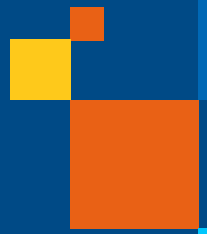
엔터프라이즈 AI

생성형 AI & 엔터프라이즈를 위한 분야별 모델

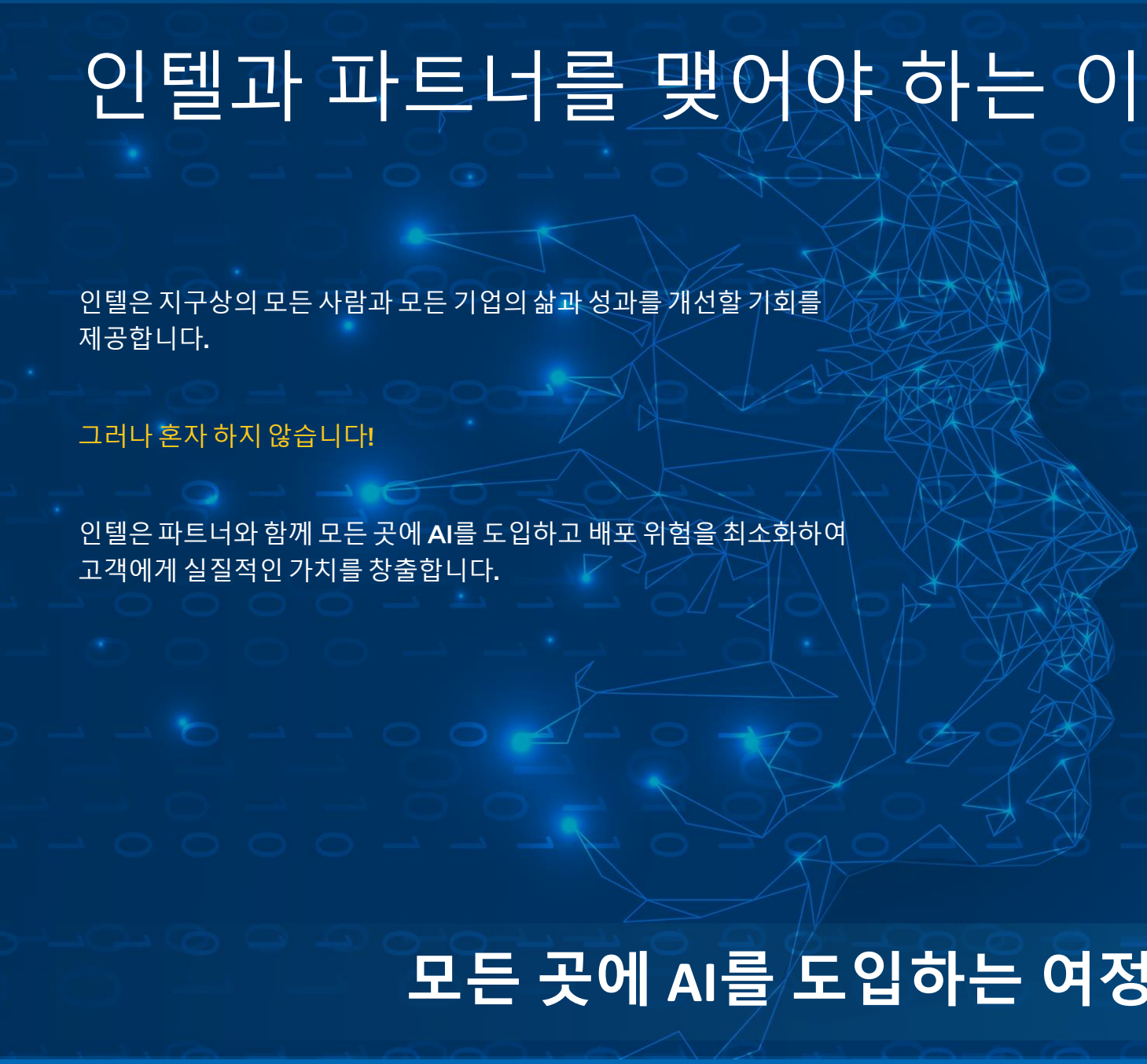
특수 설계된 인텔® AI 하드웨어 및 소프트웨어로 훈련 및 배포를 최적화하여 비즈니스 혁신을 지원하십시오.



intel®

목차

인텔과 파트너를 맺어야 하는 이유



인텔은 지구상의 모든 사람과 모든 기업의 삶과 성과를 개선할 기회를 제공합니다.

그러나 혼자 하지 않습니다!

인텔은 파트너와 함께 모든 곳에 AI를 도입하고 배포 위험을 최소화하여 고객에게 실질적인 가치를 창출합니다.

인텔과 파트너 관계를 맺으면 완벽한 AI 생태계와 파트너 관계를 맺는 것입니다.

인텔의 광범위한 AI 지원 기술 포트폴리오와 하드웨어, 소프트웨어, 시스템 통합업체와의 파트너십으로 긴밀히 협력하며 산업과 기업, 커뮤니티에 차별화된 비즈니스 성과를 제공하는 실제 솔루션을 구축합니다.

여러분의 비즈니스 성장을 지원합니다.

모든 곳에 AI를 도입하는 여정에 참여하십시오.

인텔® AI 솔루션으로 고객 가치 창출

인텔의 접근 방식은 AI 분야의 다양한 참여자가 기업별 GenAI 요구를 충족하는 솔루션을 제공할 수 있는 광범위한 개방형 생태계를 지원합니다.

NAVER

클라우드에서 온디바이스까지 전 세계적으로 고급 AI 서비스를 배포하기 위해 강력한 거대 언어 모델(LLM)을 개발하였습니다. 네이버는 뛰어난 와트당 성능으로 대규모 트랜스포머 모델을 위한 컴퓨팅 작업을 실행하는 인텔® Gaudi®의 기본 기능을 확인했습니다.

seekr

신뢰할 수 있는 AI의 리더로서, LLM 개발 및 운영 배포 지원을 위해 인텔® Tiber™ AI 클라우드의 인텔® Gaudi® 2, 인텔® Data Center GPU Max Series 및 인텔® 제온® 프로세서에서 운영 워크로드를 실행합니다.

BOSCH

강력하고 균등하게 분산된 훈련 세트(예: 자동 광학 검사)를 제공하기 위해 제조 이상 현상에 대한 합성 데이터 세트를 생성하는 기본 모델을 포함하여 스마트 제조를 위한 추가 기회를 탐색합니다.

IFF

식품, 음료, 향기 및 생명과학 분야의 글로벌 리더로서, GenAI 및 디지털 트윈 기술을 활용하여 고급 효소 설계 및 발효 프로세스 최적화를 위한 통합 디지털 생물학 워크플로를 구축합니다.

IBM

watsonx.data™ 데이터 스토어에 5세대 인텔® 제온® 프로세서를 사용하고 있으며, 인텔®과 긴밀히 협력하여 인텔® Gaudi® 가속기를 위한 watsonx™ 플랫폼을 검증하고 있습니다.

airtel

최첨단 인텔 기술의 성능을 채택한 Airtel은 풍부한 통신사 데이터를 활용하여 AI 기능을 강화하고 고객의 경험을 향상합니다. 이번 배치는 기술 혁신의 최전선을 유지하고 빠르게 진화하는 디지털 환경에서 새로운 수익원을 창출하겠다는 Airtel의 약속에 따른 것입니다.

OLA

10개 언어의 생성형 기능을 갖춘 인도 최초의 기초 모델을 사전 훈련하고 미세 조정하여 시장 솔루션 대비 업계 최고의 가격/성능을 달성하였습니다. Krutrim은 현재 인텔® Gaudi® 2 클러스터에서 더 큰 기초 모델을 사전 훈련하고 있습니다.

Infosys

차세대 디지털 서비스 및 컨설팅 분야의 글로벌 리더로서, 4세대 및 5세대 인텔® 제온® 프로세서, 인텔® Gaudi® 2 AI 가속기, 인텔® 코어™ Ultra 프로세서를 포함한 인텔® 기술을 Infosys Topaz에 도입하겠다는 전략적 협력을 발표했습니다. Infosys Topaz는 생성형 AI 기술을 사용하여 비즈니스 가치를 가속하는 AI 우선 서비스, 솔루션 및 플랫폼 세트입니다.

엔터프라이즈 AI용 개방형 플랫폼 개발을 위한 생태계 경주

엔터프라이즈 AI 가치 제안

엔터프라이즈 AI를 통한 비즈니스 혁신

현대의 고도로 경쟁적인 환경에서는 **AI를 도입하는 기업이 앞서 나갑니다.**

다양한 산업의 기업이 AI로 워크플로를 보강하거나 심지어 자동화하는 방법을 이해하기 위해 운영의 모든 측면을 재구상하고 있습니다.

인텔은 기업 조직에 AI를 내장하는 독보적인 전문성을 갖추고 있습니다.

생산성을 혁신하는 AI PC에서 가장 큰 가치를 창출하는 사용 사례를 이해하는 수년간의 전문성까지, 인텔®은 모든 곳에 안전하고 책임감 있게 AI를 적용하는 신뢰할 수 있는 파트너입니다.

모든 기업이 규모와 상관없이 인터넷 시대, 모바일 시대 또는 클라우드 시대보다 더 빠르게 생성형 AI(GenAI) 혁신을 채택할 것으로 예상됩니다.

차세대 AI 플랫폼은 경제적이고 유연한 방식으로 이러한 흥미로운 현실을 수용할 것입니다.

이제 엔터프라이즈 AI에 대해 다르게 생각해야 할 때입니다.

이 지원 패키지로 다양한 시장의 기업이 장기적인 성공을 위해 생성형 AI, 특히 분야별 모델에서 엄청난 가치를 얻는 방법을 이해할 수 있을 것입니다.

생성형 AI 및 거대 언어 모델이란?

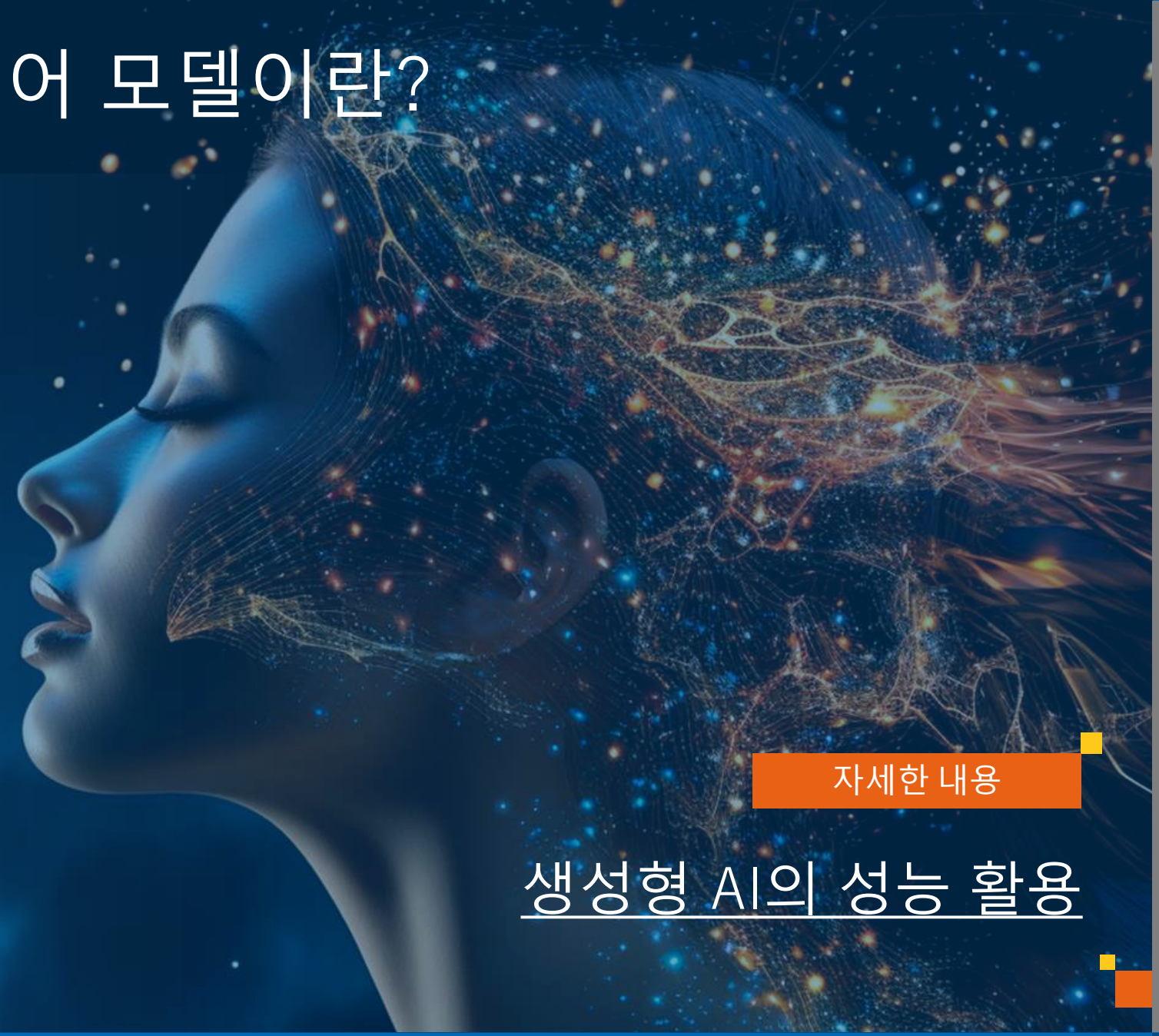
생성형 AI(GenAI)는 새롭고 독창적인 콘텐츠를 만드는 데 초점을 맞춘 AI의 하위 세트입니다.

여기에는 훈련 데이터 세트의 예제와 매우 유사한 이미지, 텍스트, 오디오 등의 데이터를 생성하는 AI 모델의 훈련 및 배포가 포함됩니다.

GenAI 알고리즘은 딥 러닝 및 신경망과 같은 고급 기술을 사용하여 이미지 합성, 텍스트 생성, 심지어 창의적인 예술 작품과 같은 응용을 가능하게 하는 현실적이고 일관된 출력을 생성합니다.

거대 언어 모델(LLM)은 특정 유형의 자연어 처리 모델로, 심층 신경망을 사용하여 텍스트를 처리하고 생성합니다. LLM은 방대한 양의 텍스트 데이터로 훈련하며, 일관되고 의미 있는 결과를 도출하도록 설계됩니다.

자세한 정보



자세한 내용

생성형 AI의 성능 활용

엔터프라이즈는 GenAI를 어떻게 사용할까요?

소비재 및 소매

- 가상 피팅룸
- 배송 및 설치
- 매장 내 제품 검색 지원
- 수요 예측 및 재고 계획 수립
- 참신한 제품 디자인

건강 관리 및 의학

- 바쁜 일선 직원 지원
- 의료 기록을 전사 및 요약
- 의학적 질문에 답하는 챗봇
- 예측 분석을 통한 진단 및 치료 정보 제공

제조

- 기술자를 위한 전문 코파일럿
- 기계와의 대화 상호작용
- 규범적 및 선제적 현장 서비스
- 자연어 문제 해결
- 보증 상태 및 문서화
- 프로세스 병목 지점의 파악, 복구 전략 고안

미디어 & 엔터테인먼트

- 지능적인 검색, 맞춤형 콘텐츠 검색
- 헤드라인 및 카피 개발
- 콘텐츠 품질에 대한 실시간 피드백
- 맞춤형 재생목록, 뉴스 요약, 추천
- 시청자가 선택하는 대화형 스토리텔링
- 맞춤 제안, 구독 요금제

금융 서비스

- 거래 신호 파악, 트레이더에게 취약한 포지션 경고
- 빠른 증권/보험 인수 결정
- 레거시 시스템의 최적화 및 재구축
- 은행 및 보험 모델 역설계
- 잠재적 금융 범죄 및 사기 모니터링
- 규정 준수를 위한 데이터 수집 자동화
- 기업 공시로부터 인사이트 추출

출처: MIT Technology Review Insights에서 편집, "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," 및 "Large Language Models in Media",
12 Databricks, 2023년 4월~6월.

생성형 AI 및 거대 언어 모델의 사용 사례

챗봇 및 가상 도우미

고객 지원

코드 생성 및 LLM 디버깅

회사 문서 기반 훈련

감정 분석

고객 만족도 평가

텍스트 분류 및 클러스터링

방대한 데이터를 분류하여 동향 파악

언어 번역

회사 웹페이지를 다른 언어로 전환

요약 및 의역

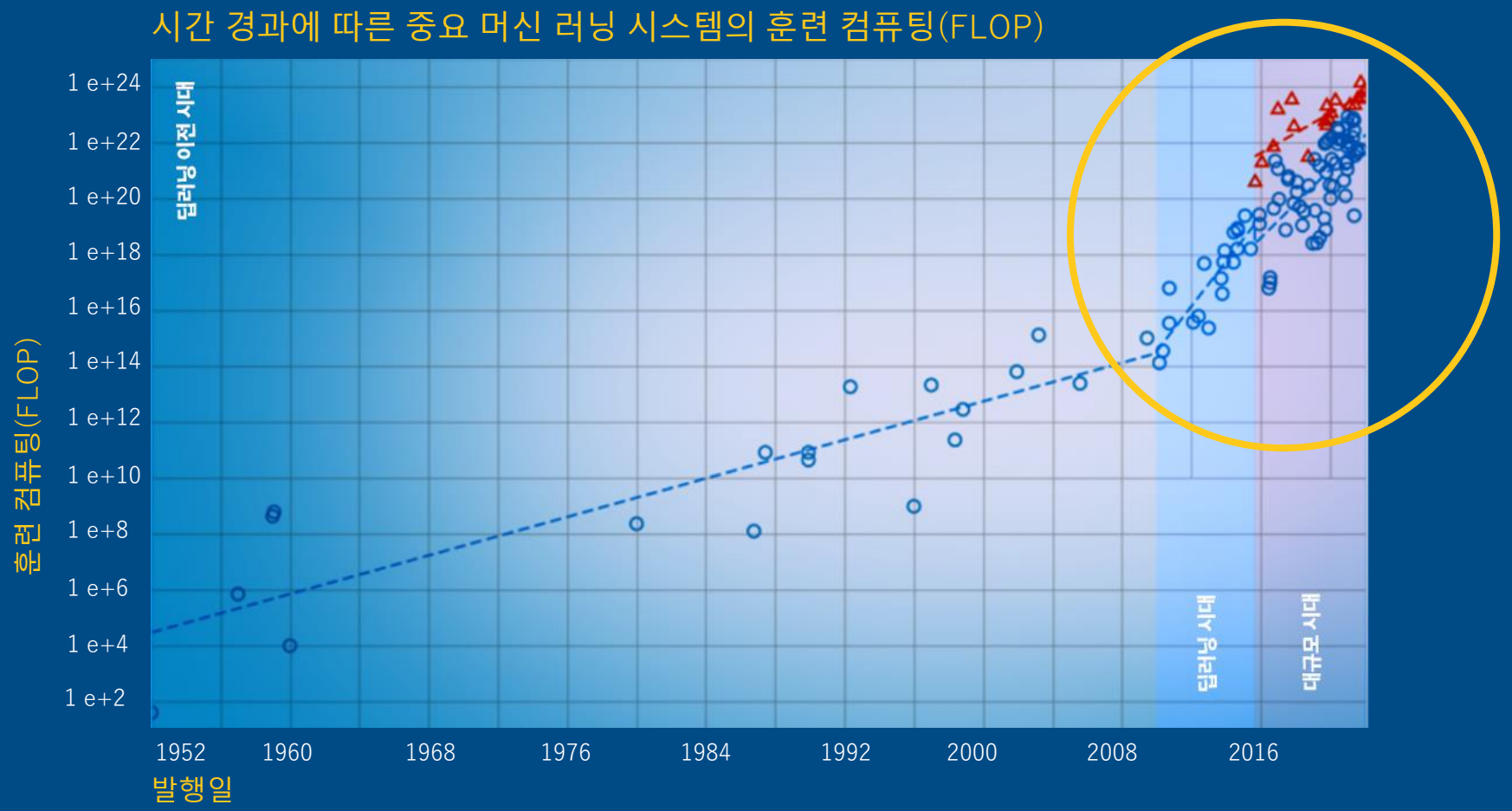
회의록 요약

콘텐츠, 이미지, 비디오 생성

이메일 초고, 아이디어 생성, 마케팅 시각 자료, 짧은 비디오

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

모델의 크기가 커지면 컴퓨팅도 커집니다



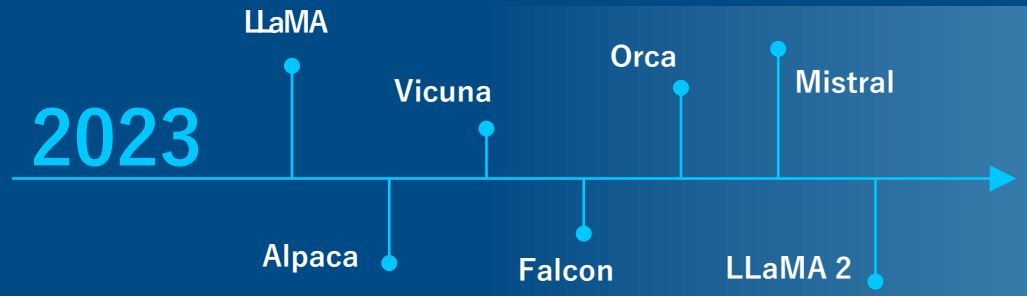
Epoch 연구, University of Aberdeen, Center for the Governance of AI, University of St. Andrews, MIT, Eberhard Karls Universitat Tubingen, Universidad Complutense

거대 모델 뿐만이 아닙니다

	거대 모델(타사)	VS.	작고 민첩한 모델(10~100배)
설명 가능성	독점 모델	VS.	오픈 소스 기반 모델
정확성	올인원 범용	VS.	타겟팅, 분야별, 맞춤형
위치	클라우드 기반(서비스형)	VS.	로컬 추론 실행. 에지, 클라이언트 및 온프레미스
비용	영구적인 비용 증가	VS.	비용 관리
시장 출시 속도	빠른 설치(초)	VS.	구축 시간(시간/일)

여러 소형 모델의 성장

6개월 이내에 수천억에서 200억 개 미만 파라미터로



databricks

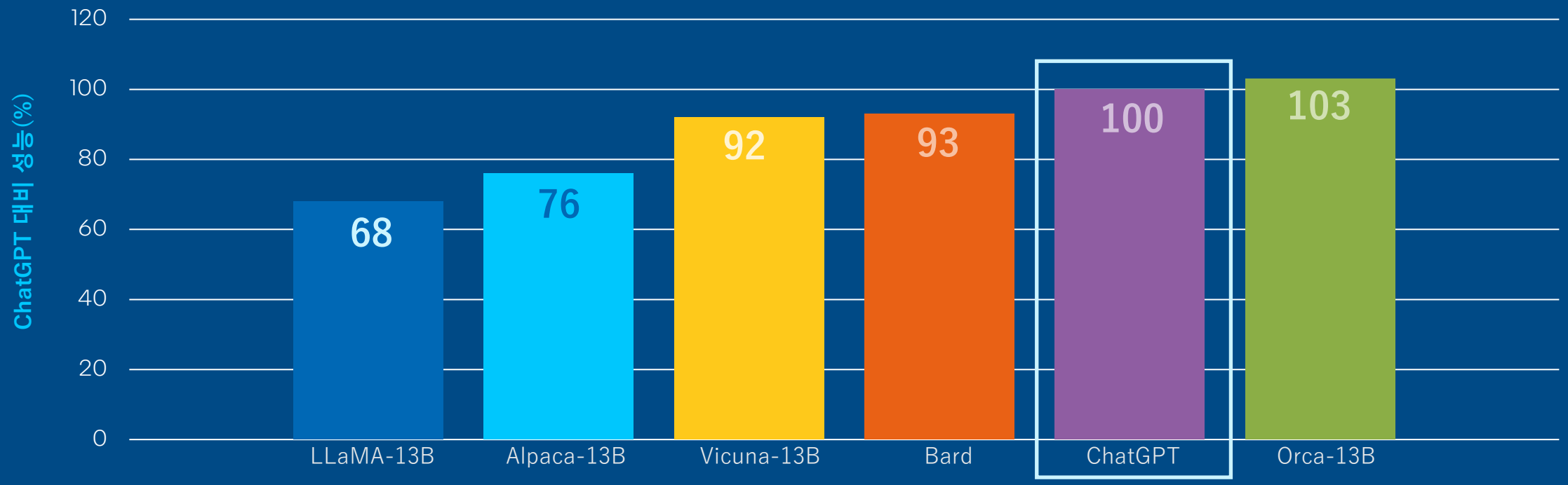
- 매주 수십 개의 소형 모델 등장
- 상용 및 오픈 소스 라이선스
- 세밀하게 준비된 데이터로 훈련받는 경우 소형 모델이 더 큰 모델의 정확도를 재현할 수 있음을 나타냄

- 수천 개의 분야별 상용 모델 및 AI 플랫폼이 입증되고 있음
- 몇 가지 프로세서에서 분야별 데이터로 모델을 미세 조정할 수 있음

ChatGPT 대비 우수한 성능을 보인 소형 모델

더 작은 모델이 실행 가능한 옵션이며 ChatGPT와 같은 대규모 모델과 비교하여 여전히 우수한 성능을 발휘한다는 증거

GPT-4를 사용한 평가



Vicuna 평가 세트에서 GPT-4로 평가한 경우와 같이, Orca는 OpenAI ChatGPT를 포함하여 광범위한 기본 모델보다 뛰어난 성능을 발휘합니다.

출처: Microsoft Research(2023). Orca: Progressive Learning from Complex Explanation Traces of GPT-4

분야별 모델 구축

1

타사 플랫폼

API

거대 기본 모델

엔터프라이즈 데이터베이스

엔터프라이즈 벡터 데이터베이스

미세 조정

검색 증강 생성

2

엔터프라이즈 플랫폼

일반 개방형 모델

엔터프라이즈 데이터베이스

엔터프라이즈 벡터 데이터베이스

미세 조정

검색 증강 생성

분야별 모델

금융 챗봇

개인화된 마케팅 콘텐츠

코드 생성기

임상 기록 전사

텍스트 기반 음성 생성기

분야별 모델은 기업에 다양한 이점을 제공합니다

특정 대상을 목표로 하는 소형 모델은 동등하거나 우수한 성능을 제공하여 시간과 비용 투자를 줄여 ROI를 높일 수 있습니다.

더 정확한 결과

엔터프라이즈 데이터를 사용하여 분야별 정확도 향상

비용 절감

사전 학습된 모델 미세 조정 및/또는 RAG 사용 및 더 작은 모델 추론

원하는 플랫폼 어디에나 배포

로컬 추론 실행. 에지, 클라이언트 및 온프레미스

보안 & 개인정보 보호

데이터 보안 및 규제 요건 충족

책임 있는 AI

미세 조정 및 RAG를 통해 모델에 데이터 소스를 인용하는 기능 제공

미래

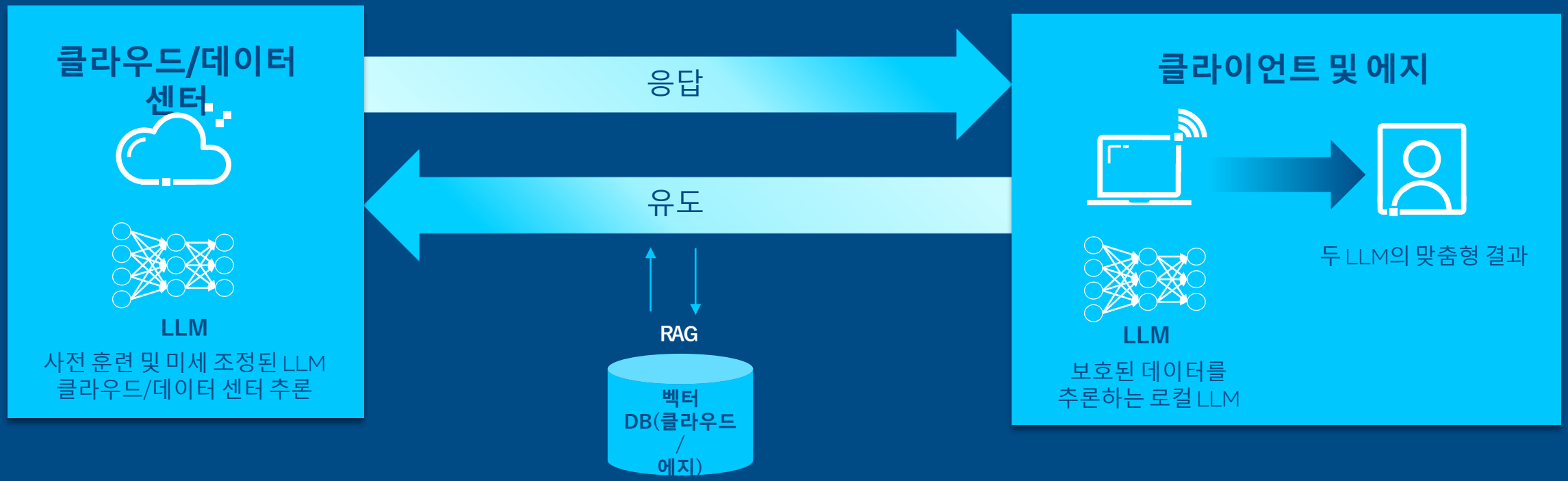
앞으로는 소수의 거대 모델만 남고, 셀 수 없이 많은 작고 더 민첩한 AI 모델이 수많은 응용 프로그램에 내장될 것입니다.[1]

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

원활한 클라우드-에지 AI 플랫폼

Intel AI Portfolio

클라우드에서의 훈련 및 추론, RAG를 사용한 분야 정확도 향상



intel GAUDI intel XEON intel XEON intel XEON intel CORE ULTRA

생성형 AI - 1년간의 운영

분야 특화되어 있으면서도, 고도로 지능적인 모델의 사용이 증가하고 있습니다

2022 실험 → **2023** 파일럿 → **2024** 운영

거대 모델이 길을 개척

- 일반적 목적에 매우 효과적
- 높은 훈련 및 배포 비용
- 대규모 공용 데이터 세트를 기반으로 구축
- 사용 편이성

분야별 소형 모델

- 비즈니스별 결과에 개인 데이터 사용
- 보유한 하드웨어에 배포
- 효율성, 정확도, 보안 및 추적 가능성 향상
- 구축 시간

블로그 읽기

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

분야별 모델에 대한 인텔의 접근 방식

분야별 모델

장점
+ 정확도를 유지/개선하면서도 10~100배 더 작은 모델
+ 범용 컴퓨팅에서 경제적
+ 정확성, 소스 속성, 설명 가능성
+ 프라이빗/엔터프라이즈 데이터 활용
+ 지속적으로 업데이트되는 정보

과제
- 작업 범위 축소
- 몇 번의 미세 조정 및 인덱싱 필요

인텔의 목표

업계 프레임워크, 사전 훈련된 모델, 인텔 AI 소프트웨어 및 도구를 사용하여 인텔 하드웨어에서 수만 개 모델을 미세 조정 및 배포하는 가장 비용 효율적이고 보편화된 접근 방식을 구현합니다.

자세한 내용

내 손 안의 생성형

E-book • 인포그래픽

엔터프라이즈 AI: 진입 장벽 극복 지원

요구사항	인텔®과의 파트너십이 유리한 이유
시장 출시 속도	인텔 및 Hugging Face의 개발자 리소스와 Gaudi 개발자 허브 및 5가지 참조 키트를 사용하여 유리한 위치에서 생성형 AI를 시작할 수 있습니다.
사용자 경험(정확도/대기 시간)	인텔® Gaudi® 가속기에서 파라미터가 100억 개 이상인 모델 또는 인텔® AMX를 탑재한 인텔® 제온® 프로세서에서 파라미터가 200억 개 미만인 소형 모델의 추론을 수행할 때 사용자에게 실시간 경험을 제공합니다.[1]
컴퓨팅 가용성	인텔® 제온® CPU + 가속기는 글로벌 GPU 부족 현상에 비용 효율적인 대안을 제공합니다. **인텔® Gaudi® 2는 이제 SuperMicro를 통해 사용할 수 있으며, 인텔® Gaudi® 3의 가용성은 더욱 늘어날 것입니다.**
익숙한 기술	이미 컴퓨팅 설정에 포함되어 있을 수 있는 유비쿼터스 솔루션을 포함하여 모든 하드웨어에서 소형 모델의 추론을 실제로 수행할 수 있습니다.[2]
규모에 맞는 운영	인텔® Gaudi® 2는 모든 가속기에 24개 100GbE 포트를 통합하여 거의 선형에 가까운 확장성을 제공합니다. 인텔® 제온®은 이미 데이터 센터와 클라우드에서 에지까지 현장에 사용됩니다. **데이터 센터 추론의 65%가 인텔® 제온®에서 실행되고 있습니다.[3]**
비용 효율성	인텔®은 실제 작업 환경에 더 나은 성능과 더 저렴한 가격, 그리고 AI 추론을 위한 보다 균형 잡힌 플랫폼을 제공하여 업계의 혁신을 주도하고 AI를 대중화하고 있습니다. 참고 자료: NVIDIA shows Intel® Gaudi 2 is 4x better performance per dollar than its H100

[1]출처: Four Roadblocks to Implementing Generative AI

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

[3]2022년 12월 기준 AI 추론 워크로드를 실행하는 데이터 센터 서버의 전 세계 설치 기반에 대한 인텔® 시장 모델링을 기반으로 합니다.

생성형 AI 학습 및 배포를 단순화하는 소프트웨어 리소스

오픈 소스 모델

BLOOM (a BigScience initiative) — 176B params, 59 languages, Open-access — 176B

BioGPT — 분야 15억

Stable Diffusion — 이미지

Llama2 GPT-J MPT Falcon — 70~650억 LLM

Stanford Alpaca — 미세 조정된 70억 LLM

지식 기반

개방형 소프트웨어

OpenVINO™

인텔® Extension for PyTorch (IPEX) — PyTorch

인텔® Extension for Transformers (ITREX)

인텔® Extension for DeepSpeed(IDEX) — DeepSpeed

haystack

fastRAG

GenAI 플랫폼

intel. Gaudi

intel. Data Center GPU Max Series

intel. Xeon

intel. Core Ultra

자세한 내용 [유비쿼터스 하드웨어 및 개방형 소프트웨어로 생성형 AI 활용]

가치 극대화

공급업체 종속 방지
오픈 소스 표준 기반 소프트웨어

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

인텔의 하드웨어 포트폴리오 활용
AI 사용 사례에 최적화

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

미래의 AI를 위해 소프트웨어 및 개방형 표준에 의해 최적화된 하드웨어를 통해 클라이언트와 에지에서 데이터 센터 및 클라우드까지 새로운 기회를 창출합니다.

인텔® AI 소프트웨어 포트폴리오

데이터 엔지니어링 → 모델 생성 → 최적화 및 배포

MODIN, SciPy, pandas, NumPy, Spark, Numba

대규모 데이터 분석†

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

머신 및 딥 러닝 프레임워크, 최적화 및 배포 도구†

oneAPI

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

CPU, GPU 및 기타 가속기를 위한 개방형 교차 아키텍처 프로그래밍 모델

클라우드 및 엔터프라이즈

클라이언트 및 워크스테이션

에지

intel AI ANALYTICS TOOLKIT

엔드투엔드 데이터 과학 및 AI 가속

인텔® Tiber™ AI 클라우드 및 인텔® Developer Catalog

최신 인텔 도구 및 하드웨어를 사용해 보고 최적화된 AI 모델에 액세스

cnvrg.io

전체 스택 ML 운영 체제

인텔® Geti

주석/훈련/최적화 플랫폼

Hugging Face

인텔 최적화 및 레시피 미세 조정, 최적화된 추론 모델 및 모델 서빙

참고: 스택의 각 계층의 구성 요소는 예상 AI 사용 모델에 따라 다른 계층의 대상 구성 요소에 최적화되었으며, 모든 구성 요소가 맨 오른쪽 열의 솔루션에서 사용되는 것은 아닙니다.
† 이 목록에는 인텔 하드웨어에 최적화된 인기 있는 오픈 소스 프레임워크가 포함되어 있습니다.

엔터프라이즈 생성 AI 도입을 단순화하고 신뢰할 수 있는 강화 솔루션의 운영까지 시간을 단축합니다.

OPEA:

엔터프라이즈 생성 AI 도입을 단순화하고
신뢰할 수 있는 강화 솔루션의 운영까지
시간을 단축합니다.

OPEA 파트너

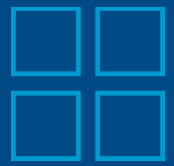
anyscale
CLOUDERA
DATASTAX
kx
DOMINO
Hugging Face
Red Hat
drant
MariaDB
MINIO
sas
Yellowbrick
vmware

OPEA 값

- 엔터프라이즈가 생성형 AI(LLM, RAG)를 더 빠르고 쉽게 사용하여 데이터에서 가치를 창출하도록 지원합니다
- 파편화된 생태계의 복잡성을 줄이고 솔루션의 운영 확장을 지원합니다.
- Linux Foundation과 파트너십을 맺고 업계 리더 간의 협업과 기여를 촉발합니다.

효율성
기존 인프라, AI 가속기 또는 선택한 다른 하드웨어를 활용합니다.

원활함
시스템 및 네트워크 전반에서 다양한 지원과 안정성을 갖춘 엔터프라이즈 소프트웨어와 통합합니다.



개방성
최고의 혁신을 결합하고 독점 공급업체 종속이 없습니다.

유비쿼터스
클라우드, 데이터 센터, 에지 및 PC를 위해 설계된 유연한 아키텍처를 통해 모든 곳에서 실행됩니다.

신뢰할 수 있음
책임감, 투명성, 추적성을 위한 안전한 엔터프라이즈 파이프라인과 도구를 갖추고 있습니다.

확장 가능
역동적인 파트너 생태계에 액세스하여 솔루션을 구축하고 확장합니다.

생성형 AI를 위한 Hugging Face 파트너십

Hugging Face

인텔은 생성형 AI와 언어 AI 훈련 및 혁신을 촉진하기 위해 AI 모델과 데이터 세트를 공유하는 데 널리 사용되는 플랫폼인 Hugging Face와 협력했습니다. 특히, Hugging Face는 NLP용으로 구축된 트랜스포머 라이브러리로 유명합니다.

intel XEON

인텔®은 Hugging Face와 협력하여 트랜스포머 모델을 훈련, 미세 조정 및 예측할 수 있는 최첨단 하드웨어 및 소프트웨어 가속을 구축해 오고 있습니다.

하드웨어 가속은 인텔® 제온® 스케일러블 프로세서가 주도하며, 소프트웨어 가속은 최적화된 AI 소프트웨어 도구, 프레임워크, 라이브러리 포트폴리오를 통해 가능합니다.

intel GAUDI

또한, 인텔® Gaudi® 딥 러닝 가속기는 Optimum Habana Library를 통해 Hugging Face 오픈 소스 소프트웨어와 결합되므로 개발자가 Hugging Face 커뮤니티에서 최적화된 수천 개 모델을 손쉽게 사용할 수 있습니다.

Hugging Face는 Stable Diffusion, T5-3B, BLOOMZ 176B 및 7B, 새로운 BridgeTower 모델 등 생성형 AI 모델에 대한 인텔® Gaudi® 2 성능의 평가 결과도 발표했습니다.

인텔®, Articul8 및 BCG는 협력을 통해 엔터프라이즈급 안전한 생성형 AI를 제공

인텔 AI 슈퍼컴퓨터 기반의 선구적인 솔루션은 높은 수준의 보안과 데이터 개인정보 보호를 유지하면서 맞춤형 데이터세트로 비즈니스 가치를 실현합니다.

Articul8는 속도, 보안 및 비용 효율성을 제공하는 턴키 GenAI 소프트웨어 플랫폼을 제공하여 대규모 엔터프라이즈 고객이 AI를 조작하고 확장하도록 지원합니다. 이 플랫폼은 인텔® 제온® 스케일러블 프로세서와 인텔® Gaudi® 가속기를 비롯한 인텔® 하드웨어 아키텍처를 기반으로 실행 및 최적화되었지만, 다양한 하이브리드 인프라 대안을 지원할 예정입니다.

intel GAUDI

intel XEON

이 팀은 Boston Consulting Group(BCG)에서 기술을 조기 배포한 후 금융 서비스, 항공 우주, 반도체, 통신을 비롯하여 높은 수준의 보안과 전문 분야 지식이 필요한 산업 부문의 엔터프라이즈 고객으로 플랫폼을 확장해 왔습니다.

자세한 내용

Articul8 발표

Articul8 웹 사이트

엔터프라이즈를 위한 책임 있는 AI

과제:

생성형 AI 모델은 인터넷상에서 이용할 수 있는 방대한 데이터로 학습하며, 따라서 사회에 존재하는 편견을 포함할 수 있고 이러한 편견을 실수로 적용할 수 있습니다. LLM은 조작을 통해 허위 정보, 피싱 이메일, 소셜 엔지니어링 공격을 생성하거나 확산할 수 있습니다.

LLM은 종종 '환각'을 일으키고 부정확한 정보를 생성할 수 있는데, 이는 모델이 진단 및 치료 결정에 영향을 미칠 수 있고, 나아가 환자에게 해를 끼칠 수 있는 의료 서비스와 같은 분야에서 특히 문제가 될 수 있습니다.

자세한 정보

생성형 AI의 위험 최소화

솔루션:

AI 기술로 작업하는 회사와 개인은 윤리적 AI 원칙에 따라 소프트웨어를 개발하고 배포해야 합니다.

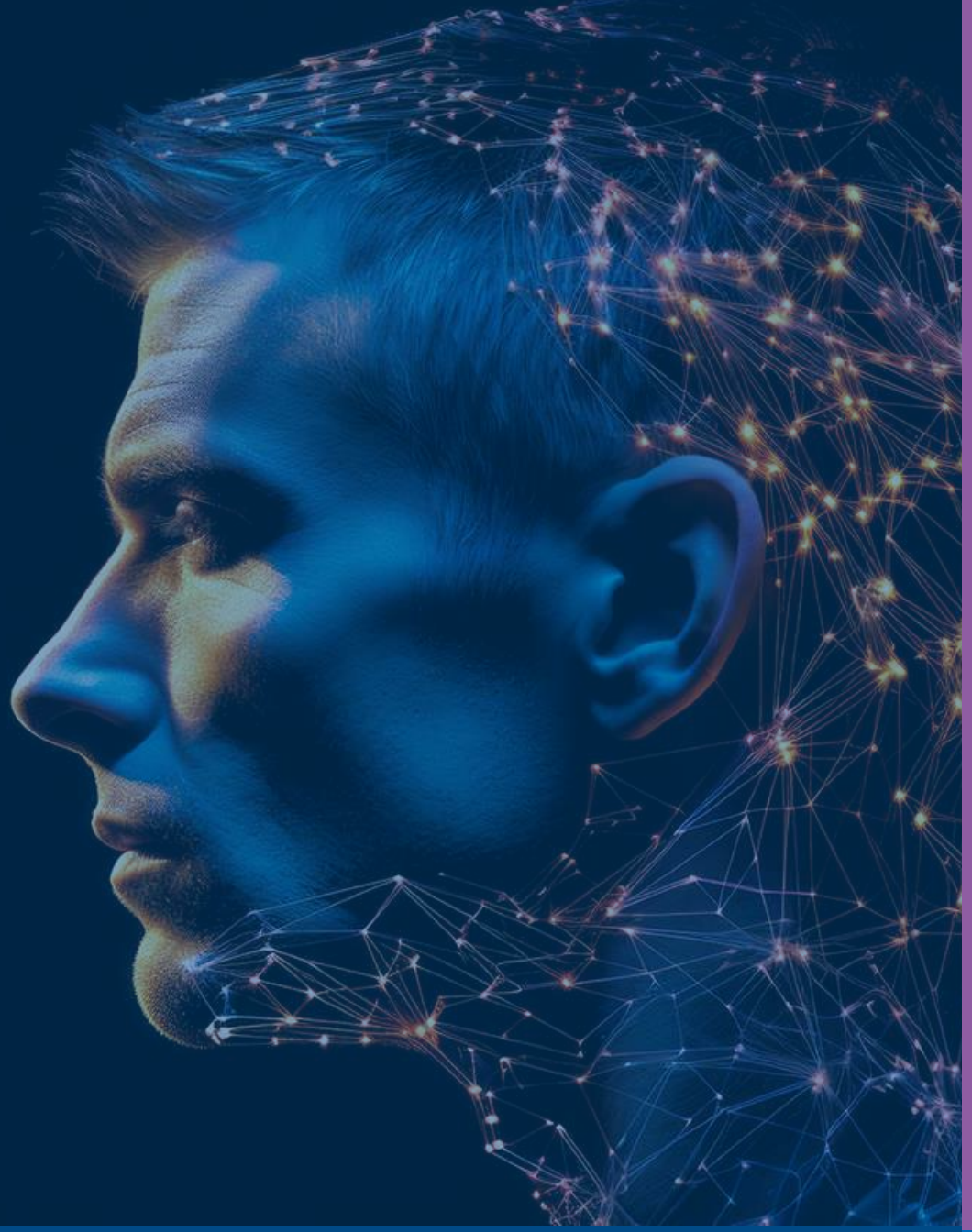
오픈 소스 인텔® Explainable AI Tools를 통해 사용자는 사후 모델 추출 및 시각화를 실행하여 TensorFlow* 및 모델의 예측 행동을 검사할 수 있습니다.

LLM은 일반적으로 대규모 공개 데이터 세트를 사용하여 훈련한 후 잠재적으로 민감할 수 있는 데이터(예: 금융 및 의료)를 사용하여 미세 조정합니다.

인텔의 Open Federated Learning(OpenFL)과 같은 기술은 기밀 컴퓨팅을 통합하므로, LLM을 민감한 데이터로 안전하게 미세 조정하여 환각과 편견을 줄이면서 모델의 일반화 가능성을 향상할 수 있습니다.

생성형 AI를 위한 인텔® 제품

모든 곳에
AI 도입



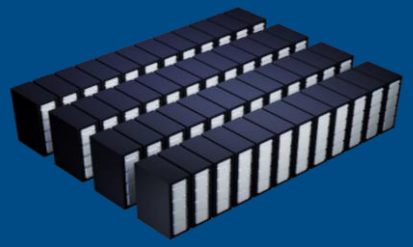
AI를 위한 확장 가능 시스템

훈련 및 미세 조정

훈련 | 최대 추론 | 메인스트림 추론/미세 조정 | 베이스라인 추론 | 엔드포인트 추론

추론 및 배포

클라우드 데이터 센터 | 에지 | 클라이언트

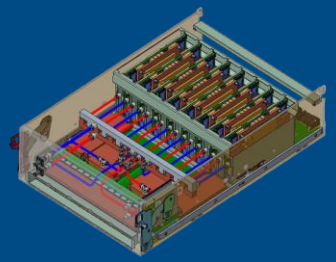


클러스터 및 데이터 센터 규모

intel. GAUDI

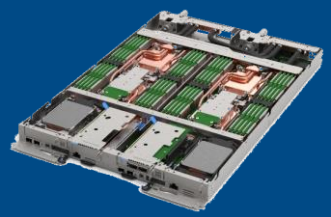
랙당 다중 노드 배포

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



다중 GPU 또는 다중 소켓 CPU

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



단일 소켓 CPU 또는 단일 GPU

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

클라이언트 CPU

intel CORE ULTRA | intel. ARC GRAPHICS

intel. ETHERNET

NLP/LLM용 인텔® 제품

훈련 추론

GAUDI®2

인텔® Gaudi® 2 AI 가속기는 LLM, NLP와 같은 대규모 모델의 훈련 및 추론을 가속하도록 특별히 설계되었습니다.

인텔® Gaudi® 2를 통한 생성형 AI 및 대규모 언어 모델의 가속화

인텔® Gaudi® 2는 AI 훈련을 위한 최고의 성능과 최적의 비용 절감을 제공합니다.

보도 자료

지금 보기

생성형 AI 및 대규모 언어 모델(LLM)의 잠재력을 포착하는 인텔® Gaudi® 2 AI 프로세서의 최첨단 기능에 관해 알아보는 인텔® 웨비나 녹화본

인텔® Gaudi® 2 딥 러닝 가속기는 **NVIDIA A100보다 최대 2.4배 빠른 성능[1]을** 제공하는 등 딥 러닝 훈련 및 추론에서 경쟁력 있는 성능을 발휘합니다.

뉴스룸 ▪ 기술 문서

인텔® Gaudi® 2는 NV H100에 대한 생성형 AI 성능의 유일한 벤치마크 대안입니다.

[1]성능은 사용, 구성 및 기타 요소에 따라 다릅니다. 워크로드 및 구성 세부 정보는 intel.com/performanceindex에서 확인할 수 있습니다. 결과는 다를 수 있습니다.

Gaudi2: 기반 모델의 효율적인 훈련 및 추론에 적합

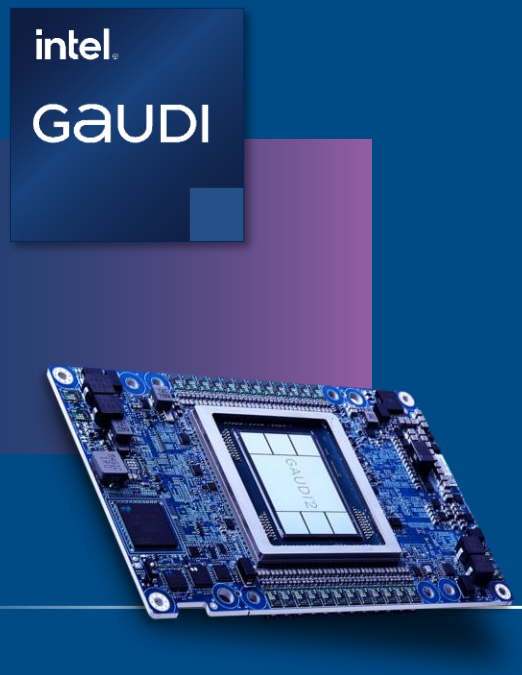
Gaudi2는 LLM(GPT) 및 GAI(Stable Diffusion)와 같은 대규모 기반 모델의 요구 사항을 충족하는 딥 러닝 성능, 효율성 및 확장성을 위해 설계되었습니다

요구사항	Gaudi2
속도	훈련 및 추론 모두에서 A100보다 **1.5~2배 빠름**
메모리	각 Gaudi2 장치는 **96GB 온칩 고대역폭 메모리**를 갖추고 있어 대규모 기반 모델의 메모리에 쉽게 장착하고 대규모로 훈련 및 배포할 수 있습니다.
확장성	**24x100GbE 포트 통합형 온칩**으로 효율성 확장, 서버 내 8개 카드에 모두 직접 연결, 서버 내부와 서버 간의 개방형 ROCEv2 기반 통신
사용 용이성	SynapseAI, PyTorch 및 DeepSpeed를 통한 **최소한의 코드 변경**으로 모델 이전 또는 구축
전원 효율성	**A100 대비 ~1.8배 더 높은 처리량/와트**
비용 효율성	Amazon 클라우드에서 A100보다 **최대 40% 향상된 가격 성능**을 제공하는 특수 목적 1세대 Gaudi 아키텍처 기반

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

수많은 LLM의 미세 조정

Hugging Face 평가를 통해 NVIDIA A100 및 H100 대비 인텔® Gaudi® 2 가속기 LLM 성능이 입증되었습니다.



워크로드 및 구성은 https://habana.ai/habana-claims-validation에서 확인하십시오. 결과는 다를 수 있습니다.
https://huggingface.co/blog/habana-gaudi-2-benchmark
https://huggingface.co/blog/bridgetower

GPT-J: 인텔® Gaudi® 2 결과

GPT-J에 대한 인텔® Gaudi® 2의 추론 성능 결과는 경쟁력 있는 성능을 강력하게 입증합니다.

- 서버 쿼리 및 오프라인 샘플의 경우 GPT-J-99 및 GPT-J-99.9에서 인텔® Gaudi® 2 추론 성능은 각각 **초당 78.58 및 초당 84.08**입니다.[1]
- **인텔® Gaudi® 2는 NVIDIA의 H100과 비교하여 경쟁력 있는 성능을 제공합니다.** H100은 Gaudi 2보다 1.09배(서버) 및 1.28배(오프라인) 높은 성능으로 약간의 우위를 보였습니다.[1]
- **인텔® Gaudi® 2의 성능은 NVIDIA의 A100보다 2.4배(서버) 및 2배(오프라인) 뛰어납니다.[1]**
- 인텔® Gaudi® 2 제출은 FP8을 채택했으며, 이 새로운 데이터 유형에서 **99.9%의 정확도**를 달성했습니다.[1]

자세한 내용

매 6~8주마다 릴리스되는 인텔® Gaudi® 2 소프트웨어 업데이트를 통해 인텔®은 계속해서 성능을 향상하고 MLPerf 벤치마크에서 모델 커버리지를 확장할 수 있을 것으로 기대합니다.

뉴스룸 기사

MLCommons 발표

[1]성능은 사용, 구성 및 기타 요소에 따라 다릅니다. 워크로드 및 구성 세부 정보는 https://mlcommons.org/2023/09/mlperf-results-highlight-growing-importance-of-generative-ai-and-storage/에서 확인할 수 있습니다. 결과는 다를 수 있습니다.

인텔® Gaudi® 2: 벤치마크 결과

Supermicro에서 제공한 벤치마크 결과, 업계 최초의 인텔® Gaudi® 2 OEM

Gaudi 성능 정보

인텔® Gaudi® 2 AI 가속기를 통한 LLM 훈련 및 추론

벤치마크

더 빠른 훈련 및 추론: 인텔® Gaudi® 2 vs. NVIDIA A100 80GB

벤치마크

결과는 다를 수 있습니다.

인텔® Gaudi® 2: 기초 모델 훈련 및 추론

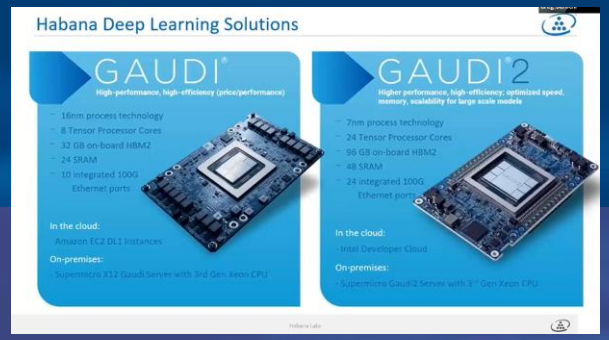
Gaudi 지원 모델 액세스:

개발자 카탈로그

GAUDI®2

인텔® Gaudi® 개발자 교육

[시작하기: Gaudi 기반 딥 러닝 및 추론]



[인텔® Gaudi® 2의 성능 극대화: 생성형 AI 및 거대 언어 모델의 가속화]

[인텔® Gaudi® 프로세서를 통한 모델 성능 극대화: 최적의 결과를 위한 고급 도구 및 전략]

인텔® Gaudi® 소프트웨어 (SynapseAI® 소프트웨어 제품군)

개발 단순화: 원하는 대로 개발하기

목표: 기존 소프트웨어에서 인텔® Gaudi® AI 가속기로의 손쉬운 마이그레이션, 소프트웨어 투자 보전, 용이한 새로운 모델 구축 — 딥 러닝, 생성형 AI, 거대 언어 모델을 정의하는, 수많은 성장 중인 모델의 훈련 및 배포

데이터 과학자, 개발자, IT 및 시스템 관리자에 대한 폭넓은 지원:

- 개발자 사이트
- GitHub

인텔® Gaudi® AI 가속기

딥 러닝을 위한 소프트웨어 생태계는 주요 소프트웨어 공급업체, 도구 및 코드를 한데 모아 PyTorch, TensorFlow, PyTorch Lightning, DeepSpeed 프레임워크를 기반으로 최첨단 딥 러닝 모델의 개발을 가속합니다.

cnvrg.io

PyTorch Lightning

인텔® Gaudi® 소프트웨어를 사용할 준비가 되셨습니까?

인텔® Gaudi® 2 AI 가속기 사용 가능! Denvr Cloud 전용

인텔® Gaudi® 2
소프트웨어 생태계

인텔® Gaudi® 2
AI 가속기(7nm)

인텔® Gaudi® 2 – 생성형 AI 요구 사항에 이상적

- 지금 바로 사용 가능! Denvr Cloud의 Gaudi 2 클러스터
- Gaudi 2 노드 최대 8개까지 드라이브 테스트
- 인텔 고객을 위한 VIP 우선 가격
- Denvr Dataworks 하이터치 상용 서비스 및 지원
- Denvr Cloud의 Gaudi 2 클러스터로 원활하게 마이그레이션
- Denvr Cloud의 Gaudi 3 클러스터에 대한 독점 우선순위 지정 – 곧 출시 예정!

지금 시작하기

공개 예정

intel GAUDI

훈련 추론

인텔® Gaudi® 3

성능, 확장성 및 효율성을 통해 더 많은 고객에게 더 많은 선택권을 제공하는 인텔® Gaudi® 3 가속기는 엔터프라이즈가 인사이트, 혁신 및 수익을 창출하는 데 도움이 됩니다.

출시 예정 - 인텔® Gaudi® 3 AI 가속기

성능, 확장성 및 효율성을 갖춘 GenAI에 대한 선택권 제공

인텔® Gaudi® 3는 GenAI를 대규모로 배포하려는 글로벌 기업에 비약적으로 향상된 AI 훈련 및 추론 성능을 제공할 것입니다.

보도 자료

인텔® Gaudi® 3 가속기 성능 vs. NVIDIA H100

인텔® Gaudi® 3는 70억 및 130억 개 파라미터의 Llama2 모델과 GPT-3 1,750억 개 파라미터 모델에서 평균 **50% 더 빠른 훈련 시간**[3]을 제공할 것으로 예상됩니다.

인텔® Gaudi® 3는 다음과 같이 H100을 능가할 것으로 예상됩니다.

50% (가속기 추론 처리량에서)[1]

40% (추론 전력 효율성에서)[2]

Llama 70억 및 700억 개 파라미터와 Falcon 1,800억 개 파라미터 모델 대상

자세한 내용

백서

인텔® Gaudi® 3는 2024년 2분기부터 OEM에 공급됩니다.

Lenovo DELL Hewlett Packard Enterprise SUPERMICRO

[1]NV H100 비교는 https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8을 기반으로 하며, 보고된 수치는 GPU당입니다. LLAMA2-7B, LLAMA2-70B 및 Falcon 180B 대상 인텔® Gaudi® 3와의 비교 예측. 결과는 다를 수 있습니다.
[2]NV H100 비교는 https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8을 기반으로 하며, 보고된 수치는 GPU당입니다. LLAMA2-7B, LLAMA2-70B 및 Falcon 180B 대상 인텔® Gaudi® 3와의 비교 예측. NVIDIA와 Gaudi 3의 전원 효율성은 내부 추정치에 기반합니다. 결과는 다를 수 있습니다.
[3]NV H100 비교는 https://developer.nvidia.com/deep-learning-performance-training-inference/training, '거대 언어 모델' 탭을 기반으로 합니다. 2024년 3월 28일 기준 LLAMA2-7B, LLAMA2-13B 및 GPT3-175B 대상 인텔® Gaudi® 3와의 비교 예측. 결과는 다를 수 있습니다.

NLP/LLM용 인텔® 제품

추론

4세대 및 5세대 인텔® 제온® 스케일러블 프로세서는 인텔® DL Boost, 인텔® AMX, 인텔® AVX-512를 통해 NLP를 가속합니다. 이 제품은 고성능 컴퓨팅용으로 설계되어 NLP 워크로드를 가속하는 데 사용할 수 있습니다. 다수의 스레드, 대용량 메모리, 높은 메모리 대역폭을 다룰 수 있어 언어 번역, 텍스트 요약, 문자 음성 변환과 같은 NLP 워크로드에 적합합니다.

5세대 인텔® 제온®: AI를 위해 설계된 프로세서

모든 코어에서 AI 가속화를 지원하는 5세대 인텔® 제온® 프로세서는 고객이 개별 가속기를 추가해야 하기 전에 까다로운 엔드투엔드 AI 워크로드를 처리합니다.

AI 추론에서 성능 향상

최대 **42%**

이전 세대 대비[1]

일반적인 컴퓨팅 성능 향상

평균 **21%**

이전 세대 대비[1]

더 빠른 자연어 처리

최대 **23%**

이전 세대 대비[1]

인텔 데이터 센터 및 AI 그룹 수석 부사장 겸 총괄 매니저, Sandra Rivera

"AI용으로 설계된 5세대 인텔® 제온® 프로세서는 클라우드, 네트워크 및 에지 사용 사례 전반에서 AI 기능을 배포하는 고객에게 더 높은 성능을 제공합니다. 고객, 파트너 및 개발자 생태계와 오랫동안 협력한 결과, 인텔은 낮은 TCO로 빠르게 채택하고 확장할 수 있는 검증된 기반에서 5세대 인텔® 제온®을 출시하고 있습니다."

추가 정보:

웹 사이트

제품 요약

워크로드 및 구성의 경우, intel.com/processorclaims 에서 5세대 인텔® 제온 스케일러블® 프로세서를 확인하십시오. 결과는 다를 수 있습니다.

인텔® 제온®: 실전 AI 적용에서의 CPU 성능 리더십

인텔은 실제 실제 업무 상황 적용에서 다음과 같이 더 나은 성능, 낮은 가격, AI 추론을 위한 더 균형 잡힌 플랫폼을 제공하여 업계를 혁신하고 AI를 대중화하고 있습니다.

- 데이터 지역성을 지원하는 더 큰 캐시와 더 큰 문제를 해결할 수 있는 대용량 메모리
- 더 높은 코어 주파수, 다중 스칼라 포트 및 단일 또는 다중 스레드이지만 스칼라인 컴퓨팅을 가속하는 비순차적 실행
- 비DL 벡터 컴퓨팅에 도움이 되는 인텔® Advanced Vector Extensions 512(인텔® AVX-512)
- AI 가속화를 위한 내장형 하드웨어 지원 인텔® Advanced Matrix Extensions(인텔® AMX)

기술 문서 전문

인포그래픽

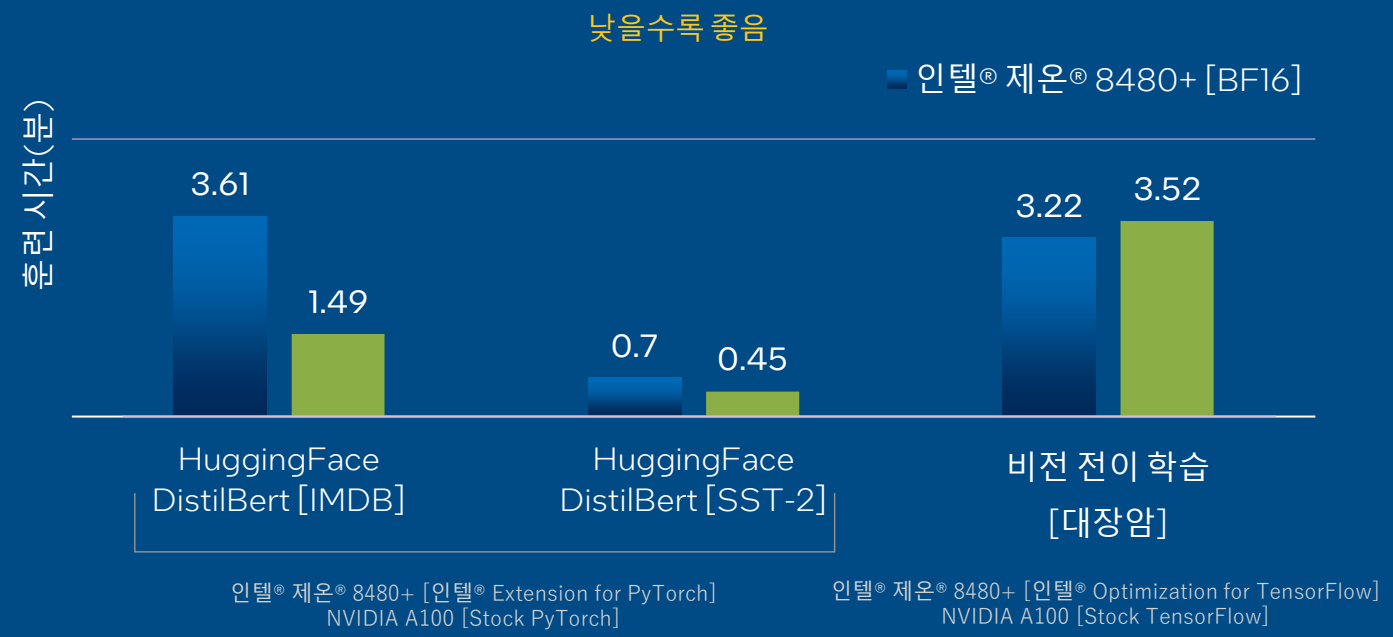
intel XEON

GPU 신화 밝히기: 내장 가속기를 갖춘 CPU로 AI를 혁신하는 방법

4분 이내에 모델을 미세 조정하는 인텔® 제온® 스케일러블 프로세서[1]

Hugging Face

미세 조정 및 훈련 시간 성능, 인텔® 제온® Platinum 8480+ 프로세서 vs. NVIDIA A100 GPU



추가 참고 자료:

더 우수한 성능: 인텔® CPU 기반 Numenta vs. NVIDIA GPU

[1] 4세대 인텔 제온 스케일러블 프로세서의 성능 지수에서 [A221]을 참조하십시오. 결과는 다를 수 있습니다.

4세대 인텔® 제온® 프로세서 기반 LLM

인공 지능(AI) 챗봇 기술은 고객과 상호 작용하고 고객 서비스를 개선하는 방법으로 기업과 조직에서 점점 더 인기를 얻고 있지만, 특정 사용 사례에 맞게 챗봇을 구축, 최적화 및 유지 관리하는 데는 비용이 많이 들어 다수의 조직은 재정적으로 엄두를 내지 못할 수 있습니다.

4세대 인텔® 제온® 프로세서가 **인텔® Advanced Matrix Extensions(AMX)**를 제공하는 향상된 데이터 관리와 효율적인 계산을 인텔® Extension for PyTorch를 통해 사용할 수 있는 **Auto Mixed Precision(AMP)** 기능과 결합하면, 이 기술 스택은 전이 학습과 처음부터 중소형 모델을 훈련하는 것과 같은 워크로드에서 큰 경쟁력을 발휘합니다.

추가 정보:

4세대 인텔® 제온® 스케일러블 프로세서 기반 AI를 위한 튜닝 가이드

가이드 링크 ›

방법을 알려주는 기술 문서

생성형 AI용 5세대 및 4세대 인텔® 제온® 프로세서를 탑재한 Cisco UCS

작을수록 좋습니다, 인텔® 제온® 프로세서에서 효율적인 생성형 AI 경험을 제공하는 Q8-Chat LLM

LLM은 검색이나 대화형 응용 프로그램과 같은 저지연 사용 사례에서 충분히 빠르게 예측하기 위해 일반적으로 고급 GPU에서 발견되는 대량의 컴퓨팅 성능이 필요합니다. 아쉽게도, 다수의 조직은 관련 비용이 터무니없이 높아 응용 프로그램에 최첨단 LLM을 사용하는 것이 어려울 수 있습니다.

Hugging Face

"훈련 및 운영 비용이 저렴한 소규모 특정 모델에 집중하는 것이 더 많은 회사에 유리할 것입니다."

LLM 크기 및 추론 대기 시간을 줄여 인텔® CPU에서 LLM의 효율적인 실행을 지원하는 최적화 기술에 관해 알아보십시오.

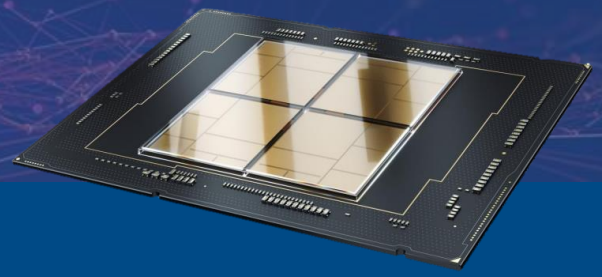
방법을 알려주는 기술 문서 ›

Hugging Face 및 4세대 인텔® 제온® 시작하기

LLM을 위한 인텔® 제온® 프로세서

요약

- 전문 분야 LLM을 추론하기 적합한 포지셔닝
- 전이 학습 사용 사례에서 만족스러운 결과 제공
- 오픈 소스 SW를 갖춘 인텔® 제온®에 LLM을 배포하여 최적의 성능을 쉽게 제공



LLM을 위한 인텔® 제온® 스케일러블 프로세서

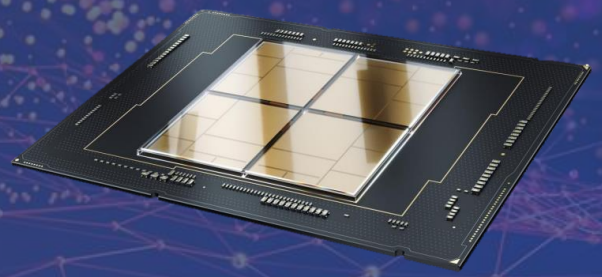
가장 인기 있는 AI 프레임워크 및 라이브러리를 갖추어 범용 AI 워크로드를 구축하고 배포하는 데 이상적입니다.

- 기존 인프라를 활용하여 분야별 LLM 추론
- 전이 학습 사용 사례에서 만족스러운 결과 제공
- 오픈 소스 SW를 갖춘 인텔® 제온®에 LLM을 배포하여 최적의 성능을 쉽게 제공

인텔® 제온®

실제 AI 응용 프로그램에서 CPU 성능 리더쉽 유지

기술 문서 ▪ 인포그래픽



GPT-J

4세대 인텔® 제온® 결과

2 문단/초 - 오프라인 모드[1]

뉴스룸 기사

1 문단/초 - 실제 서버 모드[1]

MLCommons 발표

GPU 신화 밝히기: 내장 가속기를 갖춘 CPU로 AI를 혁신하는 방법

인텔® AMX를 탑재한 4세대 인텔® 제온®의 Alibaba NLP 사례 연구

자세한 내용

[1]성능은 사용, 구성 및 기타 요소에 따라 다릅니다. 워크로드 및 구성 세부 정보는 https://mlcommons.org/2023/09/mlperf-results-highlight-growing-importance-of-generative-ai-and-storage/에서 확인할 수 있습니다. 결과는 다를 수 있습니다.

NLP/LLM용 인텔® 제품

클라이언트에서 소규모 추론 수행

AI PC 시대를 열어가는 인텔® 코어™ Ultra

3D 성능 하이브리드 아키텍처, 고급 AI 기능을 특징으로 하며 내장 인텔® Arc™ GPU와 함께 사용할 수 있는 인텔® 코어™ Ultra 프로세서는 얇고 강력한 프리미엄 노트북에 최적화되어 있습니다. 새로운 인텔® 4 프로세스를 사용하여 제작된 인텔® 코어™ Ultra 프로세서는 이동 중 게이밍, 콘텐츠 제작 및 생산성을 위해 성능과 전원 효율성 사이에서 최적의 균형을 제공합니다.

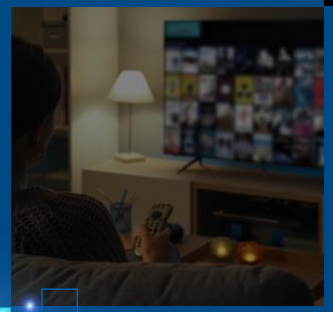
사용 사례: PC에서의 AI

크리에이터: 사진 및 비디오 검색 및 편집

더 빠르고 자연스러운 필터, 더 높은 품질의 미리보기, 더 빠른 자동 내보내기, 더 빠른 검색

협업/스트리밍

차세대 비디오 회의, 스트리밍 및 협업을 위한 새로운 AI 기능으로 배터리 수명 보존



메인스트림 게임

인게임, 더한 현실감을 위한 3D 애니메이션, 전사 및 채팅 번역을 위한 새로운 AI 기능

PC에서의 AI

"평범함에 도전하다"

생산성

글쓰기, 제작, 코딩 및 오프라인 기능을 위한 AI 지원, 텍스트 및 문법 예측 등

접근성

다양한 사용자 요구 사항을 충족하는 AI 지원 오디오-비주얼 기능으로 PC에서 더 쉽게 제작하고 생산성을 높일 수 있습니다.

크리에이터: 텍스트를 이미지로

마케팅, 광고, 디자인을 위해 몇 가지 수식어로 이미지를 만들 수 있는 새로운 AI 효과 및 기능

*인텔 2023년 12월 14일, 모든 분야의 AI 이벤트 보도 자료

생성형 AI를 위한 인텔® 코어™ Ultra

인텔의 가장 전력 효율적인 클라이언트 프로세서가 AI PC 시대를 선도합니다

효율성과 성능의 주요 개선 사항

AI 효율성
최대 **70%**
더 빠른 생성형 AI 성능[2]

절전
최대 **25%**
전력 소비 감소[3]

자세한 내용 발표 ▪ 제품 요약 ▪ 웹 사이트

인텔® 코어™ Ultra는 인텔 최초의 클라이언트 온칩 AI 가속기인 신경망 처리 장치(NPU)를 갖추어 새로운 수준의 전원 효율적인 AI 가속을 실현하며, 이전 세대보다 **2.5배 더 높은 전원 효율성**을 제공합니다[1]

인텔® 코어™ Ultra H 및 U 세대 칩에는 저강도 워크로드를 위한 2개의 새로운 Low Power Island(LP-E) 코어가 포함되어 있으며, **생성형 AI 추론을 처리**하도록 설계된 인텔 AI NPU 내에 2개의 신경망 컴퓨팅 엔진이 포함되어 있습니다..

[1]인텔® 코어™ Ultra 7 165H NPU와 인텔® 코어™ i7-1370P GPU에서 int8 모델을 실행하는 동안 UL Procyon AI 벤치마크에서 와트당 성능으로 측정한 결과입니다.
[1,2,3]워크로드 및 구성은 www.intel.com/PerformanceIndex를 참조하십시오. 결과는 다를 수 있습니다.

AI 혁신 가속

인텔®은 업계 최고의 ISV와 협력하여 AI 사용 경험을 최적화하고 있습니다.

AI PC 가속화 프로그램은 독립 하드웨어 공급업체(IHV)와 독립 소프트웨어 공급업체(ISV)를 인공 지능(AI) 툴체인, 훈련, 공동 엔지니어링, 소프트웨어 최적화, 하드웨어, 설계 리소스, 기술 전문 지식, 공동 마케팅, 판매 기회를 포함한 인텔® 리소스와 연결하는 것을 목표로 합니다.

자세한 정보

인텔® Tiber™ AI 클라우드로 엔터프라이즈 AI 개발 가속

최신 인텔® 하드웨어 및 소프트웨어의 클러스터에서 응용 프로그램 및 워크로드를 학습하고 프로토타이핑하고 테스트하고 실행해 보십시오.

이 개발 환경에서 최신 하드웨어 및 소프트웨어 혁신을 활용하여 **AI를 가속하고 확장**해 보십시오. **소프트웨어** 및 **생성형 AI를 미세 조정**할 때 **더 많은 컴퓨팅** 성능 및 선택권을 누리십시오.

인텔과 함께 시작하기

최신 인텔 제품을 실제로 경험해 보십시오. 인텔로 AI 기술을 강화하십시오.

조기 기술 액세스

출시 전 인텔 플랫폼과 관련 인텔 최적화 소프트웨어 스택을 평가해 보십시오.

대규모의 AI 배포

인텔의 최신 머신 러닝 툴킷과 인텔® Tiber™ AI 클라우드에서 호스팅되는 라이브러리로 AI 배포 속도를 높이십시오.

기술 문서 읽기 ›
시작하기 ›

콜 투 액션

교육

인텔® 기술이 생성형 AI 및 분야별 모델에 어떻게 사용될 수 있는지, 그리고 어떤 인텔® 제온® 및 인텔® Gaudi® 제품군이 비즈니스에 도움이 될 수 있는지 알아보십시오.

시작하기

참여

시작하기

인텔® Tiber™ AI 클라우드

이 개발 환경에서 최신 하드웨어 및 소프트웨어 혁신을 활용하여 AI를 가속하고 확장해 보십시오.

&

AI 참조 키트 사용

문의

자세한 내용은 **인텔® 담당자**에게 문의하십시오.

How to Access Intel® Partner Alliance Customer Support

인텔 Virtual Assistant

각 파트너 얼라이언스 웹 페이지에서 오른쪽 하단에 있는 이 채팅 봇은 대부분 질문에 셀프 도움말 또는 빠른 실시간 지원 담당자 연결을 제공합니다.

'도움 받기' 블레이드

온라인 지원 요청을 제출하십시오. 파트너 얼라이언스 웹 사이트 내 대부분의 페이지 꼬릿말에 이 링크가 표시되어 있습니다.

파트너 얼라이언스 '도움말' 페이지

도움말 페이지는 파트너 얼라이언스 회원이 사용할 수 있는 대부분의 도구 및 혜택에 대한 자세한 셀프 도움말 가이드를 제공합니다.

AI 활성화 존

중요한 리소스, 도구 및 혜택을 선별해 놓은 디지털 퍼스트 AI 작업 공간 - 인텔® 기술을 기반으로 솔루션을 구축, 마케팅 및 판매할 수 있도록 파트너를 지원합니다.

기술 지원

영업 및 마케팅 지원

기술 지원

영업 및 마케팅 지원

기술 지원

영업 및 마케팅 지원

AI 참조 키트

조직은 **이러한 참조 키트를 활용하여** 솔루션 개발 **시간을 크게 단축**하고 **상당한 성능 향상**을 경험할 수 있습니다.

금융 및 보험

사기 감지

GitHub ▪ 블로그 ▪ 청사진

의료 및 생명 과학

질병 방지

GitHub ▪ 블로그

제조 및 유틸리티

이상 탐지

GitHub ▪ 블로그

차량 관리

예측적 유지보수

GitHub

프로세스 자동화

문서 자동화

GitHub ▪ 블로그 ▪ 청사진

워크플로

- DL 전이 학습
- HF 미세 조정 및 추론 최적화
- DL-분산 압축
- 분산된 일반 ML 워크플로
- 인텔® 가속기를 통한 DL 사전 훈련
- DGL 및 PyG를 통한 그래프 분석 및 GNN
- Big-DL 기반 분산 훈련/추론
- Ray 기반 LLM 사전 훈련 및 미세 조정

도구

- 인텔® Distribution for Python
- 인텔® Optimized Modin
- 인텔® Optimized XGBoost
- 인텔® Extension for Scikit-learn
- 인텔® Optimized Tensorflow(ITEX)
- 인텔® Optimized PyTorch(stock & IPEX)
- 인텔® Neural Compressor
- SigOpt Python SDK & CLI
- CNVRG Python SDK & CLI
- 인텔 최적화 Horovod
- DeepSpeed

분야 키트

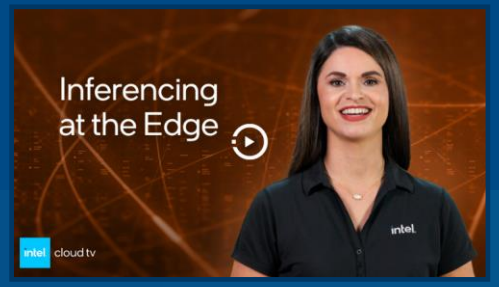
- 시계열
- PPML
- 전이 학습
- Transformer/NLP

참조 키트는 **컨테이너**로 제공되며 **온프레미스뿐만 아니라 주요 클라우드**에서도 사용할 수 있습니다. **참조 키트**는 **워크플로** 및 **분야 툴킷**에 계층화되어 있으며, **여러 산업의 더 다양한 사용 사례**를 지원하기 위해 독립적으로 활용할 수 있습니다.

Cloud TV

인텔® Cloud TV에서는 성공을 촉진하는 클라우드 컴퓨팅 뉴스, 트렌드 및 전략을 살펴볼 수 있습니다

인텔® Gaudi® AI 가속기와 함께하는 GenAI 기회



에지에서 데이터 추론을 사용하여 인사이트 확보

클라우드에서 AI로 경쟁 우위 창출

클라우드 기술을 사용한 AI 추론

클라우드에서의 AI

모든 곳에 AI를 확장하는 가장 빠른 경로

교육

모든 곳에 AI 도입 - 생성형 AI 엔터프라이즈 사용 사례

생성형 AI는 인터넷 챗봇만을 위한 것이 아닙니다. 수많은 기업이 생성형 AI 및 거대 언어 모델의 힘을 사용하여 일상 운영을 지원할 방법을 모색하고 있습니다. 이 세션에서는 엔터프라이즈의 생성형 AI 사용 사례를 살펴보고 조직이 일상 운영에 이를 적용할 수 있는 방법과 관련하여 고려 사항을 알려드립니다.

등록 ›

데이터 생성 및 거대 언어 모델을 위한 AI 간소화

등록 ›

AI를 조직의 워크로드에 통합하거나 이미 존재하는 인프라를 확장하는 것은 기술 및 컴퓨팅 집약적인 작업으로, 방대한 데이터 세트로 훈련된 강력한 모델을 개발하고 이를 적절히 실행할 수 있는 강력한 GPU를 갖춰야 합니다. 모든 조직이 이 작업을 수행하는 데 필요한 리소스를 갖춘 것은 아닙니다.

이 세션에서는 조직이 AI를 더 쉽게 이용하도록 설계되었으며 훈련 및 추론 시간을 개선하도록 최적화된 Accenture* 및 인텔®의 오픈 소스 AI 참조 키트 모음에 초점을 맞춥니다.

추가 교육

기술

자산 유형	제목 및 링크
역량	클라우드 내 AI 역량
웨비나	Hugging Face를 통한 인텔® 하드웨어용 AI 최적화
웨비나	클라우드 기반 분산 훈련을 설정하여 LLM을 미세 조정하는 방법
교육 과정	신속한 경제화 및 인컨텍스트 학습을 통한 LLM 개선
교육 과정	데이터 생성 및 거대 언어 모델을 위한 AI 간소화
교육 과정	자연어 처리
교육 과정	TensorFlow*를 통한 응용 딥 러닝
교육 과정	작고 민첩한 모델 – 엔터프라이즈 GenAI를 향한 빠른 경로
교육 과정	차세대 GenAI - 분야별 LLM
가이드	생성형 AI 시작을 위한 개발자 가이드: 사용 사례별 접근 방식
교육 과정	인텔® 제온® 프로세서 기반 AI를 솔루션 공간에서 활용하기

추가 교육

비기술

자산 유형	제목 및 링크
비디오 시리즈	생성형 AI의 수용
교육 과정	작고 민첩한 모델 – 엔터프라이즈 GenAI를 향한 빠른 경로
교육 과정	차세대 GenAI - 분야별 LLM
교육 과정	AI Everywhere 역량의 원칙
교육 과정	AI 소프트웨어 및 생태계 역량의 원칙
교육 과정	AI 생태계 참여: 소프트웨어로 성공, SI로 확장하고 솔루션 판매하기
교육 과정	실제 세계를 위한 생성형 AI 및 거대 언어 모델

추가 리소스

자산 유형	제목 및 링크
웨비나	생성형 AI 웨비나 시리즈
웨비나	모든 곳에 GenAI 도입
팟캐스트	How Copilot, ChatGPT, Stable Diffusion and Generative AI Will Change How We Develop, Work and Live
비즈니스 요약	어디에서나 AI 배포
블로그 시리즈	4세대 인텔 제온 프로세서로 생성형 AI를 위한 조정 및 추론 수행하기
솔루션 요약	Lenovo ThinkSystem SR650 V3 및 4세대 인텔 제온 프로세서를 통한 생성형 AI 추론의 배포 및 확장 Lenovo ThinkAgile VX V3 시스템을 크게 강화하는 인텔과 VMware의 신기술
기술 문서	인텔® AI 하드웨어 및 소프트웨어 최적화를 통한 Llama 2 가속
연구 PR	조사 대상 조직의 10%가 2023년 운영에 GenAI 솔루션 도입
대담 비디오	생성형 AI의 컴퓨팅 및 지속 가능성 과제 해결하기
팟캐스트	Hugging Face 및 인텔 - 실용적이고 더 빠르며, 민주적이고 윤리적인 AI 솔루션
Twitter/X 대화	민주화된 대규모 언어 모델이 AI 개발을 촉진하는 방법
Supermicro 벤치마크	Habana 성능 정보 검증
Hugging Face 벤치마크	벤치마크
교육/웨비나	클라우드 솔루션 아키텍트(CSA) 기술 강연: Habana를 갖춘 AI
백서	'Enterprise AI is all about the Developer'
인포그래픽	CPU는 엔터프라이즈 AI의 핵심입니다.

추가 리소스

자산 유형	제목 및 링크
솔루션 요약	Red Hat® OpenShift® AI와 Intel Enterprise AI를 사용하여 AI 채택 및 배포 간소화
가이드	AI 가이드
참조 키트	AI 비정형 텍스트 데이터 생성
백서	'Zoho is Optimizing and Accelerating Video AI Workloads'
백서	'Seekr Develops Trustworthy AI Screening System'
솔루션 요약	교육 분야 보안: 안전한 원격 시험을 현실화하는 AI와 기밀 컴퓨팅
사례 연구 및 비디오	씨앗부터 매장까지 AI 활용하는 Nature Fresh Farms
사례 연구	조기암 발견율을 주도하는 QMed Asia
사례 연구 및 비디오	AI 기반 추천 시스템을 개선한 MetaApp
솔루션 요약	자동 광학 검사(AOI)를 위한 AI 모델 훈련 및 개선 최적화
블로그	LLM을 위한 프롬프트 기반 효율성

법적 고지 및 면책 사항

고지 및 면책 조항.


intel