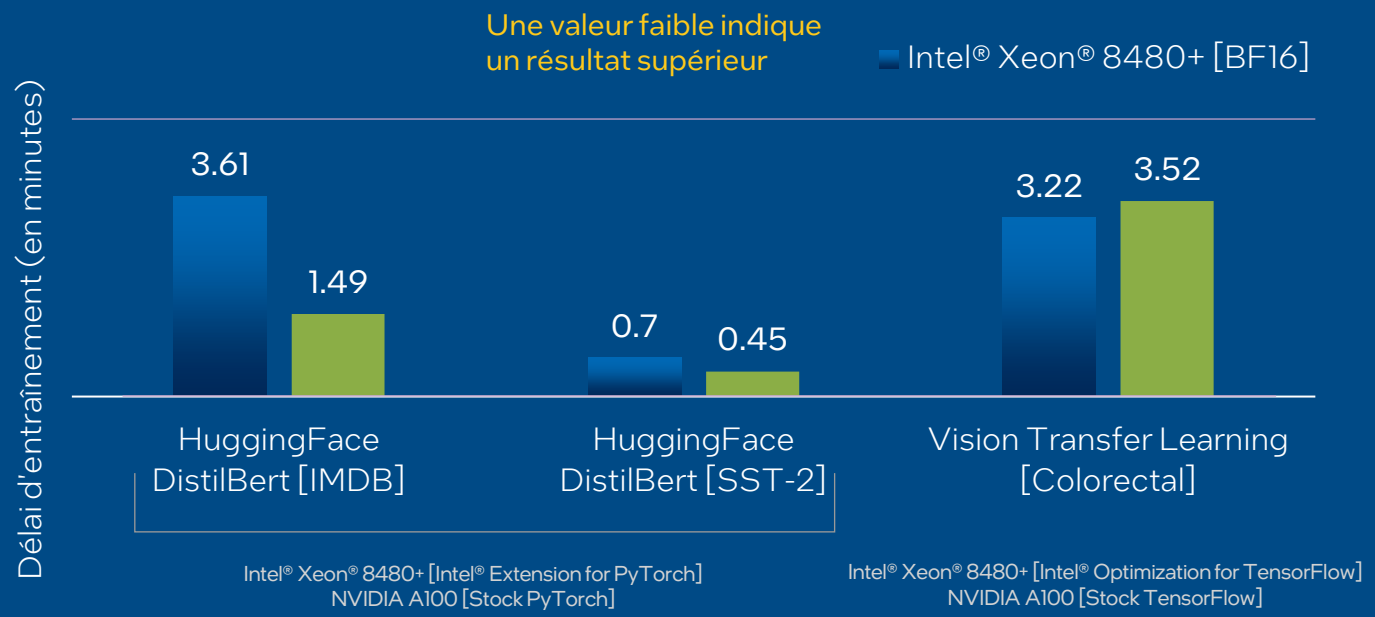
Infographie

Au-delà du GPU : découvrez comment les processeurs avec accélérateurs intégrés révolutionnent l'IA.

Mettez au point les modèles en moins de 4 minutes avec les processeurs Intel® Xeon® Scalable[1]

Hugging Face

Mise au point : performances en délais d'entraînement du processeur Intel® Xeon® Platinum 8480+ par rapport au GPU NVIDIA A100



Voir également :

Meilleures performances : Numenta sur les CPU Intel® par rapport aux GPU NVIDIA

[1]Voir la section [A221] de l'Indice de performance des processeurs Intel Xeon Scalable de 4e génération Les résultats effectifs peuvent varier.

LLM sur les processeurs Intel® Xeon® de 4e génération

La technologie des chatbots pilotés par l'intelligence artificielle (IA) gagne en popularité dans les entreprises et les organisations comme mode d'interaction avec les clients et dans le cadre de l'amélioration de leur service à la clientèle. La création, l'optimisation et la maintenance de chatbots spécialisés est toutefois coûteuse et peut être financièrement prohibitive pour de nombreuses organisations

EN SAVOIR PLUS

Guide de mise au point de l'IA sur les processeurs Intel® Xeon® Scalable de 4e génération

Lien vers le guide ›

Les processeurs Intel® Xeon® de 4e génération offrent une gestion des données améliorée et un traitement efficace grâce à **Intel® AMX (Advanced Matrix Extensions)**. Lorsqu'elle est combinée à la fonctionnalité **Auto Mixed Precision (AMP)** disponible par le biais d'Intel® Extension for PyTorch, cette pile technologique devient une alternative concurrentielle sur les charges de travail comme l'apprentissage par transfert et l'entraînement complet de modèles de petite taille ou de taille moyenne

Article technique pratique

Cisco UCS avec processeurs Intel® Xeon® de 5e génération et de 4e génération pour l'IA générative

Petite taille, meilleurs résultats : LLM Q8-Chat est une expérience d'IA générative efficace sur les processeurs Intel® Xeon®

Les LLM nécessitent une grande puissance de traitement, généralement présente dans les GPU haut de gamme, pour offrir des prédictions suffisamment rapides dans les cas d'utilisation à faible latence comme les applications de recherche ou de conversation. Malheureusement, les coûts associés peuvent être prohibitifs pour de nombreuses organisations et empêcher l'utilisation de LLM de pointe dans leurs applications.

Hugging Face

« Nombre d'entreprises ont tout à gagner à se tourner vers des modèles plus petits et plus spécifiques, moins coûteux à entraîner et à exécuter. »

Découvrez les techniques d'optimisation qui permettent de réduire la taille des LLM et la latence d'inférence pour les exécuter avec plus d'efficacité sur les CPU Intel®.

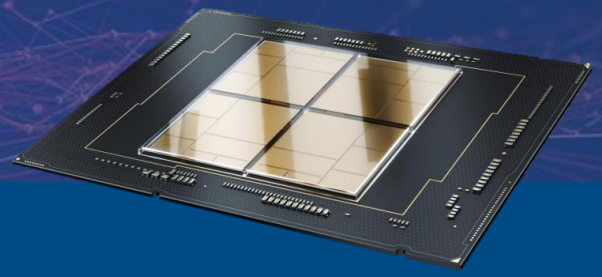
Article technique pratique ›

Démarrer sur les processeurs Intel® Xeon® de 4e génération avec Hugging Face

Processeurs Intel® Xeon® pour LLM

RÉSUMÉ

- Bien placés pour l'inférence des LLM spécialisés
- Fournissent de bons résultats dans les cas d'utilisation d'apprentissage par transfert
- Déployez des LLM sur des processeurs Intel® Xeon® avec des logiciels Open Source pour obtenir facilement des performances optimales



Processeurs Intel® Xeon® Scalable pour LLM

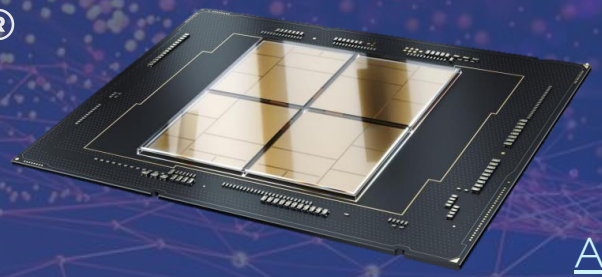
Parfaits pour créer et déployer des charges de travail d'IA généralistes avec les infrastructures et les bibliothèques d'IA les plus populaires

- Utilisent l'infrastructure existante pour l'inférence des LLM spécialisés
- Fournissent de bons résultats dans les cas d'utilisation d'apprentissage par transfert
- Déployez des LLM sur des processeurs Intel® Xeon® avec des logiciels Open Source pour obtenir facilement des performances optimales

Processeur Intel® Xeon®

Leader des performances du CPU dans les applications d'IA réelles

Article technique ▪ Infographie



GPT-J

Résultats sur le processeur Intel® Xeon® de 4e génération

2 paragraphes par seconde en mode hors ligne[1]

1 paragraphe par seconde en mode serveur en temps réel[1]

Article de la salle de presse ▪ Annonce de MLCommons

Démystifier le mythe du GPU : comment les processeurs avec accélérateurs intégrés révolutionnent l'IA

Étude de cas du NLP Alibaba sur les processeurs Intel® Xeon® de 4e génération avec Intel® AMX

LIRE LA SUITE

[1]Les performances varient en fonction de l'utilisation, de la configuration et d'autres facteurs ; les charges de travail et les détails de la configuration sont disponibles à l'adresse <g id="1269">: <g id="1266">https://mlcommons.org/2023/09/mlperf-results-highlight-growing-importance-of-generative-ai-and-storage/</g> Les résultats effectifs peuvent varier.</g>

Produits Intel® pour les NLP/LLM

Inférence à petite échelle au niveau du client

Intel® Core™ Ultra inaugure l'ère du PC accéléré par l'IA

Les processeurs Intel® Core™ Ultra sont optimisés pour les PC portables haut de gamme fins et puissants. Ils sont dotés d'une architecture 3D hybride hautes performances, de capacités d'IA avancées et sont disponibles avec un GPU Intel® Arc™ intégré. Créés à l'aide du nouveau processus de gravure Intel® 4, les processeurs Intel® Core™ Ultra offrent un équilibre optimal entre performances et efficacité énergétique pour le jeu, la création de contenu et la productivité en déplacement.

Cas d'utilisation : l'IA sur le PC

Créateur : recherche et retouche de photos/vidéos

Des filtres plus rapides et plus naturels, des aperçus de meilleure qualité et des temps d'exportation plus courts avec des recherches automatisées et plus rapides.

Jeux courants

Nouvelles fonctionnalités d'IA en jeu, animation 3D pour plus de réalisme, transcription et traduction des conversations.

Créateur : conversion texte-image

Nouveaux effets et nouvelles fonctionnalités d'IA pour créer des images en quelques mots, dans les domaines du marketing, de la publicité et du design.

L'IA sur le PC

« Déverrouillez le quotidien »

Collaboration/streaming

Nouvelles capacités d'IA pour la vidéoconférence, le streaming et la collaboration de nouvelle génération, tout en préservant l'autonomie de la batterie.

Productivité

Assistants IA pour l'écriture, la création, le codage, plus des fonctionnalités hors ligne telles que la prédiction de texte et de grammaire.

Accessibilité

Capacités audiovisuelles assistées par l'IA pour répondre aux divers besoins des utilisateurs, facilitant la création et la productivité sur PC.

*Intel 14 décembre 2023, Communiqué de presse sur les événements AI partout

Intel® Core™ Ultra pourl'IA générative

Le processeur client le plus écoénergétique d'Intel inaugure l'ère du PC accéléré par l'IA

Améliorations majeures en matière d'efficacité et de performances

EFFICACITÉ DE L'IA

jusqu'à **70 %**

plus rapide pour l'IA générative[2]

ÉCONOMIES D'ÉNERGIE

jusqu'à **25 %**

de réduction de la consommation d'énergie[3]

LIRE LA SUITE Annonce ▪ Fiche produit ▪ Site Web

Le processeur Intel® Core™ Ultra est doté du premier accélérateur d'IA client sur puce d'Intel, l'unité de traitement neuronal (NPU), qui permet d'atteindre un niveau d'accélération de l'IA écoénergétique sans précédent avec **une efficacité énergétique 2,5 fois supérieure** à celle de la génération précédente[1]

Les puces Intel® Core™ Ultra de génération H et U comprennent deux nouveaux cœurs Low Power Island (LP-E) pour les charges de travail de faible intensité, avec deux moteurs de calcul neuronal dans le NPU IA d'Intel, conçus pour traiter l'**inférence de l'IA générative**.

[1]Mesures basées sur le rapport Performance/Watt du banc d'essai UL Procyon AI lors de l'exécution d'un modèle int8 sur le NPU de l'Intel® Core™ Ultra 7 165H par rapport au GPU Intel® Core™ i7-1370P.
[1,2,3] Voir www.intel.com/PerformanceIndex pour connaître les charges de travail et les configurations. Les résultats effectifs peuvent varier.

Accélération de l'innovation dans le domaine de l'IA

Intel® collabore avec des éditeurs de logiciels indépendants de premier plan pour optimiser votre expérience avec l'IA.

Le programme d'accélération des PC boostés par l'IA vise à permettre aux fournisseurs de matériel indépendants (IHV) (IHV) et aux éditeurs de logiciels indépendants (ISV) de bénéficier des ressources Intel®, notamment des chaînes d'outils d'intelligence artificielle (IA), des formations, de la co-ingénierie, de l'optimisation des logiciels, du matériel, des ressources de design, de l'expertise technique, du co-marketing et des opportunités commerciales.

En savoir plus

Accélérer le développement de l'IA en entreprise avec le Cloud Intel® Tiber™ AI Cloud pour les développeurs (anciennement Intel® Developer Cloud)

Découvrez, prototypez, testez et exécutez des applications et des charges de travail sur un cluster équipé du matériel et des logiciels Intel® les plus récents

Accélérez et **faites évoluer l'IA** grâce aux dernières innovations matérielles et logicielles dans cet environnement de développement. **Profitez d'une puissance de calcul supérieure** et d'un plus grand choix pour **perfectionner vos logiciels** et **l'IA générative.**

Lancez-vous avec Intel

Profitez d'une expérience concrète avec les derniers produits Intel. Développez vos connaissances en matière d'IA avec Intel.

Accès aux technologies en avant-première

Évaluez les plateformes Intel en pré-version et les piles logicielles associées optimisées par Intel.

Déploiement de l'IA à grande échelle

Accélérez le déploiement d'applications d'IA avec les derniers kits d'outils de Machine Learning d'Intel et les bibliothèques hébergées sur Cloud Intel® Tiber™ pour les développeurs.

Lire l'article technique ›

Commencer ›

Appel à l'action

FORMATION

Découvrez comment la technologie Intel® peut être utilisée pour l'IA générative et les modèles spécialisés et dans quelle mesure les gammes de produits Intel® Xeon® et Intel® Gaudi® peuvent vous aider à développer vos activités

Démarrer

ENGAGEMENT

Lancez-vous avec

Intel® Tiber™ AI Cloud

Accélérez et faites évoluer l'IA grâce aux dernières innovations matérielles et logicielles dans cet environnement de développement

+

Utilisez les kits de référence de l'IA

CONTACT

Contactez votre **représentant Intel®** pour obtenir plus d'informations

Accès à l'assistance client de l'Alliance partenaire Intel®

Intel® Virtual Assistant

Ce chat bot, situé dans l'angle inférieur droit de chaque page Web de l'Alliance partenaire, fournit des réponses à la plupart des questions ou un lien rapide pour contacter un conseiller en direct.

Lame « Obtenir de l'aide »

Soumettez une demande d'assistance en ligne.

Ce lien se trouve en bas de la plupart des pages du site Web de l'Alliance partenaire.

Page d'assistance de l'Alliance partenaire

La page d'assistance offre des guides détaillés en libre-service sur la plupart des outils et sur les avantages disponibles aux membres de l'Alliance partenaire.

Zones d'activation de l'IA

Des espaces de travail d'IA axés sur le numérique qui comprennent des ressources, des outils et des avantages essentiels pour aider les partenaires à créer, commercialiser et vendre des solutions basées sur la technologie Intel®

Habilitation technique

Habilitation à la vente et au marketing

Habilitation technique

Habilitation à la vente et au marketing

Habilitation technique

Habilitation à la vente et au marketing

Kits de référence sur l'IA

En tirant parti de ces kits de référence, les organisations peuvent **réduire considérablement les délais** de mise en œuvre de leur solution et bénéficier **de gains de performances notoires**

Finance et assurance

Détection de fraudes

GitHub ▪ Blog ▪ Plan détaillé

Santé et sciences de la vie

Protection contre les maladies

GitHub ▪ Blog

Fabrication et services publics

Détection d'anomalies

GitHub ▪ Blog

Gestion de parc

Maintenance prédictive

GitHub

Automatisation des processus

Automatisation des documents

GitHub ▪ Blog ▪ Plan détaillé

Flux de travail

- Apprentissage DL par transfert
- Mise au point HF et optimisation de l'inférence
- Compression distribuée DL
- Flux de travail ML classique distribué
- Pré-entraînement DL avec accélérateurs Intel®
- Analytique graphique et GNN avec DGL et PyG
- Trang/inférence distribuée sur Big-DL
- Pré-entraînement et mise au point de LLM sur Ray

Outils

- Intel® Distribution for Python
- Intel® Optimized Modin
- Intel® Optimized XGBoost
- Intel® Extension for Scikit-Learn
- Intel® Optimized Tensorflow (ITEX)
- Intel® Optimized PyTorch (stock & IPEX)
- Intel® Neural Compressor
- SigOpt Python SDK & CLI
- CNVRG Python SDK & CLI
- Horovod optimisé par Intel
- DeepSpeed

Kits de domaine

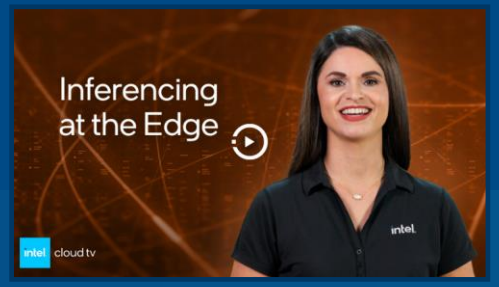
- Série chronologique
- PPML
- Apprentissage par transfert
- Transformateur/NLP

Les **kits de référence** sont livrés sous forme **de conteneurs** et peuvent être utilisés sur les **principaux Clouds ainsi que sur site.** Les **kits de référence** sont superposés sur des **flux de travail** et **des kits d'outils de domaine** qui peuvent être exploités de manière indépendante pour prendre en charge une **plus grande variété de cas d'utilisation dans plusieurs secteurs**.

Cloud TV

Intel® Cloud TV explore l'actualité, les tendances et les stratégies du cloud computing pour favoriser votre réussite

Votre opportunité d'IA générative avec les accélérateurs d'IA Intel® Gaudi®



Obtenir des informations grâce à l'inférence des données à l'Edge

Créer un avantage concurrentiel grâce à l'IA dans le Cloud

L'inférence d'IA à l'aide des technologies Cloud

IA dans le Cloud

Prenez la voie express vers l'IA évolutive partout

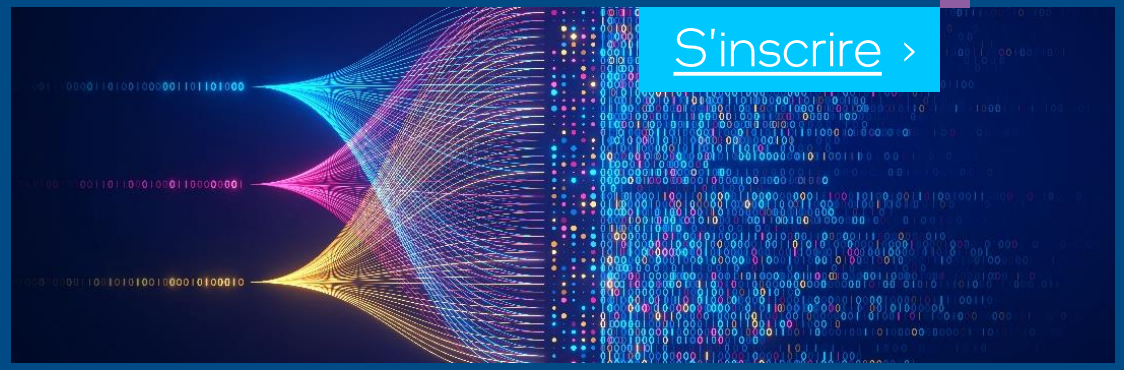
Formations

Intégrer l'IA partout – Cas d'utilisation en entreprise de l'IA générative

L'IA générative ne se limite pas aux chatbots sur Internet. De nombreuses entreprises envisagent d'exploiter la puissance de l'IA générative et des grands modèles de langage dans le cadre de leurs opérations quotidiennes. Cette formation porte sur les cas d'utilisation de l'IA générative en entreprise et sur la manière dont celle-ci peut être appliquée dans les opérations quotidiennes de votre organisation.

S'inscrire ›

Rationaliser l'IA pour la génération de données et les grands modèles de langage



S'inscrire ›

L'intégration de l'IA dans les charges de travail d'une organisation ou la mise à l'échelle d'une infrastructure déjà existante exige beaucoup de compétences et de puissance de calcul. Il est nécessaire de développer des modèles robustes, formés sur des ensembles de données de très grande taille, et des GPU puissants sur lesquels ils peuvent être exécutés correctement. Toutes les organisations ne disposent pas des ressources nécessaires pour accomplir cette tâche.

Cette formation se concentre sur une solution : une collection de kits de référence d'IA Open Source d'Accenture* et d'Intel, conçus pour rendre l'IA plus accessible aux organisations et optimisés pour améliorer l'entraînement et les délais d'inférence.

Formations supplémentaires

Technique

Type de ressource	Titre et lien
Compétence	Compétence IA dans le Cloud
Webinaire	Optimiser l'IA pour le matériel Intel® avec Hugging Face
Webinaire	Comment mettre en place un entraînement distribué basé sur le Cloud pour mettre au point un LLM
Formation	Améliorer les LLM grâce aux économies d'invites et à l'apprentissage dans le contexte
Formation	Rationaliser l'IA pour la génération de données et les grands modèles de langage
Formation	Traitement du langage naturel
Formation	Deep Learning appliqué avec TensorFlow*
Formation	Petit et agile : le raccourci vers l'IA générative d'entreprise
Formation	La prochaine vague d'IA générative : les LLM spécialisés
Guide	Guide des développeurs pour se lancer avec l'IA générative : une approche spécifique à chaque cas d'utilisation
Cours de formation	L'IA sur les processeurs Intel® Xeon® dans l'espace consacré aux solutions

Formations supplémentaires

Non technique

Type de ressource	Titre et lien
Série de vidéos	Adopter l'IA générative
Formation	Petit et agile : le raccourci vers l'IA générative d'entreprise
Formation	La prochaine vague d'IA générative : les LLM spécialisés
Cours de formation	Compétence Principes de l'IA partout
Cours de formation	Compétence Principes des logiciels et de l'écosystème d'IA
Cours de formation	Engager l'écosystème de l'IA : Gagner avec les logiciels, mettre à l'échelle avec les intégrateurs de systèmes et vendre la solution
Cours de formation	Applications concrètes de l'IA générative et des grands modèles de langage

Autres ressources

Type de ressource	Titre et lien
Webinaire	Série de webinaires sur l'IA générative
Webinaire	Intégrer l'IA générative partout
Podcast	Comment Copilot, ChatGPT, Stable Diffusion et l'IA générative changeront la façon dont nous développons, travaillons et vivons
Fiche stratégie	Déployez l'IA partout
Série d'articles de blog	Mise au point et inférence de l'IA générative avec les processeurs Intel Xeon de 4e génération
Fiche de solution	Déployer et développer l'inférence de l'IA générative avec Lenovo ThinkSystem SR650 V3/les processeurs Intel Xeon de 4e génération Les nouvelles technologies Intel et VMware boostent les systèmes Lenovo ThinkAgile VX V3
Article technique	Accélérer Llama 2 grâce aux optimisations matérielles et logicielles d'IA Intel®
Communiqué de presse recherche	10 % des organisations interrogées ont lancé la production de solutions d'IA générative en 2023
Vidéo de conversation informelle	Relever les défis de l'IA générative en matière de calcul et de durabilité
Podcast	Hugging Face et Intel : vers des solutions d'IA pratiques, plus rapides, démocratisées et éthiques
Discussion sur Twitter/X	Comment les grands modèles de langage démocratisés stimulent le développement de l'IA
Bancs d'essai Supermicro	Validation des déclarations Habana
Bancs d'essai Hugging Face	Bancs d'essai
Formation/Webinaire	Cloud Solution Architect (CSA) Tech Talk: l'IA avec Habana
Livre blanc	Livre blanc : Enterprise AI is all about the Developer (L'IA en entreprise concerne les développeurs)
Infographie	Les processeurs sont la clé de l'IA en entreprise

Autres ressources

Type de ressource	Titre et lien
Fiche de solution	Rationaliser l'adoption et le déploiement de l'IA en entreprise avec la plateforme OPEA d'Intel et Red Hat® OpenShift® AI
Guide	Le guide de l'IA
Kit de référence	Génération de données non structurées par l'IA
Livre blanc	Zoho optimise et accélère les charges de travail d'IA vidéo
Livre blanc	Seekr développe un système de dépistage par IA fiable
Fiche de solution	Sécurité dans l'éducation : l'IA et l'informatique confidentielle permettent de faire des examens à distance sécurisés une réalité
Étude de cas et vidéo	Nature Fresh Farms utilise l'IA des semis au magasin
Étude de cas	QMed Asia favorise la détection précoce du cancer
Étude de cas et vidéo	MetaApp réorganise le système de recommandation basé sur l'IA
Fiche de solution	Optimisation de l'entraînement et de l'amélioration des modèles d'IA pour l'inspection optique automatisée (AOI)
Blog	Invites pour améliorer l'efficacité des LLM

Avertissements et informations importantes

- Avis et avertissements.
- © Intel Corporation. Intel, le logo Intel et les autres marques Intel sont des marques commerciales d'Intel Corporation ou de ses filiales. Les autres noms et marques peuvent être revendiqués comme la propriété de tiers.


intel