

Generative AI and Intel Offerings

Ashish Sharma ashish3.sharma@intel.com
Puneesh Khanna puneesh.khanna@intel.com
Ankur Shukla ankur.shukla@intel.com

Copyright © 2023 Intel Corporation.
This document is intended for personal use only.
Unauthorized distribution, modification, public performance,
public display or copying of this material via any medium is strictly prohibited.



intel®

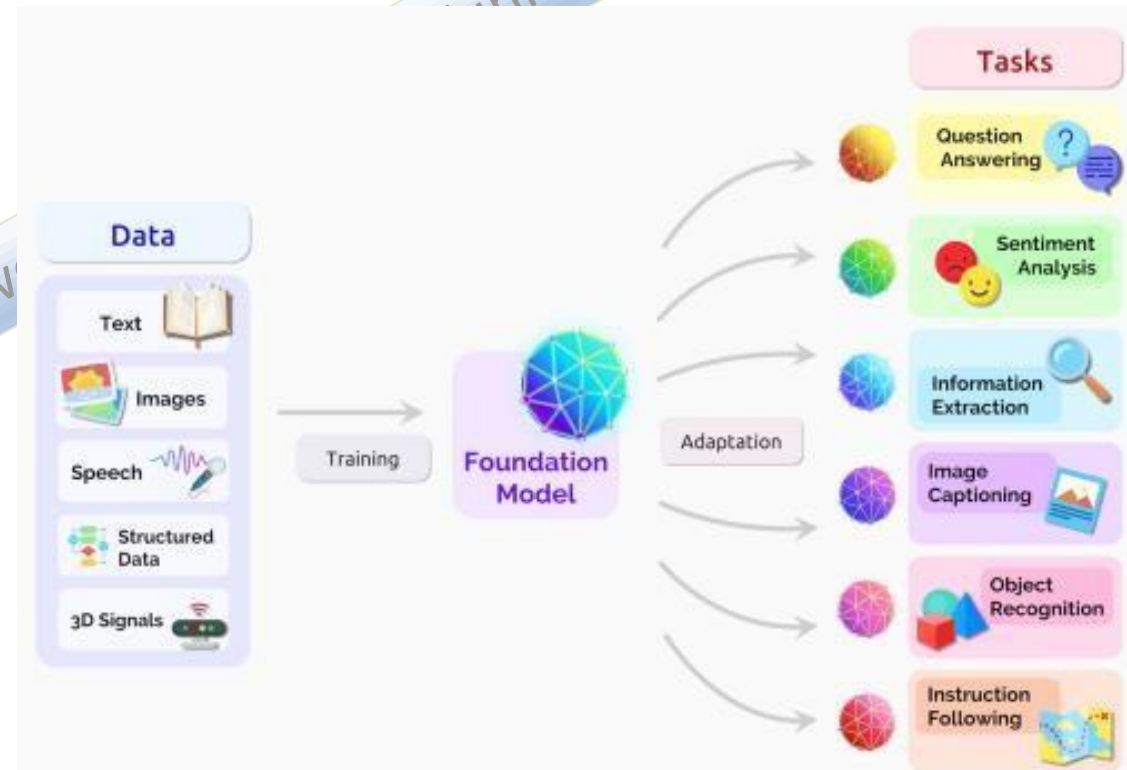
Generative AI – Market is exploding!

The Generative AI Market Map v3

A work in progress



Empower



Innovation

Generative AI - Definition

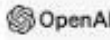









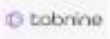


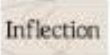










- Generative AI refers to AI solutions that generate content—whether it's a demand generation email, a fantastic landscape, or a dynamic chatbot reply—in response to a user prompt
- Solutions built using these technologies, such as ChatGPT, Stable Diffusion, and Dall-E, are making headlines every day, and organizations everywhere are seeking ways to operationalize them and capture their game-changing value
- Generative AI is trained on sets of unstructured data using transformer models that require data scientists and developers to fine-tune the output or experience that their business needs
- Organizations looking to apply generative AI to their business challenges have the option to train models from scratch or select a pretrained model that can be fine-tuned to the needs of their business
- Generative AI is built on and deployed in conjunction with language AI and natural language processing (NLP), which allow AI to process and understand human language. Together, generative AI and NLP can understand a user prompt to generate an appropriate response, whether it's text, video, imagery, or audio

Generative AI - <https://genai.works/>

Text	ChatGPT OpenAI ChatGPT is fine-tuned from GPT-3.5, a language model trained to produce text. Free	GPT-4 OpenAI GPT-4 is OpenAI's most advanced system, producing safer and more useful responses. Free	Notion AI Notion A connected assistant which produces text responses based on user's questions and the current page context. Free
Image	Stable Diffusion Stability AI A deep learning, text-to-image model, used to generate detailed images, conditioned on text descriptions. Free	Image Creator Microsoft Bing Image Creator generates AI images based on your text. Free	Midjourney Midjourney Midjourney is an AI program which generates images from natural language descriptions/prompts. Free
Video	Runway Runway A multi-modal AI system that can generate novel videos with text, images, or video clips. Free	Fliki Fliki Fliki is a text to video and text to speech creator powered by generative AI. Free	Wonder... Wonder Studio An AI tool that automatically animates, lights and composes CG characters into a live-action scene. Free
Audio	JukeBox OpenAI A neural net that generates music, including rudimentary singing, as raw audio in a variety of genres and artist... Free	MusicLM Google Research MusicLM is an AI model that can generate high-fidelity music from text. Free	Text to Speech Microsoft Azure AI voice generators to speak naturally using synthesized speech from input text. Free
3D	Point-E OpenAI A system for generating 3D point clouds from complex prompts. Free	Magic3D NVIDIA Magic3D is a new text-to-3D content creation tool that creates 3D mesh models with unprecedented quality. Free	Imagine 3D Luma AI An early experiment to prototype and create 3D with text. Free
Coding	Copilot X Github An AI pair programmer with an early adoption of OpenAI's GPT-4. Free	Codex OpenAI An AI AI system that translates natural language to code. Free	CodeWhisperer Amazon CodeWhisperer can generate code suggestions ranging from snippets to full functions based on your comments and... Free

Text	Perplexity AI perplexity.ai Perplexity AI is an answer engine that aims to deliver accurate answers to questions using large language models. Free	Humata humata.ai Humata is ChatGPT for your files. Ask questions about your data (i.e. technical paper, report) and get instant answers... Free	Tripnotes AI tripnotes.ai Tripnotes is an intelligent travel planner that uses a custom recommendation engine, their own data, and GPT to help... Free
Image	Segment... Meta AI A new AI model from Meta AI that can "cut out" any object, in any image, with a single click. Free	Cariyon V2 Cariyon AI model that can draw images from any text prompt. Free	PhotoRoom PhotoRoom AI to generate a infinite choice of backgrounds from the description you made. Free
Video	HeyGen HeyGen HeyGen is a video platform that help you create engaging business videos with generative AI. Free	Pictory Pictory Automatically create short, highly-shareable branded videos from your long form content. Free	Vidyo AI Vidyo.ai AI platform helps you convert long form podcasts and videos automatically into shorter shareable clips for Tik Tok, Reels... Free
Audio	Koe Recast koe.ai Voice transformation tool that allows users to change their voice into different styles such as a narrator, female, or... Free	Soundraw Soundraw AI music generator that allows you to create and compose original, royalty-free music. Free	Krisp Krisp AI-powered Voice Clarity and Meeting Assistant that eliminate all background noise with a single click. Free
3D	Magic3D NVIDIA Magic3D is a new text-to-3D content creation tool that creates 3D mesh models with unprecedented quality. Free	Spline AI Spline Design Generate 3d objects, animations, and textures using prompts. Free	Unity AI Unity AI ecosystem that will put AI-powered game-development tools in the hands of millions of creators. Soon they'll be able... Free
Coding	Ask Codi askcodi AskCodi simplifies your development process by giving you the power to create prototypes and applications faster. Free	Safurai Safurai Safurai is the AI Coding Assistant that saves you time in changing, optimizing, and searching code. Free	FlutterFlow AI... Flutterflow IO AI-powered tool that can generate a app design with code from text or prompt. Free

Generative AI - Companies

 OpenAI	▼	 Cohere	▼	 Anthropic	▼
 Hugging Face	▼	 Synthesia	▼	 Jasper AI	▼
 Glean	▼	 IBM	▼	 DeepMind	▼
 AI21 Labs	▼	 Tabnine	▼	 Meta	▼
 Inworld AI	▼	 Inflection AI	▼	 Adobe	▼
 AQEMIA	▼	 Syntheticaic	▼	 Rephrase House	▼
 MOSTLY AI	▼	 Tavus	▼	 Aleph Alpha	▼
 Copy.ai	▼	 Kinetix	▼	 Aimi	▼

Generative AI – Applications



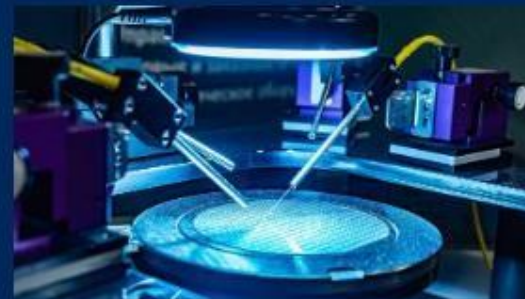
Healthcare and Life Sciences

Accelerate research and patient outcomes with faster, more accurate analysis across precision medicine, medical imaging, lab automation, and more.



Finance

Mitigate risk, identify fraud, automate processes, and reveal business optimization opportunities before your competition does.



Manufacturing

Transform data into insights that help you optimize plant performance, minimize downtime, improve safety, and drive profitability.



Media and Entertainment

Streamline workflows and bring amazing experiences to life more easily while you reduce costs, simplify archiving, and better understand viewer preferences.



Energy

Tap into connected operations data to balance supply and demand, enable predictive maintenance, identify issues, and discover usage trends.



Retail

Harness your ever-growing data to gain real-time understanding of customer behavior, inventory, internal loss, and other critical metrics.



Telecommunications

Improve efficiency while rapidly revealing opportunities for cost optimization, service enhancement, or new technologies such as AR and VR.



Government

Securely harness AI resources to unlock new possibilities, from scientific research to defense, mapping, and disaster response.

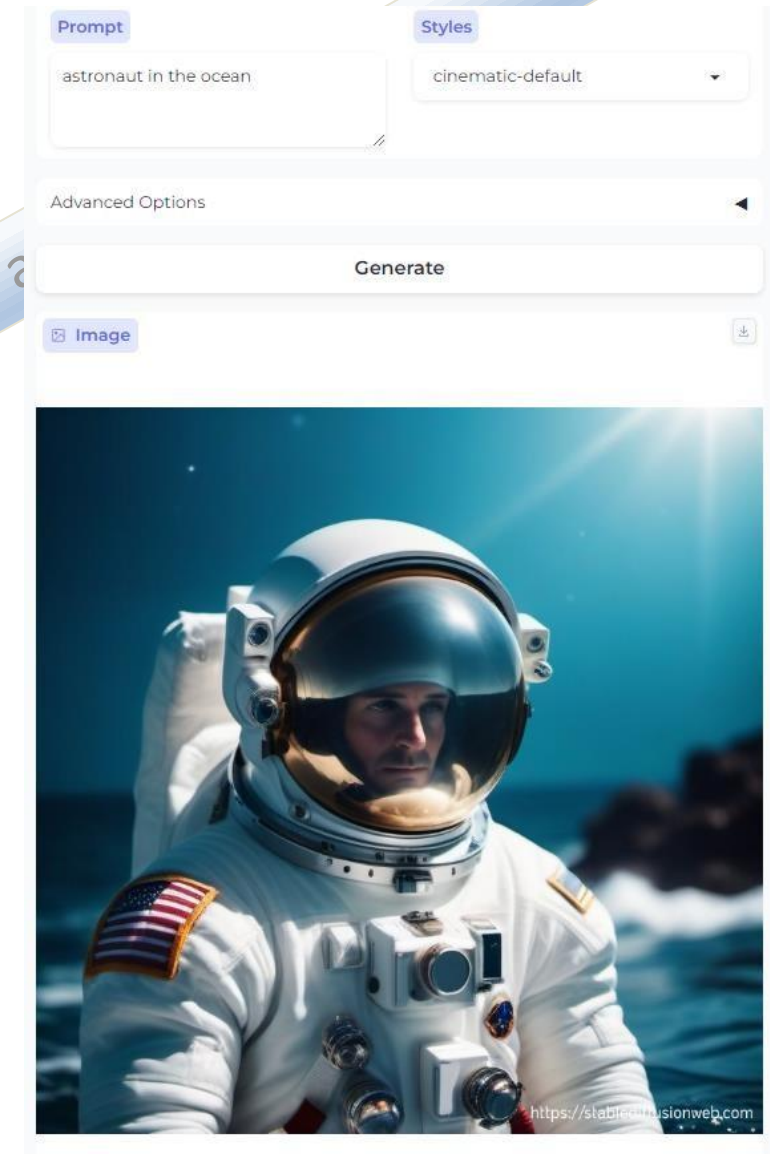
<https://www.intel.com/content/www/us/en/artificial-intelligence/industries.html>

Agenda - Part 1

- What is Generative AI and Large Language Models (LLMs)
- Use Cases
- Prompt Engineering
- Transformer Architectures
- LLM fine tuning
- PEFT (Parameter Efficient Finetuning)
- RLHF (Reinforcement Learning with Human Feedback)
- Retrieval Augmented Generation (RAG)
- LLM deployment techniques
- Q and A

What is Generative AI

- Informally
 - **Generative** models can generate new data instances
 - **Discriminative** models discriminate between different kinds of data instances
- Formally, given a set of data instances X and a set of labels Y :
 - **Generative** models capture the joint probability $p(X, Y)$, or just $p(X)$ if there are no labels
 - **Discriminative** models capture the conditional probability $p(Y | X)$
- Generative artificial intelligence (AI) describes algorithms (such as ChatGPT and DALL-E) that can be used to create new content, including audio, code, images, text, simulations, and videos

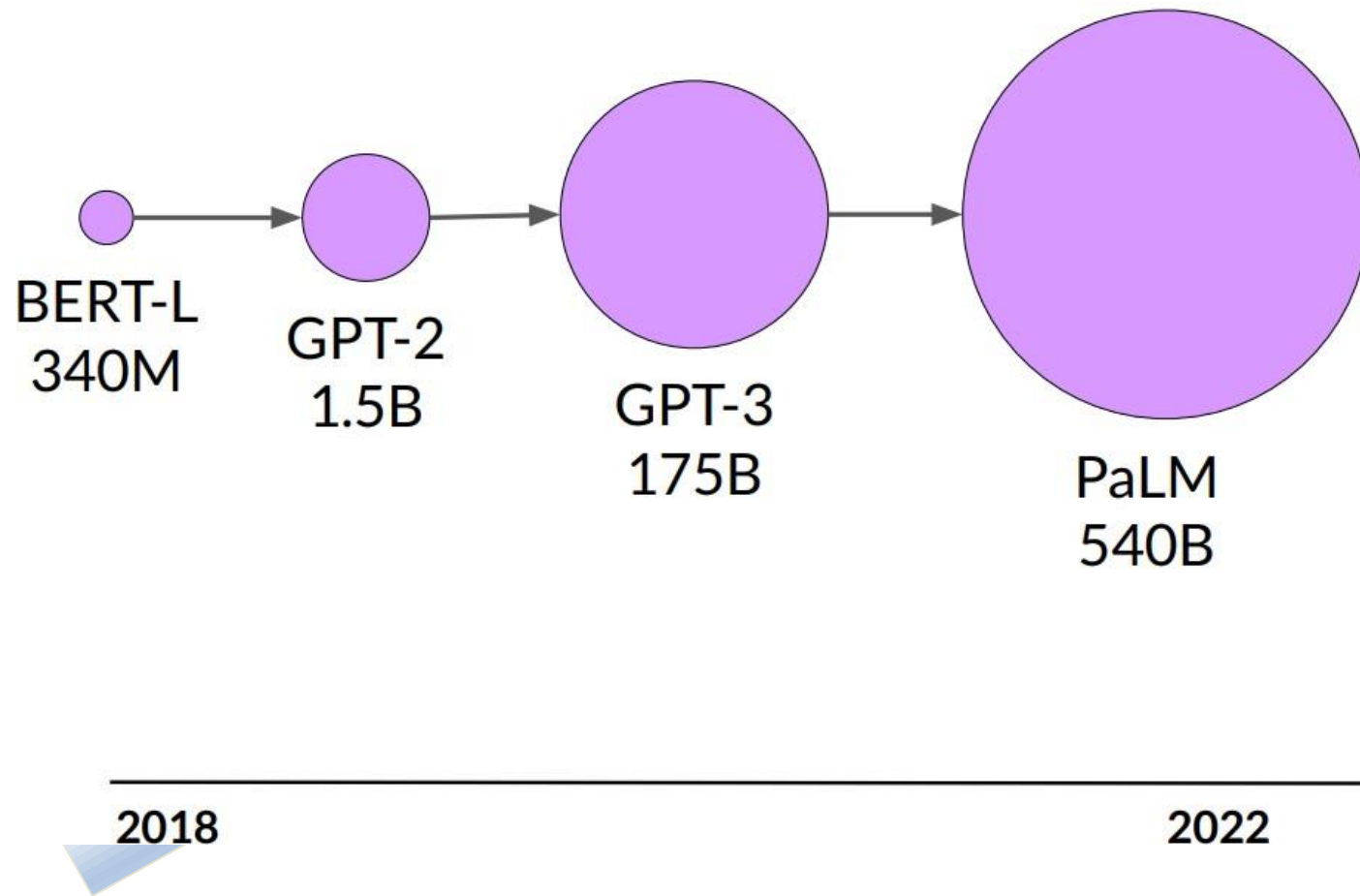


What is Generative AI and LLMs

- Estimating the probability of a token or sequence of tokens occurring within a longer sequence of tokens
 - When I hear rain on my roof, I _____ in my kitchen.
 - cook soup 9.4%
 - warm up a kettle 5.2%
 - relax 2.2%
- "Large" in Large Language Models (LLMs) can refer either to the number of parameters in the model, or the number of words in the dataset

Intel India Education Conclave: Empowering educators and innovation

Model Size vs Time



- Growth powered by:
- Introduction of transformer
 - Access to massive datasets
 - More powerful compute resources

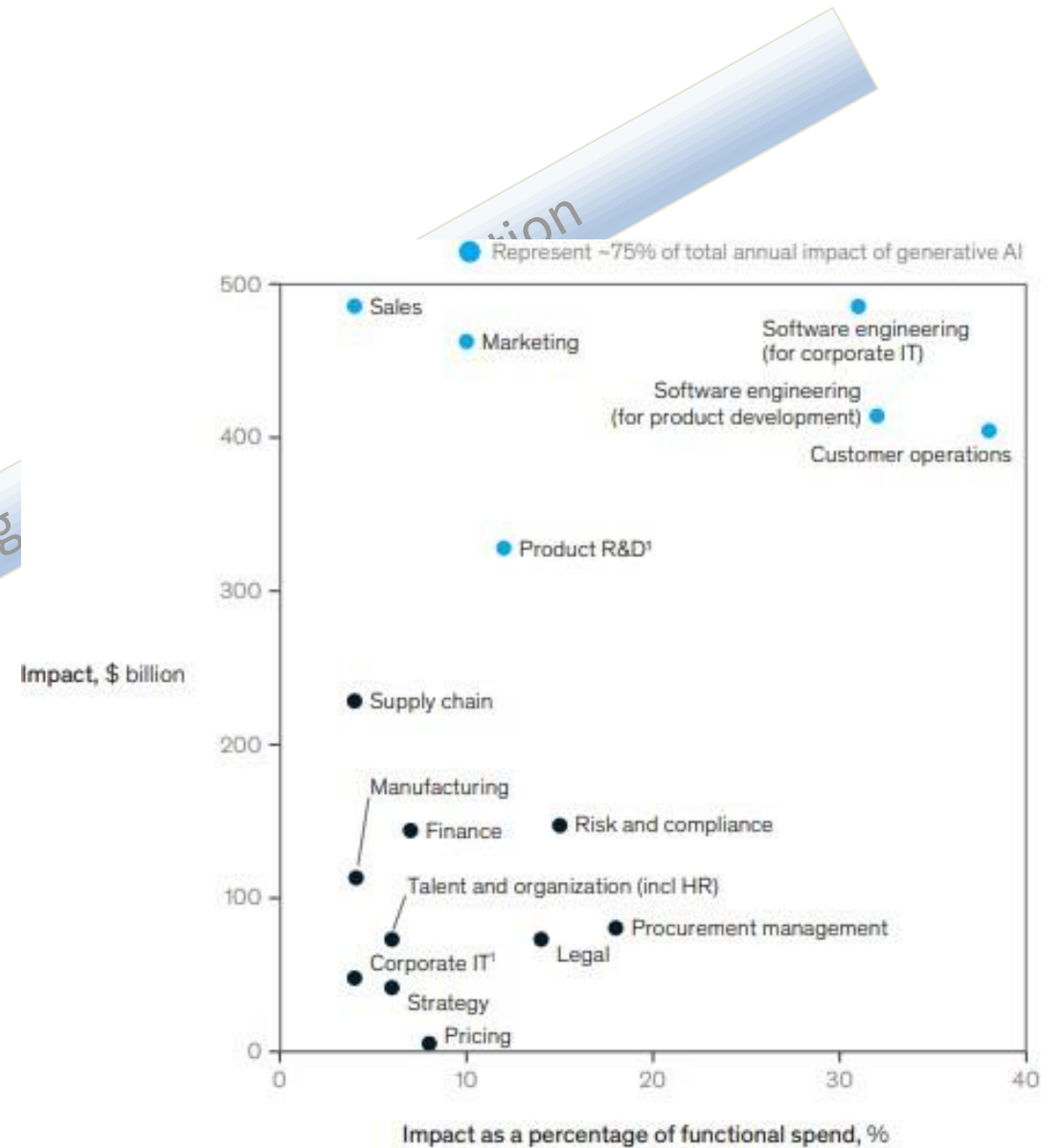
Source - <https://www.deeplearning.ai/courses/generative-ai-with-llms/>

Generative AI – Use Cases

- **Content generation:** Automatically create articles, blog posts, product descriptions, and other written materials
- **Chatbots:** Power dynamic and intelligent conversational AI models that your customers can interact with through text or speech
- **Image, video, and audio generation:** Create new visuals and sounds by examining preexisting materials and working against a user prompt
- **Language translation:** Translate text from one language to another
- **Data augmentation:** Create synthetic data for other machine learning models to help improve their accuracy and performance
- **Text summarization:** Summarize large pieces of text into a concise format so readers can quickly understand the main points and ideas

Generative AI Global Impact

- In a McKinsey study, they believe that Generative AI can add an equivalent of \$2.6 to \$4.4 trillion annually to the global economy. This is more than the entire GDP of the United Kingdom
- 75% of the use cases fall under the four areas of Marketing and sales, software engineering, R&D, and Customer Operations
- They identified 63 use cases total include retail and packaging, and banking
- Source: <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#business-value>



Generative AI LLM tasks

Translate

French:
J'aime l'apprentissage automatique.

German:
Ich liebe maschinelles Lernen.

Generate

Code AI

Prompt:
Write some python code that will return the mean of every column in a dataframe.

Generate


Code:

```
import pandas as pd

df = pd.DataFrame({
    'A': [1, 2, 3, 4, 5],
    'B': [2, 3, 4, 5, 6],
    'C': [3, 4, 5, 6, 7]
})

mean_values = df.mean()
```

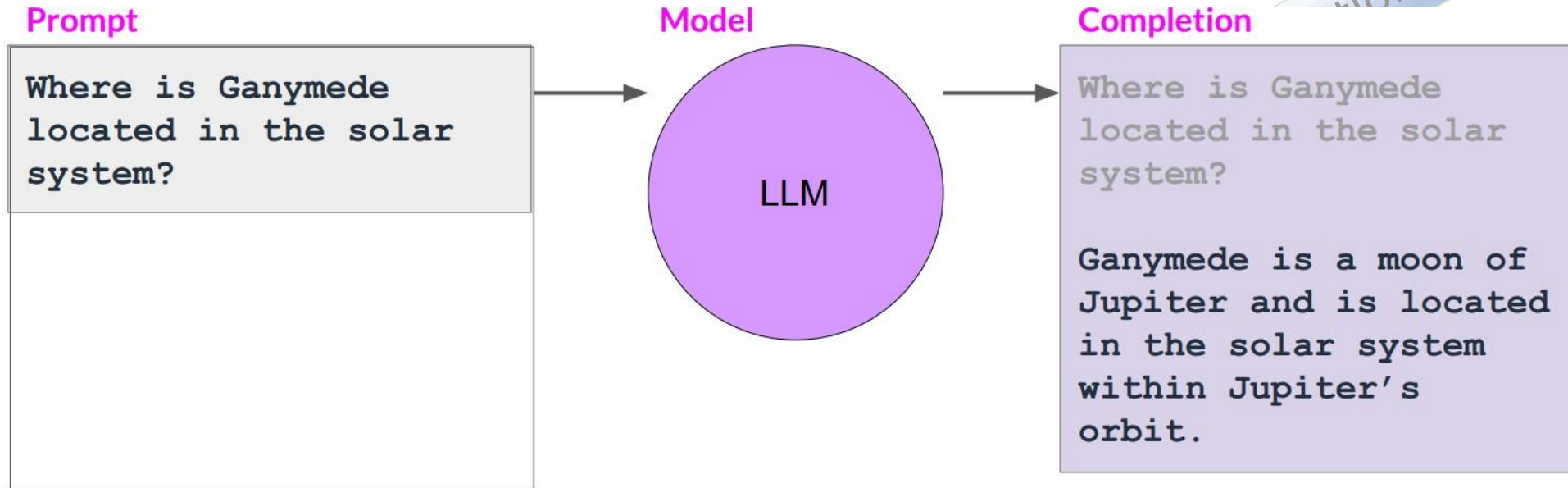
Summarize

Text file:  support.txt

Generate

In the chat session, Support efficiently and effectively assists Alex, who was initially unable to access their account due to issues with a password reset email, leading to a positive customer service experience.

Prompts and Completions



Context window: typically a few thousand words

Writing well structured prompts is an essential part of ensuring accurate, high-quality responses from a language model

Source - <https://www.deeplearning.ai/courses/generative-ai-with-llms/>

Prompt Engineering (In-context learning during inference)



Source - <https://www.deeplearning.ai/courses/generative-ai-with-llms/>

Chain-of-Thought Prompting

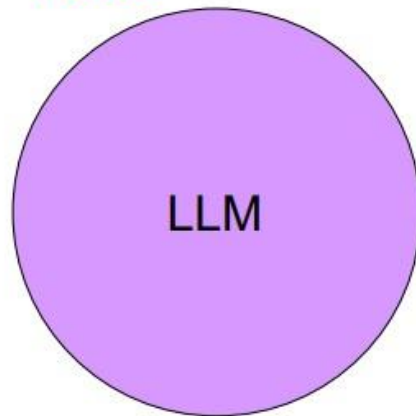
Prompt

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model



Completion

Q: Roger has 5 tennis balls.
...
...
...
how many apples do they have?

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✓

Generative AI - Inference

Evolution

Enter your prompt here...

Max new tokens

Sample top K

Sample top P

Temperature

Submit

Inference configuration parameters

$$\sigma(z_i) = \frac{e^{\frac{z_i}{\theta}}}{\sum_{j=0}^N e^{\frac{z_j}{\theta}}}$$

Source - <https://lukesalamone.github.io/posts/what-is-temperature/>

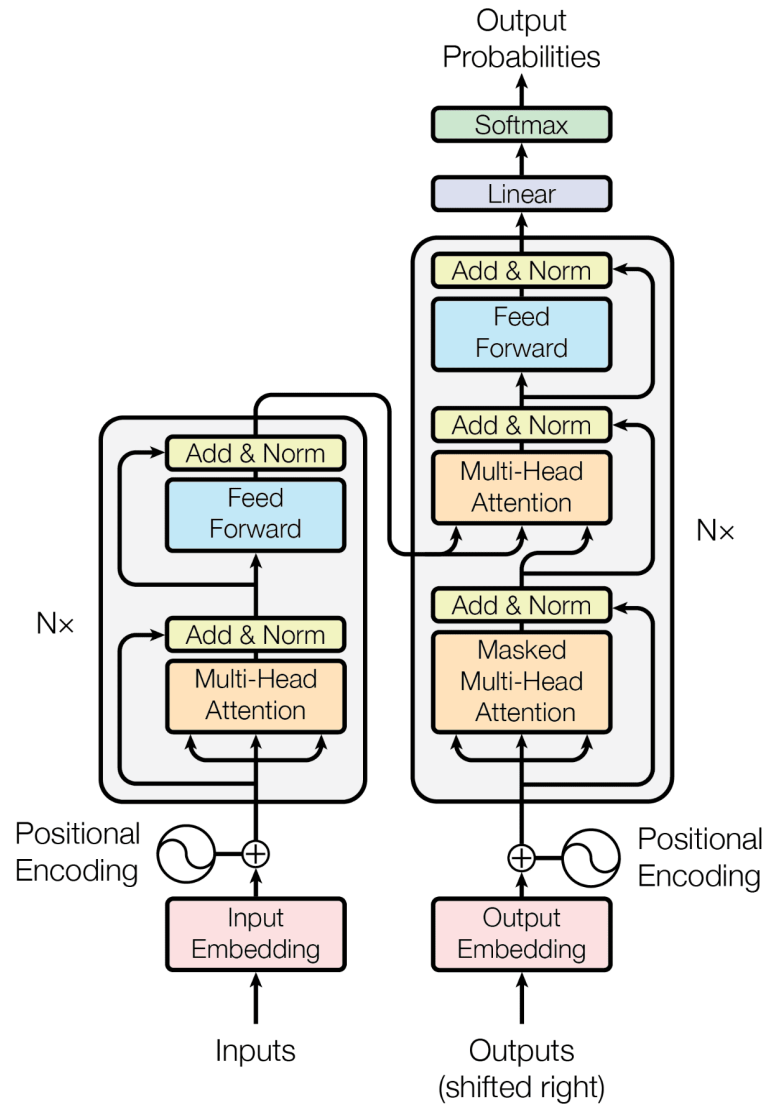
Transformer – Attention Is All you Need

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

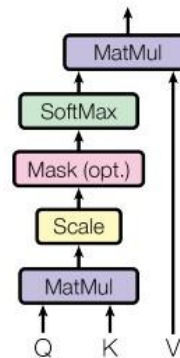
$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$

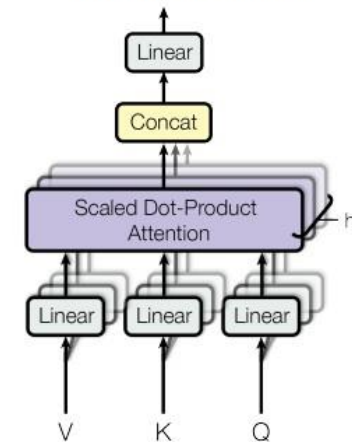
Empowering educators and learners



Scaled Dot-Product Attention

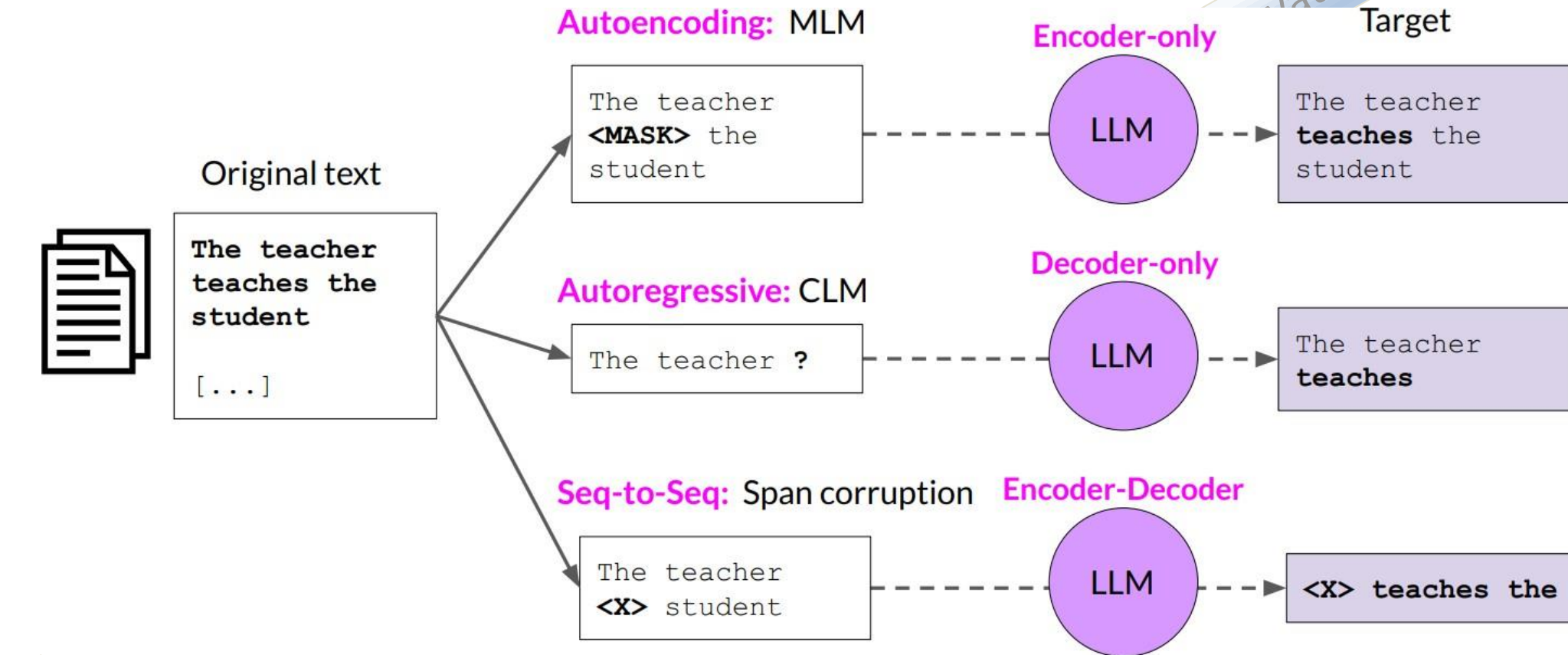


Multi-Head Attention



Source - <https://arxiv.org/abs/1706.03762>

Model Architectures and Training Objectives



Source - <https://www.deeplearning.ai/courses/generative-ai-with-llms/>

Computational Challenges for training LLMs

1 billion parameters model

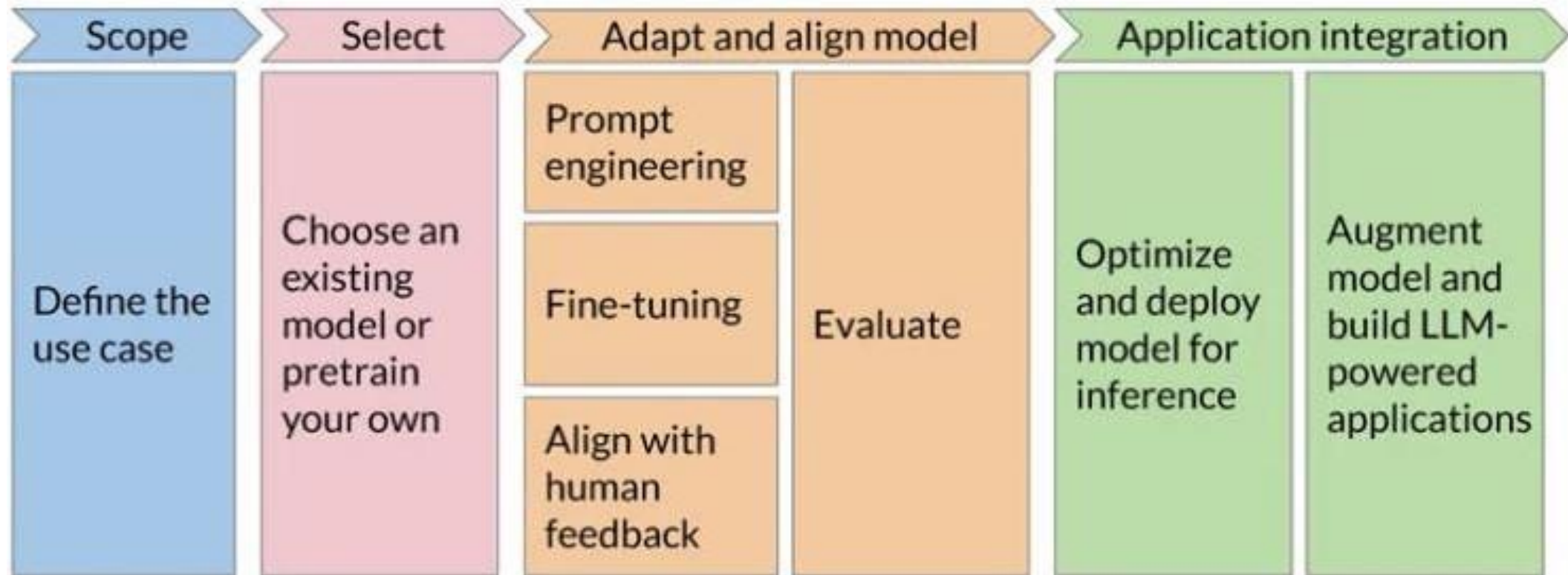
	Bytes per parameter
Model Parameters (Weights)	4 bytes per parameter
Adam optimizer (2 states)	+8 bytes per parameter
Gradients	+4 bytes per parameter
Activations and temp memory (variable size)	+8 bytes per parameter (high-end estimate)
TOTAL	=4 bytes per parameter +20 extra bytes per parameter

~20 extra bytes per parameter

Memory needed to train model

80GB @ 32-bit full precision

Generative AI – Project Lifecycle



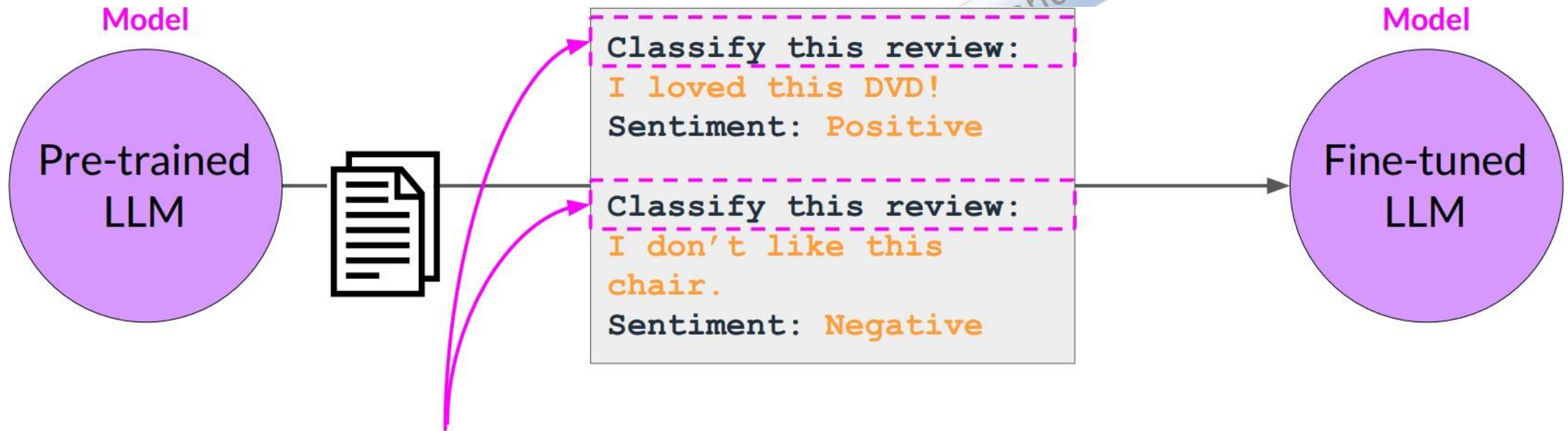
Source - <https://www.deeplearning.ai/courses/generative-ai-with-llms/>

Why finetune a model?



- Steers the model to more consistent outputs
- Reduces hallucinations
- Customizes the model to a specific use case
- Process is similar to the model's earlier training

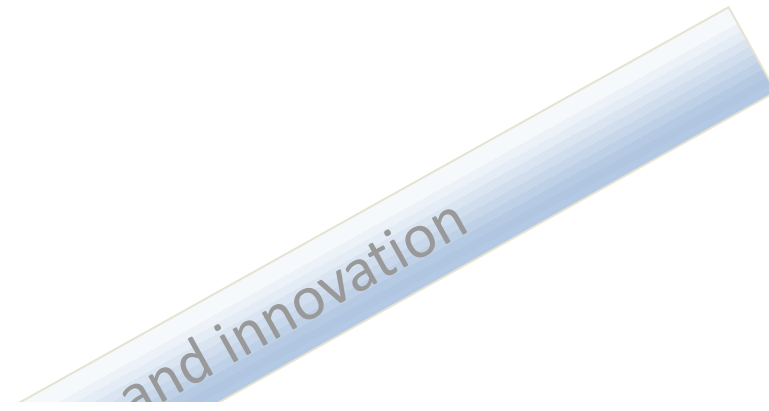
LLM Instruction fine tuning



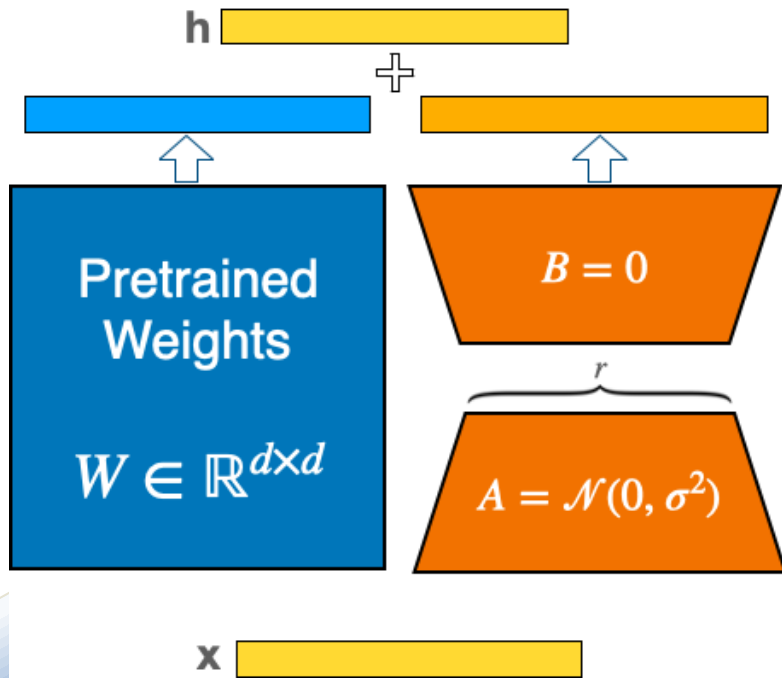
Each prompt/completion pair includes a specific "instruction" to the LLM

PEFT LoRA

Re parameterize model weights using a low-rank representation



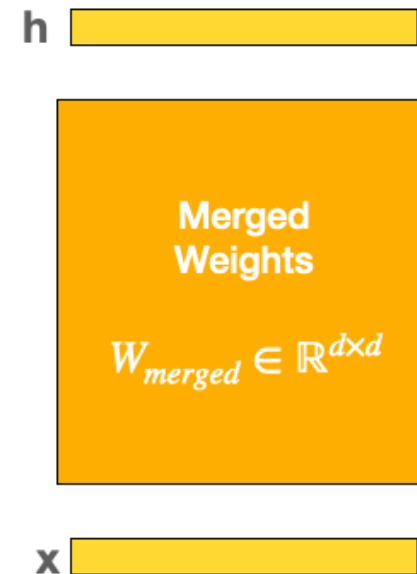
During training



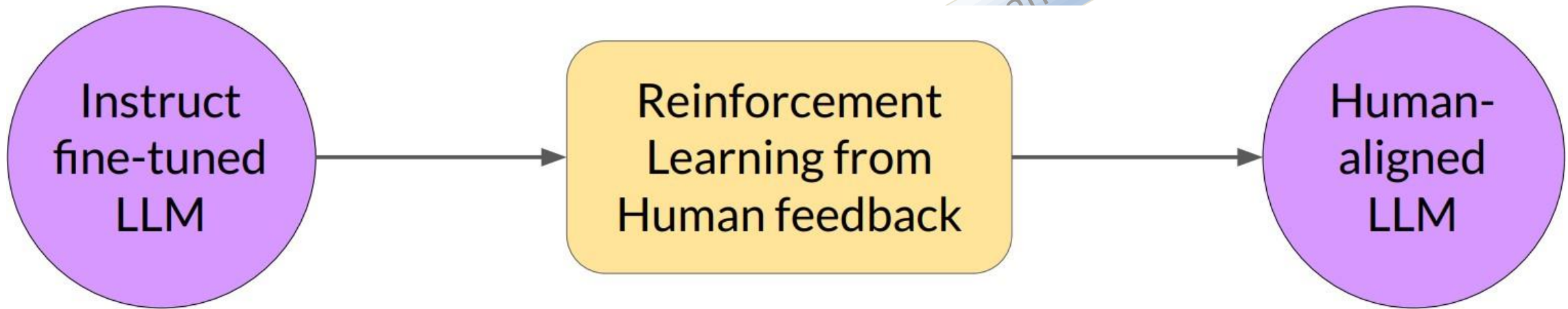
$$h = Wx + BAx$$

$$h = \underbrace{(W + BA)}_{W_{merged}}x$$

After training



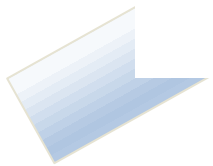
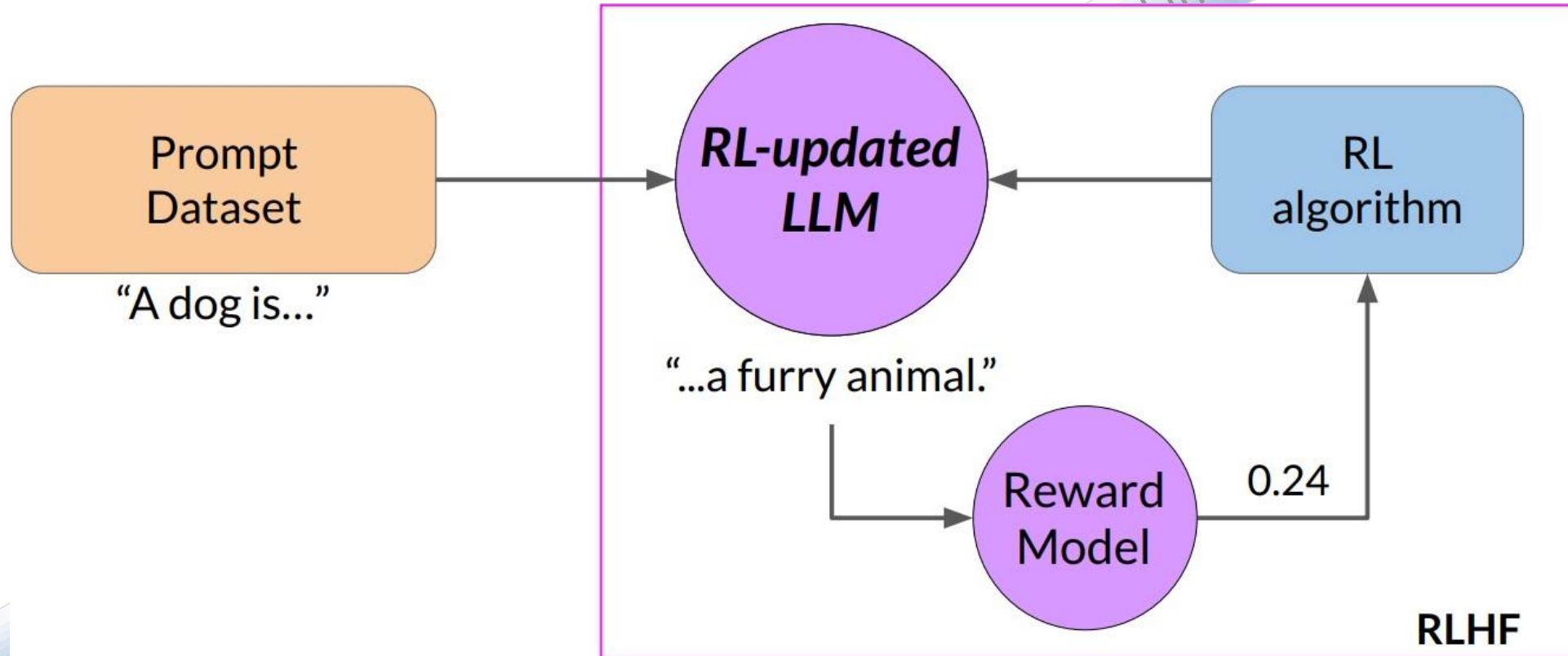
Reinforcement Learning from Human feedback (RLHF)



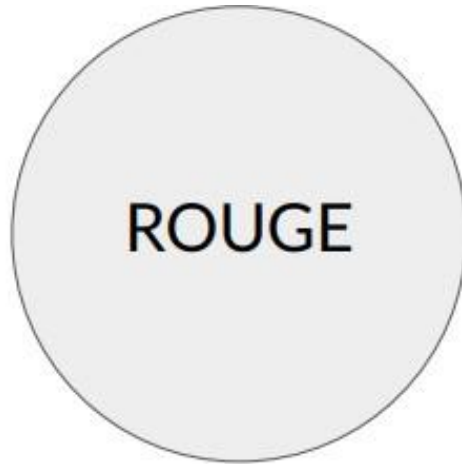
- Maximize helpfulness, relevance
- Minimize harm
- Avoid dangerous topics

Use reward model to fine-tune LLM

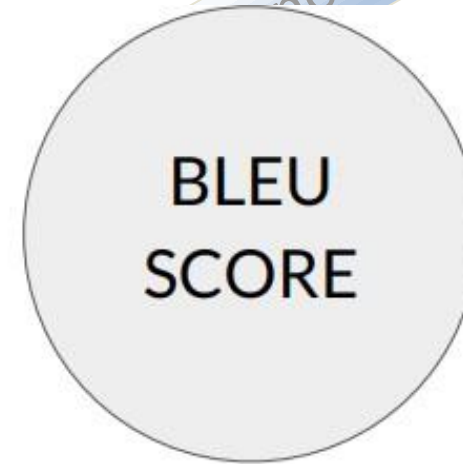
Innovation



LLM Evaluation Metrics



- Used for text summarization
- Compares a summary to one or more reference summaries



- Used for text translation
- Compares to human-generated translations

LLM Evaluation Metrics – ROUGE-1

innovation

Reference (human):

It is cold outside.

Generated output:

It is very cold outside.

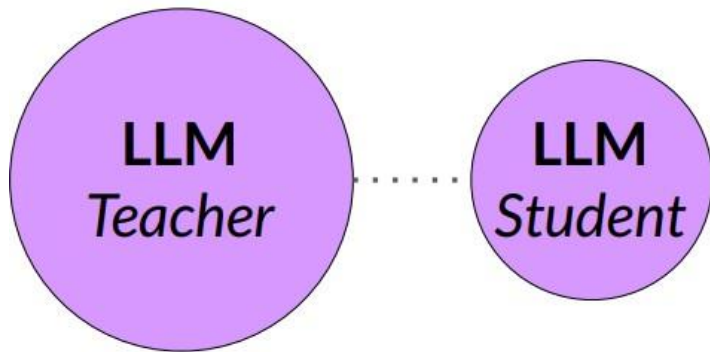
$$\text{ROUGE-1 Recall} = \frac{\text{unigram matches}}{\text{unigrams in reference}} = \frac{4}{4} = 1.0$$

$$\text{ROUGE-1 Precision} = \frac{\text{unigram matches}}{\text{unigrams in output}} = \frac{4}{5} = 0.8$$

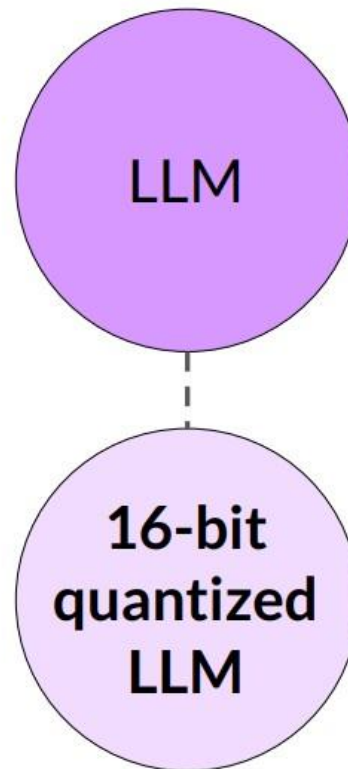
$$\text{ROUGE-1 F1} = 2 \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = 2 \frac{0.8}{1.8} = 0.89$$

LLM Optimization Techniques for deployment/inference

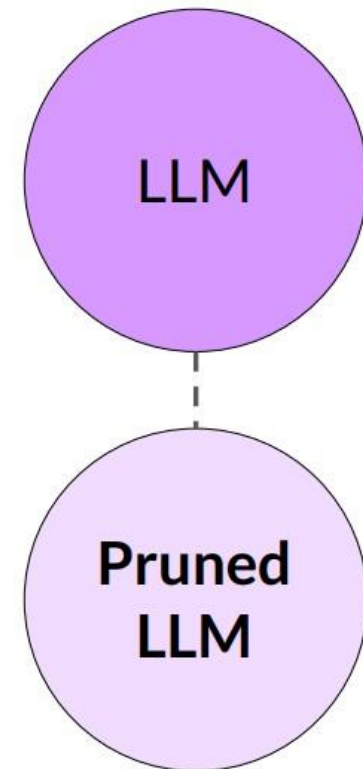
Distillation



Quantization

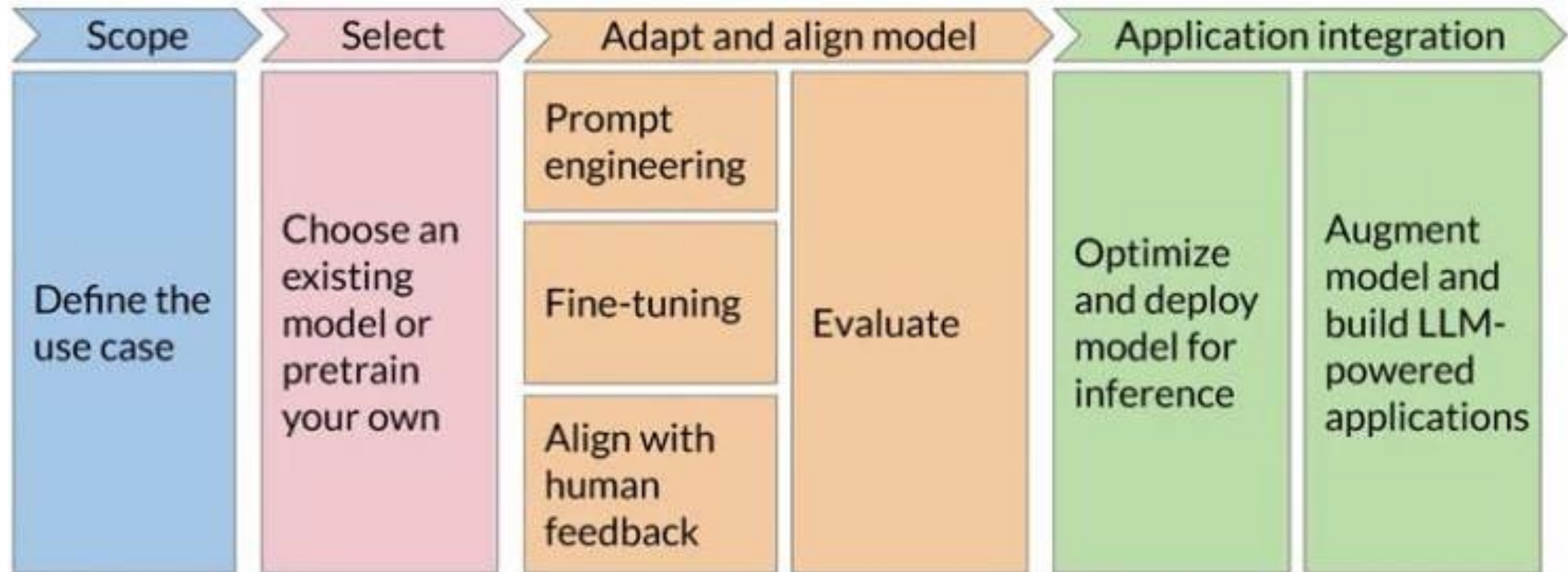


Pruning



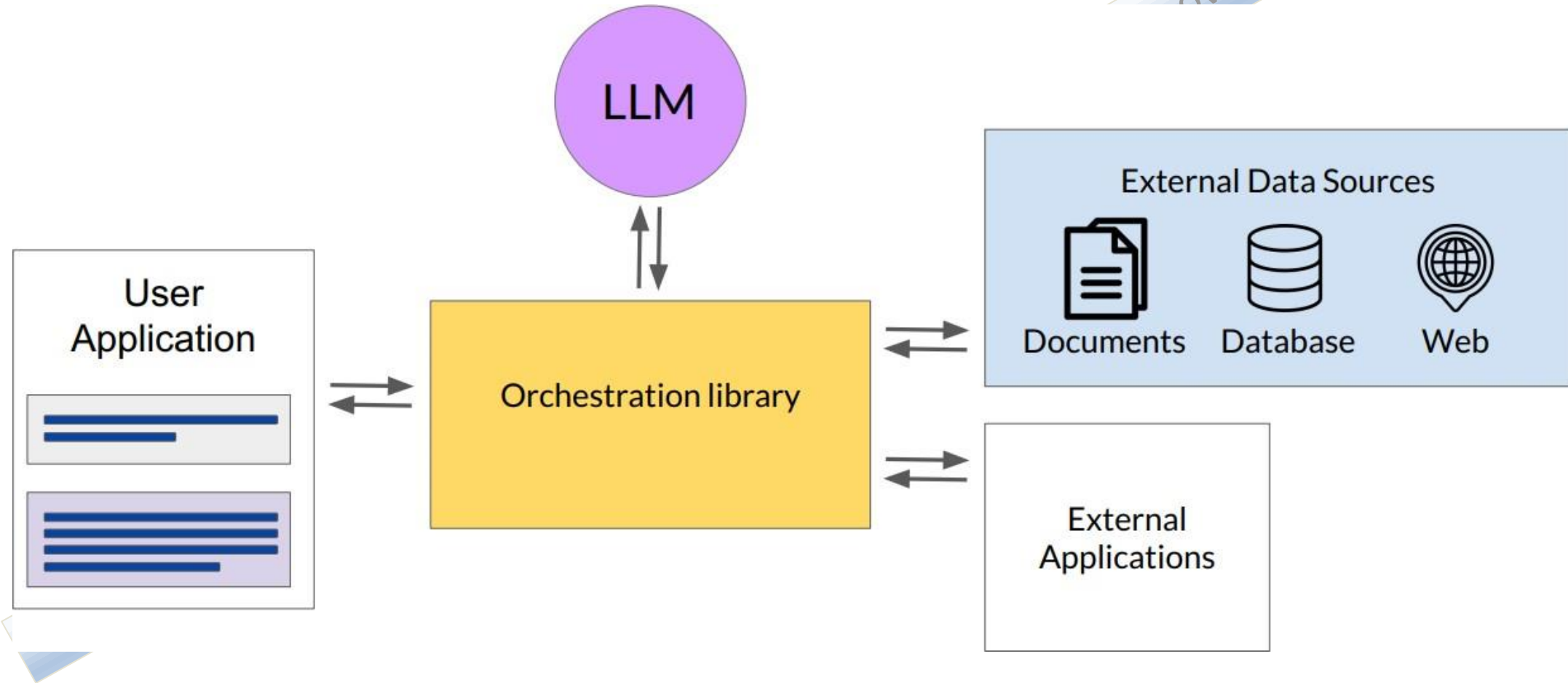
Source - <https://www.deeplearning.ai/courses/generative-ai-with-llms/>

Generative AI – Project Lifecycle



Source - <https://www.deeplearning.ai/courses/generative-ai-with-llms/>

Retrieval Augmented Generation (RAG)



LLM Lifecycle summary

	Pre-training	Prompt engineering	Prompt tuning and fine-tuning	Reinforcement learning/human feedback	Compression/optimization/deployment
Training duration	Days to weeks to months	Not required	Minutes to hours	Minutes to hours similar to fine-tuning	Minutes to hours
Customization	<p>Determine model architecture, size and tokenizer.</p> <p>Choose vocabulary size and # of tokens for input/context</p> <p>Large amount of domain training data</p>	<p>No model weights</p> <p>Only prompt customization</p>	<p>Tune for specific tasks</p> <p>Add domain-specific data</p> <p>Update LLM model or adapter weights</p>	<p>Need separate reward model to align with human goals (helpful, honest, harmless)</p> <p>Update LLM model or adapter weights</p>	<p>Reduce model size through model pruning, weight quantization, distillation</p> <p>Smaller size, faster inference</p>
Objective	Next-token prediction	Increase task performance	Increase task performance	Increase alignment with human preferences	Increase inference performance
Expertise	High	Low	Medium	Medium-High	Medium

Source - <https://www.deeplearning.ai/courses/generative-ai-with-llms/>

References

- <https://developers.google.com/machine-learning/resources/intro-llms>
- What is a Generative Model? <https://developers.google.com/machine-learning/gan/generative>
- What is generative AI? <https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai>
- Prompt Engineering for Generative AI <https://developers.google.com/machine-learning/resources/prompt-eng>
- Parameter-efficient fine-tuning of large-scale pre-trained language models <https://www.nature.com/articles/s42256-023-00626-4>
- Parameter-Efficient Fine-Tuning of Large Language Models with LoRA and QLoRA <https://www.analyticsvidhya.com/blog/2023/08/lora-and-qlora/>
- Attention is All You Need: <https://research.google/pubs/pub46201/>
- Solving a machine-learning mystery: <https://news.mit.edu/2023/large-language-models-in-context-learning-0207>
- <https://medium.com/beaucoupdata/generative-ai-101-whats-up-with-chatgpt-1bfbb3f04ed>
- <https://www.deeplearning.ai/short-courses/>
- <https://learnprompting.org/docs/intro>
- What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization <https://arxiv.org/pdf/2204.05832.pdf>
- Language Models are Few-Shot Learners <https://arxiv.org/pdf/2005.14165.pdf>
- Guide to Parameter-Efficient Fine-Tuning <https://arxiv.org/pdf/2303.15647.pdf>
- ROUGE - <https://aclanthology.org/W04-1013.pdf>
- FLAN T5 model - <https://blog.research.google/2021/10/introducing-flan-more-generalizable.html>
- Llama model - <https://ai.meta.com/blog/large-language-model-llama-meta-ai/>
- LLM prompting - <https://www.kaggle.com/competitions/llm-prompting-with-makersuite>

Agenda – Part 2

- Generative AI Models Training and Deployment @ Intel
- Intel AI Software Tools – OneAPI, OpenVINO
- Intel Hardware Offerings
- Habana Gaudi2 – Hardware/Software Overview
- Large Language Models on Gaudi2
- Hugging Face – Optimum Habana
- Gaudi2 Performance - MLPerf Results
- Demo – LLaVA Model <https://llava-vl.github.io/>
- How to Access Habana Gaudi2
- Q and A

Generative AI Models Training @Intel

- Intel offers a purpose-built portfolio of both hardware and software technologies that combine to help streamline the business initiative and accelerate ROI
- Our mission is to enable AI innovators to deploy AI anywhere it is needed—from the edge to the cloud and data center—with optimal performance, scalability, and cost
- Software Resources to Simplify Generative AI Training and Deployment
 - Intel offers developers and data scientists [a wide range of software tools and optimizations](#) that can help maximize performance and dramatically boost productivity both during training and deployment

Generative AI Training – Software Tools

- oneAPI unified programming language, [Intel® oneAPI Deep Neural Network Library](#) with highly optimized implementations of deep learning building blocks
 - The [oneAPI® unified programming model](#) can also be used to support heterogeneous hardware platforms with less effort from development teams
- [Intel® Extension for Transformers](#) is another critical tool that helps to accelerate transformer-based models on Intel® platforms
 - This toolkit features a seamless user experience for model compression, advanced software optimizations, a unique compression-aware runtime, and optimized model packages, including Stable Diffusion, GPT-J-6BM, and BLOOM-176B
- Accenture partnership offers a range of [reference kits](#) that can help kick-start the generative or language AI project

Intel AI Software Tools



Engineer Data

MODIN, SciPy, pandas, NumPy, Apache Spark, Numba

Data Analytics at Scale*

Create Models

dmlc XGBoost, PyTorch, ONNX RUNTIME, OpenVINO, learn, TensorFlow, Direct ML, SigOpt, Intel® Neural Compressor

Machine & Deep Learning Frameworks, Optimization and Deployment Tools*

1 oneAPI

Intel® oneAPI Deep Neural Network Library (oneDNN), Intel® oneAPI Collective Communications Library (oneCCL), Intel® oneAPI Math Kernel Library (oneMKL), Intel® oneAPI Data Analytics Library (oneDAL)

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



Intel AI ANALYTICS TOOLKIT

Accelerate End-to-End Data Science and AI

Intel® Developer Cloud, Intel® Developer Catalog

Try Latest Intel Tools and Hardware, and access optimized AI Models

cnvrg.io

Full stack ML Operating System

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

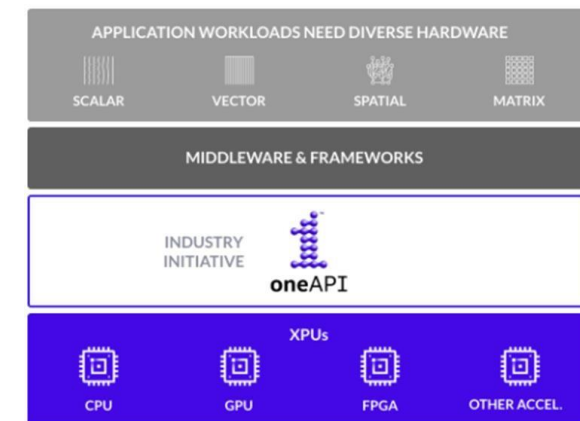
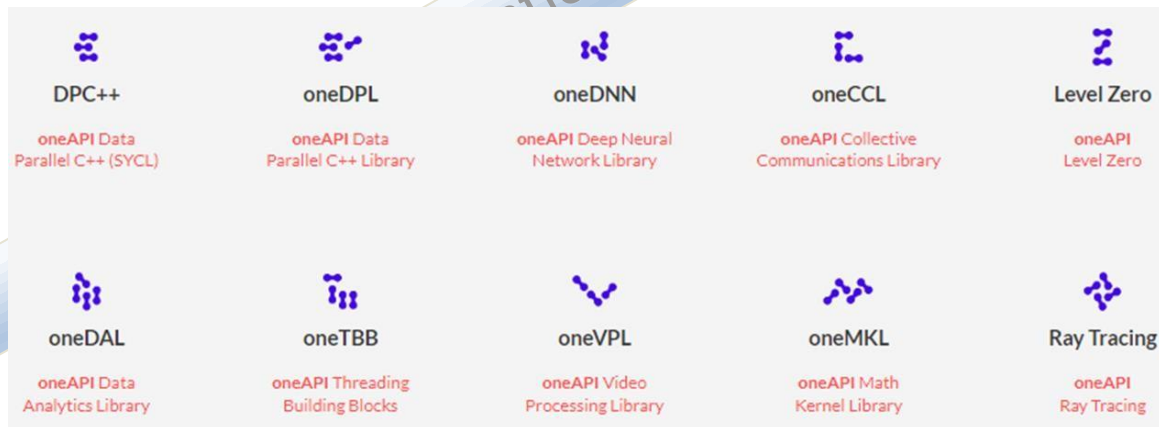
* This list includes popular open source frameworks that are optimized for Intel hardware

<https://cnvrg.io>

<https://www.intel.com/content/www/us/en/developer/topic-technology/artificial-intelligence/overview.html>

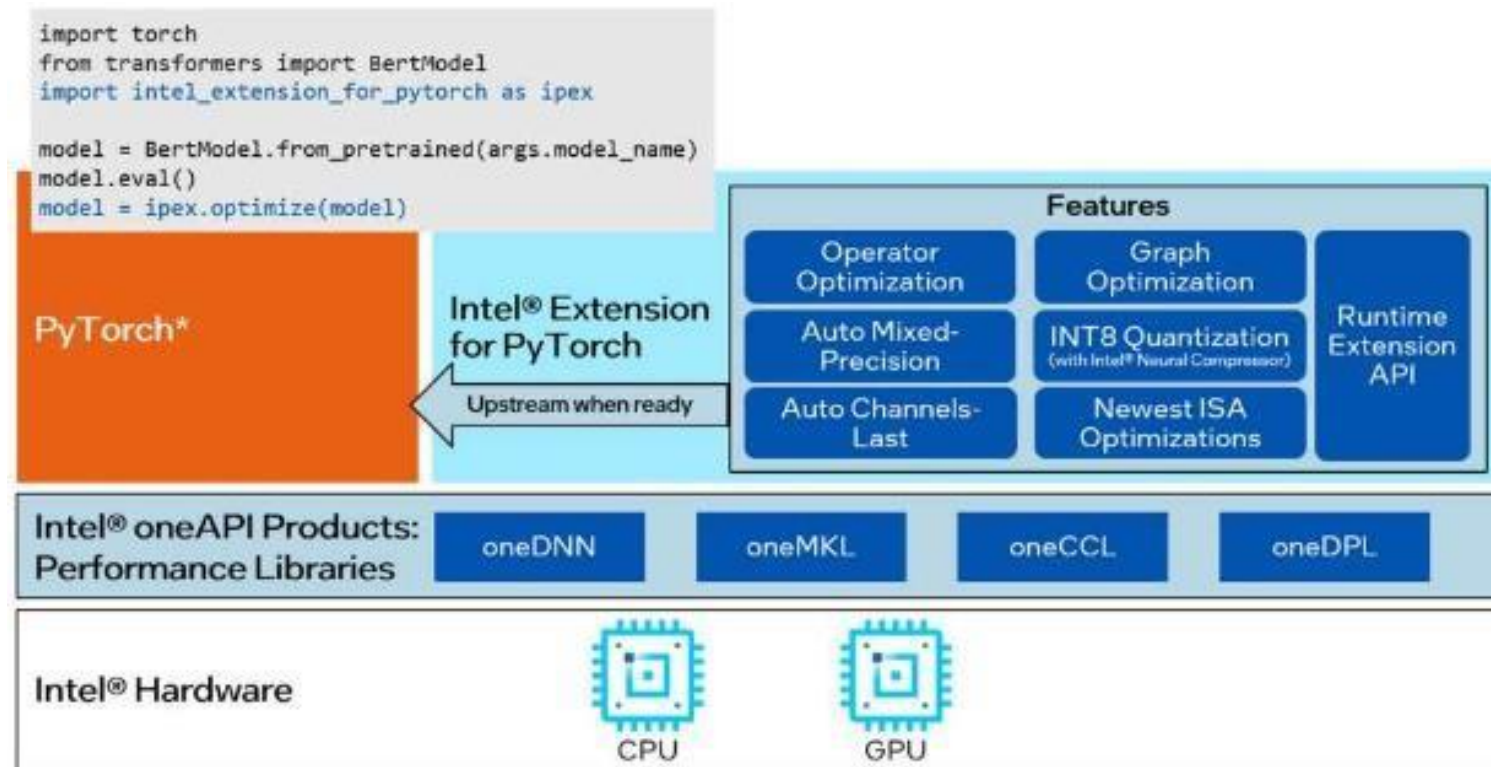


- oneAPI consists of a language and libraries for creating parallel applications
 - **SYCL**: oneAPI's core language for programming accelerators and multiprocessors. SYCL allows developers to reuse code across hardware targets (CPUs and accelerators such as GPUs and FPGAs) and tune for a specific architecture
 - **oneDPL**: A companion to the DPC++ Compiler for programming oneAPI devices with APIs from C++ standard library, Parallel STL, and extensions
 - **oneDNN**: High performance implementations of primitives for deep learning frameworks
 - **oneCCL**: Communication primitives for scaling deep learning frameworks across multiple devices
 - **Level Zero**: System interface for oneAPI languages and libraries
 - **oneDAL**: Algorithms for accelerated data science/Analytics
 - **oneTBB**: Library for adding thread-based parallelism to complex applications on multiprocessors
 - **oneVPL**: Algorithms for accelerated video processing
 - **oneMKL**: High performance math routines for science, engineering, and financial applications
 - **Ray Tracing**: A set of advanced ray tracing and high-fidelity rendering and computation routines for use in a wide variety of 3D graphics uses including, film and television photorealistic visual effects and animation rendering, scientific visualization, high-performance computing computations, gaming, and more



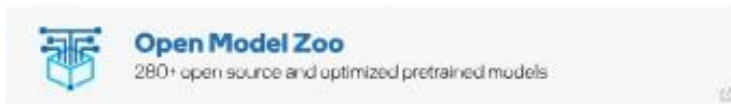
PyTorch Optimizations from Intel

- Intel releases its newest optimizations and features in Intel Extension for PyTorch before upstreaming them into open source PyTorch



<https://www.intel.com/content/www/us/en/developer/tools/oneapi/optimization-for-pytorch.html>

OpenVINO Toolkit



Model Optimizer
Converts and optimizes trained model using a supported framework

Read, load, infer

OpenVINO format (intermediate representation file) (.xml, .bin)

Post-Training Quantization with NNCF
Reduces model size into low precision without retraining

Intel® Developer Cloud for the Edge
Test within your own sandbox

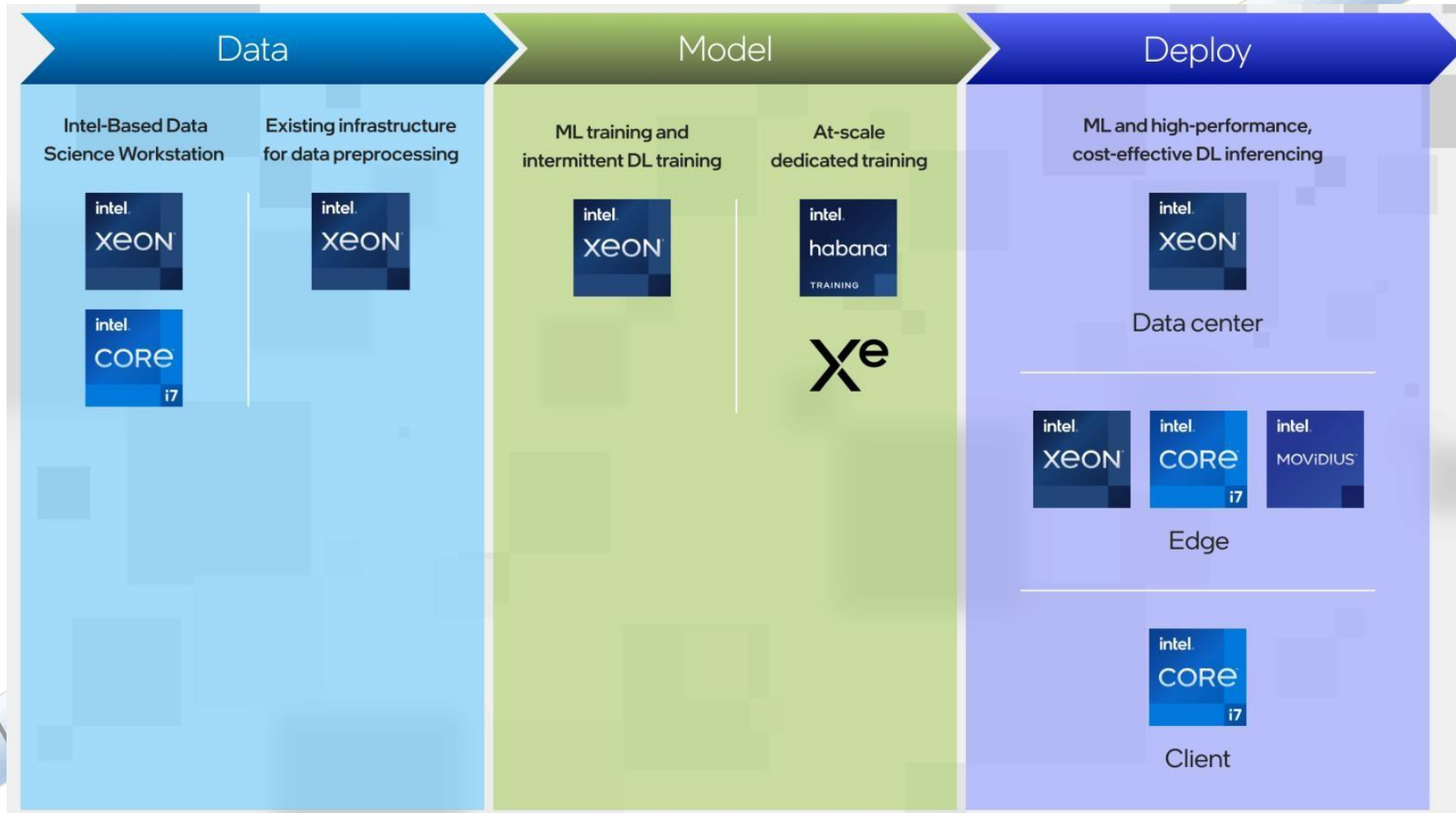
OpenVINO Model Server
Serve models over gRPC, REST, or C API endpoints

OpenVINO Runtime
Common Python, C and C++ APIs that abstracts low-level programming for each device below

intel XEON intel CORE intel GPU intel ARC GRAPHICS
intel ATOM intel MOVIDIUS intel IRIS XE MAX GRAPHICS intel FPGA AI Suite

<https://www.intel.com/content/www/us/en/developer/tools/opencvino-toolkit/overview.html>

AI Hardware @ Intel



<https://www.intel.com/content/www/us/en/artificial-intelligence/hardware.html>

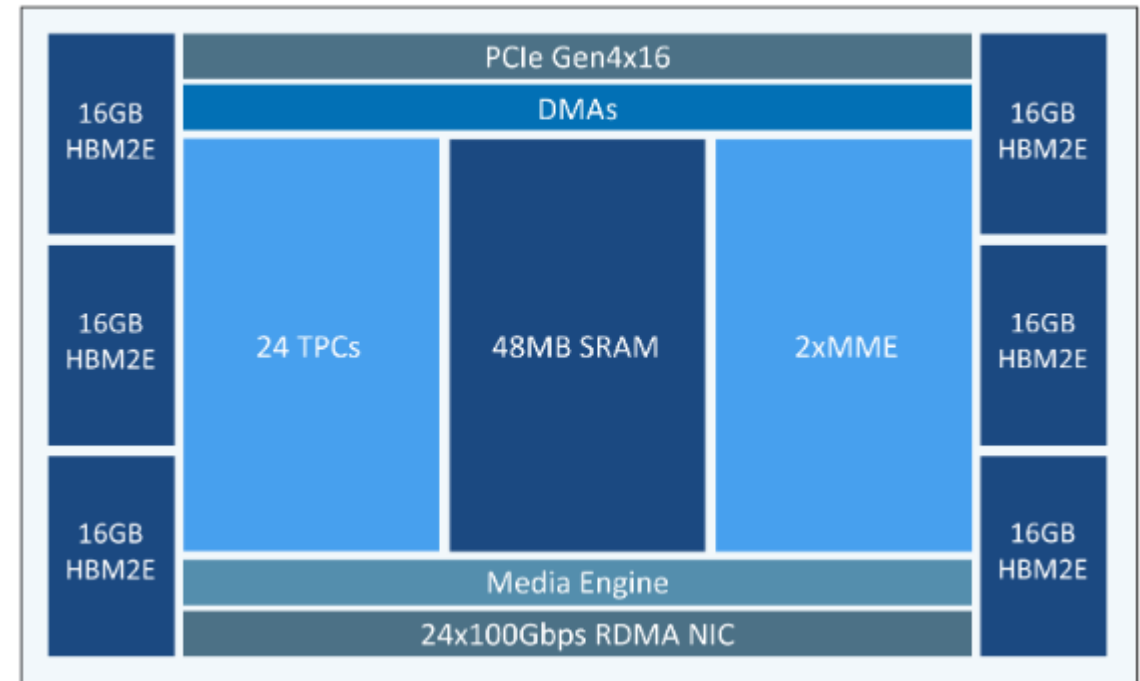
Habana Gaudi2 – Hardware Architecture

- 7nm process technology
- Heterogeneous compute
- 24 Tensor Processor Cores
- Dual matrix multiplication engines
- 24 100 Gigabit Ethernet integrated on chip
- 96 GB HBM2E memory on board
- 48 MB SRAM
- Integrated Media Control



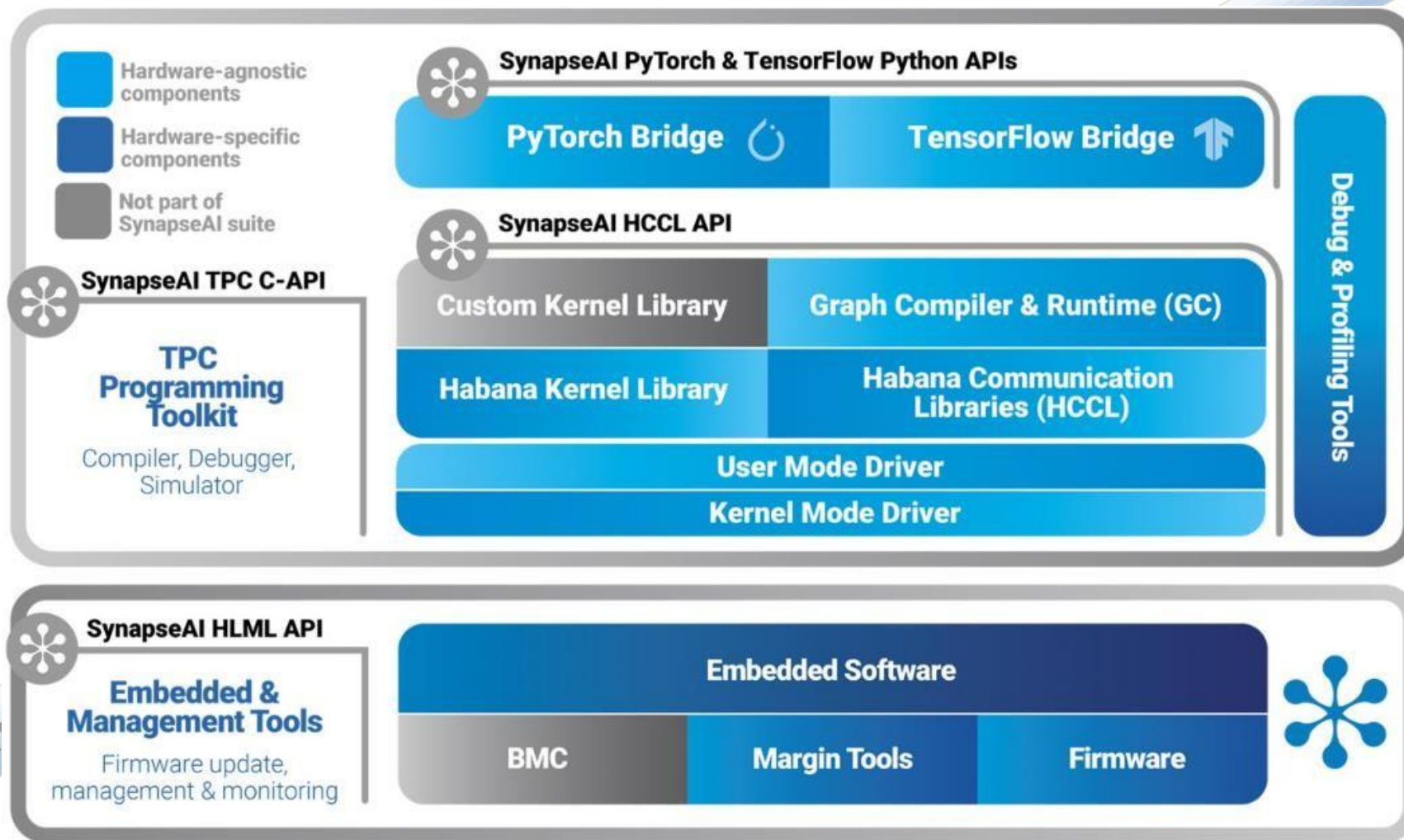
Chip Architecture Diagram

GAUDI²



<https://habana.ai/products/gaudi2/>

Habana Gaudi2 – Software Architecture

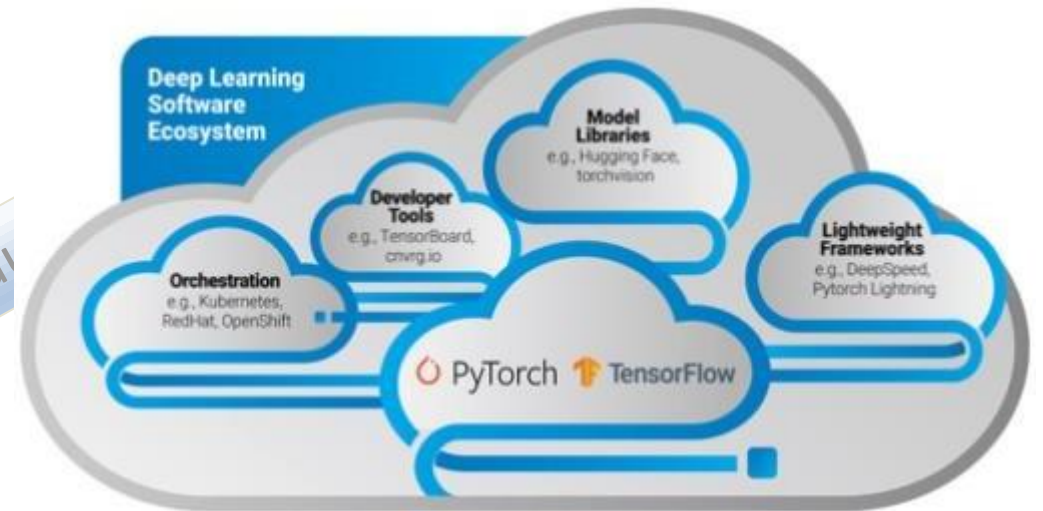


<https://habana.ai/training-software/>

Habana Gaudi2 – Software Ecosystem

Habana Deep Learning Software Ecosystem

brings together leading software providers, tools and code to accelerate development of state-of-the-art deep learning models based on [PyTorch](#), [TensorFlow](#), [PyTorch Lightning](#) and [DeepSpeed](#) frameworks



Hugging Face: over 50,000 AI models and 90,000+ GitHub stars. Habana Optimum library on Hugging Face provides customers using Gaudi and Gaudi2 hardware access to the entire Hugging Face model universe. Checkout a list of Hugging Face Habana optimized models [here](#).



Lightning: acceleration of PyTorch deep learning workloads



deepspeed

Deep Speed: easy-to-use deep learning optimization software that enables scale and speed with particular focus on large scale models

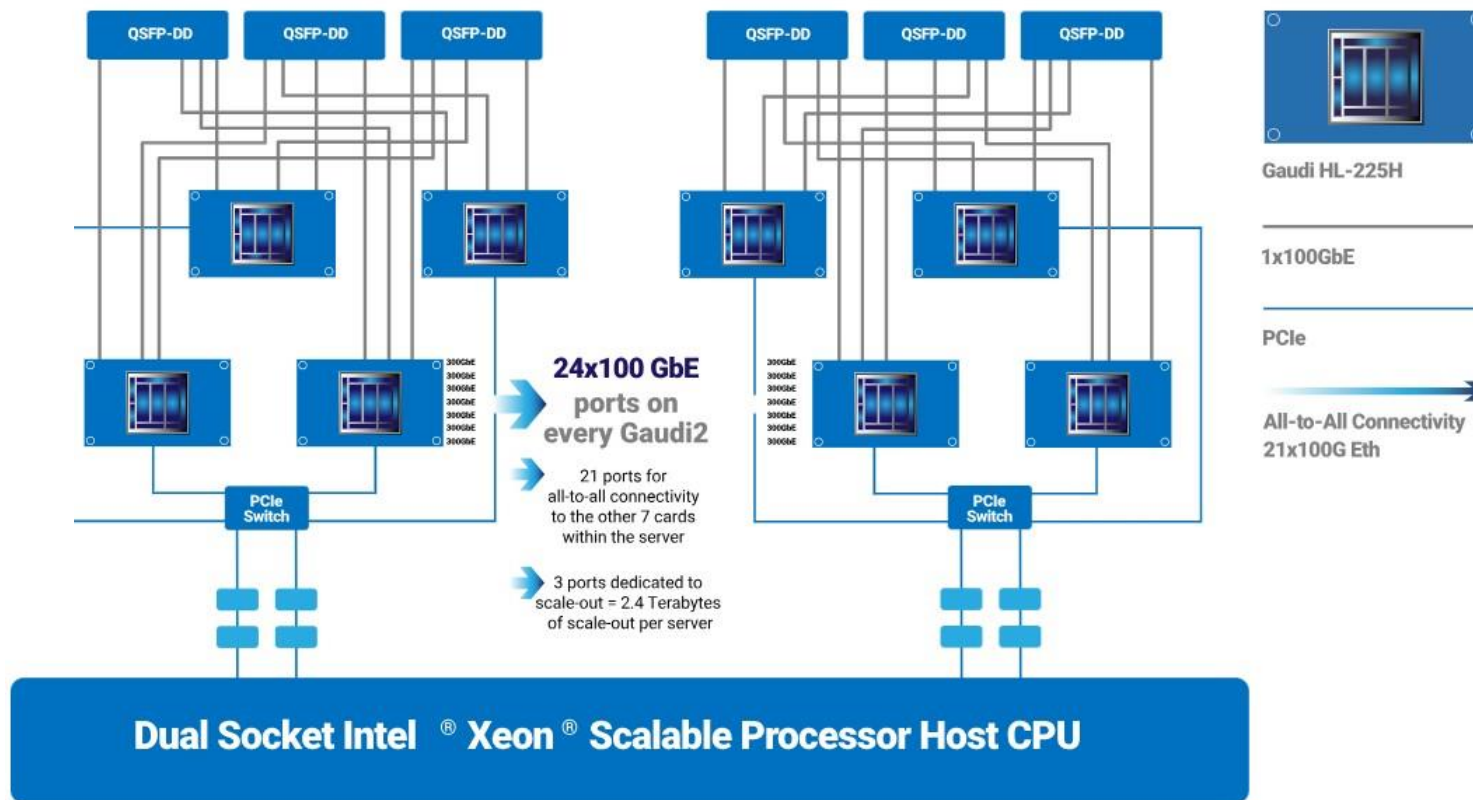
cnvrg.io

Cnvrg.io: MLOps support for customers implementing Habana processor solutions

<https://habana.ai/training-software/>

Habana Gaudi2 - Server Architecture

Gaudi2 Server Architecture



Habana Gaudi2 - Scale Out Scenarios

Small Pod Architecture

16-40 Gaudi2s (2-5 servers, with 8x Gaudi2 each)



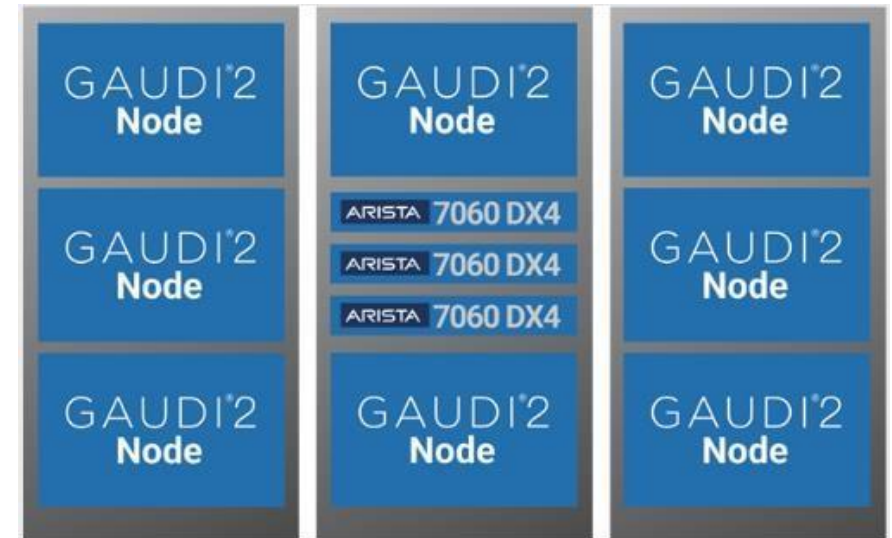
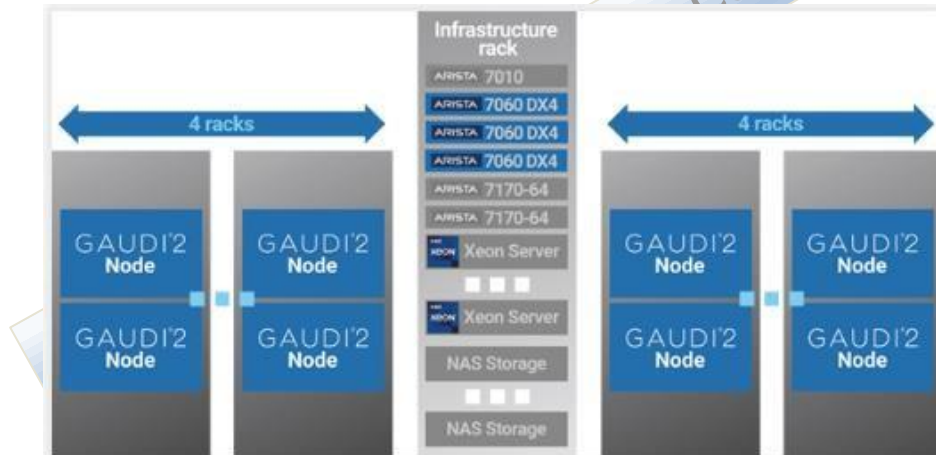
Gaudi2 MegaPod Architecture

Large clusters can be easily built using multiple MegaPods

- ▶ 8 Gaudi2 servers & 3 switches
- ▶ 8x Gaudi2 per node (server)
- ▶ Leaf switches are placed close to Gaudi2 nodes to minimize communication latency
- ▶ Copper cables can be used within pod

Large Pod Architecture

- ▶ Supports variable ratio of nodes-to-switches
- ▶ 3x 400G switches + up to 128 Gaudi2s (16 servers with 8x Gaudi2 each)



ive: Empowering e

Diffusion and Large Language Models on Gaudi2

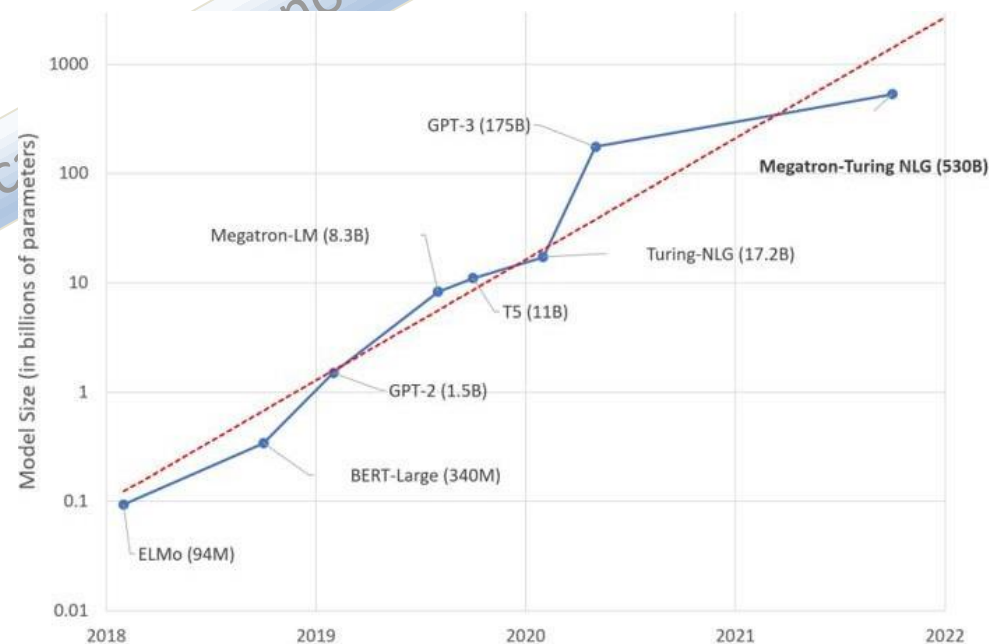
- **Diffusion Models**

- [Stable Diffusion Training](#)
- [Stable Diffusion Inference](#)
- [Hugging Face Stable Diffusion](#)

- **Large Language Models (LLMs)**

- Inference
 - [BLOOM Models \(7B & 176B\)](#)
- Training
 - [Megatron-DeepSpeed LLaMA 13B](#)
 - [MLPerf GPT3](#)
 - [Megatron-DeepSpeed BLOOM 13B](#)
 - [DeepSpeed BERT Models](#)
- [Hugging Face Models](#)

<https://developer.habana.ai/resources/generative-ai-and-large-language-models/>



DeepSpeed Architecture

- **Efficiency features**

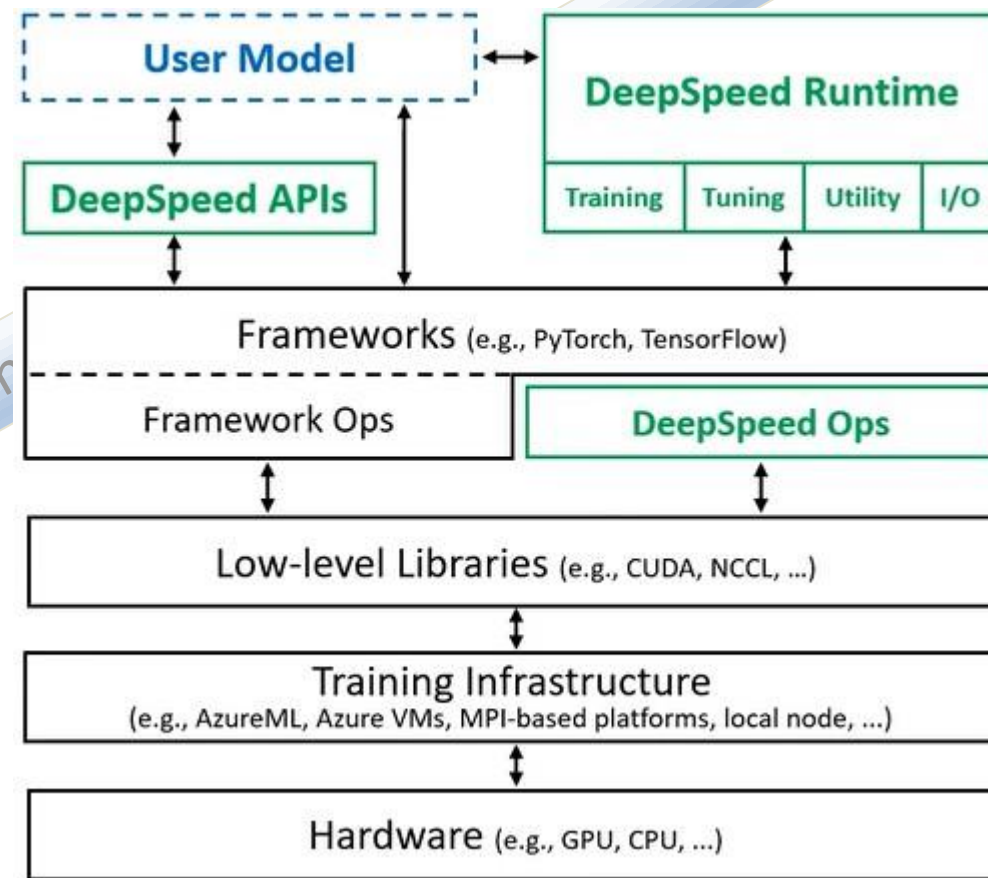
- Memory: Zero Redundancy Optimizer (ZeRO)
- Compute: fastest transformer kernels
- Communication: sparse gradient
- IO: overlapping IO with compute
- Parallelism: ZeRO-powered data parallelism, combination of data + model parallelism

- **Effectiveness features**

- Adaptive hyperparameter tuning
- Optimizers for large-batch training

- **Usability features**

- Distributed training with mixed precision, gradient accumulation, etc.
- IO: simplified data loader with auto batch creation
- Training agnostic checkpoint / recompute
- Performance profiling



- Deploy training job across distributed devices
- Data partitioning
- Model partitioning
- System optimizations
- Tuning for effectiveness
- Utility, e.g., failure detection and checkpointing.

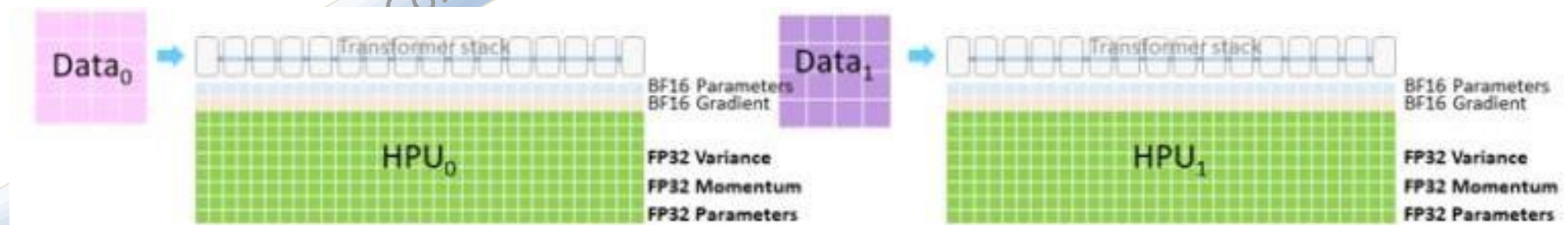
Intel India Education Conclave Empowering

<https://youtu.be/IQCG4zUCYao>

<https://github.com/microsoft/DeepSpeed>

LLMs - Memory Efficient Training using DeepSpeed

- Memory is used for the following
 - BF16 parameters
 - BF16 gradients
 - FP32 optimizer states which includes FP32 momentum of the gradients, FP32 variance of the gradients and FP32 Parameters
- DeepSpeed includes ZeRO (Zero Redundancy Optimizer), a memory-efficient approach for distributed training
 - ZeRO-1 stage - partitions the optimizer states alone across the data parallel processes
 - ZeRO-2 stage - partitions both the optimizer states and gradients across the data parallel processes



<https://youtu.be/IQCG4zUCYao>

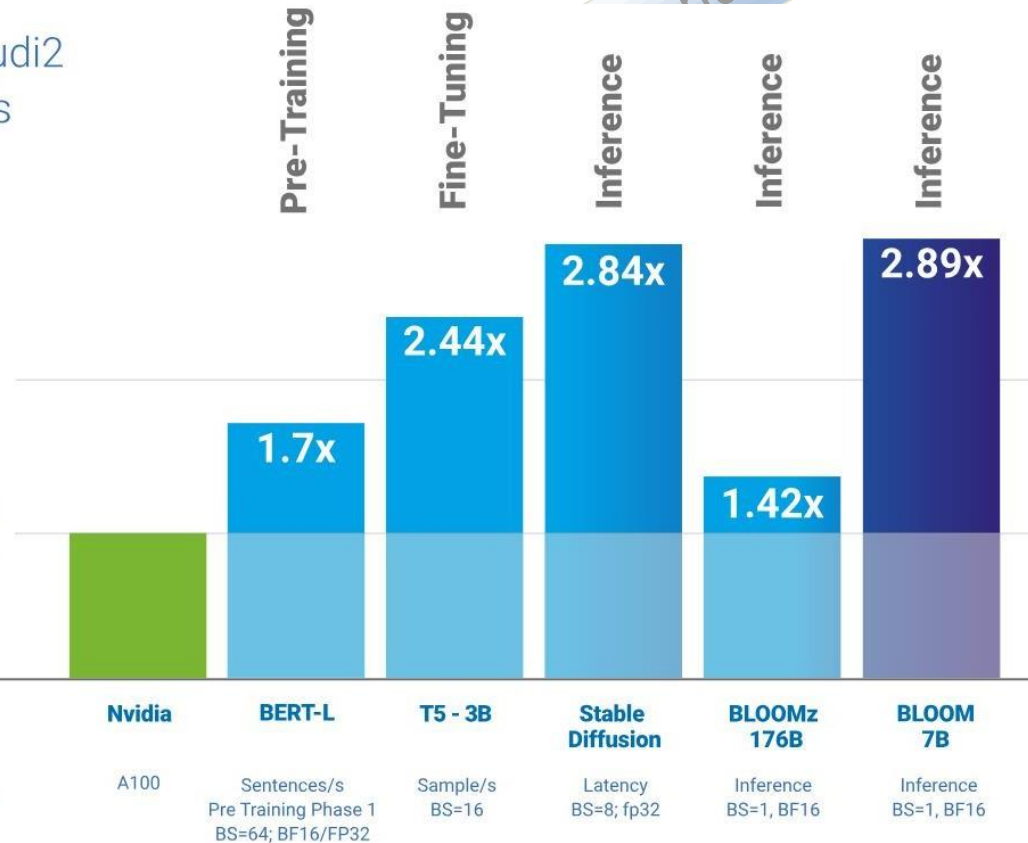
<https://docs.habana.ai/en/latest/PyTorch/DeepSpeed/index.html>

<https://developer.habana.ai/blog/memory-efficient-training-on-habana-gaudi-with-deepspeed/>

Habana Gaudi2 – LLMs using Optimum Habana

- Optimum Habana is the interface between the 🤖 Transformers and 🤖 Diffusers libraries and Habana's Gaudi processor (HPU)

Hugging Face shows Gaudi2 advantage across models



Gaudi2 - Hugging Face – Optimum Habana Models

Transformers

Architecture	Training	Inference	Tasks
BERT	✓	✓	<ul style="list-style-type: none"> text classification question answering language modeling
RoBERTa	✓	✓	<ul style="list-style-type: none"> question answering language modeling
ALBERT	✓	✓	<ul style="list-style-type: none"> question answering language modeling
DistilBERT	✓	✓	<ul style="list-style-type: none"> question answering language modeling
GPT2	✓	✓	<ul style="list-style-type: none"> language modeling text generation
BLOOM(Z)	✗	<ul style="list-style-type: none"> DeepSpeed 	<ul style="list-style-type: none"> text generation
StarCoder	✗	<ul style="list-style-type: none"> Single card 	<ul style="list-style-type: none"> text generation
GPT-J	<ul style="list-style-type: none"> DeepSpeed 	<ul style="list-style-type: none"> Single card DeepSpeed 	<ul style="list-style-type: none"> language modeling text generation
GPT-NeoX	<ul style="list-style-type: none"> DeepSpeed 	<ul style="list-style-type: none"> DeepSpeed 	<ul style="list-style-type: none"> language modeling text generation
OPT	✗	<ul style="list-style-type: none"> DeepSpeed 	<ul style="list-style-type: none"> text generation

Llama 2 / CodeLlama	<ul style="list-style-type: none"> DeepSpeed LoRA 	<ul style="list-style-type: none"> DeepSpeed LoRA 	<ul style="list-style-type: none"> language modeling text generation
StableLM	✗	<ul style="list-style-type: none"> Single card 	<ul style="list-style-type: none"> text generation
Falcon	✗	<ul style="list-style-type: none"> Single card 	<ul style="list-style-type: none"> text generation
CodeGen	✗	<ul style="list-style-type: none"> Single card 	<ul style="list-style-type: none"> text generation
MPT	✗	<ul style="list-style-type: none"> Single card 	<ul style="list-style-type: none"> text generation
T5	✓	✓	<ul style="list-style-type: none"> summarization translation question answering
ViT	✓	✓	<ul style="list-style-type: none"> image classification
Swin	✓	✓	<ul style="list-style-type: none"> image classification
Wav2Vec2	✓	✓	<ul style="list-style-type: none"> audio classification speech recognition
CLIP	✓	✓	<ul style="list-style-type: none"> contrastive image-text training
BridgeTower	✓	✓	<ul style="list-style-type: none"> contrastive image-text training
ESMFold	✗	<ul style="list-style-type: none"> Single card 	<ul style="list-style-type: none"> protein folding

Diffusers

Architecture	Training	Inference	Tasks
Stable Diffusion	✗	<ul style="list-style-type: none"> Single card 	<ul style="list-style-type: none"> text-to-image generation
LDM3D	✗	<ul style="list-style-type: none"> Single card 	<ul style="list-style-type: none"> text-to-image generation

Copyright

and inno

<https://huggingface.co/docs/optimum/habana/index>

Gaudi2 - MLPerf 3.0 Training Benchmark Results – Jun'23

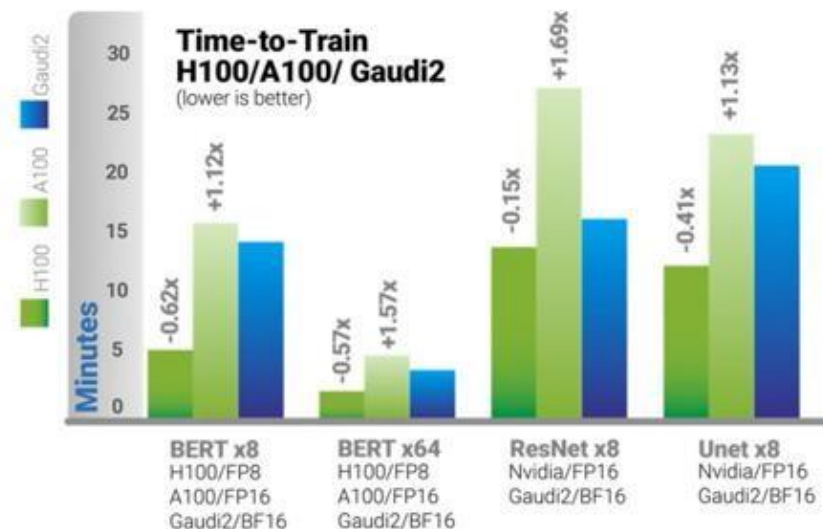
Gaudi2 performance > A100 today;
 Gaudi2 price/performance with FP8 expected > H100*

- Gaudi2 software coverage expands: GPT-3 and Unet
- Gaudi2 performance > A100 on 4 models

	H100	A100	Gaudi2
GPT-3	64.3/512 GPU _s	Not reported	311.9/384 G2 _s
		Not reported	442.6/256 G2 _s
Bert x8	5.4	15.8	14.1
Bert x64	0.9	3.3*	2.1
ResNet x8	13.5	27.0	16.0
Unet x8	12.0	23.2	20.5

* Nvidia reported A100 BERT x64 with 40 GB; other Nvidia metrics cited here are 80 GB.

Performance metrics based on MLPerf Training 3.0 benchmark. For configuration details, see the results provided under embargo by MLPCOMMONS. Results may vary. Performance expectations for FP8 based on Intel internal evaluation.



<https://mlcommons.org/en/training-normal-30/>

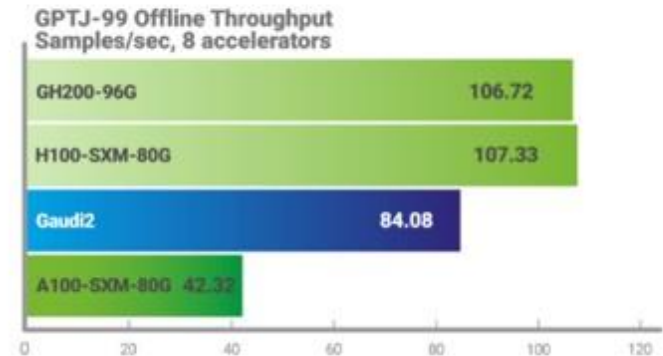
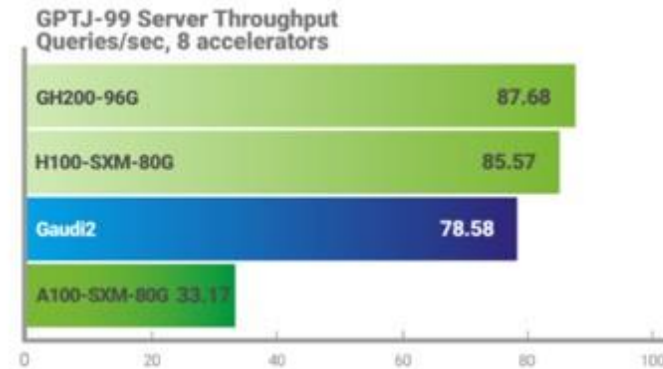
<https://habana.ai/blog/new-mlcommons-results-highlight-impressive-competitive-ai-gains-for-intel/>

<https://www.forbes.com/sites/karlfreund/2023/09/11/intel-gaudi2-looked-to-be-a-credible-alternative-to-nvidia-until/?sh=36c922db2882>

Gaudi2 - MLPerf 3.1 LLM Inference Performance – Jun'23

GPT-J Benchmark: Outstanding Gaudi2 LLM Inference Performance

- H100 slightly outperforms Gaudi2 on GPT-J with 1.09x (Server) and 1.28x (Offline)
- Gaudi2 outperforms A100 by 2.4x (Server) and 2x (offline)
- Gaudi2 employed FP8 and reached 99.9% accuracy
- Gaudi2 offers significantly competitive price-performance as compared to A100, H100 and GH-200



Submitter	a#	Accelerator	GPTJ-99				GPTJ-99.9			
			Server		Offline		Server		Offline	
			Queries/Sec	Vs. Gaudi2	Samples/Sec	Vs. Gaudi2	Queries/Sec	Vs. Gaudi2	Samples/Sec	Vs. Gaudi2
Intel-Habana Labs	8	Habana Gaudi2	78.58	1	84.08	1	78.58	1	84.08	1
Supermicro	8	NVIDIA H100-SXM-80GB	85.57	x1.09	107.33	x1.28	85.43	x1.09	107.06	x1.27
Dell	8	NVIDIA A100-SXM-80GB CTS	33.17	x0.42	42.32	x0.5	Not Submitted			
Nvidia	1	NVIDIA GH200-96G	10.96	x1.12	13.34	x1.27	10.96	x1.12	13.34	x1.27

Performance metrics based on MLPerf Inference 3.1 benchmark. For configuration details, see the [results published by MLCommons](#). Results may vary.

<https://www.forbes.com/sites/karlfreund/2023/09/11/intel-gaudi2-looked-to-be-a-credible-alternative-to-nvidia-until/?sh=36c922db2882>

Gaudi2 - Hugging Face - Visual-Language AI Models

- A new fine-tuning performance benchmark for BridgeTower, a Vision-Language (VL) AI model, has shown that there's life to the AI acceleration camp other than Nvidia's green
- While Nvidia does dominate the AI acceleration market (through exceptional foresight, a well-thought-out and documented software stack, and pure processing performance), other players are eager to take a piece of the AI market for themselves
- BridgeTower, Intel's own Gaudi 2 silicon [has been shown by Hugging Face](#) to outperform Nvidia's A100 80 GB by a staggering 2.5x - and it even beats Nvidia's prodigy-child H100 by 1.4x

Device	dataloader_num_workers=0	dataloader_num_workers=1	dataloader_num_workers=2	dataloader_num_workers=2 + mediapipe_dataloader
Gaudi 2 HPU				
H100 GPU				
A100 80 GB GPU				

<https://www.tomshardware.com/news/intel-habana-gaudi-beats-nvidias-h100-in-visual-language-ai-models-hugging-face>

Gaudi2 – MLPerf Training 3.1 Results – Nov' 23

Intel® Gaudi®2 Accelerator Performance Doubled with FP8

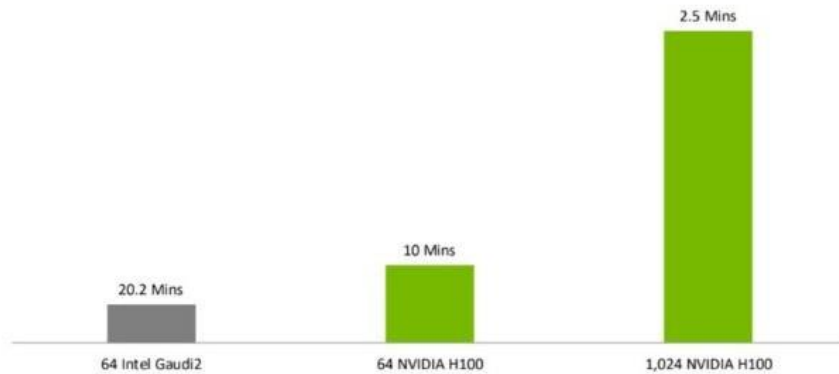
- Intel Gaudi team projected to customers +90% performance gain with FP8
- Delivered more than promised: 103% on GPT-3 industry benchmark



Intel Gaudi 2 MLPerf V3.1 GPT 3 FP8 Performance Boost

MLPerf Training Stable Diffusion

Relative Performance, Higher is Better



MLPerf Training V3.1 NVIDIA Showing The Price Performance Advantage Of Intel Gaudi 2

Outstanding Intel® Gaudi®2 AI Accelerator performance on MLPerf v3.1 Inference Benchmark (June)

Intel Gaudi2 Accelerator with FP8: near-parity performance on GPT-J (Server) with H100

- Gaudi 2 inference performance on GPT-J: -9% (Server) and -28% (Offline) vs H100
- Gaudi 2 outperformed A100 by 2.4x (Server) and 2x (Offline)
- Gaudi 2 employed FP8 and reached 99.9% accuracy

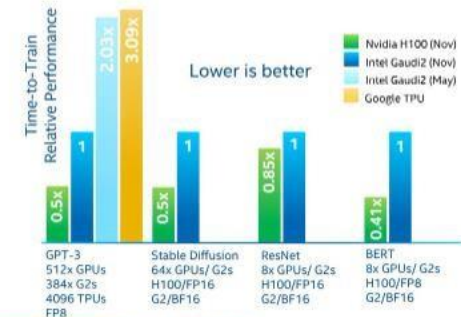


For complete results information and configurations, see MLCommons publication: <https://mlcommons.org/en/>

Intel Gaudi 2 MLPerf Inference V3.1 GPT 3 FP8 Performance Boost

Intel® Gaudi®2 performance advances strengthen competitive price-performance vs. H100

- Gaudi2 performance on ResNet near that of H100.
- H100 with FP8 outperformed Gaudi2 with BF16 on BERT.
- Vs. TPU, Gaudi2 delivered 3x performance on GPT-3.
- Given its significantly lower server cost vs. H100 server cost, Intel Gaudi2 delivers price-performance advantage vs. H100 across models.



For complete results information and configurations, see MLCommons publication: <https://mlcommons.org/en/>

Intel Data Center and AI Group

See backup for workloads and configurations. Results may vary.

intel

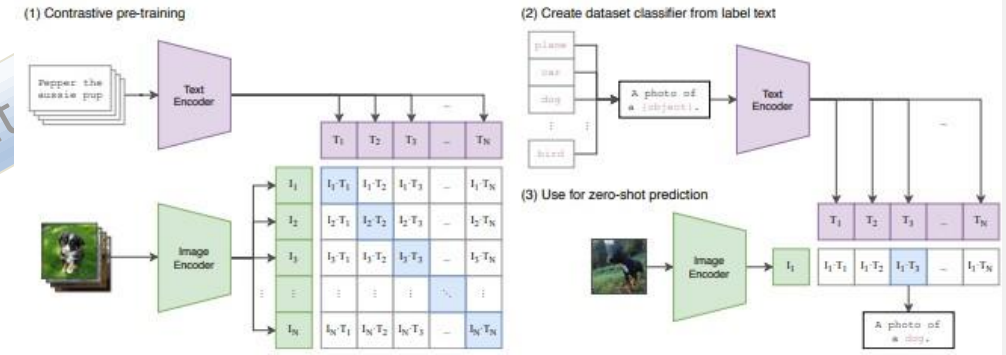
Gaudi2 - Demo – LLaVA - <https://llava-vl.github.io/>

LLaVa connects pre-trained [CLIP ViT-L/14](#) visual encoder and large language model [Vicuna](#), using a simple projection matrix

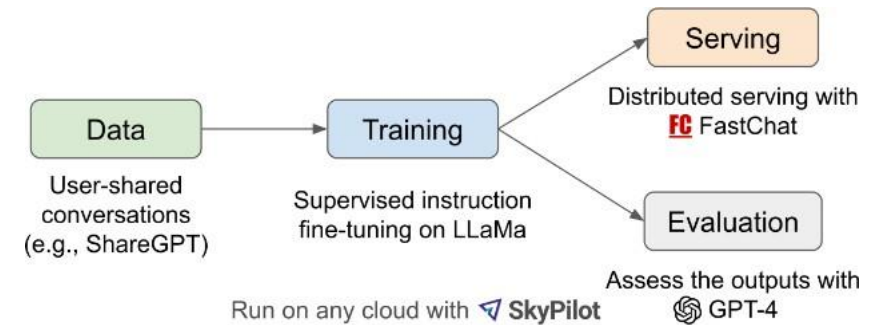
LLaVA: Large Language and Vision Assistant

[Project Page](#) | [Paper](#) | [Code](#) | [Model](#)

The screenshot shows the LLaVA web interface. At the top, it says "LLaVA: Large Language and Vision Assistant" with links to Project Page, Paper, Code, and Model. Below this is a dropdown menu for the model version, currently set to "LLaVA-v1-13B-336px". There is an "Image" section with a "Drop Image Here" area and a "Click to Upload" button. Below that are options to "Preprocess for non-square image" with buttons for "Crop", "Resize", and "Pad". An "Examples" section shows two sample prompts: "What is unusual about this image?" and "What are the things I should be cautious about when I visit here?". The main chat area shows a prompt: "What is unusual about this image?" and a response: "The unusual aspect of this image is that a man is ironing a shirt on the back of a parked car, which is unusual because it is not a typical place to iron clothes. Ironing a shirt on a car is not only unconventional but also potentially dangerous, as it could lead to damage to the car's paint or bodywork. Additionally, ironing on a car is not a practical or efficient". At the bottom, there are buttons for "Upvote", "Downvote", "Flag", "Regenerate", and "Clear history".



CLIP



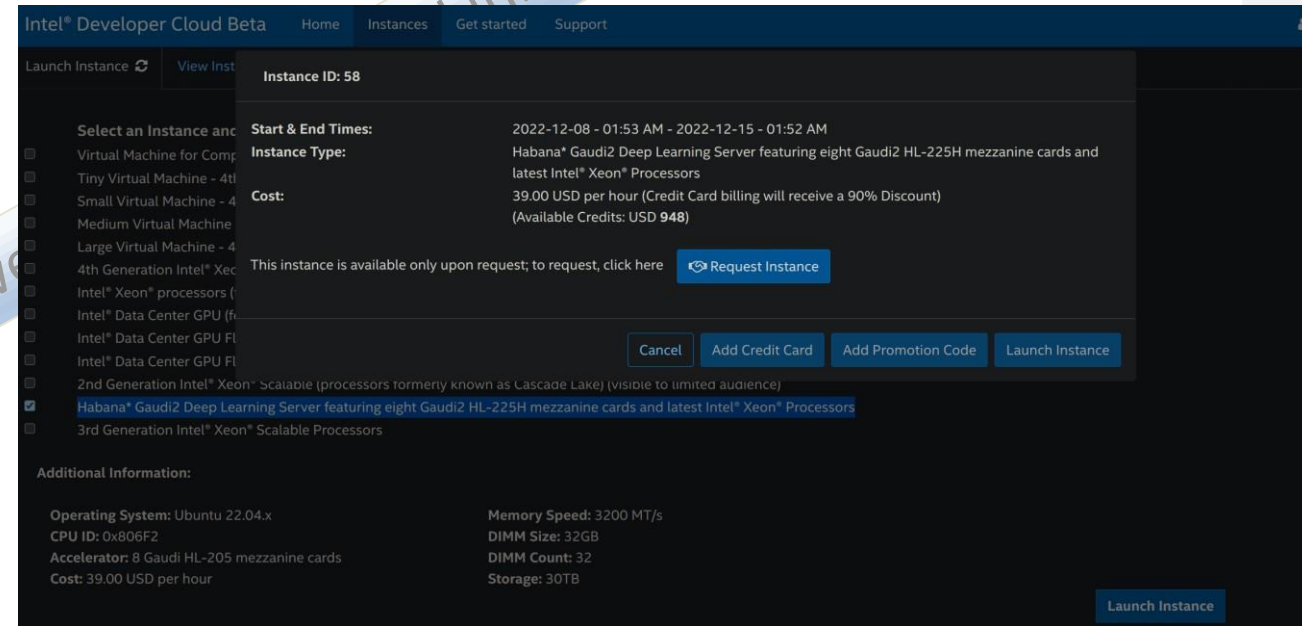
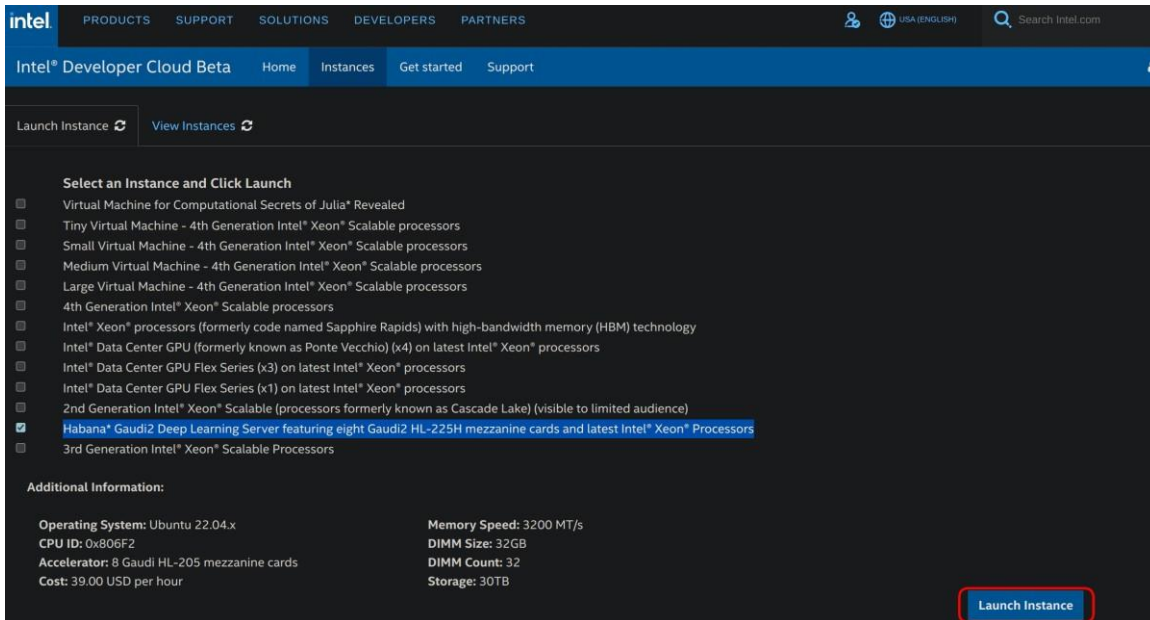
Vicuna

<https://openai.com/research/clip>

<https://lmsys.org/blog/2023-03-30-vicuna/>

Habana Gaudi2 Access

1. Go to the [Intel Developer Cloud landing page](#) and sign into your account or register if you do not have one
2. Go to the [Intel Developer Cloud management console](#)
3. Select Habana Gaudi2 Deep Learning Server featuring eight Gaudi2 HL-225H mezzanine cards and latest Intel® Xeon® Processors and click on Launch Instance in the lower right corner as shown below, and request an instance



- Once your request is validated, re-do step 3 and click on Add OpenSSH Publickey to add a payment method (credit card or promotion code) and a SSH public key that you can generate with `ssh-keygen -t rsa -b 4096 -f ~/.ssh/id_rsa`. You may be redirected to step 3 each time you add a payment method or a SSH public key.
- Re-do step 3 and then click on Launch Instance. You will have to accept the proposed general conditions to launch the instance
- Go to the Intel Developer Cloud management console and click on the tab called View Instances

https://scheduler.cloud.intel.com/public/Intel_Developer_Cloud_Getting_Started.html

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

References

- <https://www.intel.com/content/www/us/en/artificial-intelligence/industries.html>
- <https://www.intel.com/content/www/us/en/developer/topic-technology/artificial-intelligence/overview.html>
- https://docs.habana.ai/en/latest/Gaudi_Overview/index.html
- <https://habana.ai/products/gaudi2/>
- <https://huggingface.co/docs/optimum/main/habana/index>
- <https://community.intel.com/t5/Blogs/Tech-Innovation/Artificial-Intelligence-AI-Survival-of-the-Fittest-Compact-Generative-AI-Models-Are-the/post/1508220>
- https://docs.habana.ai/en/v1.6.0/Release_Notes/GAUDI_Release_Notes.html
- <https://www.deepspeed.ai/>
- <https://www.microsoft.com/en-us/research/blog/zero-deepspeed-new-system-optimizations-enable-training-models-with-over-100-billion-parameters/>
- <https://www.microsoft.com/en-us/research/blog/zero-2-deepspeed-shattering-barriers-of-deep-learning-speed-scale/>
- [ZeRO: Memory Optimizations Toward Training Trillion Parameter Models. Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase and Yuxiong He.](#)
- [DeepSpeed Usage Guide](#)
- [Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model](#)

Intel India Education Conclave: Empowering educators and innovation

Thank You