Enterprise AI

Generative AI & Domain Specific Models for Enterprise

Optimize training and deployment with purpose-built Intel® AI hardware and software to help transform your business

intel®

Contents

- **Why Partner with Intel on Generative AI**
- **Generative AI Landscape**
  - What is Generative AI and Large Language Models
  - What are some of the GenAI challenges today?
- **Domain Specific Models**
  - Why Domain Specific Models for Enterprise
  - Benefits of domain specific models for Enterprise and how partnering with Intel can help
- **Intel AI Software and Hardware Overview**
- **Intel Products for Large Language Models**
  - Intel® Gaudi® AI Accelerator
  - Intel® Xeon® Scalable Processors
  - Intel® Core™ Ultra
- **Call to Action**
- **Resources**

Why Partner With **Intel**?

At Intel, our goal is to improve lives and outcomes for everyone and every enterprise on this planet

But we aren't doing this alone!

Together with our partners, we are creating real value for our customers by **bringing AI everywhere** and minimizing the risks in AI solution deployment

When you partner with Intel, you partner with a complete AI ecosystem

Our broad portfolio of AI-enabling technologies and collaboration with hardware, software, and solution ecosystem partners delivers real world solutions and differentiated business outcomes for industries, companies, and communities.

Helping you to grow your business.

Join Us On the Journey to Bring Enterprise AI Everywhere

Generating Value for Customers with Intel AI Solutions

Intel's approach enables a broad, open ecosystem of AI players to offer solutions that satisfy enterprise-specific GenAI needs

NAVER

To develop a powerful large language model (LLM) for the deployment of advanced AI services globally, from cloud to on-device. NAVER has confirmed Intel Gaudi's foundational capability in executing compute operations for large-scale transformer models with outstanding performance per watt.

BOSCH

To explore further opportunities for smart manufacturing, including foundational models generating synthetic datasets of manufacturing anomalies to provide robust, evenly-distributed training sets (e.g., automated optical inspection).

IBM

Using 5th Gen Intel® Xeon® processors for its watsonx.data™ data store and working closely with Intel to validate the watsonx™ platform for Intel Gaudi accelerators.

OLA

To pre-train and fine-tune its first India foundational model with generative capabilities in 10 languages, producing industry-leading price/performance versus market solutions. Krutrim is now pre-training a larger foundational model on an Intel® Gaudi® 2 cluster.

seekr

Leader in trustworthy AI runs production workloads on Intel Gaudi 2, Intel® Data Center GPU Max Series and Intel® Xeon® processors in the Intel® Tiber™ Developer Cloud for LLM development and production deployment support.

IFF

Global leader in food, beverage, scent and biosciences will leverage GenAI and digital twin technology to establish an integrated digital biology workflow for advanced enzyme design and fermentation process optimization.

airtel

Embracing the power of Intel's cutting-edge technology, Airtel plans to leverage its rich telecom data to enhance its AI capabilities and turbo charge the experiences of its customers. The deployments will be in line with Airtel's commitment to stay at the forefront of technological innovation and help drive new revenue streams in a rapidly evolving digital landscape.

Infosys

Global leader in next-generation digital services and consulting announced a strategic collaboration to bring Intel technologies including 4th and 5th Gen Intel Xeon processors, Intel Gaudi 2 AI accelerators and Intel® Core™ Ultra to Infosys Topaz – an AI-first set of services, solutions and platforms that accelerate business value using generative AI technologies.

Enterprise AI Value Proposition

Transforming your business with Enterprise AI

In today's hypercompetitive environment, **enterprises that embrace AI are pulling ahead.**

Businesses across industries are reimaging every aspect of operations to understand how AI can augment or even automate workflows.

At Intel, embedding AI into the fabric of the enterprise is our unique expertise.

From AI PCs that transform productivity, to years of expertise in understanding which use cases return the most value, Intel is your trusted partner to bring AI everywhere, securely and responsibly.

Generative AI (GenAI) innovations are expected to be adopted by enterprises of all sizes at a rate faster than the internet era, the mobile era, or the cloud era.

The next wave of AI platforms will embrace these exciting realities in a way that is affordable and flexible.

It's time to think differently about your Enterprise AI.

This Enablement Package will help you understand how businesses across markets can gain significant value from Generative AI, in particular domain-specific models, for long-term success

What is Generative AI and Large Language Models?

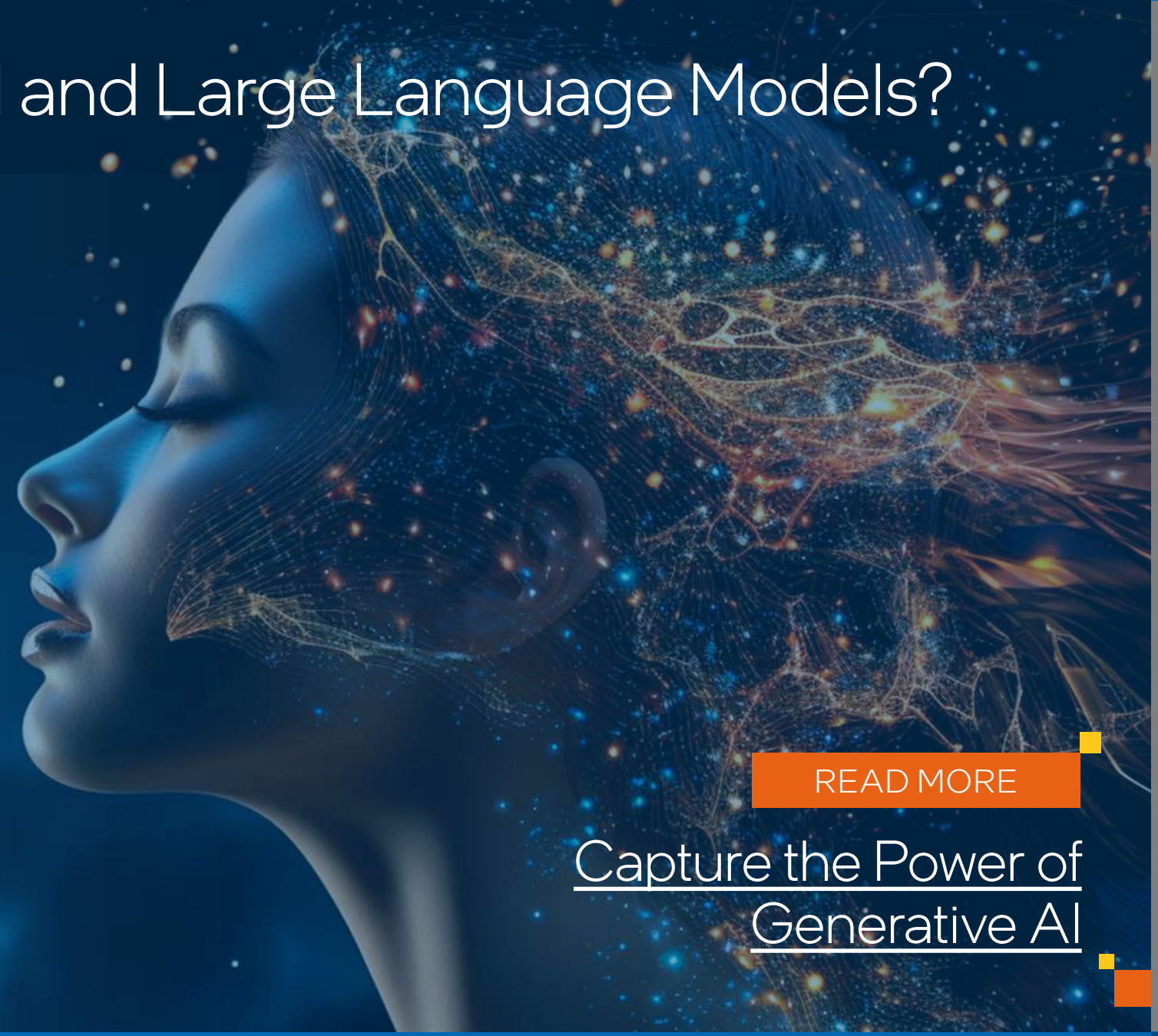
Generative AI (GenAI) is a subset of AI that focuses on creating new, original content.

It involves the training and deployment of AI models to generate data such as images, text, or audio that closely resemble examples from the training dataset.

GenAI algorithms use advanced techniques like deep learning and neural networks to produce realistic and coherent outputs that enable applications like image synthesis, text generation, and even creative artwork.

Large Language Model (LLM) is a specific type of Natural Language Processing model that uses deep neural networks to process and generate text. LLMs are trained on massive amounts of text data and are designed to generate coherent and meaningful outputs.

Learn More



READ MORE

Capture the Power of Generative AI

How will Enterprises use GenAI?

Consumer Goods & Retail

- Virtual fitting rooms
- Delivery and installation
- In-store product-finding assistance
- Demand prediction and inventory planning
- Novel product designs

Healthcare & Medicine

- Assist busy front-line staff
- Transcribe and summarize medical notes
- Chatbots to answer medical questions
- Predictive analytics to inform diagnosis and treatments

Manufacturing

- Expert copilot for technicians
- Conversational interactions with machines
- Prescriptive and proactive field service
- Natural language troubleshooting
- Warranty status and documentation
- Understanding process bottlenecks, devising recovery strategies

Media & Entertainment

- Intelligent search, tailored content discovery
- Headline and copy development
- Real-time feedback on content quality
- Personalized playlists, news digests, recommendations
- Interactive storytelling via viewer choices
- Targeted offers, subscription plans

Financial Services

- Uncovering trading signals, alerting traders to vulnerable positions
- Accelerating underwriting decisions
- Optimizing and rebuilding legacy system
- Reverse-engineering banking and insurance models
- Monitoring for potential financial crimes and fraud
- Automating data gathering for regulatory compliance
- Extracting insights from corporate disclosures

Source: Compiled by MIT Technology Review Insights, based on data from "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," and "Large Language Models in Media & Entertainment,"12 Databricks, April–June 2023.

Generative AI and Large Language Model Use Cases

Chatbots & virtual assistants

Customer support

Code generation & debugging LLMs

Trained on company's documents

Sentiment analysis

Assess customer satisfaction

Text classification & clustering

Categorize large volumes of data to identify trends

Language translation

Transition company web pages into other languages

Summarization & paraphrasing

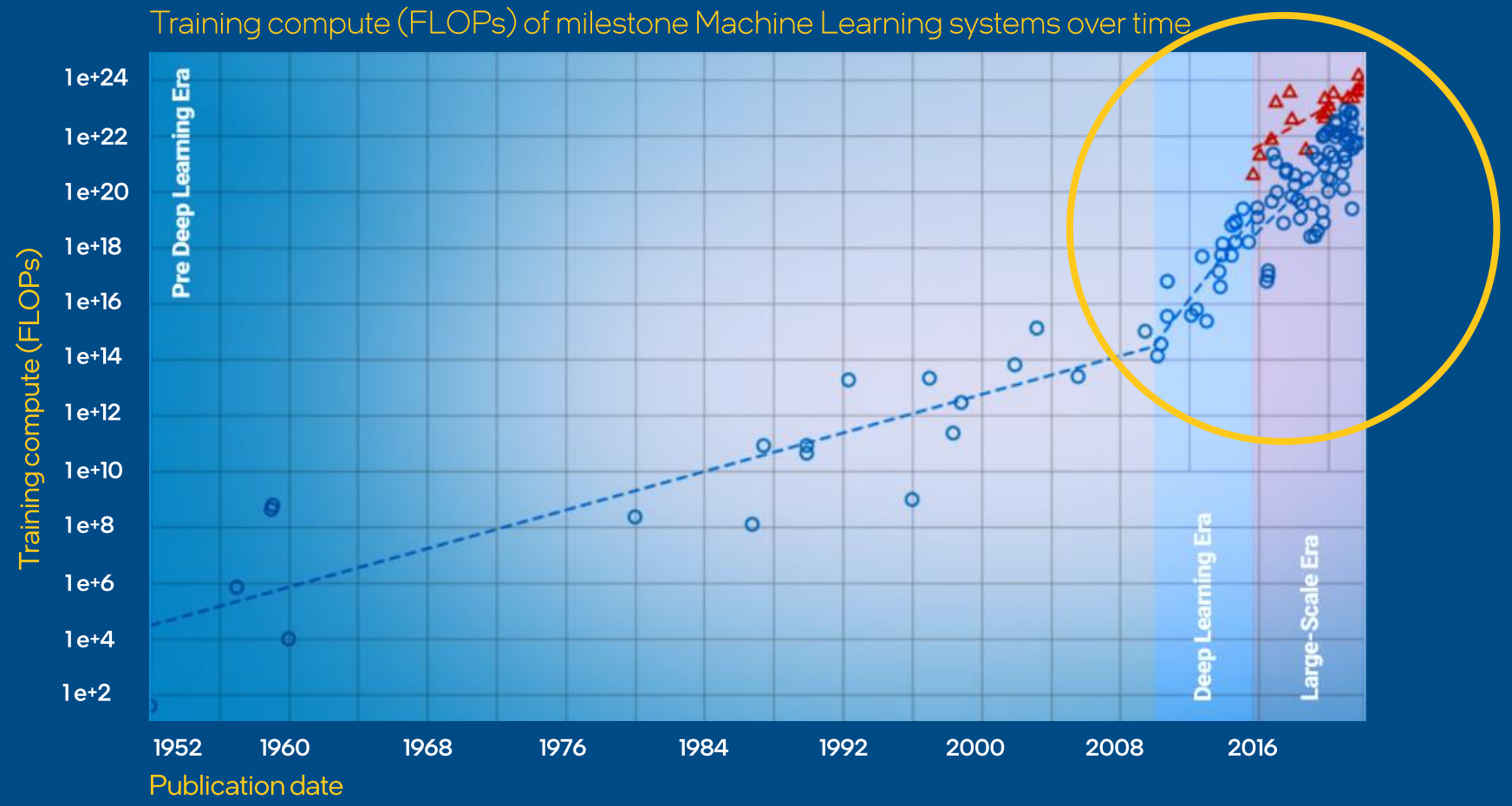
Meeting notes summarized

Content, image, video generation

First drafts of emails, idea generation, marketing visuals, short video

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

As Models Grow in Size, Compute Also Grows



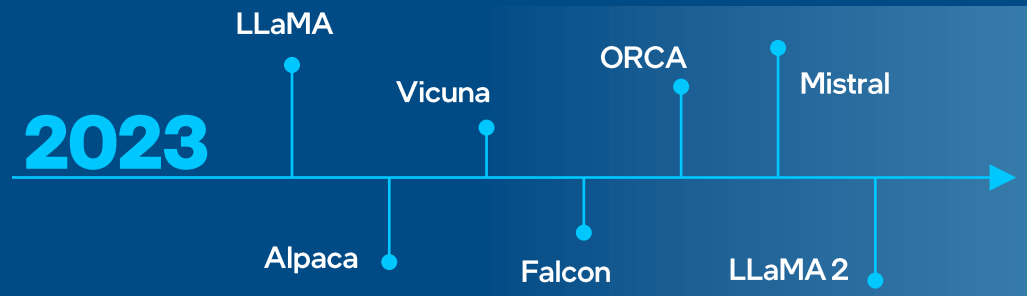
Study by Epoch, University of Aberdeen, Center for the Governance of AI, University of St. Andrews, MIT, Eberhard Karls Universitat Tubingen, Universidad Complutense

Not Just About Giant Models

	Giant (3rd party)	VS.	**Small and Nimble** (by 10-100X)
Explainability	Proprietary model	VS.	Open Source based model
Accuracy	All-in-one general purpose	VS.	Targeted, domain-specific, customized
Location	Cloud-based (as-a-service)	VS.	Locally run inference; edge, client & on-prem
Cost	Scaling cost in perpetuity	VS.	Cost management
Speed to Market	Fast setup (seconds)	VS.	Time to build (hours/days)

Growth of Many Smaller Models

100's of billions to <20B parameters in 6 months



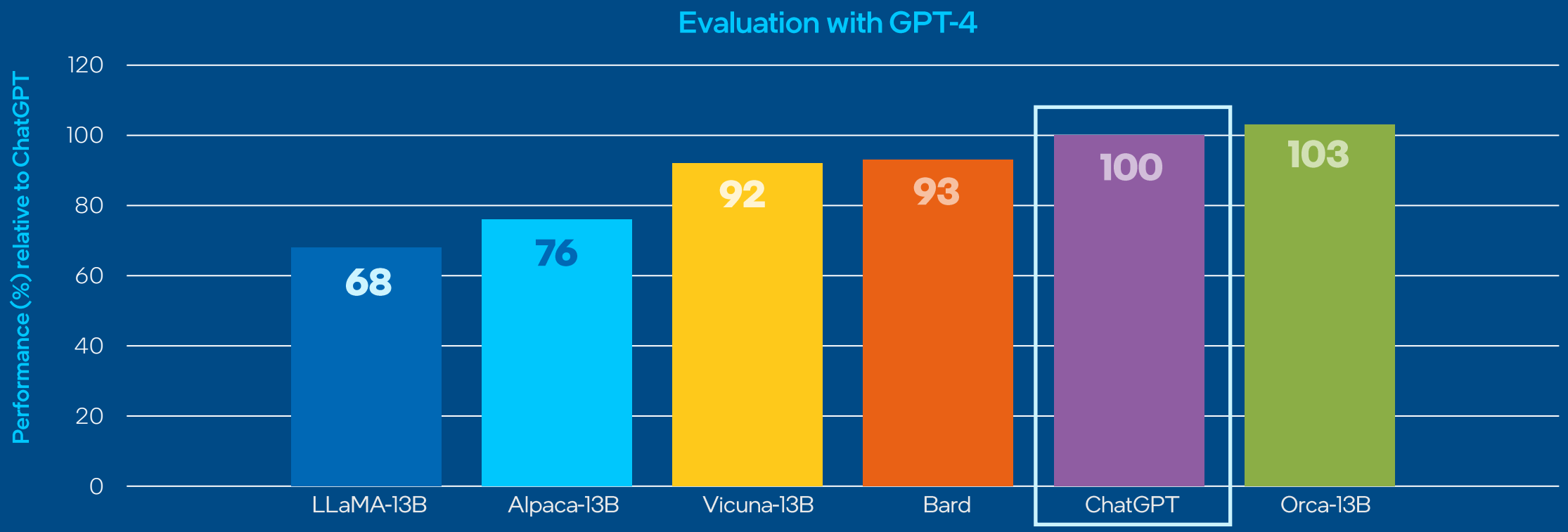
databricks

- Dozens of smaller models emerging weekly
- Commercial and open source licenses
- Indication that smaller models can replicate the accuracy of larger models if trained on carefully sourced data

- Thousands of domain-specific commercial models and AI platforms being demonstrated
- Models can be fine-tuned on a few processors in domain-specific data

Smaller Models Performed Well vs. ChatGPT

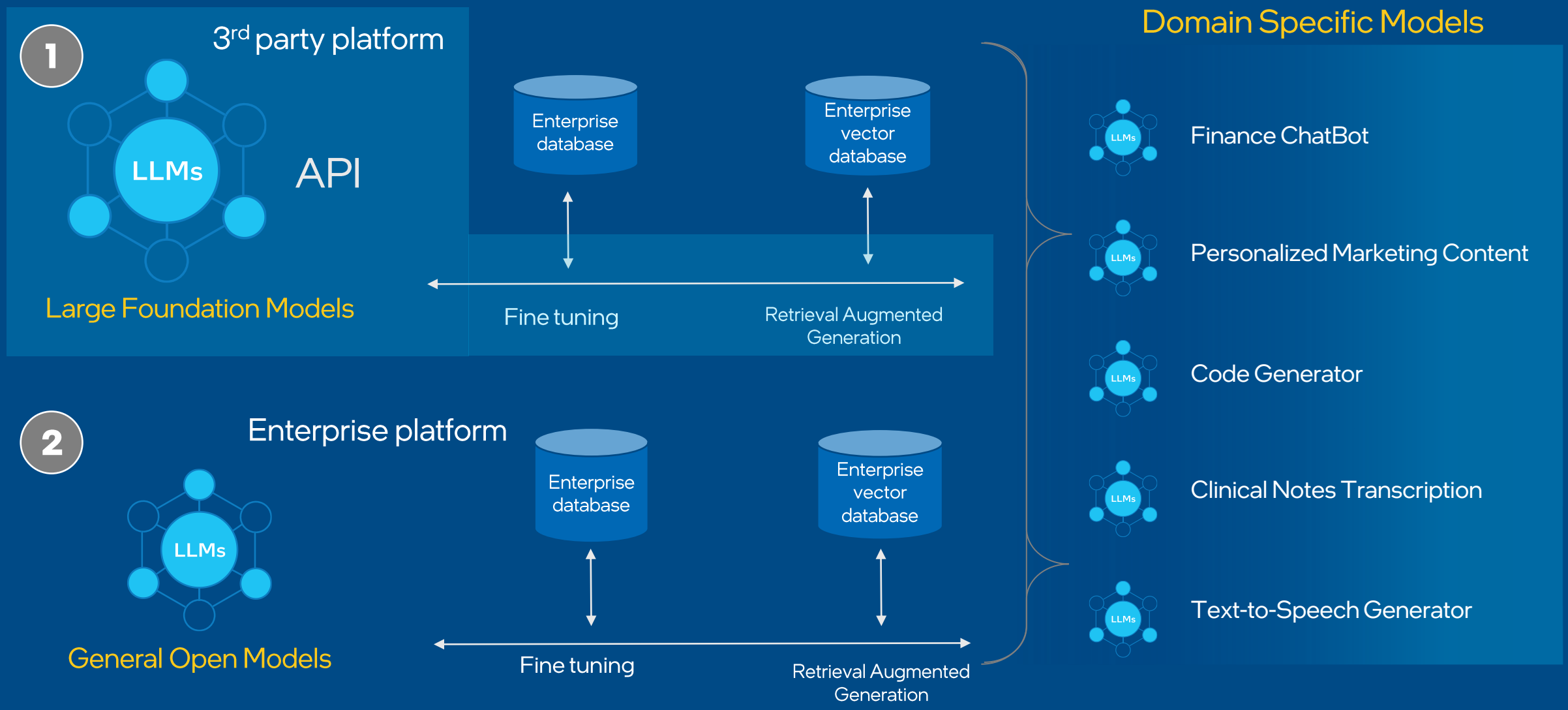
Proof that smaller models are a viable option and still perform well in comparison to large models like ChatGPT



Orca outperforms a wide range of foundation models including OpenAI ChatGPT as evaluated by GPT-4 in the Vicuna evaluation set

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

Build Domain Specific Models



Domain Specific Models Have Many Benefits for Enterprise

Smaller, targeted models can provide equivalent or superior performance, increasing ROI by decreasing time and cost investment

More Accurate Output

Use your enterprise data for more domain specific accuracy

Lower Cost

Fine-tuning a pre-trained model, and/or use RAG, and inferencing smaller model

Deploy Anywhere on Platform of Choice

Locally run inference; edge, client & on-prem

Secure & Private

Meets data security and regulatory requirements

Responsible AI

Giving model the ability to cite source of data with fine-tuning and RAG

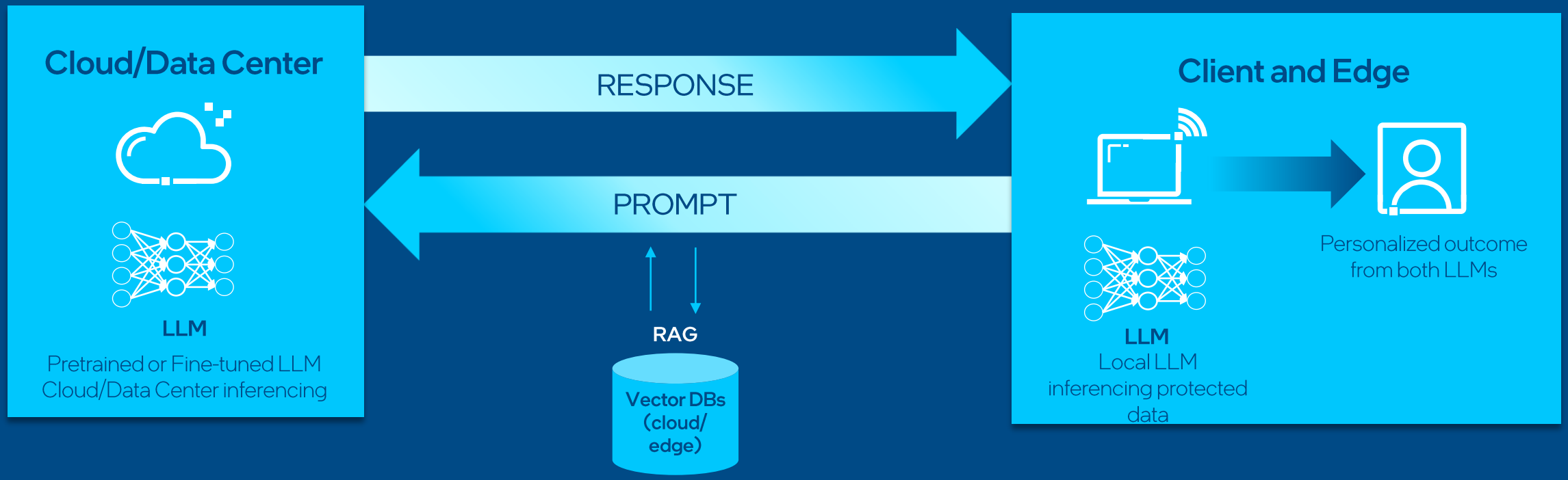
THE FUTURE

There will be a small number of giant models and a giant number of small, more nimble AI models embedded in countless applications[1]

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

Seamless cloud to edge AI platform

Train and inference in the cloud. Use RAG to improve domain accuracy.



Generative AI - A Year in Production

The use of domain specific, yet highly intelligent models, is rising

2022 EXPERIMENTATION → **2023** PILOTS → **2024** PRODUCTION

Huge models paved the way

- Very effective for general purpose
- Expensive to train and deploy
- Built on large public data sets
- Easy to use

Smaller, domain specific models

- Use your private data for business specific results
- Deploy on the hardware you have
- Increased efficiency, accuracy, security, and traceability
- Time to build

READ THE BLOG

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

Intel's Approach to Domain Specific Models

DOMAIN SPECIFIC MODELS

Advantages

+ 10-100x smaller models while maintaining/improving accuracy
+ Economical on general-purpose compute
+ Correctness; Source attribution; Explainability
+ Utilizing private/enterprise data
+ Continuously updated information

Challenges

- Reduced range of tasks
- Requires few-shot fine-tuning and indexing

INTEL GOAL

Enable the most cost-effective and ubiquitous approach to fine tune and deploy 10,000s of models on Intel hardware using industry frameworks, pre-trained models, and Intel AI SW and tools

READ MORE

Generative AI at Our Fingertips

E-book ▪ Infographic

Enterprise AI: Helping to Overcome Barriers to Entry

Requirements	How partnering with Intel can help
Speed to market	Use Developer resources from Intel and Hugging Face, the Gaudi Developer Hub and 5 Reference Kits to get a running head start in generative AI
User experience (accuracy/latency)	Inference on models greater than 10B parameters on Intel® Gaudi® accelerator and small models <20B parameters on Intel® Xeon® processors with Intel® AMX, giving users a real-time experience[1]
Compute availability	Intel® Xeon® CPU + accelerators offers a cost-effective alternative to the global GPU shortage. **Gaudi® 2 is available now through SuperMicro, with greater availability for Gaudi® 3.**
Familiar technology	Inference of smaller models can be done practically on any hardware, including ubiquitous solutions that might already be part of your compute setting[2]
Operationalize at scale	Gaudi® 2 offers near-linear scalability with 24 100 GbE ports integrated onto every accelerator. Xeon is already in your data center, out in the field; cloud to edge. **65% of data center inferencing runs on Xeon**[3]
Cost effective	In real work applications, Intel is disrupting the industry and democratizing AI by delivering better performance, lower pricing and a more balanced platform for AI inference. See Nvidia shows Intel® Gaudi 2 is 4x better performance per dollar than its 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] Based on Intel market modeling of the worldwide installed base of data center servers running AI Inference workloads as of December 2022.

Software Resources to Simplify Generative AI Training and Deployment

Open-Source Model

- BLOOM — a BigScience initiative — 1768 params 59 languages Open-access — 176B
- BioGPT — Domain 1.5B
- Stable Diffusion — Image
- Llama2 GPT-J MPT Falcon — 7-65B LLM
- Stanford Alpaca — Fine-tuned 7B LLM
- Knowledge base

Open Software

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

GenAI Platform

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

READ MORE [Unlock Generative AI with Ubiquitous Hardware and Open Software]

Why Intel's Open-Source SW Approach is Suited to Your AI Business Needs

Avoid Vendor Lock-in

open-source standards-based software

MODIN · SciPy · dmlc XGBoost · PyTorch

pandas · NumPy · scikit learn · TensorFlow

Spark · Numba · SigOpt AutoML · oneAPI

Leverage Intel's Hardware portfolio

optimized for AI use cases

intel XEON · intel CORE ULTRA · intel ARC GRAPHICS

intel DATA CENTER GPU · intel GAUDI · intel ETHERNET

For tomorrow's AI, create new opportunities from the client and edge, to the data center and cloud, with **software optimized hardware and open standards**

Intel® AI Software Portfolio

Engineer Data → **Create Models** → **Optimize & Deploy**

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

Data Analytics at Scale†

Machine & Deep Learning Frameworks, Optimization and Deployment Tools†

oneAPI

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

Open, cross-architecture programming model for CPUs, GPUs, and other accelerators

CLOUD & ENTERPRISE

CLIENT & WORKSTATION

EDGE

intel AI ANALYTICS TOOLKIT

Accelerate end-to-end data science and AI

Intel® Tiber™ Developer Cloud
(formerly; Intel® Developer Cloud)
and Intel® Developer Catalog

Try the latest Intel tools and hardware, and access optimized AI Models

Intel® Tiber™ AI Studio

Full stack ML operating system

Intel® Geti

Annotation/training/ optimization platform

Hugging Face

Intel optimizations and fine-tuning recipes, optimized inference models, and model serving

Note: components at each layer of the stack are optimized for targeted components at other layers based on expected AI usage models, and not every component is utilized by the solutions in the rightmost column
† This list includes popular open-source frameworks that are optimized for Intel hardware

Simplify enterprise generative AI adoption and reduce the time to production of hardened, trusted solutions

OPEA:

Simplify Enterprise Generative AI adoption and reduce the time to production of hardened, trusted solutions

Open Platform for Enterprise AI

OPEA Partners

anyscale

CLOUDERA

DATASTAX

kx

DOMINO

Hugging Face

Red Hat

drant

MariaDB

MINIO

sas

Yellowbrick

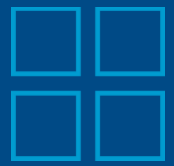
vmware

OPEA Value

- Helps Enterprises unlock value from their data using Generative AI (LLM, RAG) faster and easier
- Reduces complexities of fragmented ecosystem and helps solutions to scale in production
- Ignites collaboration and contribution across industry leaders partnering with the Linux Foundation

Efficient
Harnesses existing infrastructure, the AI accelerator or other hardware of your choosing.

Seamless
Integrates with enterprise software, with heterogeneous support and stability across system & network.



Open
Brings together best of breed innovations and is free from proprietary vendor lock-in.

Ubiquitous
Runs everywhere through a flexible architecture built for cloud, data center, edge and PC.

Trusted
Features a secure enterprise-ready pipeline and tools for responsibility, transparency, and traceability.

Scalable
Provides access to a vibrant ecosystem of partners to help build and scale your solution.

Hugging Face Partnership for Generative AI

Hugging Face

To facilitate generative AI and language AI training and innovation, Intel has teamed up with Hugging Face, a popular platform for sharing AI models and data sets. Most notably, Hugging Face is known for its transformers library built for NLP.

intel. XEON

Intel has worked with Hugging Face to build state-of-the-art hardware and software acceleration to train, fine-tune, and predict with transformer models.

The hardware acceleration is driven by Intel® Xeon® Scalable processors, while the software acceleration is enabled by our portfolio of optimized AI software tools, frameworks, and libraries.

intel. GAUDI

Intel® Gaudi® deep learning accelerators are also paired with Hugging Face open-source software through the Optimum Habana Library to enable developer ease of use on thousands of models optimized by the Hugging Face community.

Hugging Face has also published several evaluations of Intel® Gaudi® 2 performance on generative AI models: Stable Diffusion, T5-3B, BLOOMZ 176B and 7B, and the new BridgeTower model.

Intel, Articul8 and BCG Collaborate to Deliver Enterprise-Grade, Secure Generative AI

Pioneering solution powered by Intel AI supercomputer unlocks business value with custom datasets while maintaining high levels of security and data privacy

Articul8* offers a turnkey GenAI software platform that delivers speed, security, and cost-efficiency to help large enterprise customers operationalize and scale AI. The platform was launched and optimized on Intel hardware architectures, including Intel® Xeon® Scalable processors and Intel® Gaudi® accelerators, but will support a range of hybrid infrastructure alternatives.

intel. GAUDI intel. XEON

Following the technology's early deployment at Boston Consulting Group (BCG), the team has scaled the platform to enterprise customers in industry segments requiring high levels of security and specialized domain knowledge, including financial services, aerospace, semiconductors, and telecommunications.

READ MORE

Articul8 Announcement

Articul8 Website

Articul8 Training

Responsible AI for Enterprise

CHALLENGE:

Generative AI models learn from vast amounts of data available on the internet, which can contain biases present in society and may inadvertently apply these biases. LLMs can be manipulated to generate or spread misinformation, phishing emails, or social engineering attacks.

LLMs can often have "hallucinations" and generate inaccurate information, which can be particularly problematic in industries like healthcare, where models can influence diagnostic and therapeutic decisions and potentially harm patients.

Learn More

Minimizing the Risks of Generative AI

SOLUTIONS:

Companies and individuals working on AI technology need to make sure their software is developed and deployed according to ethical AI principles

The open-source Intel® Explainable AI Tools allow users to run post hoc model distillation and visualization to examine the predictive behavior of both TensorFlow* and PyTorch* models

LLMs are typically trained on large public datasets and then fine-tuned on potentially sensitive data (e.g. financial and healthcare)

Technologies like Intel's Open Federated Learning (OpenFL) incorporate confidential computing so that LLMs can be safely fine-tuned on sensitive data, which in turn improves the generalizability of models while reducing hallucinations and bias

Intel Products for NLP / LLMs

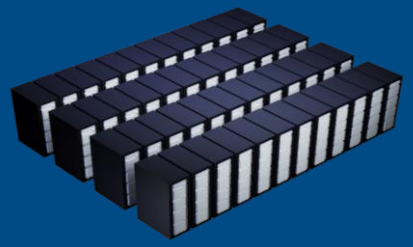
Scalable Systems for AI

Training and Fine-Tuning

Training | Peak Inference | Mainstream Inference/ Fine-Tuning | Baseline Inference | Endpoint Inference

Inference and Deployment

Cloud Data Center | Edge | Client

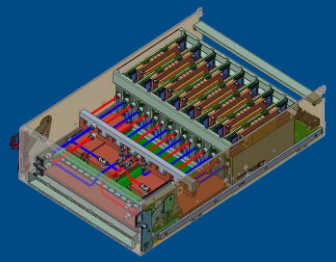


Cluster and Data Center Scale

intel. GAUDI

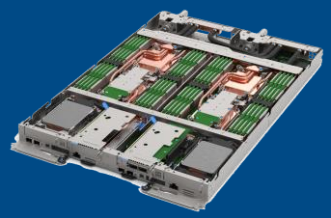
Multi-node Deployment per Rack

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



Multi-GPU or Multi-socket CPU

intel. XEON | intel. XEON MAX SERIES



Single-Socket CPU or Single GPU

intel. XEON | intel. CORE ULTRA

Client CPU

intel. CORE ULTRA | intel. ARC GRAPHICS

intel. ETHERNET

Intel Products for NLP/LLMs

Training/ Inference

Intel® Gaudi® AI accelerators are specifically designed to accelerate training and inference of large-scale models, such as LLMs and NLPs, greater than 10B parameters.

Accelerating Generative AI and Large Language Models with Intel® Gaudi® 2

Intel® Gaudi® 2 Remains Only Benchmarked Alternative to NV H100 for GenAI Performance

Intel® Gaudi® 2 delivers leading performance and optimal cost savings for AI training

Press Release
Newsroom

The Gaudi® 2 deep learning accelerator performs competitively on deep learning training and inference, with up to **2.4x faster performance than Nvidia A100**[1]

Newsroom
Tech Article

Gaudi® 2 delivers compelling performance vs. Nvidia's H100[2,3] **for GPT-3 and GPT-J**

Newsroom
MLCommons Announcement

WATCH NOW ›

Intel webinar recording discussing the cutting-edge capabilities of the Intel® Gaudi® 2 AI processor in capturing the potential of Generative AI and Large Language Models (LLMs)

CASE STUDY ›

AWS instances featuring Intel® AI acceleration technologies, with Optimum Intel and Optimum Habana libraries, give companies powerful tools for generative AI implementation

[1] Performance varies by use, configuration, and other factors; workloads and configuration details available at: intel.com/performanceindex Results may vary.

[2 3] Performance varies by use, configuration, and other factors; workloads and configuration details available at: https://mlcommons.org/2023/09/mlperf-results-highlight-growing-importance-of-generative-ai-and-storage/ Results may vary.

COMING SOON - Intel® Gaudi® 3 AI accelerator

Bringing Choice to GenAI with Performance, Scalability and Efficiency

intel GAUDI

Intel® Gaudi® 3 will deliver a significant leap in AI training and inference for global enterprises looking to deploy GenAI at scale

Press Release

Intel® Gaudi® 3 accelerator performance vs Nvidia H100

Intel Gaudi 3 is projected to deliver

50% faster time-to-train on average[3] across Llama2 models with 7B and 13B parameters, and GPT-3 175B parameter model

Intel Gaudi 3 is projected to outperform H100 by:

50% for accelerator inference throughput[1]

40% for inference power-efficiency[2]

across Llama 7B and 70B parameters, and Falcon 180B parameter models

READ MORE ›

Intel Breaks Down Proprietary Walls to Bring Choice to Enterprise GenAI Market

WHITE PAPER

[1]NV H100 comparison based on https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8 , Reported numbers are per GPU. Vs Intel® Gaudi® 3 projections for LLAMA2-7B, LLAMA2-70B & Falcon 180B projections. Results may vary.
[2]NV H100 comparison based on https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8 , Reported numbers are per GPU. Vs Intel® Gaudi® 3 projections for LLAMA2-7B, LLAMA2-70B & Falcon 180B. Power efficiency for both Nvidia and Gaudi 3 based on internal estimates. Results may vary.
[3]NV H100 comparison based on: https://developer.nvidia.com/deep-learning-performance-training-inference/training, "Large Language Model" tab vs. Intel® Gaudi® 3 projections for LLAMA2-7B, LLAMA2-13B & GPT3-175B as of 3/28/2024. Results may vary.

Intel Products for NLP/LLMs

Inference

4th and 5th Gen Intel® Xeon® Scalable processors accelerate NLP with Intel® DL Boost, Intel® AMX, and Intel® AVX-512. It is designed for high-performance computing and can be used to accelerate NLP workloads. They can handle large number of threads, large memory capacity, and high memory bandwidth, which is suitable for NLP workloads such as language translation, text summarization, and text-to-speech.

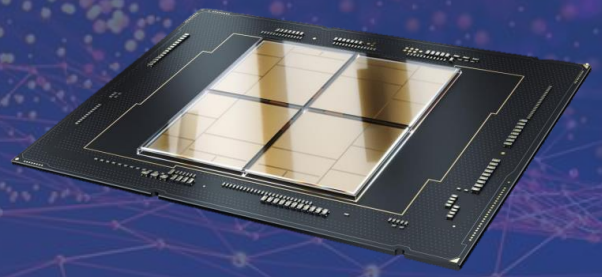
Intel® Xeon® Scalable Processors for LLMs

Ideal for building and deploying general-purpose AI workloads with the most popular AI frameworks and libraries

- Utilize existing infrastructure for inferencing domain specific LLMs
- Delivers for transfer learning use cases
- Deploy LLMs on Intel® Xeon® with open-source SW for ease of delivering optimal performance

Intel® Xeon®
CPU Performance Leadership in Real World AI Applications

Tech Article ▪ Infographic



GPT-J
4th Gen Intel® Xeon® Results

2 paragraphs per second in offline mode[1]

1 paragraph per second in real-time server mode[1]

Newsroom Article ▪ MLCommons Announcement

Debunking the GPU Myth: How CPUs with Built-In Accelerators Revolutionize AI

Alibaba NLP Case Study on 4th Gen® Xeon® with Intel® AMX

READ MORE

[1]Performance varies by use, configuration, and other factors; workloads and configuration details available at: https://mlcommons.org/2023/09/mlperf-results-highlight-growing-importance-of-generative-ai-and-storage/ Results may vary.

Intel Products for NLP / LLMs

Small Scale Inference on Client

Intel® Core™ Ultra ushers in the age of the AI PC

Intel® Core™ Ultra processors are optimized for premium thin and powerful laptops, featuring 3D performance hybrid architecture, advanced AI capabilities, and available with built-in Intel® Arc™ GPU. Created using the new Intel 4 process, Intel® Core™ Ultra processors deliver an optimal balance of performance and power efficiency for gaming, content creation, and productivity on the go.

Intel® Core™ Ultra for Generative AI

Intel's most power-efficient client processor ushers in the age of the AI PC

Major Improvements in Efficiency and Performance

AI EFFICIENCY
up to **70%**
faster generative AI performance[2]

POWER SAVINGS
up to **25%**
reduction in power consumption[3]

READ MORE Announcement ▪ Product Brief ▪ Website

Intel® Core™ Ultra features Intel's first client on-chip AI accelerator — the neural processing unit, or NPU — to enable a new level of power-efficient AI acceleration with **2.5x better power efficiency** than the previous generation[1]

Both the Intel® Core™ Ultra H and U generation of chips include two new Low Power Island (LP-E) cores for low intensity workloads, with two Neural Compute Engines within the Intel AI NPU designed to tackle **generative AI inferencing**.

[1]As measured by Perf/Watt on UL Procyon AI benchmark while running an int8 model on Intel® Core ™ Ultra 7 165H NPU vs. Intel® Core ™ i7-1370P GPU.
[1,2,3]See www.intel.com/PerformanceIndex for workloads and configurations. Results may vary.

Accelerating AI Innovation

Intel is working with leading industry ISVs to optimize your experience with AI.

The AI PC Acceleration Program aims to connect independent hardware vendors (IHVs) and independent software vendors (ISVs) with Intel resources including artificial intelligence (AI) toolchains, training, co-engineering, software optimization, hardware, design resources, technical expertise, co-marketing, and sales opportunities.

Learn More

Accelerate Enterprise AI Development with Intel® Tiber™ Developer Cloud (formerly; Intel® Developer Cloud)

Learn, prototype, test, and run applications and workloads on a cluster of the latest Intel® hardware and software

Accelerate and **scale AI** with the latest hardware and software innovations in this development environment. **Gain more compute** power and choices to **fine-tune your software** and **generative AI.**

Get Started with Intel

Get hands-on experience with the latest Intel products. Empower your AI skills with Intel.

Early Technology Access

Evaluate prerelease Intel platforms and associated Intel-optimized software stacks.

Deploy AI at Scale

Speed up AI deployments with the latest machine learning toolkits from Intel and libraries hosted on Intel Developer Cloud.

Read the Technical Article ›

Get Started ›

Call to Action

EDUCATION

Understand how Intel® technology can be used for Generative AI & Domain-Specific Models, and the scope upon which Intel® Xeon® and Intel® Gaudi® product lines can help you win more business

Get Started

ENGAGEMENT

Get started with

Intel® Tiber™ Developer Cloud

Accelerate and scale AI with the latest hardware and software innovations in this development environment

CONTACT

Reach out to your Intel Representative for more information

How to Access Intel® Partner Alliance Customer Support

Intel Virtual Assistant

This Chat Bot, located in the bottom-right corner of each Partner Alliance webpage, provides self-help to most questions or a quick link to a live support agent.

Get Help "Blade"

Submit an online support request.

This link is found on the footer of most pages within the Partner Alliance website.

Partner Alliance "Get Help" page

The Get Help page provides detailed self-help guides on most of the tools and benefits available to Partner Alliance members.

AI Activation Zones

Digital-first AI workspaces that curate critical resources, tools and benefits - activating partners to build, market, and sell solutions based on Intel technology

Technical Enablement

Sales & Marketing Enablement

Technical Enablement

Sales & Marketing Enablement

Technical Enablement

Sales & Marketing Enablement

AI Reference Kits

Leveraging these reference kits, organizations can **significantly reduce time** to solution and experience **substantial performance gain**

Finance & Insurance

Fraud Detection

GitHub ▪ Blog ▪ Blueprint

Health & Life Sciences

Disease Protection

GitHub ▪ Blog

Manufacturing & Utilities

Anomaly Detection

GitHub ▪ Blog

Fleet Management

Predictive Maintenance

GitHub

Process Automation

Document Automation

GitHub ▪ Blog ▪ Blueprint

Workflows

- DL Transfer Learning
- HF Fine-tuning & Inference Optimization
- DL-Distributed Compression
- Distributed classical ML workflow
- DL pre-training with Intel accelerators
- Graph analytics & GNN with DGL & PyG
- Distributed Trang/Inference on Big-DL
- LLM Pre-Training & Fine-Tuning on Ray

Tools

- Intel® Distribution of 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
- Intel-optimized Horovod
- DeepSpeed

Domain Kits

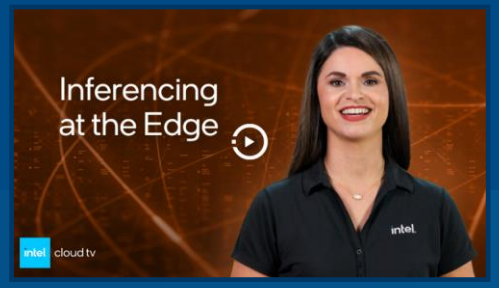
- Time Series
- PPML
- Transfer Learning
- Transformer/NLP

The **reference kits** are delivered as **containers** and can be used on **major clouds as well as on-prem.** The **reference kits** are layered on **workflows** and **domain-toolkits** which can be independently leveraged to support a **wider variety of use-cases in multiple industries.**

Cloud TV

Intel® Cloud TV explores cloud computing news, trends, and strategies to drive your success

Your GenAI Opportunity with Intel® Gaudi® AI Accelerators



Gain Insights Using Data Inferencing at the Edge

Creating Competitive Advantage with AI in the Cloud

AI Inferencing Using Cloud Technologies

AI in the Cloud

Get on the Fast Path to Scale AI Everywhere

Training

Bringing AI Everywhere - Generative AI Enterprise Use Cases

Generative AI is not just for internet chatbots. A myriad of enterprises are considering ways to use the power of generative AI and large language models to assist in the day to day of operations. This session will explore the use cases for Generative AI in Enterprise and will provide considerations as to how your organization might apply it in your day-to-day operations.

Enroll ›

Streamline AI for Data Generation and Large Language Models

Enroll ›

Incorporating AI into an organization's workloads or scaling up an already existing infrastructure is skill-heavy and computationally intensive, requiring the development of robust models trained on massive datasets and powerful GPUs on which to run them adequately. Not every organization has the necessary resources to accomplish this task.

This session focuses on a solution: A collection of open source AI reference kits from Accenture* and Intel® designed to make AI more accessible to organizations and optimized for improved training and inference time.

Additional Trainings

Technical

Asset Type	Title and Link
Training Course	Improving LLMs with Prompt Economization and In-Context Learning
Training Course	Streamline AI for Data Generation and Large Language Models
Training Course	Applied Deep Learning with TensorFlow*
Training Course	Small and Nimble – the Fast Path to Enterprise GenAI
Training Course	The Next Wave of GenAI - Domain-Specific LLMs
Guide	A Developer's Guide to Getting Started with Generative AI: A Use-Case-Specific Approach
Training Course	Taking AI on Intel® Xeon® Processors Into the Solution Space

Additional Trainings

Non-Technical

Asset Type	Title and Link
Video Series	Embracing Generative AI
Training Course	Small and Nimble – the Fast Path to Enterprise GenAI
Training Course	The Next Wave of GenAI - Domain-Specific LLMs
Training Course	Principles of AI Everywhere Competency
Training Course	Principles of AI Software & Ecosystem Competency
Training Course	Engaging the AI Ecosystem: Win with Software, Scale with SIs and Sell the Solution
Training Course	Generative AI and Large Language Models for the Real World

Additional Resources

Asset Type	Title and Link
Webinar	Generative AI Webinar Series
Webinar	Bringing GenAI Everywhere
Podcast	How Copilot, ChatGPT, Stable Diffusion and Generative AI Will Change How We Develop, Work and Live
Business Brief	Deploy AI Everywhere
Blog Series	Tuning and Inference for Generative AI with 4th Generation Intel Xeon Processors
Solution Brief	Deploy and Scale Generative AI Inference with Lenovo ThinkSystem SR650 V3 / 4th Gen Intel Xeon Processors
Solution Brief	New Intel and VMware Technologies Turbocharge Lenovo ThinkAgile VX V3 Systems
Tech Article	Accelerate Llama 2 with Intel® AI Hardware and Software Optimizations
Research PR	10% of Organizations Surveyed Launched GenAI Solutions to Production in 2023
Fireside Chat Video	Taking on the Compute and Sustainability Challenges of Generative AI
Podcast	Hugging Face and Intel - Driving Towards Practical, Faster, Democratized and Ethical AI solutions
Twitter / X Conversation	How Democratized Large Language Models Boost AI Development
Supermicro Benchmarks	Habana Claims Validation
Hugging Face Benchmarks	Benchmarks
Training / Webinar	Cloud Solution Architect (CSA) Tech Talk: AI with Habana
White Paper	Enterprise AI is all about the Developer
Infographic	CPUs are Key to Enterprise AI

Additional Resources

Asset Type	Title and Link
Solution Brief	Streamline AI Adoption and Deployment Using Intel Enterprise AI with Red Hard OpenShift AI
Guide	The AI Guide
Reference Kit	AI Unstructured Text Data Generation
White Paper	Zoho is Optimizing and Accelerating Video AI Workloads
White Paper	Seekr Develops Trustworthy AI Screening System
Solution Brief	Security in Education: AI and Confidential Computing Help Make Secure Remote Exams a Reality
Case Study & Video	Nature Fresh Farms Utilizes AI from Seed to Store
Case Study	QMed Asia Drives Early-Stage Cancer Detection Rate
Case Study & Video	MetaApp Revamps AI-Based Recommendation System
Solution Brief	Optimizing AI Model Training and Refinement for Automated Optical Inspection (AOI)
Blog	Prompt-Driven Efficiencies for LLMs

Notices and Disclaimers

Performance varies by use, configuration and other factors. Learn more on the Performance Index site.
Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.
Your costs and results may vary.
Intel technologies may require enabled hardware, software or service activation.


intel