Enterprise AI

IA generativa e modelos específicos de domínio para empresas

Otimize o treinamento e a implantação com hardware e software Intel® AI de finalidade específica para ajudar a transformar sua empresa

intel®

Conteúdo

Por que fazer parceria com a Intel?

Na Intel, temos a oportunidade de melhorar vidas e resultados para todas as pessoas e empresas neste planeta

Mas não estamos fazendo isso sozinhos!

Juntamente com nossos parceiros, estamos criando valor real para nossos clientes **levando a IA a todos os lugares** e minimizando os riscos de implantação

Ao fazer parceria com a Intel, você faz parceria com um ecossistema completo de IA

Nosso amplo portfólio de tecnologias para habilitar IA e nossas parcerias com integradores de hardware, software e sistemas trabalhando em estreita colaboração estão criando soluções reais que oferecem resultados comerciais diferenciados para setores, empresas e comunidades.

Ajudando você a expandir sua empresa.

Junte-se a nós na jornada para levar a IA a todos os lugares

Gerando valor para clientes com as soluções Intel® AI

A abordagem da Intel permite que um ecossistema aberto e amplo de empresas de IA ofereça soluções que atendam às necessidades de IA generativa específicas para empresas

Para desenvolver um grande modelo de linguagem (LLM) poderoso para a implantação de serviços de IA avançados globalmente, da nuvem ao dispositivo. A NAVER confirmou a capacidade fundamental do Intel® Gaudi® de executar operações de computação para modelos transformadores de grande escala com um desempenho por watt excepcional.

Para explorar mais oportunidades de fabricação inteligente, incluindo modelos fundamentais gerando conjuntos de dados sintéticos de anomalias de fabricação para fornecer conjuntos de treinamento robustos e distribuídos de maneira uniforme (por exemplo, inspeção óptica automatizada).

Utilizando processadores Intel® Xeon® da 5ª Geração para seu armazenamento de dados watsonx.data™ e trabalhando em estreita colaboração com a Intel® para validar a plataforma watsonx™ para aceleradores Intel® Gaudi®.

Para pré-treinar e ajustar seu primeiro modelo fundamental da Índia com capacidades generativas em 10 idiomas, produzindo um preço/desempenho líder do setor em comparação com soluções do mercado. A Krutrim está pré-treinando um modelo fundamental maior em um cluster do Intel® Gaudi® 2.

A líder em IA confiável executa cargas de trabalho de produção no Intel® Gaudi® 2, Intel® Data Center GPU Max series e processadores Intel® Xeon® na Intel® Tiber™ Developer Cloud para desenvolvimento e suporte de implantação de produção de LLMs.

Líderes globais em alimentos, bebidas, aromas e biociências aproveitarão a IA generativa e a tecnologia de gêmeos digitais para estabelecer um fluxo de trabalho de biologia digital integrado para design avançado de enzimas e otimização de processos de fermentação.

Incorporando o poder da tecnologia de ponta da Intel, a Airtel planeja aproveitar seus ricos dados de telecomunicações para aprimorar suas capacidades de IA e impulsionar as experiências de seus clientes. As implantações estarão alinhadas ao compromisso da Airtel de permanecer na vanguarda da inovação tecnológica e ajudar a impulsionar novos fluxos de receita em um cenário digital em rápida evolução.

A líder global em serviços digitais e consultoria de última geração anunciou uma colaboração estratégica para trazer tecnologias Intel®, incluindo processadores Intel® Xeon® da 4ª e 5ª Gerações, aceleradores Intel® Gaudi 2 e processadores Intel® Core™ Ultra para a Infosys Topaz – um conjunto de serviços, soluções e plataformas IA-first para acelerar o valor de negócios utilizando tecnologias de IA generativa.

Proposta de valor da IA empresarial

Transformando sua empresa com a IA empresarial

No ambiente hipercompetitivo atual, **as empresas que adotam a IA estão assumindo a liderança.**

Empresas de todos os setores estão reinventando cada aspecto das operações para entender como a IA pode ampliar ou até mesmo automatizar fluxos de trabalho.

Na Intel, a incorporação da IA à estrutura da empresa é nosso conhecimento especializado exclusivo.

Desde PCs com IA que transformam a produtividade até anos de experiência em entender quais casos de uso resultam no maior valor, a Intel® é sua parceira confiável para levar a IA a todos os lugares, de forma segura e responsável.

Espera-se que inovações de IA generativa (GenAI) sejam adotadas por empresas de todos os tamanhos a uma taxa mais rápida que na era na internet, na era móvel ou na era da nuvem.

A próxima onda de plataformas de IA abraçará essas realidades interessantes de uma maneira acessível e flexível.

É hora de pensar diferente sobre sua IA empresarial.

Este Pacote de capacitação o ajudará a entender como empresas de todos os setores podem obter valor significativo da IA generativa, especialmente de modelos específicos de domínio, para sucesso de longo prazo

O que são IA generativa e grandes modelos de linguagem?

A IA generativa (GenAI) é um subconjunto da IA focado em criar conteúdos novos e originais.

Ela envolve o treinamento e a implantação de modelos de IA para gerar dados, como imagens, textos ou áudios, muito parecidos com exemplos do conjunto de dados de treinamento.

Os algoritmos de IA generativa utilizam técnicas avançadas, como aprendizado profundo e redes neurais, para produzir resultados realistas e coerentes que permitem aplicações como síntese de imagens, geração de textos e até mesmo artes criativas.

Os grandes modelos de linguagem (LLMs) são um tipo específico de modelo de processamento de linguagem natural que utiliza redes neurais profundas para processar e gerar textos. Os LLMs são treinados com quantidades maciças de dados de texto e são projetados para gerar resultados coerentes e significativos.

Saiba mais

LEIA MAIS

Capture o poder da IA generativa

Como as empresas utilizarão a IA generativa?

Bens de consumo e varejo

- Provadores virtuais
- Entregas e instalações
- Assistência de busca de produtos em lojas
- Previsão de demanda e planejamento de inventário
- Projetos de novos produtos

Saúde e medicina

- Assistência a funcionários sobrecarregados no atendimento
- Transcrição e resumo de notas médicas
- Chatbots para responder a perguntas médicas
- Análise preditiva para informar diagnósticos e tratamentos

Manufatura

- Copiloto especializado para técnicos
- Interações de conversação com máquinas
- Serviço de campo prescritivo e proativo
- Resolução de problemas com linguagem natural
- Status e documentação de garantia
- Compreender gargalos de processos, elaborar estratégias de recuperação

Mídia e entretenimento

- Pesquisa inteligente, descoberta de conteúdo personalizada
- Composição de títulos e textos
- Feedback em tempo real da qualidade do conteúdo
- Listas de reprodução personalizadas, resumos de notícias, recomendações
- Narração interativa de histórias por meio de escolhas do espectador
- Ofertas direcionadas, planos de assinatura

Serviços financeiros

- Descoberta de sinais de negociações, alertando corretores de ações sobre posições vulneráveis
- Aceleração de decisões de subscrição
- Otimização e reconstrução de sistemas legados
- Engenharia reversa de modelos bancários e de seguros
- Monitoramento de possíveis crimes e fraudes financeiras
- Automação da coleta de dados para conformidade regulatória
- Extração de percepções de divulgações corporativas

Fonte: compilado pela MIT Technology Review Insights, com base em dados 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" e "Large Language Models in Media & Entertainment",12 Databricks, abril-junho de 2023.

Casos de uso de IA generativa e grandes modelos de linguagem

Chatbots e assistentes virtuais

Suporte ao cliente

Geração de código e depuração de LLMs

Treinado com documentos da empresa

Análise de sentimentos

Avaliação da satisfação do cliente

Classificação e agrupamento de texto

Categorizar grandes volumes de dados para identificar tendências

Tradução de idiomas

Transição de páginas da web da empresa para outros idiomas

Resumos e paráfrases

Notas de reunião resumidas

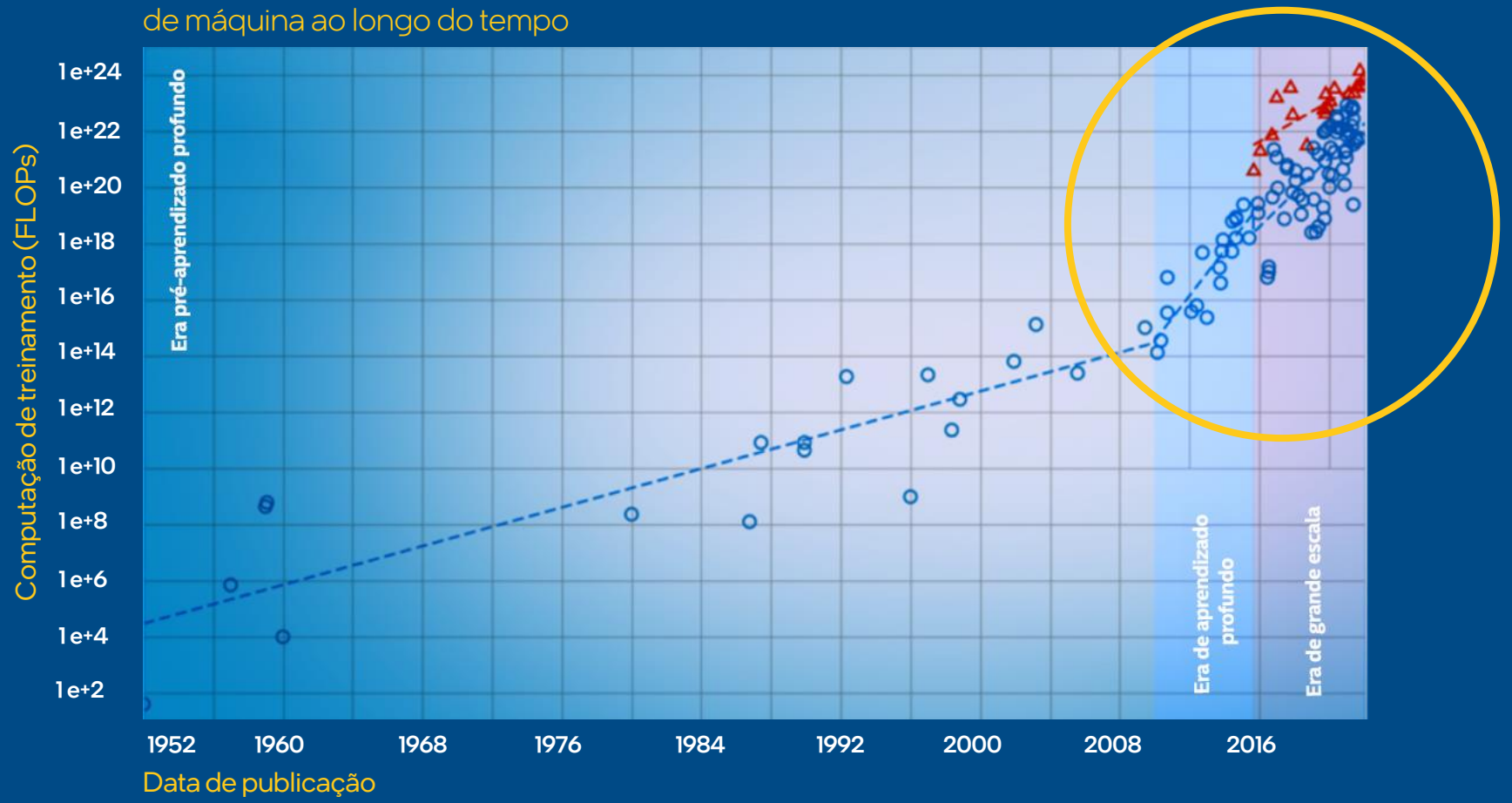
Geração de conteúdos, imagens e vídeos

Primeiros esboços de e-mails, geração de ideias, visuais de marketing, vídeos curtos

Fonte: A Compact Guide to Large Language Models, Databricks, 2023.

Conforme os modelos crescem, a computação também cresce

Computação de treinamento (FLOPs) dos principais sistemas de aprendizado de máquina ao longo do tempo



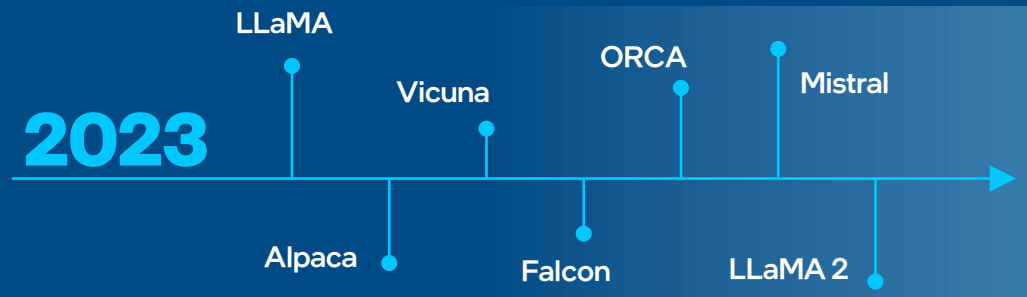
Estudo da Epoch, University of Aberdeen, Center for the Governance of AI, University of St. Andrews, MIT, Eberhard Karls Universitat Tubingen, Universidad Complutense

Não se trata apenas de modelos gigantes

	Gigante (terceiro)	VS.	**Pequeno e rápido** (em 10-100X)
Explicabilidade	Modelo proprietário	VS.	Modelo baseado em código aberto
Precisão	Propósito geral all-in-one	VS.	Direcionado, específico de domínio, personalizado
Local	Baseado em nuvem (como um serviço)	VS.	Inferência executada localmente; borda, cliente e no local
Custo	Dimensionamento do custo em perpetuidade	VS.	Gerenciamento de custos
Velocidade de entrada no mercado	Configuração rápida (segundos)	VS.	Tempo de construção (horas/dias)

Crescimento de muitos modelos menores

Centenas de bilhões para parâmetros <20 B em seis meses



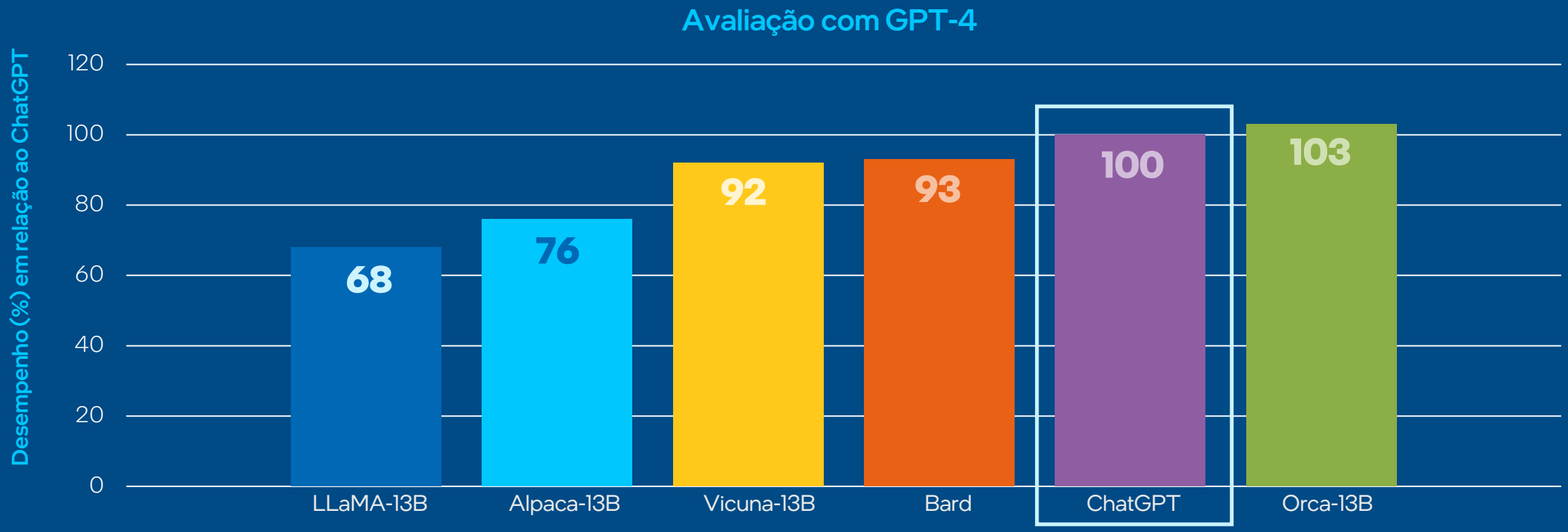
databricks

- Dúzias de modelos menores surgindo semanalmente
- Licenças comerciais e de código aberto
- Indicação de que modelos menores podem replicar a precisão de modelos maiores se treinados com dados obtidos de forma cuidadosa

- Milhares de modelos comerciais específicos de domínio e plataformas de IA sendo demonstrados
- Modelos podem ser ajustados em alguns processadores em dados específicos de domínio

Modelos menores tiveram um bom desempenho em comparação com o ChatGPT

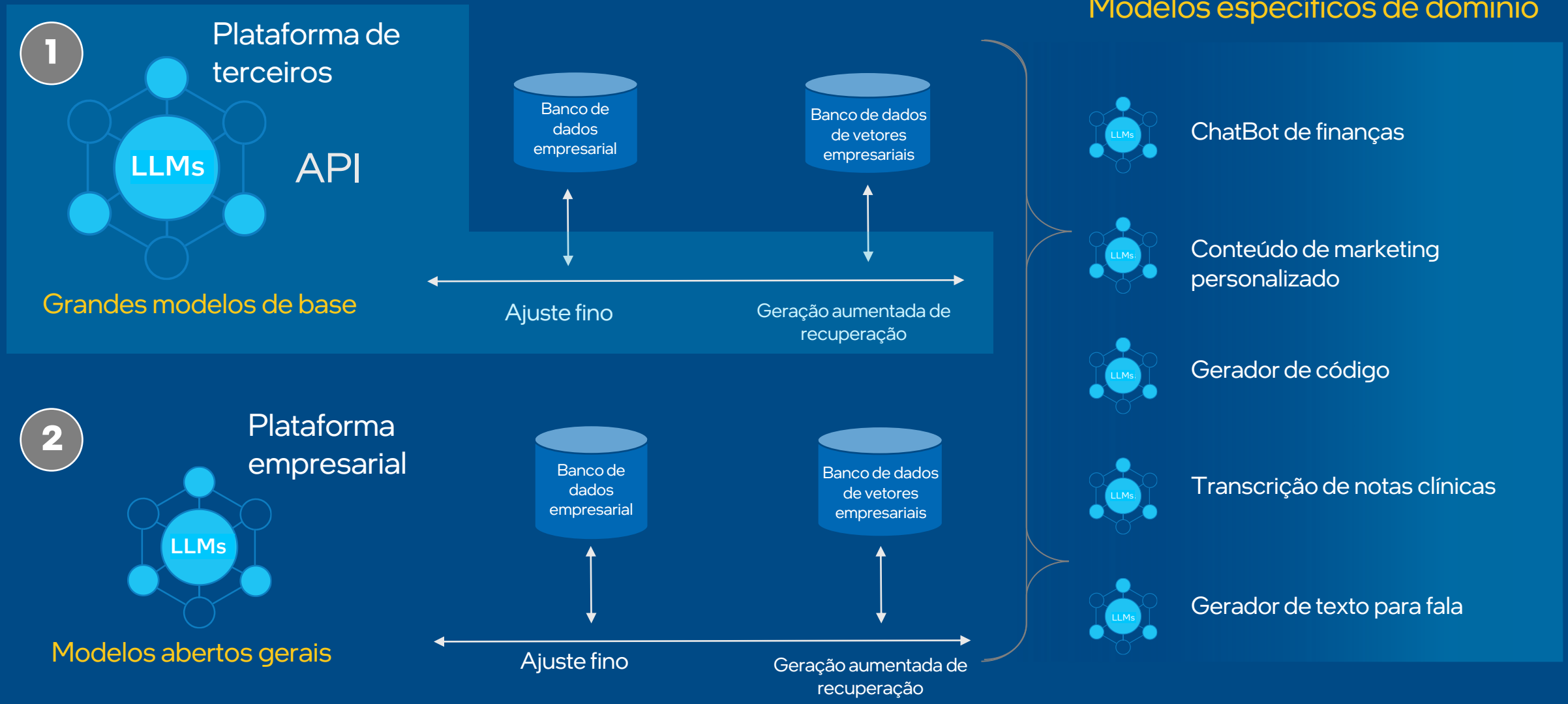
Prova de que modelos menores são uma opção viável e ainda têm um bom desempenho em comparação com modelos grandes, como o ChatGPT



O Orca supera uma série de modelos de base, incluindo o ChatGPT da OpenAI, conforme avaliado pelo GPT-4 no conjunto de avaliação Vicuna

Fonte: pesquisa da Microsoft (2023). Orca: aprendizado progressivo a partir de traços de explicação complexos do GPT-4

Construir modelos específicos de domínio



Modelos específicos de domínio têm muitos benefícios para empresas

Modelos menores e direcionados podem fornecer desempenho equivalente ou superior, aumentando o ROI ao reduzir o investimento de tempo e custo

Resultados mais precisos

Utilize seus dados empresariais para maior precisão específica de domínio

Custo menor

Ajuste fino de um modelo pré-treinado, e/ou uso de RAG, e inferência de modelos menores

Implantação em qualquer lugar na plataforma escolhida

Inferência executada localmente; borda, cliente e no local

Seguros e privados

Atende a requisitos regulatórios e de segurança de dados

IA responsável

Dar ao modelo a capacidade de citar fontes de dados com ajuste fino e RAG

O FUTURO

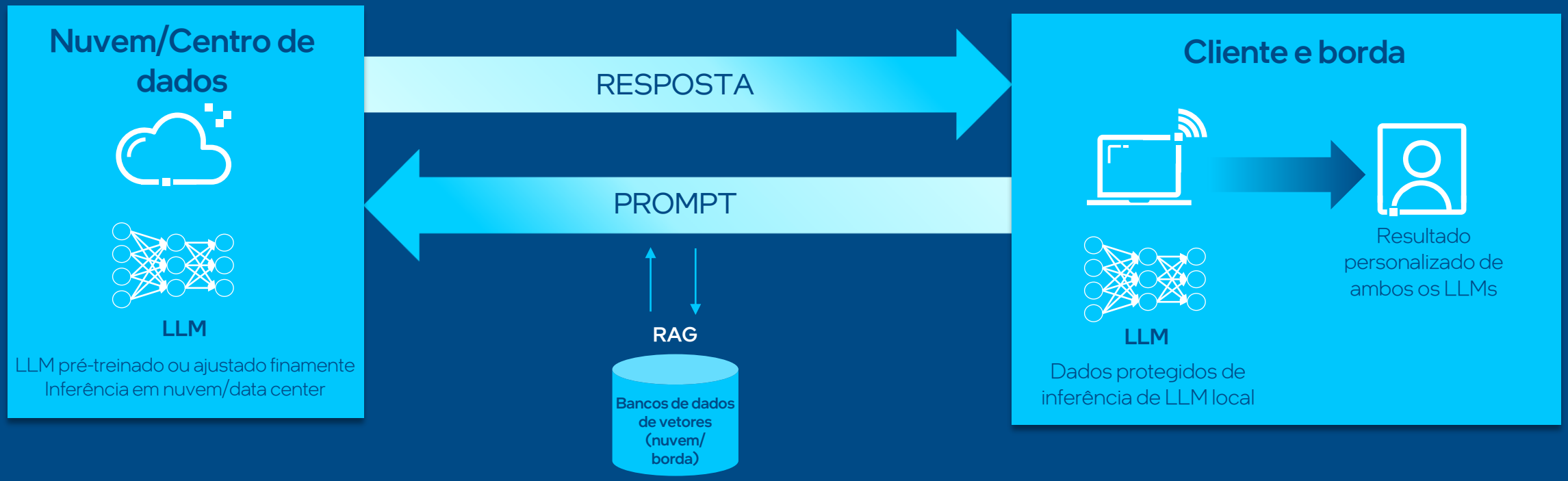
Haverá um pequeno número de modelos gigantes e um número gigante de modelos de IA menores e mais ágeis incorporados em inúmeras aplicações[1]

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

Plataforma de IA de nuvem para borda **perfeita**

Intel AI Portfolio

Treinamento e inferência na nuvem. Utilize RAG para melhorar a precisão de domínio.



IA generativa - Um ano em produção

O uso de modelos específicos de domínio, porém altamente inteligentes, está aumentando

2022 EXPERIMENTAÇÃO → **2023** PILOTOS → **2024** PRODUÇÃO

Modelos enormes abriram o caminho

- Muito eficazes para uso geral
- Caros para treinar e implantar
- Construídos com grandes conjuntos de dados públicos
- Fáceis de usar

Modelos menores, específicos de domínio

- Utilize seus dados privados para resultados empresariais específicos
- Implante no seu hardware existente
- Maior eficiência, precisão, segurança e rastreabilidade
- Tempo de construção

LEIA O BLOG

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

Abordagem da Intel sobre modelos específicos de domínio

MODELOS ESPECÍFICOS DE DOMÍNIO

Vantagens

+ Modelos de 10 a 100 vezes menores enquanto mantêm/melhoram a precisão
+ Econômicos em computação de uso geral
+ Exatidão; atribuição de fonte; explicabilidade
+ Utilização de dados privados/empresariais
+ Informações atualizadas continuamente

Desafios

- Gama de tarefas reduzida
- Requer pouco ajuste fino e indexação

META DA INTEL

Habilitar a abordagem mais econômica e universal para ajustar e implantar dezenas de milhares de modelos em hardware Intel utilizando estruturas do setor, modelos pré-treinados e softwares e ferramentas Intel AI

LEIA MAIS

IA generativa na ponta dos dedos

E-book ▪ Infográfico

IA empresarial: ajudando a superar as barreiras de entrada

Requisitos	Como uma parceria com a Intel® pode ajudar
Velocidade de entrada no mercado	Utilize recursos de desenvolvedor da Intel e da Hugging Face, o Gaudi Developer Hub e cinco kits de referência para obter uma grande vantagem na IA generativa
Experiência do usuário (precisão/latência)	Inferência em modelos maiores que 10 bilhões de parâmetros no acelerador Intel® Gaudi® e modelos menores que 20 bilhões de parâmetros em processadores Intel® Xeon® com Intel® AMX, dando aos usuários uma experiência em tempo real[1]
Disponibilidade de computação	Uma CPU Intel® Xeon® + aceleradores oferecem uma alternativa econômica à escassez global de CPUs. **O Intel® Gaudi® 2 agora está disponível por meio da SuperMicro, com maior disponibilidade para o Intel® Gaudi® 3.**
Tecnologia familiar	A inferência de modelos menores pode ser feita praticamente em qualquer hardware, incluindo soluções universais que podem já fazer parte da sua configuração de computação[2]
Operacionalize em escala	O Intel® Gaudi® 2 oferece escalabilidade quase linear com 24 portas de 100 GbE integradas a cada acelerador. O Intel® Xeon® já está no seu data center, no campo; da nuvem à borda. **65% das inferências de data center são executadas no Intel® Xeon®[3]**
Boa relação custo/benefício	Em aplicações reais, a Intel® está revolucionando o setor e democratizando a IA ao oferecer melhor desempenho, preços mais baixos e uma plataforma mais equilibrada para a inferência de IA. Veja a NVIDIA mostrar como o Intel® Gaudi 2 tem um desempenho por dólar quatro vezes melhor do que seu H100

[1]Fonte: Four Roadblocks to Implementing Generative AI

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

[3]Com base na modelagem de mercado da Intel® da base mundial instalada de servidores de data center que executam cargas de trabalho de inferência de IA em dezembro de 2022.

Recursos de software para simplificar o treinamento e a implantação de IA generativa

Modelo de código aberto

- BLOOM (a BigScience initiative) — 176 B
- BioGPT — Domínio 1,5 B
- Stable Diffusion — Imagem
- Llama2, GPT-J, MPT, Falcon — LLM de 7-65 B
- Stanford Alpaca — Preparado para LLM de 7 B
- Conhecimento base

Software aberto

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

Plataforma de GenAI

- Intel Gaudi
- Intel Data Center GPU Max Series
- Intel Xeon
- Intel Core Ultra

LEIA MAIS

Libere a IA generativa com hardware universal e software aberto

Maximize o valor

Por que a abordagem de IA aberta da Intel é adequada às suas necessidades de negócios de IA

Evite o bloqueio de fornecedores

software baseado em padrões de código aberto

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

Aproveite o portfólio de hardware da Intel

otimizado para casos de uso de IA

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

Crie novas oportunidades, desde o cliente e a borda até o data center e a nuvem, com hardware otimizado para software e padrões abertos para a IA do futuro

Portfólio de software Intel® AI

Projete dados → **Crie modelos** → **Otimize e implante**

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

Análise de dados em escala †

Frameworks de aprendizado de máquina e profundo, otimização e ferramentas de implantação†

oneAPI

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

Modelo de programação aberto e de arquitetura cruzada para CPUs, GPUs e outros aceleradores

NUVEM E EMPRESA

CLIENTE E ESTAÇÕES DE TRABALHO

BORDA

intel AI ANALYTICS TOOLKIT

Acelere a ciência de dados e a IA de ponta a ponta

Intel® Tiber™ AI Cloud and Intel® Developer Catalog

Experimente as ferramentas e hardware Intel mais recentes e acesse modelos de IA otimizados

Intel® Geti

Plataforma de anotação/treinamento/otimização

Hugging Face

Otimizações e receitas de ajuste fino da Intel, modelos de inferência otimizados e distribuição de modelos

Observação: os componentes em todas as camadas da pilha são otimizados para componentes direcionados em outras camadas com base nos modelos de uso de IA esperados, e nem todos os componentes são utilizados pelas soluções na coluna mais à direita

† Esta lista inclui frameworks de código aberto populares que são otimizados para o hardware Intel

Simplifique a adoção de IA generativa empresarial e reduza o tempo de produção de soluções reforçadas e confiáveis

OPEA:

Simplifique a adoção de IA generativa empresarial e reduza o tempo de produção de soluções reforçadas e confiáveis

Open Platform for Enterprise AI

Parceiros da OPEA

anyscale

CLOUDERA

DATASTAX

kx

DOMINO

Hugging Face

Red Hat

drant

MariaDB

MINIO

sas

Yellowbrick

vmware

Valor da OPEA

- Ajuda as empresas a desbloquear valor de seus dados usando a IA generativa (LLM, RAG) de forma mais rápida e fácil
- Reduz as complexidades do ecossistema fragmentado e ajuda as soluções a ampliar a produção
- Impulsiona a colaboração e a contribuição entre líderes do setor em parceria com a Linux Foundation

Eficiente
Aproveita a infraestrutura existente, o acelerador de IA ou outro hardware de sua preferência.

Transparente
Integra-se ao software empresarial, com suporte heterogêneo e estabilidade em todo o sistema e rede.

Aberto
Reúne as melhores inovações e está livre da dependência de fornecedores exclusivos.

Onipresente
Opera em todos os lugares por meio de uma arquitetura flexível construída para nuvem, data center, borda e PC.

Confiável
Apresenta um pipeline e ferramentas seguros prontos para a empresa para oferecer responsabilidade, transparência e rastreabilidade.

Escalável
Fornece acesso a um ecossistema vibrante de parceiros para ajudar a construir e ampliar sua solução.

Parceria com a Huggin Face para IA generativa

Hugging Face

Para facilitar a inovação e o treinamento de IA generativa e IA de linguagem, a Intel fez uma parceria com a Hugging Face, uma plataforma popular para o compartilhamento de modelos e conjuntos de dados de IA. Mais notavelmente, a Hugging Face é conhecida por sua biblioteca de transformadores construída para NLP.

intel Xeon

A Intel® trabalhou com a Hugging Face para criar aceleração de hardware e software de última geração para treinar, ajustar finamente e prever com modelos de transformadores.

A aceleração de hardware é impulsionada pelos processadores escaláveis Intel® Xeon®, enquanto a aceleração de software é habilitada pelo nosso portfólio de ferramentas de software, frameworks e bibliotecas de IA otimizados.

intel Gaudi

Os aceleradores de aprendizado profundo Intel® Gaudi® também estão alinhados ao software de código aberto da Hugging Face por meio da Optimum Habana Library , para pproporcionar facilidade de uso para os desenvolvedores em milhares de modelos otimizados pela comunidade Hugging Face.

A Hugging Face também publicou várias avaliações de desempenho do Intel® Gaudi® 2 em modelos de IA generativa: Stable Diffusion, T5-3B, BLOOMZ 176B e 7B, e o novo modelo BridgeTower.

A Intel®, Articul8 e BCG colaboram para oferecer IA generativa segura e de nível empresarial

A solução pioneira impulsionada por supercomputadores de IA da Intel agrega valor de negócios com conjuntos de dados personalizados enquanto mantém altos níveis de segurança e privacidade de dados

A Articul8* oferece uma plataforma de software de IA generativa pronta para uso que oferece velocidade, segurança e economia para ajudar grandes clientes empresariais a operacionalizar e escalar a IA. A plataforma foi lançada e otimizada em arquiteturas de hardware Intel®, incluindo processadores escaláveis Intel® Xeon® e aceleradores Intel® Gaudi®, mas será compatível com uma série de alternativas de infraestrutura híbrida.

intel GAUDI

intel XEON

Após a implantação inicial da tecnologia no Boston Consulting Group (BCG), a equipe escalou a plataforma para clientes empresariais em segmentos de mercado que exigem altos níveis de segurança e conhecimento de domínio especializado, incluindo serviços financeiros, aeroespaciais, semicondutores e telecomunicações.

LEIA MAIS

Anúncio da Articul8

Site da Articul8

IA responsável para empresas

DESAFIO:

Modelos de IA generativa aprendem com vastas quantidades de dados disponíveis na internet, que podem conter vieses presentes na sociedade, e podem aplicar tais vieses inadvertidamente. Os LLMs podem ser manipulados para gerar ou espalhar desinformação, e-mails de phishing ou ataques de engenharia social.

LLMs podem ter "alucinações" e gerar informações imprecisas, o que pode ser especialmente problemático em setores como a saúde, no qual modelos podem influenciar diagnósticos e decisões terapêuticas e potencialmente prejudicar pacientes.

Saiba mais

Minimizando os riscos da IA generativa

SOLUÇÕES:

Empresas e indivíduos que trabalham com tecnologia de IA devem garantir que seus softwares sejam desenvolvidos e implantados de acordo com os princípios éticos da IA

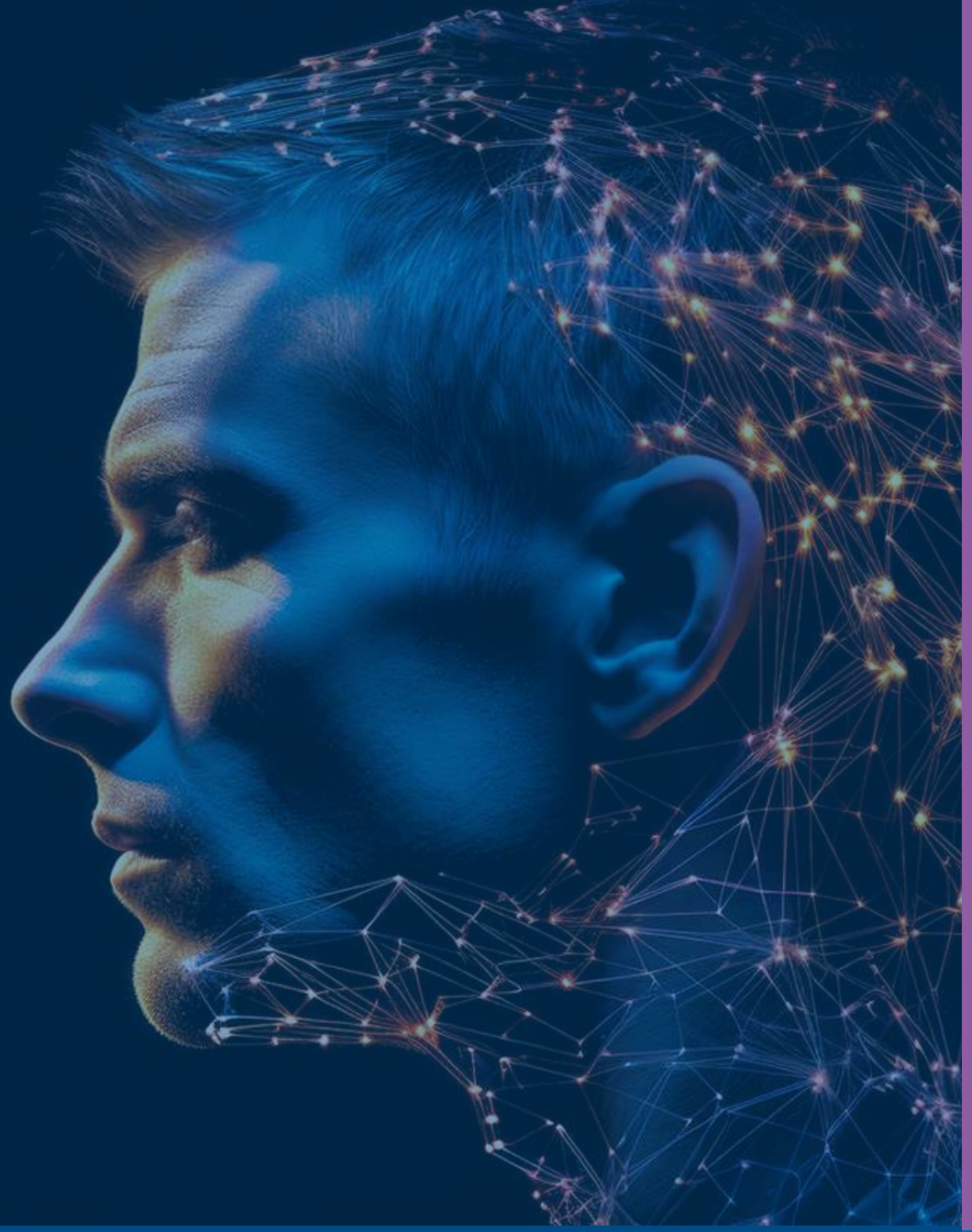
As Intel® Explainable AI Tools de código aberto permitem que usuários executem destilação e visualização de modelos post-hoc para examinar o comportamento preditivo de modelos TensorFlow* e PyTorch*

Geralmente, LLMs são treinados com grandes conjuntos de dados e ajustados finamente sobre dados potencialmente confidenciais (por exemplo, financeiros e médicos)

Tecnologias como o Open Federated Learning (OpenFL) da Intel incorporam computação confidencial para que LLMs possam ser ajustados finamente com segurança sobre dados confidenciais, o que, por sua vez, melhora a capacidade de generalização de modelos enquanto reduz alucinações e vieses.

Produtos Intel® para IA generativa

Levando a IA a todos os lugares



Sistemas escaláveis para IA

Treinamento e ajuste fino

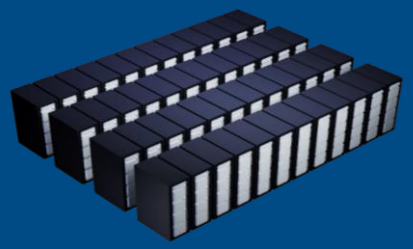
Treinamento | Inferência de pico | Inferência/ajuste fino convencionais | Inferência de linha de base | Inferência de terminais

Inferência e implantação

Nuvem
Data Center

Borda

Cliente

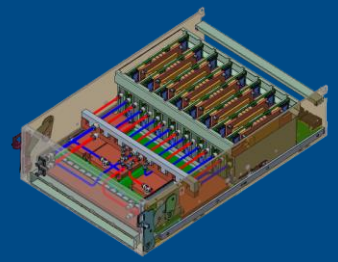


Escala de clusters e data centers

intel. GAUDI

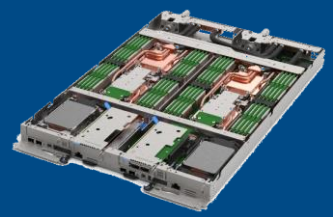
Implantação multinós por rack

intel. GAUDI | intel. XEON | intel. XEON MAX SERIES



Multi-GPU ou CPU de vários soquetes

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



CPU de soquete único ou GPU única

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

CPU cliente

intel CORE ULTRA | intel. ARC GRAPHICS

intel. ETHERNET

Produtos Intel® para NLP/LLMs

Inferência e treinamento

GAUDI®2

Os aceleradores de IA Intel® Gaudi® 2 são projetados especificamente para acelerar o treinamento e a inferência de modelos de grande escala, como LLMs e NLPs.

Acelerando a IA generativa e grandes modelos de linguagem com o Intel® Gaudi® 2

O Intel® Gaudi® 2 oferece desempenho líder e economia de custo ideal para treinamento de IA

Comunicado à imprensa

Assista agora

Gravação do webinar da Intel® sobre os recursos de ponta do processador de IA Intel® Gaudi® 2 para capturar o potencial da IA generativa e de grandes modelos de linguagem (LLMs)

O acelerador de aprendizado profundo Intel® Gaudi® 2 tem desempenho competitivo em treinamento e inferência de aprendizado profundo, com desempenho até **2,4x mais rápido que o NVIDIA A100**[1]

Sala de imprensa ▪ Artigo técnico

O Intel® Gaudi® 2 permanece como a única alternativa de referência ao NV H100 para desempenho de IA generativa

[1]O desempenho varia de acordo com o uso, a configuração e outros fatores; cargas de trabalho e detalhes de configuração disponíveis em: intel.com/performanceindex Os resultados podem variar.

Gaudi2: ideal para treinamento e inferência eficientes de modelos de base

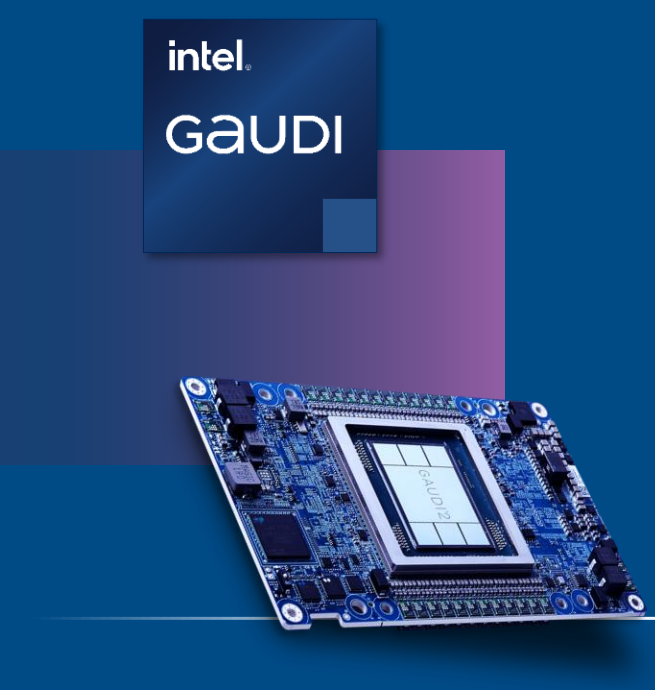
O Gaudi2 foi arquitetado para desempenho, eficiência e escalabilidade de aprendizado profundo para atender às demandas de modelos de base de grande escala, como LLMs (GPT) e GAIs (Stable Diffusion)

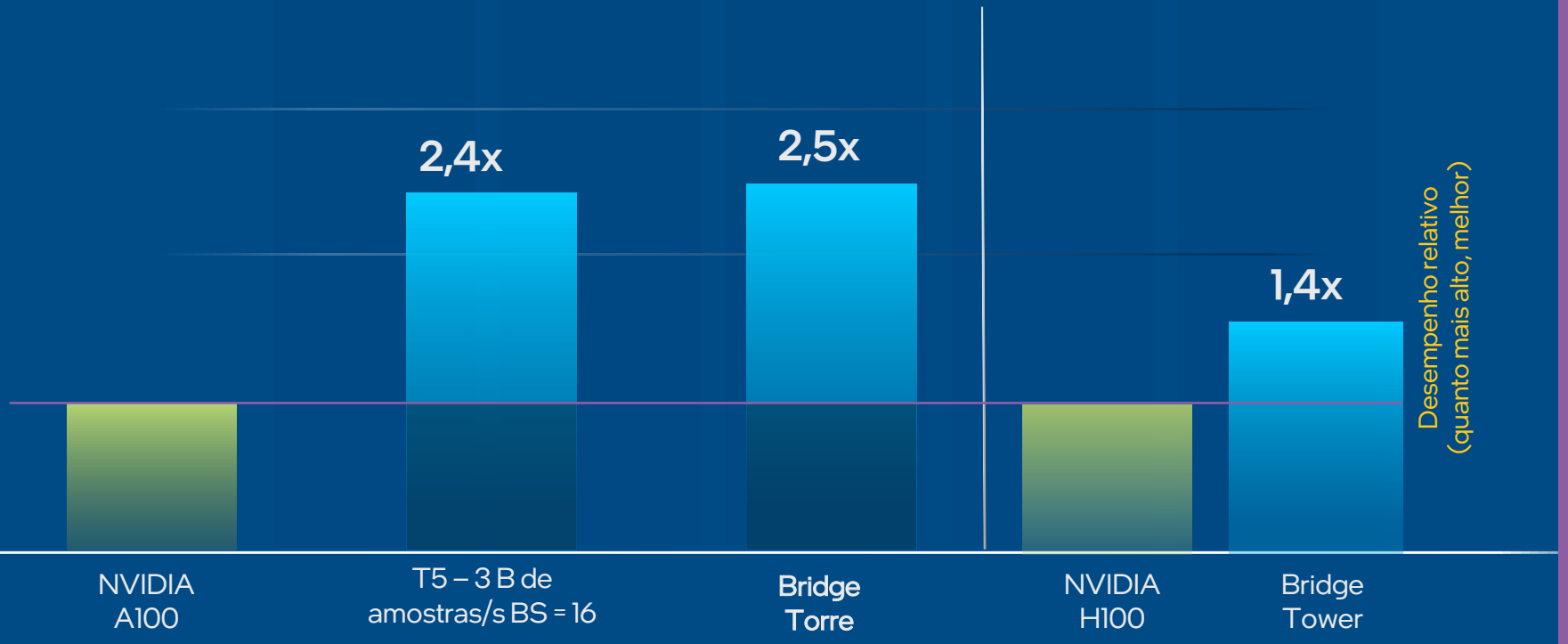
Requisitos	Gaudi2
Velocidade	**1,5-2x mais rápido** que o A100 para treinamento e inferência
Memória	Cada dispositivo Gaudi2 apresenta **96 GB de memória de alta largura de banda no chip**, facilitando encaixar grandes modelos de base na memória, além de treiná-los e implantá-los em escala
Escalabilidade	Escalando a eficiência com **24 portas de 100 GbE integradas no chip**, conectividade direta para todos entre 8 placas em um servidor e comunicação aberta baseada em ROCEv2 dentro e entre servidores.
Facilidade de uso	Migre ou crie modelos com **alterações mínimas de código** com SynapseAI, PyTorch e DeepSpeed
Eficiência no uso de energia	**taxa de transferência/Watt ~1,8x mais alta em comparação com o A100**
Eficiência de custos	Baseado na arquitetura Gaudi da 1ª Geração criada especificamente para esse fim, com **um desempenho de preços até 40% melhor** do que o A100 na nuvem da Amazon

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

Ajuste fino entre vários LLMs

Avaliações da Hugging Face comprovam o desempenho de LLM do acelerador Intel® Gaudi® 2 em relação ao NVIDIA A100 e H100





Acesse https://habana.ai/habana-claims-validation para consultar cargas de trabalho e configurações. Os resultados podem variar.

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

GPT-J: Resultados do Intel® Gaudi® 2

Os resultados de desempenho de inferência do Intel® Gaudi® 2 para GPT-J fornecem uma sólida validação de seu desempenho competitivo

- Os desempenhos de inferência do Intel® Gaudi® 2 em GPT-J-99 e GPT-J-99.9 para consultas de servidores e amostras offline são **de 78,58 por segundo e 84,08 por segundo**, respectivamente[1]
- **O Intel® Gaudi® 2 oferece desempenho irrefutável em comparação com o H100 da NVIDIA**, com o H100 mostrando uma pequena vantagem de 1,09x (servidor) e 1,28x (offline) em relação ao Gaudi 2[1]
- **O Intel® Gaudi® 2 supera o A100 da NVIDIA em 2,4x (servidor) e 2x (offline)** [1]
- O envio do Intel® Gaudi® 2 empregou o FP8 e atingiu **99,9% de precisão** nesse novo tipo de dados[1]

LEIA MAIS

Com atualizações de software do Intel® Gaudi® 2 lançadas a cada seis a oito semanas, a Intel® espera continuar oferecendo avanços de desempenho e cobertura de modelos ampliada em parâmetros de referência do MLPerf

Artigo da sala de imprensa

Anúncio da MLCommons

[1]O desempenho varia de acordo com o uso, a configuração e outros fatores; cargas de trabalho e detalhes de configuração disponíveis em: https://mlcommons.org/2023/09/mlperf-results-highlight-growing-importance-of-generative-ai-and-storage/ Os resultados podem variar.

Intel® Gaudi® 2: resultados dos parâmetros de referência

Resultados dos parâmetros de referência fornecidos pela Supermicro;
a primeira OEM do Intel® Gaudi® 2 do mercado

Validação de alegações do Gaudi

Treinamento e inferência de LLMs com os aceleradores de IA Intel® Gaudi® 2

Parâmetros de referência

Treinamento e inferência mais rápidos: Intel® Gaudi® 2 versus NVIDIA A100 80 GB

Parâmetros de referência

Os resultados podem variar.

Intel® Gaudi® 2: treinamento e inferência de modelos de base

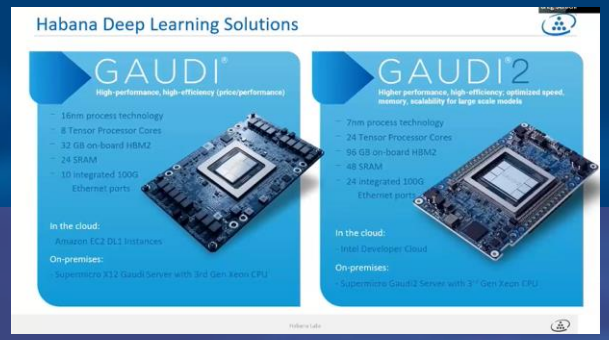
Modelos habilitados pelo Gaudi disponíveis podem ser acessados no

Catálogo do desenvolvedor

GAUDI®2

Treinamento para desenvolvedores do Intel® Gaudi®

[Primeiros passos: aprendizado profundo e inferência no Gaudi]



[Maximizando o poder do Intel® Gaudi® 2: acelerando a IA generativa e grandes modelos de linguagem]

[Maximizando o desempenho de modelos com os processadores Intel® Gaudi®: ferramentas e estratégias avançadas para resultados ideais]

Software Intel® Gaudi® (SynapseAI® Software Suite)

Desenvolvimento simplificado: a maneira como você quer desenvolver

Objetivo: facilitar a migração de softwares existentes para aceleradores de IA Intel® Gaudi®, preservando investimentos em software e facilitando a construção de novos modelos, tanto para treinamento quanto para implantação dos vários e crescentes modelos que definem o aprendizado profundo, a IA generativa e grandes modelos de linguagem.

Amplo suporte para cientistas de dados, desenvolvedores e administradores de TI e sistemas com o:

- Site do desenvolvedor
- GitHub

Acelerador de AI Intel® Gaudi®

O ecossistema de software para aprendizado profundo reúne os principais provedores de software, ferramentas e código para acelerar o desenvolvimento de modelos de aprendizado profundo de última geração baseados nos frameworks PyTorch, TensorFlow, PyTorch Lightning e DeepSpeed

cnvrg.io

PyTorch Lightning

Pronto para utilizar o software Intel® Gaudi®?

Aceleradores de IA Intel® Gaudi® 2 DISPONÍVEIS AGORA! Exclusivamente na nuvem Denvr

Ecossistema do software
Intel® Gaudi® 2

Acelerador de IA
Intel® Gaudi® 2 (7 nm)

Intel® Gaudi® 2 — ideal para as demandas de IA generativa

- Já disponível! Clusters do Gaudi 2 na nuvem Denvr
- Teste até 8 nós do Gaudi 2
- Preços VIP prioritários para clientes Intel
- Serviço e suporte comerciais de alto nível da Denvr Dataworks
- Migração perfeita para clusters do Gaudi 2 na nuvem Denvr
- Posicionamento prioritário exclusivo para clusters do Gaudi 3 na nuvem Denvr — em breve!

Comece agora mesmo

EM BREVE

intel GAUDI

Inferência e treinamento

Intel® Gaudi® 3

Com desempenho, escalabilidade e eficiência que oferecem mais opções para mais clientes, os aceleradores Intel® Gaudi® 3 ajudam as empresas a obter percepções, inovações e receita

EM BREVE - Acelerador de IA Intel® Gaudi® 3

Trazendo opções para a IA generativa com desempenho, escalabilidade e eficiência

O Intel® Gaudi® 3 oferecerá um salto significativo no treinamento e na inferência de IA para empresas globais que desejam implantar IA generativa em escala

Comunicado à imprensa

Desempenho do acelerador Intel® Gaudi® 3 em comparação com o NVIDIA H100

É estimado que o Intel® Gaudi® 3 ofereça um tempo de treinamento em média

50% mais rápido[3] em modelos Llama2 com 7 B e 13 B de parâmetros e no modelo GPT-3 de 175 B

É esperado que o Intel® Gaudi® 3 supere o H100 em:

50% na taxa transferência de inferência do acelerador[1]

40% na eficiência energética de inferência[2]

no Llama de 7 B e 70 B de parâmetros, e modelos do Falcon de 180 B de parâmetros

LEIA MAIS

PUBLICAÇÃO TÉCNICA

O Intel® Gaudi® 3 estará disponível para OEMs a partir do 2º trimestre de 2024, incluindo:

Lenovo DELL Hewlett Packard Enterprise SUPERMICRO

[1]Comparação com o NV H100 baseado em https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8; os números relatados são por GPU. versus projeções do Intel® Gaudi® 3 para projeções LLAMA2-7B, LLAMA2-70B e Falcon de 180 B. Os resultados podem variar.
[2]Comparação com o NV H100 com base em https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8; os números relatados são por GPU. versus projeções do Intel® Gaudi® 3 para LLAMA2-7B, LLAMA2-70B e Falcon de 180 B. Eficiência energética para NVIDIA e Gaudi 3 com base em estimativas internas. Os resultados podem variar.
[3]Comparação com o NV H100 com base em: https://developer.nvidia.com/deep-learning-performance-training-inference/training, guia "Grande modelo de linguagem" versus projeções do Intel® Gaudi® 3 para LLAMA2-7B, LLAMA2-13B e GPT3-175B em 28/3/2024. Os resultados podem variar.

Produtos Intel® para NLP/LLMs

Inferência

Processadores escaláveis Intel® Xeon® da 4ª e 5ª Geração aceleram o NLP com o Intel® DL Boost, Intel® AMX e Intel® AVX-512. Foram projetados para computação de alto desempenho e podem ser utilizados para acelerar cargas de trabalho de NLP. Podem lidar com um grande número de threads, grande capacidade de memória e alta largura de banda de memória, o que é adequado para cargas de trabalho de NLP, como tradução de idiomas, resumo de textos e texto para fala.

Intel® Xeon® da 5ª Geração: o processador projetado para IA

Com aceleração de IA em todos os núcleos, os processadores Intel® Xeon® da 5ª Geração atendem às exigentes cargas de trabalho de IA de ponta a ponta, antes que os clientes precisem adicionar aceleradores dedicados

Maior desempenho em inferência de IA

até 42%

em comparação com a geração anterior[1]

Ganhos de desempenho geral de computação

média de 21%

em comparação com a geração anterior[1]

Processamento de linguagem natural mais rápido

até 23%

em comparação com a geração anterior[1]

Sandra Rivera, Vice-Presidente executiva e Gerente geral do grupo de data center e IA da Intel

"Projetados para IA, nossos processadores Intel® Xeon® da 5ª Geração oferecem maior desempenho para os clientes que implantam recursos de IA em casos de uso de nuvem, rede e borda. Como resultado de nosso trabalho de longa data com clientes, parceiros e o ecossistema de desenvolvedores, estamos lançando o Intel® Xeon® da 5ª Geração com uma base comprovada que permitirá uma adoção rápida e escala com um TCO menor."

Mais informações

Site

Resumo do produto

Para cargas de trabalho e configurações, visite intel.com/processorclaims: Processadores escaláveis Intel® Xeon® da 5ª Geração. Os resultados podem variar.

Intel® Xeon®: liderança em desempenho de CPU em aplicativos de IA do mundo real

Em aplicativos de trabalho reais, a Intel está revolucionando o setor e democratizando a IA ao oferecer melhor desempenho, preços mais baixos e uma plataforma mais equilibrada para a inferência de IA com:

- Cache maior que ajuda com a localidade de dados e grande capacidade de memória que permite resolver problemas maiores
- Frequência de núcleo mais alta, várias portas escalares e execução fora de ordem que ajudam a acelerar a computação de thread única ou multithread, mas escalar
- Intel® Advanced Vector Extensions 512 (Intel® AVX-512) que ajuda com a computação vetorial não DL
- Intel® Advanced Matrix Extensions (Intel® AMX) que tem suporte de hardware integrado para aceleração de IA

Artigo técnico completo

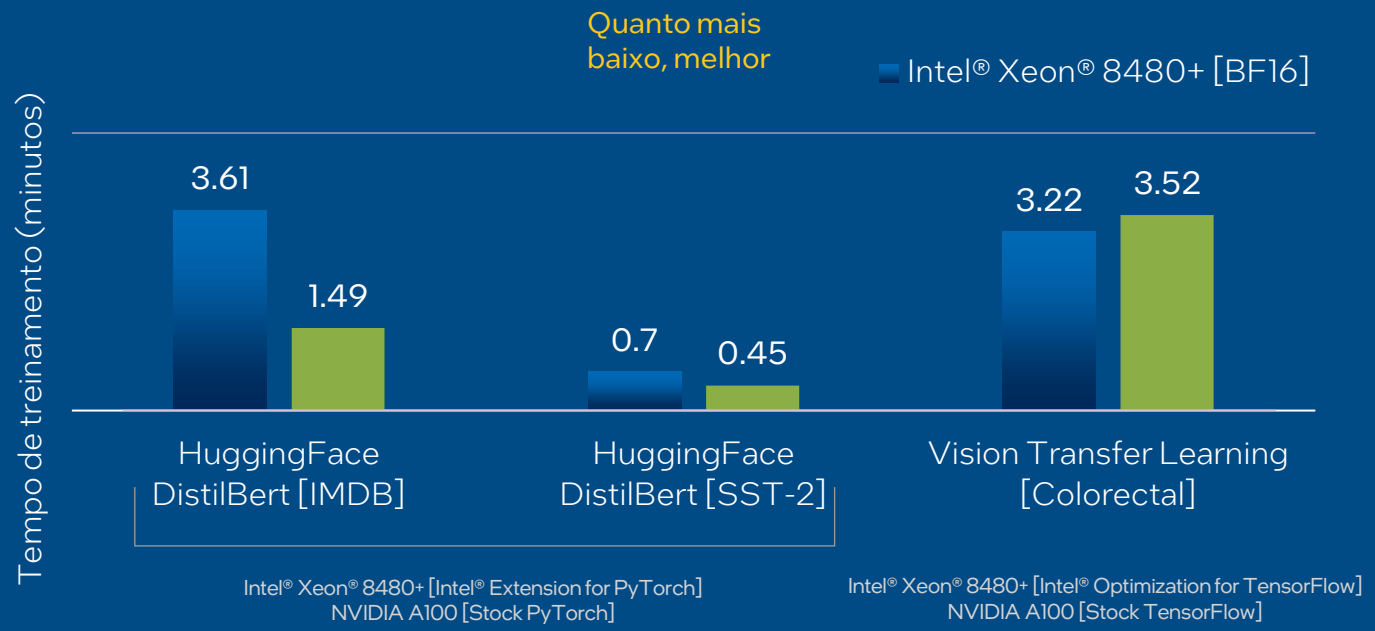
Infográfico

Desmascarando o mito da GPU: como CPUs com aceleradores integrados revolucionam a IA

Ajuste fino de modelos em menos de quatro minutos com processadores escaláveis Intel® Xeon®[1]

Hugging Face

Desempenho de tempo de treinamento de ajuste fino do processador Intel® Xeon® Platinum 8480+ em comparação com a GPU NVIDIA A100



Ver também:

Melhor desempenho: Numenta em CPUs Intel® em comparação com GPUs NVIDIA

[1]Consulte [A221] do Índice de desempenho de processadores escaláveis Intel® Xeon® da 4ª Geração Os resultados podem variar.

LLMs em processadores Intel® Xeon® da 4ª Geração

A tecnologia de chatbots de inteligência artificial (IA) está se tornando cada vez mais popular entre empresas e organizações como uma maneira de interagir com clientes e melhorar o atendimento ao cliente, mas a construção, otimização e manutenção de chatbots para casos de uso específicos são caras e podem ser financeiramente proibitivas para muitas organizações

MAIS INFORMAÇÕES

Guia de ajuste para IA nos processadores escaláveis Intel® Xeon® da 4ª Geração

Link para o Guia ›

Os processadores Intel® Xeon® da 4ª Geração oferecem gerenciamento de dados aprimorado e computações eficientes por meio das **Intel® Advanced Matrix Extensions (AMX)** e quando combinados com a funcionalidade **Auto Mixed Precision (AMP)** disponível por meio da Intel® Extension for PyTorch, essa pilha de tecnologia torna-se bastante competitiva para cargas de trabalho como transferência de aprendizado e treinamento de modelos pequenos/médios do zero

Artigo técnico instrutivo

Cisco UCS com processadores Intel® Xeon® da 5ª Geração e da 4ª Geração para IA generativa

Quanto menor, melhor: o Q8-Chat LLM é uma experiência de IA generativa eficiente em processadores Intel® Xeon®

LLMs exigem muito poder de computação, o que normalmente é encontrado em GPUs de alto nível, para prever com rapidez suficiente para casos de uso de baixa latência, como aplicativos de pesquisa ou de conversação. Infelizmente, para muitas organizações, os custos associados podem ser proibitivos e dificultam o uso de LLMs de última geração em suas aplicações.

Hugging Face

"Mais empresas seriam mais bem atendidas com foco em modelos menores e específicos que são mais baratos de treinar e executar."

Aprenda sobre técnicas de otimização que ajudam a reduzir o tamanho de LLMs e a latência de inferência, ajudando a serem executados de forma eficiente em CPUs Intel®.

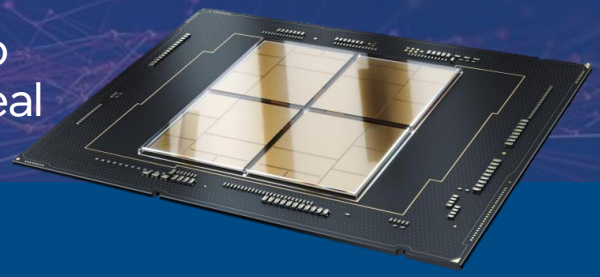
Artigo técnico instrutivo ›

Comece agora com o Intel® Xeon® da 4ª Geração com a Hugging Face

Processadores Intel® Xeon® para LLMs

RESUMO

- Bem posicionado para inferência de LLMs de domínio especializado
- Atende a casos de uso de aprendizado de transferência
- Implante LLMs no Intel® Xeon® com software de código aberto para facilitar o fornecimento de desempenho ideal



Processadores escaláveis Intel® Xeon® para LLMs

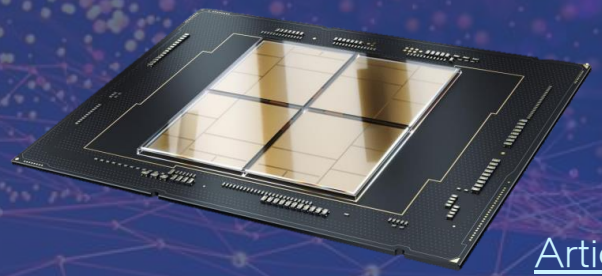
Ideal para a construção e implantação de cargas de trabalho de IA de uso geral com as estruturas e bibliotecas de IA mais populares

- Utilize a infraestrutura existente para inferência de LLMs específicos de domínio
- Atende a casos de uso de aprendizado de transferência
- Implante LLMs no Intel® Xeon® com software de código aberto para facilitar o fornecimento de desempenho ideal

Intel® Xeon®

Liderança no desempenho de CPU em aplicações de IA reais

Artigo técnico ▪ Infográfico



GPT-J

Resultados do Intel® Xeon® da 4ª Geração

2 parágrafos por segundo no modo offline[1]

1 parágrafo por segundo no modo de servidor em tempo real[1]

Artigo da sala de imprensa ▪ Anúncio da MLCommons

Desmascarando o mito da GPU: como CPUs com aceleradores integrados revolucionam a IA

Estudo de caso do Alibaba NLP no Intel® Xeon® da 4ª Geração com Intel® AMX

LEIA MAIS

[1]O desempenho varia de acordo com o uso, a configuração e outros fatores; cargas de trabalho e detalhes de configuração disponíveis em: https://mlcommons.org/2023/09/mlperf-results-highlight-growing-importance-of-generative-ai-and-storage/ Os resultados podem variar.

Produtos Intel® para NLP/LLMs

Inferência de pequena escala no cliente

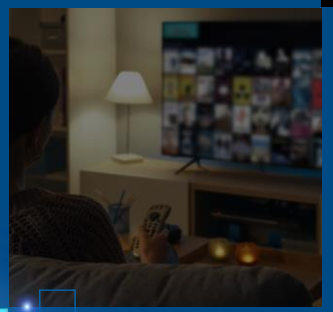
O Intel® Core™ Ultra anuncia a era do PC com IA

Os processadores Intel® Core™ Ultra são otimizados para notebooks finos e potentes de alto nível, apresentando arquitetura híbrida de desempenho 3D, recursos avançados de IA e disponíveis com a GPU Intel® Arc™ integrada. Criados utilizando o novo processo Intel® 4, os processadores Intel® Core™ Ultra oferecem um equilíbrio ideal entre desempenho e eficiência energética para jogos, criação de conteúdo e produtividade em qualquer lugar.

Casos de uso: IA no PC

Criador: pesquisa e edição de fotos e vídeos

Filtros mais rápidos e naturais, pré-visualizações de maior qualidade e tempos de exportação mais rápidos com pesquisas automatizadas e mais rápidas.



Colaboração/streaming

Novos recursos de IA para videoconferências, streaming e colaboração de última geração, preservando a autonomia da bateria.

Jogos convencionais

Novos recursos de IA para animação 3D no jogo para maior realismo, transcrição e tradução de bate-papo.

IA no PC

"Desbloqueando o mundano"

Produtividade

Assistentes de IA para escrever, criar, programar e recursos offline, como texto e gramática preditivos.

Criador: texto para imagem

Novos efeitos e recursos de IA para a criação de imagens com apenas algumas palavras descritivas — marketing, publicidade, design.

Acessibilidade

Recursos audiovisuais assistidos por IA para diversas necessidades dos usuários, facilitando a criação e a produtividade no PC.

*Intel, 14 de dezembro de 2023, comunicado de imprensa do evento sobre IA em todos os lugares

Intel® Core™ Ultra para IA generativa

O processador cliente com maior eficiência energética da Intel anuncia a era do PC com IA

Grandes melhorias em eficiência e desempenho

EFICIÊNCIA DE IA

até **70%**

maior rapidez no desempenho de IA generativa[2]

ECONOMIA DE ENERGIA

até **25%**

de redução no consumo de energia[3]

LEIA MAIS [Anúncio](#) ▪ [Resumo de produtos](#) ▪ [Site](#)

O Intel® Core™ Ultra apresenta o primeiro acelerador de IA cliente no chip da Intel — a unidade de processamento neural, ou NPU — para permitir um novo nível de aceleração de IA com **eficiência energética 2,5x melhor** que a geração anterior[1]

Tanto a geração de chips Intel® Core™ Ultra H quanto a U incluem dois novos núcleos Low Power Island (LP-E) para cargas de trabalho de baixa intensidade, com dois mecanismos de computação neural dentro da NPU Intel AI projetados para lidar com a **inferência de IA generativa**.

[1]Conforme medido pelo desempenho por watt no parâmetro de referência UL Procyon AI executando um modelo int8 na NPU Intel® Core™ Ultra 7 165H versus a GPU Intel® Core™ i7-1370P.
[1, 2, 3]Consulte www.intel.com/PerformanceIndex para cargas de trabalho e configurações. Os resultados podem variar.

Acelerando a inovação da IA

A Intel® está trabalhando com ISVs líderes do setor para otimizar sua experiência com IA.

O Programa de aceleração IA em PC visa conectar fornecedores de hardware independentes (IHVs) e fornecedores de software independentes (ISVs) com os recursos Intel®, incluindo ferramentas de inteligência artificial (IA), treinamento, engenharia conjunta, otimização de software, hardware, recursos de design, experiência técnica, marketing conjunto e oportunidades de vendas.

[Saiba mais](#)

Acelere o desenvolvimento de IA empresarial com a Nuvem para desenvolvedores Intel® Tiber™ AI Cloud (anteriormente; Intel® Developer Cloud)

Aprenda, faça protótipos, teste e execute aplicativos e cargas de trabalho em um cluster dos mais recentes hardwares e softwares Intel®

Acelere e **escale a IA** com as mais recentes inovações de hardware e software nesse ambiente de desenvolvimento. **Obtenha mais** poder de computação e opções para **ajustar seu software** e **sua IA generativa.**

Comece agora com a Intel

Obtenha experiência prática com os produtos Intel mais recentes. Capacite suas habilidades de IA com a Intel.

Acesso antecipado a tecnologias

Avalie plataformas Intel de pré-lançamento e pilhas de software associadas otimizadas pela Intel.

Implante IA em escala

Acelere as implantações de IA com os mais recentes kits de ferramentas de aprendizado de máquina da Intel e bibliotecas hospedadas na Nuvem para desenvolvedores Intel® Tiber™.

Leia o artigo técnico ›

Comece agora ›

Chamada à ação

EDUCAÇÃO

Entenda como a tecnologia Intel® pode ser utilizada para IA generativa e modelos específicos de domínio, e o escopo com o qual as linhas de produtos Intel® Xeon® e Intel® Gaudi® podem ajudá-lo a conquistar mais negócios

Comece agora

ENVOLVIMENTO

Comece a usar a

Intel® Tiber™ AI Cloud

Acelere e escale a IA com as mais recentes inovações de hardware e software nesse ambiente de desenvolvimento

&

Utilize kits de referência de IA

CONTATO

Entre em contato com o seu **representante Intel®** para obter mais informações

Como acessar o suporte ao cliente do Intel® Partner Alliance

Intel® Virtual Assistant

Esse chatbot, localizado no canto inferior direito de cada página do Partner Alliance, fornece autoajuda para a maioria das dúvidas ou um link rápido de contato com um agente de suporte ao vivo.

Faixa "Obter ajuda"

Envie uma solicitação de suporte online.

Este link pode ser encontrado no rodapé da maioria das páginas do site do Partner Alliance.

Página “Obter ajuda” do Partner Alliance

A página Obter ajuda fornece guias de autoajuda detalhados sobre a maioria das ferramentas e dos benefícios disponíveis para membros do Partner Alliance.

Zonas de ativação de IA

Espaços de trabalho de IA digitais que selecionam recursos, ferramentas e benefícios críticos — ativando parceiros para construir, comercializar e vender soluções com base em tecnologia Intel®

Habilitação técnica

Capacitação de vendas e marketing

Habilitação técnica

Capacitação de vendas e marketing

Habilitação técnica

Capacitação de vendas e marketing

Kits de referência de IA

Ao aproveitar esses kits de referência, as organizações podem **reduzir significativamente o tempo** de solução e obter **ganhos de desempenho substanciais**

Finanças e seguros

Detecção de fraudes

GitHub ▪ Blog ▪ Projeto

Saúde e ciências da vida

Proteção contra doenças

GitHub ▪ Blog

Fabricação e serviços públicos

Detecção de anomalias

GitHub ▪ Blog

Gerenciamento de frotas

Manutenção preditiva

GitHub

Automatização de processos

Automação de documentos

GitHub ▪ Blog ▪ Projeto

Fluxos de trabalho

- Aprendizado de transferência de DL
- Otimização de inferência e ajuste fino de HF
- Compressão distribuída por DL
- Fluxo de trabalho de ML clássico distribuído
- Pré-treinamento de DL com aceleradores Intel®
- Análise de gráficos e GNN com DGL e PyG
- Trang/inferência distribuída em Big-DL
- Pré-treinamento e ajuste fino de LLMs no Ray

Ferramentas

- Intel® Distribution do Python
- Intel® Optimized Modin
- Intel® Optimized XGBoost
- Intel® Extension for Scikit-Learn
- Intel® Optimized Tensorflow (ITEX)
- Intel® Optimized PyTorch (padrão e IPEX)
- Intel® Neural Compressor
- SDK e CLI do SigOpt Python
- SDK e CLI do CNVRG Python
- Horovod otimizado pela Intel
- DeepSpeed

Kits de domínio

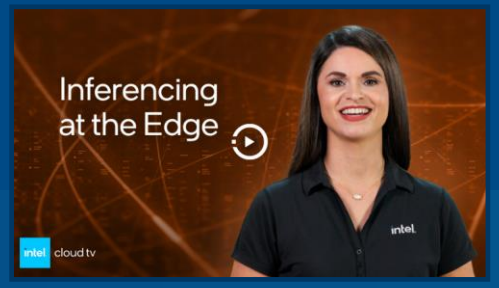
- Séries temporais
- PPML
- Transferência de aprendizagem
- Transformadores/NLP

Os **kits de referência** são entregues como **contêineres** e podem ser utilizados nas **principais nuvens, bem como no local.** Os **kits de referência** são dispostos em camadas em **fluxos de trabalho** e **kits de ferramentas de domínio,** que podem ser aproveitados de forma independente para compatibilidade com uma **variedade mais ampla de casos de uso em vários setores**.

Cloud TV

A Intel® Cloud TV explora notícias, tendências e estratégias de computação em nuvem para impulsionar o seu sucesso

Sua oportunidade de IA generativa com os aceleradores de AI Intel® Gaudi®



Obtenha insights utilizando a inferência de dados na borda

Criando uma vantagem competitiva com a IA na nuvem

Inferência de IA usando Tecnologias de nuvem

IA na nuvem

Entre no caminho rápido para escalar a IA em qualquer lugar

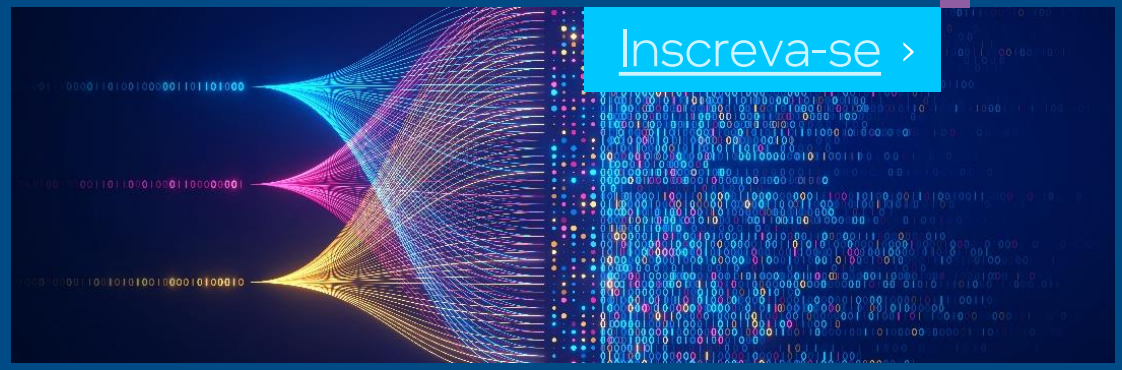
Treinamento

Levando IA a todos os lugares - Casos de uso empresariais de IA generativa

A IA generativa não serve apenas para chatbots online. Uma infinidade de empresas estão considerando maneiras de usar o poder da IA generativa e de grandes modelos de linguagem para ajudar no dia a dia das operações. Esta sessão explora os casos de uso de IA generativa em empresas e fornece considerações sobre como sua organização pode aplicá-la em suas operações diárias.

Inscreva-se ›

Simplificar a IA para geração de dados e grandes modelos de linguagem



Inscreva-se ›

Incorporar a IA às cargas de trabalho de uma organização ou ampliar uma infraestrutura já existente necessita de muita habilidade e computação, exigindo o desenvolvimento de modelos robustos treinados com conjuntos de dados maciços e GPUs poderosas para executá-los de forma adequada. Nem todas as organizações têm os recursos necessários para realizar essa tarefa.

Esta sessão foca em uma solução: uma coleção de kits de referência de IA de código aberto da Accenture* e da Intel® projetada para tornar a IA mais acessível para organizações e otimizada para tempos de treinamento e inferência aprimorados.

Treinamentos adicionais

Técnicos

Tipo de ativo	Título e link
Competência	Competência de IA na nuvem
Webinar	Otimizando a IA para hardware Intel® com a Hugging Face
Webinar	Como configurar o treinamento distribuído baseado em nuvem para ajuste fino de um LLM
Curso de treinamento	Melhorando LLMs com economia imediata e aprendizado em contexto
Curso de treinamento	Simplificar a IA para geração de dados e grandes modelos de linguagem
Curso de treinamento	Processamento de linguagem natural
Curso de treinamento	Aprendizagem profunda aplicada com TensorFlow*
Curso de treinamento	Pequenos e ágeis – o caminho rápido para a IA generativa
Curso de treinamento	A próxima onda de IA generativa - LLMs específicos de domínio
Guia	Um guia do desenvolvedor para começar a usar a IA generativa: uma abordagem específica para casos de uso
Curso de treinamento	Levando a IA nos processadores Intel® Xeon® para o espaço da solução

Treinamentos adicionais

Não técnicos

Tipo de ativo	Título e link
Série de vídeos	Adotando a IA generativa
Curso de treinamento	Pequenos e ágeis – o caminho rápido para a IA generativa
Curso de treinamento	A próxima onda de IA generativa - LLMs específicos de domínio
Curso de treinamento	Competência em Princípios de IA em todos os lugares
Curso de treinamento	Competência em Princípios de software de IA e ecossistemas
Curso de treinamento	Envolvendo o ecossistema de IA: ganhe com software, escale com SIs e venda a solução
Curso de treinamento	IA generativa e grandes modelos de linguagem para o mundo real

Recursos adicionais

Tipo de ativo	Título e link
Webinar	Série de webinars sobre IA generativa
Webinar	Levando a IA generativa para todos os lugares
Podcast	Como o Copilot, ChatGPT, Stable Diffusion e a IA generativa mudarão a maneira como desenvolvemos, trabalhamos e vivemos
Resumo comercial	Implante IA em qualquer lugar
Série de blogs	Ajuste e inferência para IA generativa com processadores Intel® Xeon® da 4ª Geração
Resumo da solução	Implante e escale a inferência de IA generativa com o Lenovo ThinkSystem SR650 V3/processadores Intel® Xeon® da 4ª Geração Novas tecnologias Intel e VMware impulsionam os sistemas Lenovo ThinkAgile VX V3
Artigo técnico	Acelere o Llama 2 com as otimizações de hardware e software da Intel® AI
Pesquisar comunicados à imprensa	Dez por cento das organizações entrevistadas lançaram soluções de IA generativa em produção em 2023
Vídeo de bate-papo Fireside	Enfrentando os desafios de computação e sustentabilidade da IA generativa
Podcast	Hugging Face e Intel — Seguindo em direção a soluções de IA práticas, rápidas, democratizadas e éticas
Conversação do Twitter/X	Como grandes modelos de linguagem democratizados impulsionam o desenvolvimento da IA
Parâmetros de referência da Supermicro	Validação de alegações da Habana
Parâmetros de referência da Hugging Face	Parâmetros de referência
Treinamento/webinar	Palestra de tecnologia do arquiteto de soluções em nuvem (CSA): IA com Habana
Relatório técnico	A IA empresarial é o tema da publicação técnica do desenvolvedor
Infográfico	As CPUs são fundamentais para a IA empresarial

Recursos adicionais

Tipo de ativo	Título e link
Resumo da solução	Simplifique a adoção e a implantação de IA usando a IA empresarial Intel com a IA Red Hat® OpenShift®
Guia	O guia de IA
Kit de referência	Geração de dados de texto não estruturados de IA
Publicação técnica	A Zoho está otimizando e acelerando as cargas de trabalho de IA de vídeo
Publicação técnica	A Seekr desenvolve sistema de triagem de IA confiável
Resumo da solução	Segurança na educação: a IA e a computação confidencial ajudam a tornar os exames remotos seguros uma realidade
Estudo de caso e vídeo	A Nature Fresh Farms utiliza IA do plantio ao mercado
Estudo de caso	A QMed Asia impulsiona a taxa de detecção de câncer em estágio inicial
Estudo de caso e vídeo	A MetaApp renova o sistema de recomendação baseado em IA
Resumo da solução	Otimizando o treinamento e o refinamento de modelos de IA para inspeção óptica automatizada (AOI)
Blog	Eficiências orientada por prompt para LLMs

Avisos legais e isenções de responsabilidade

Avisos legais e isenções de responsabilidade.


intel