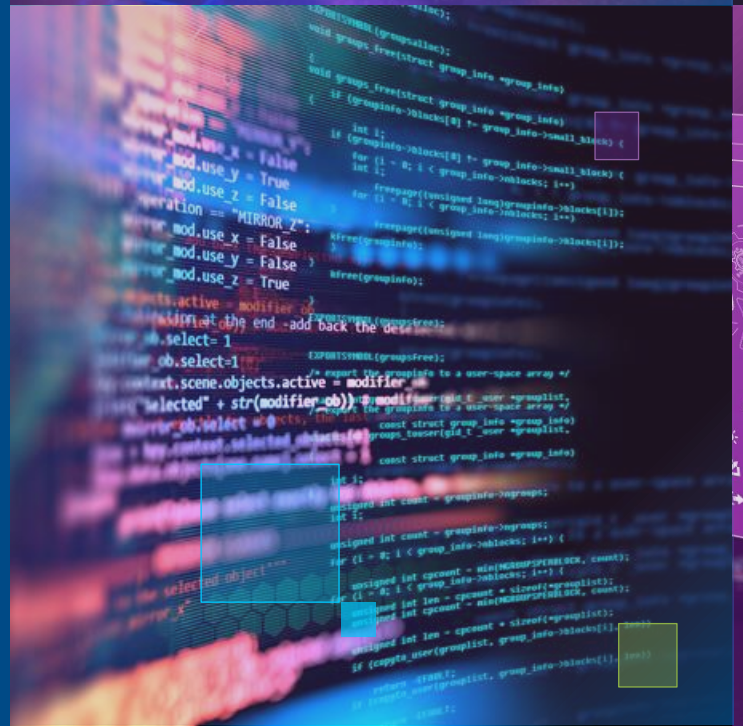
Accelerate Innovation
Generative AI with the

intel® OpenVINO™ toolkit

Supplementary Content for the
OpenVINO™ 30-3-30



Notices & Disclaimers

Performance varies by use, configuration, and other factors. Learn more at intel.com/PerformanceIndex.

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® optimizations, for Intel® compilers or other products, may not optimize to the same degree for non-Intel products.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Results have been estimated or simulated.

Intel is committed to respecting human rights and avoiding complicity in human rights abuses.
See Intel's Global Human Rights Principles. Intel® products and software are intended only to be used in applications that do not cause or contribute to a violation of an internationally recognized human right.

2023's Hottest Topic Generative AI

Generative AI broadly describes machine learning systems capable of generating text, images, code, or other content in response to a user prompt.

Technology

Powered by foundation models that can multi-task and perform out-of-the-box tasks, including summarization, Q&A, classification, and more

Economic Impact

Potential to add $4.4 trillion to global economy annually[1]

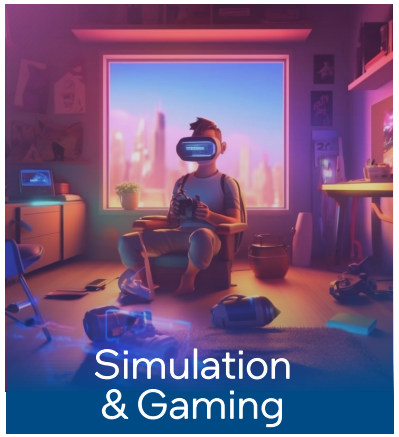
Technical Leader

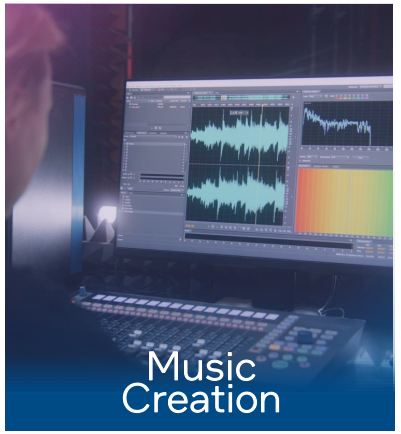
500 senior IT leaders surveyed reveals that a majority (67%) are prioritizing generative AI for their business within the next 18 months, with one-third (33%) naming it as a top priority[2]

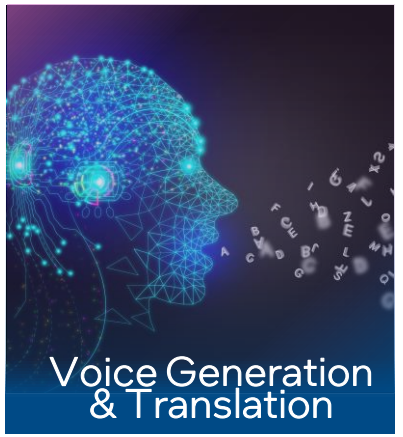
1. McKinsey & Co. Featured Insights August 2023
2. Salesforce News & Insights March 2023

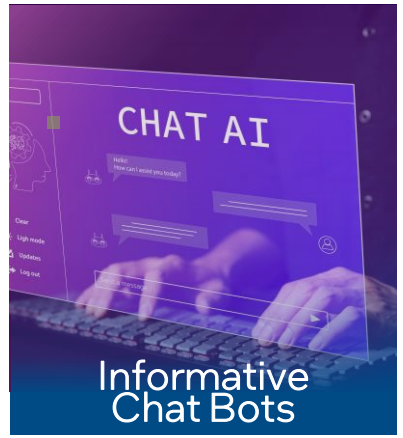
Generative AI Use Cases


Creative Content Generation


Simulation & Gaming

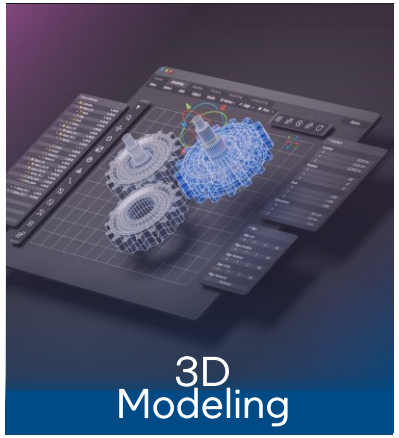

Music Creation

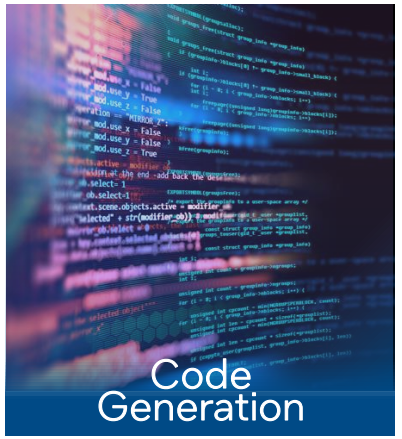

Voice Generation & Translation



Informative Chat Bots


Video Creation & Editing


3D Modeling

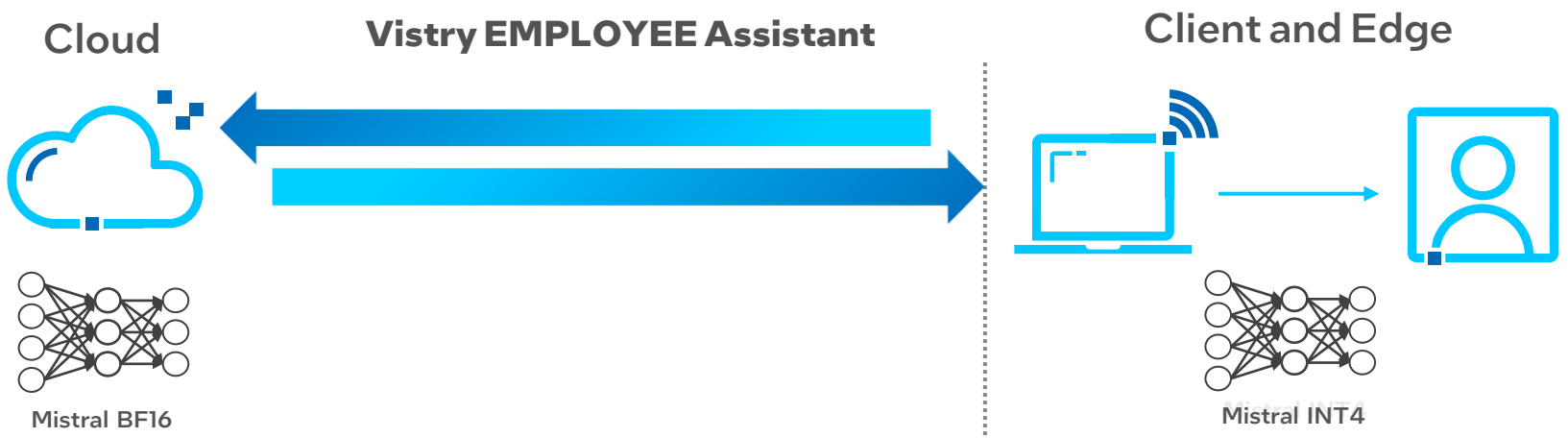

Code Generation


Sim-to-Real Robotics


Healthcare Synthetic Data

Conversational Platform for Food Commerce

Vistry.ai



Enables **consumer online ordering**, complemented by Vistry Employee Assistant

Moderate-traffic venues

High-traffic venues

Provides food service workers (quick serve restaurant, stadium events, etc.) real-time, accurate information about **operational procedures, standards, and guides** despite reliable network/connectivity

intel Xeon ← OpenVINO™ → intel Core

Improve the Quick-Service Restaurant Industry with AI

Toms Hardware[1]

"Automatic1111's Stable Diffusion WebUI now works with Intel GPU hardware, thanks to the integration of Intel's OpenVINO toolkit that takes AI models and optimizes them to run on Intel hardware."

News

Stable Diffusion Optimized for Intel Silicon Boosts Arc A770 Performance by 54%

By Aaron Klotz

Create AI-generated images even faster

(Image credit: Intel)

Automatic1111's Stable Diffusion WebUI now works with Intel GPU hardware, thanks to the integration of Intel's OpenVINO toolkit that takes AI models and optimizes them to run on Intel hardware. We've re-tested the latest release of Stable Diffusion to see how much faster Intel's GPUs are compared to our previous results, with gains of 40 to 55 percent.

1. Stable Diffusion Optimized for Intel Silicon Boosts Arc A770 Performance by 54%

Challenges with Generative AI

Time is our most precious resource. Delays increase expense.

Gen AI models are large and getting bigger

Microsoft's* Megatron Turing NLG[1] has **530 billion** parameters
Google's* DeepMind Gopher[2] has **280 billion** parameters

Google recognizes the size problem; released Gemini[3] in three sizes for device scalability; **Ultra, Pro, and Nano**

Retraining is expensive and time consuming

Can take hours, even days

Requires a large dataset

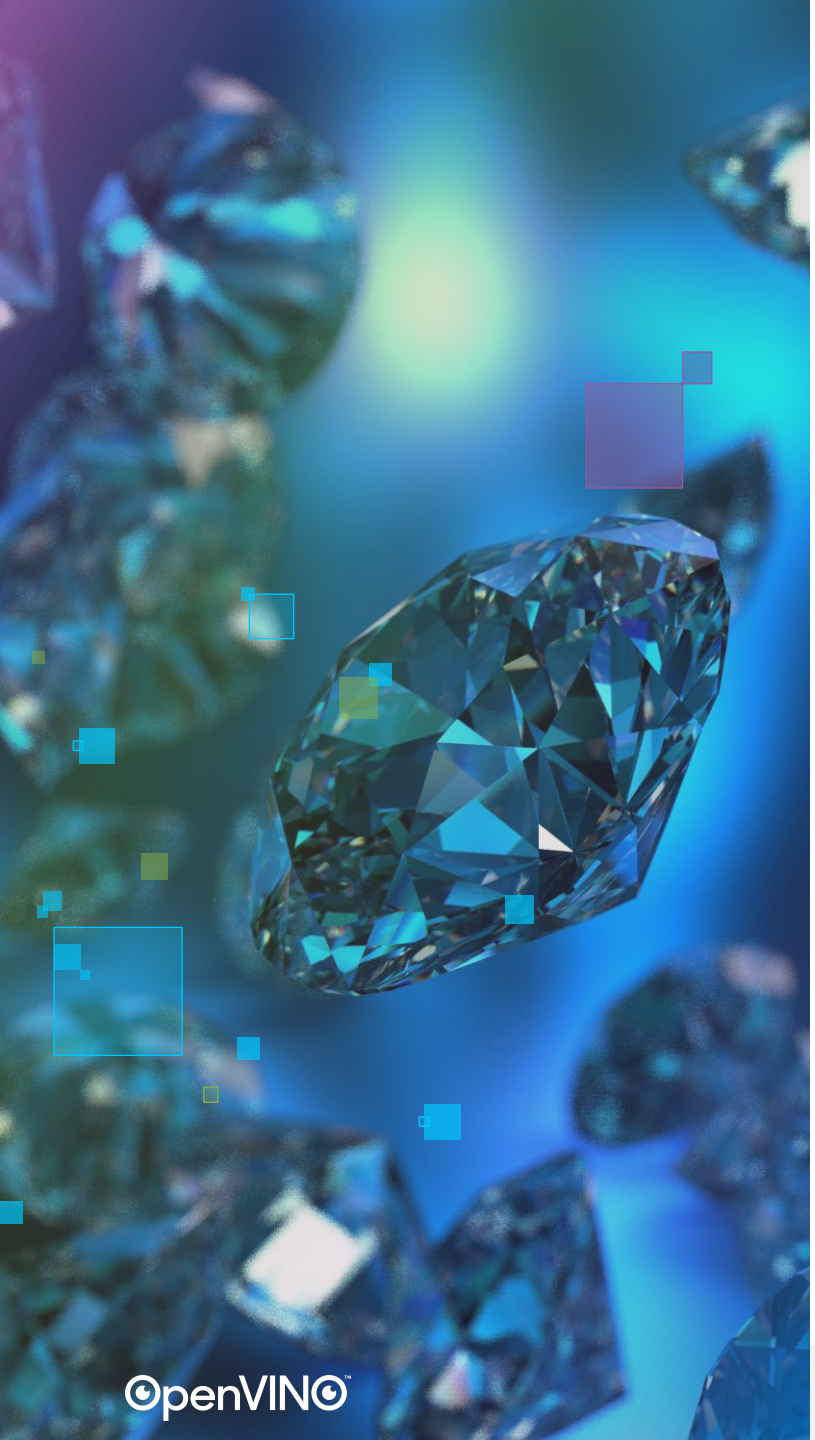
Slows time to solution/market and added expense

Compute where deployed is often limited

AI is going to come to all ranges of device sizes, with diverse requirements for myriad use cases requiring unique approaches

Inability to meet use-case-specific requirements

1. Arxiv.org
2. DeepMind Google
3. Google Blog



Solution is Compression

OpenVINO™ toolkit's Neural Network Compression Framework

Training Aware Quantization

Training-time model compression improves model performance by applying optimizations (such as quantization) during the training.

Post-Training Quantization

Post-training model optimization is the process of applying special methods that transform the model into a more hardware-friendly representation without retraining or fine-tuning.

Weight Compression

Weight compression aims to reduce the memory footprint of a model. It can also lead to significant performance improvement for large memory-bound models, such as Large Language Models (LLMs).

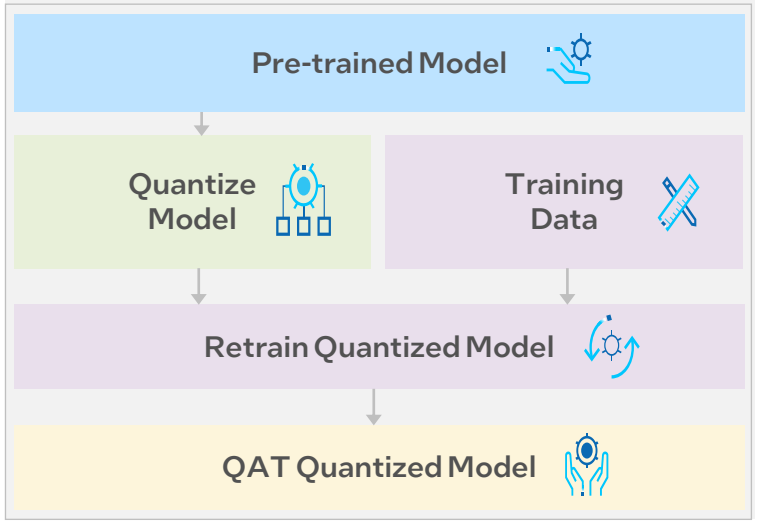
Solution is Compression

OpenVINO™ toolkit's Neural Network Compression Framework

Quantization Paths

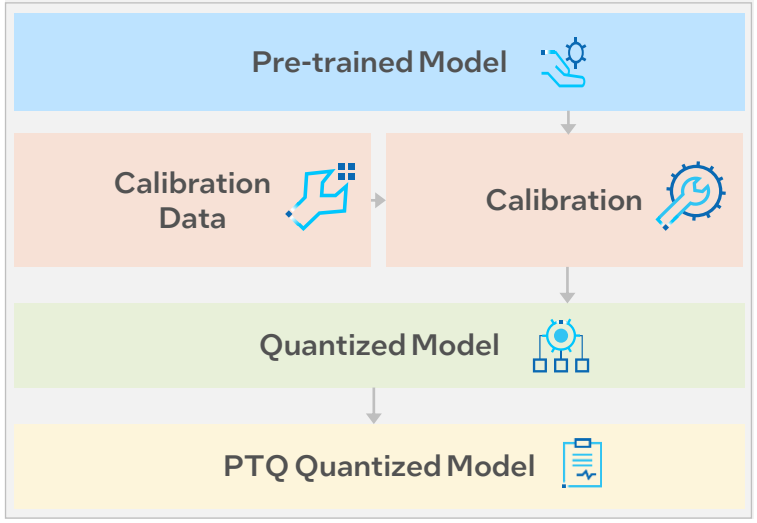
Training Aware Quantization (QAT)

Incorporates quantization-aware techniques during training to optimize the model for lower precision. Aims to preserve accuracy but requires complex and time-consuming model retraining.



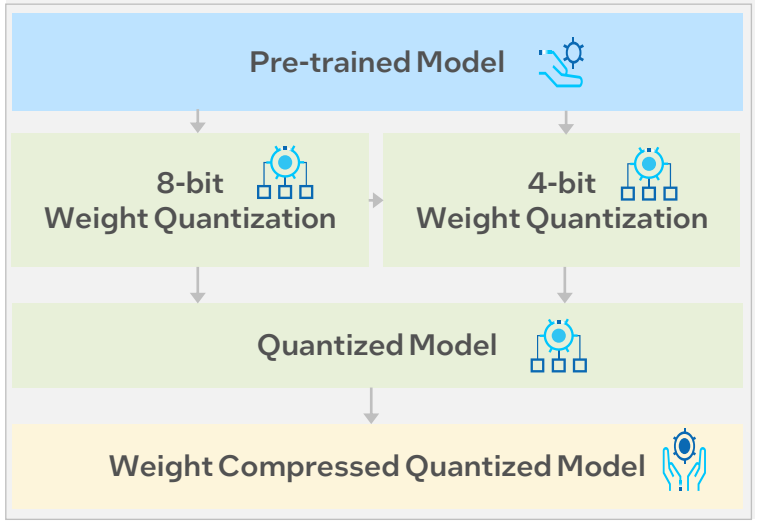
Post Training Quantization (PTQ)

Quantizes a pre-trained model after training, reducing model precision to improve memory usage and inference speed while potentially sacrificing some accuracy.



Weight Compression

An easy-to-use quantization method for Large Language Models footprint reduction and inference acceleration.

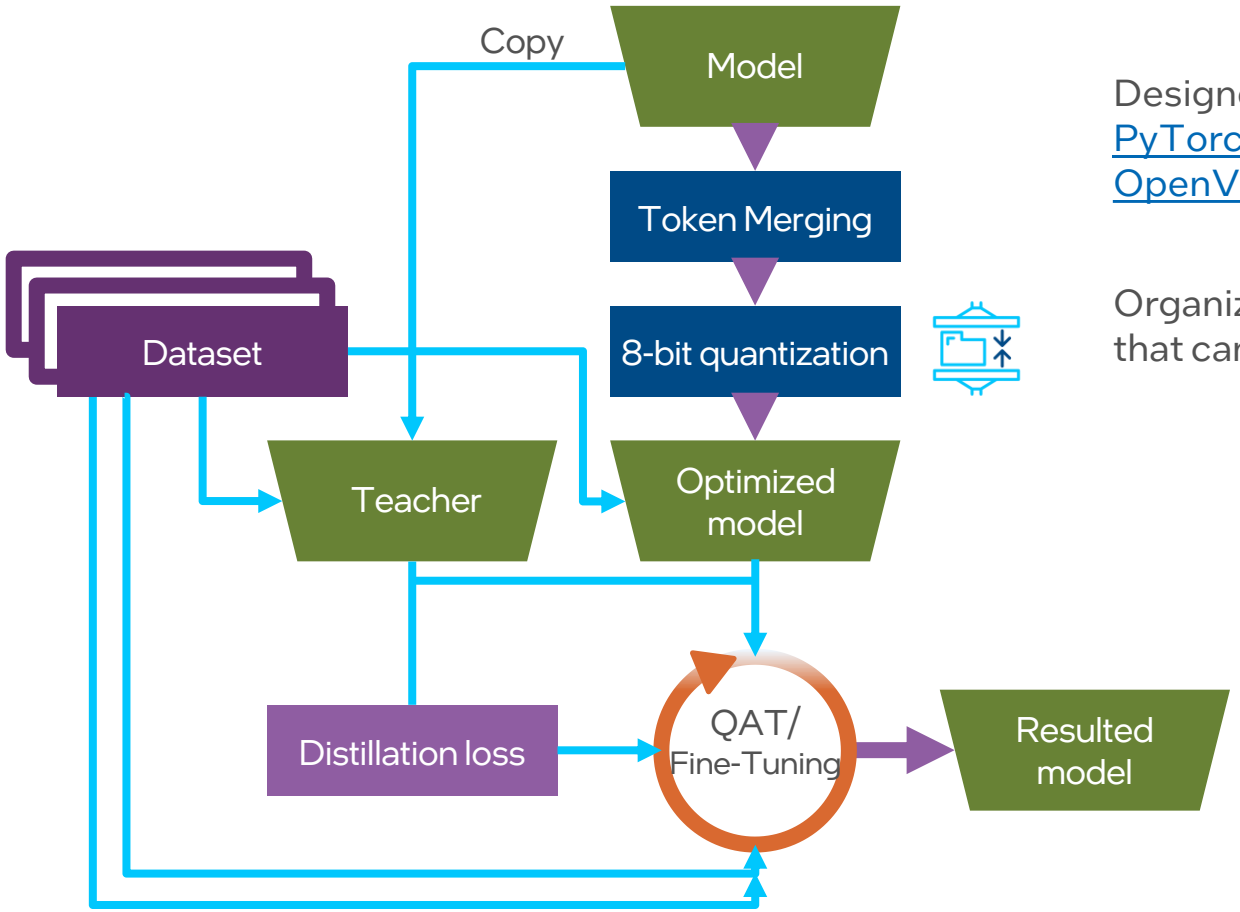


Note: Except for ONNX (.onnx model formats), all models have to be converted to an IR format to use as input to the Runtime
Development guide: https://docs.openvino.ai/2023.2/openvino_docs_model_optimization_guide.html

Solution is Model Compression

Reduce Model Size

Reduce Memory Footprint



Designed to work with models from PyTorch, TensorFlow, ONNX and OpenVINO™.

Organized as a Python* package that can be built and used standalone

Read Hugging Face blog

Stable Diffusion Quantization

Compressed Model Size Produces Real Results

FP32 Native Model

8.0K	stabilityai_cpu/feature_extractor
1.3G	stabilityai_cpu/text_encoder
8.0K	stabilityai_cpu/scheduler
1.6M	stabilityai_cpu/tokenizer
3.3G	stabilityai_cpu/unet
320M	stabilityai_cpu/vae
4.9G	stabilityai_cpu/

FP32 OpenVINO™ Model

8.0K	openvino_ir/feature_extractor
1.3G	openvino_ir /text_encoder
131M	openvino_ir /vae_encoder
8.0K	openvino_ir /scheduler
1.6M	openvino_ir /tokenizer
3.3G	openvino_ir /unet
190M	openvino_ir /vae_decoder
4.9G	openvino_ir /

FP16 OpenVINO™ Model

8.0K	modelSD21_dGPU_0V/feature_extractor
652M	modelSD21_dGPU_0V/text_encoder
8.0K	modelSD21_dGPU_0V/scheduler
1.6M	modelSD21_dGPU_0V/tokenizer
1.7G	modelSD21_dGPU_0V/unet
96M	modelSD21_dGPU_0V/vae_decoder
2.4G	modelSD21_dGPU_0V/

Use Case Models

- Video Generation
- 3D Modeling
- Image Generator
- Image Segmentation

Visual*

- CLIP
- BLIP
- FILM
- Pix2Pix
- Riffusion
- ControlNet
- Zero Scope
- QR code monster
- Segment Anything Model (SAM)
- Latent Consistency Models (LCM)
- Würstchen
- DeepFloyd IF
- DeciDiffusion
- Stable Diffusion

- Chat Bot
- Code Generation
- Search
- Text Classification
- Content Creation
- Instruction Following

Language*

- GPT J
- Notus
- LLaVa
- Llama 2
- BLOOM
- chatGLM
- chatGLM3
- Baichuan 2
- Neural Chat
- StableLM-Epoch-3B
- StableLM-tuned-alpha-3b
- MPT
- Dolly
- Youri
- Qwen
- Mistral
- Zephyr
- RedPajama
- LLM chatbot

- Music Generation
- Text to Audio
- Audio to Text
- Single Voice Conversion

Audio*

- BARK
- VITS
- SoftVC
- Whisper
- MusicGen
- AudioLDM
- Distil-Whisper

Example model support includes, but not limited to:

Introducing the OpenVINO Generative AI Github* Repository

Generative AI Roadmap with OpenVINO™

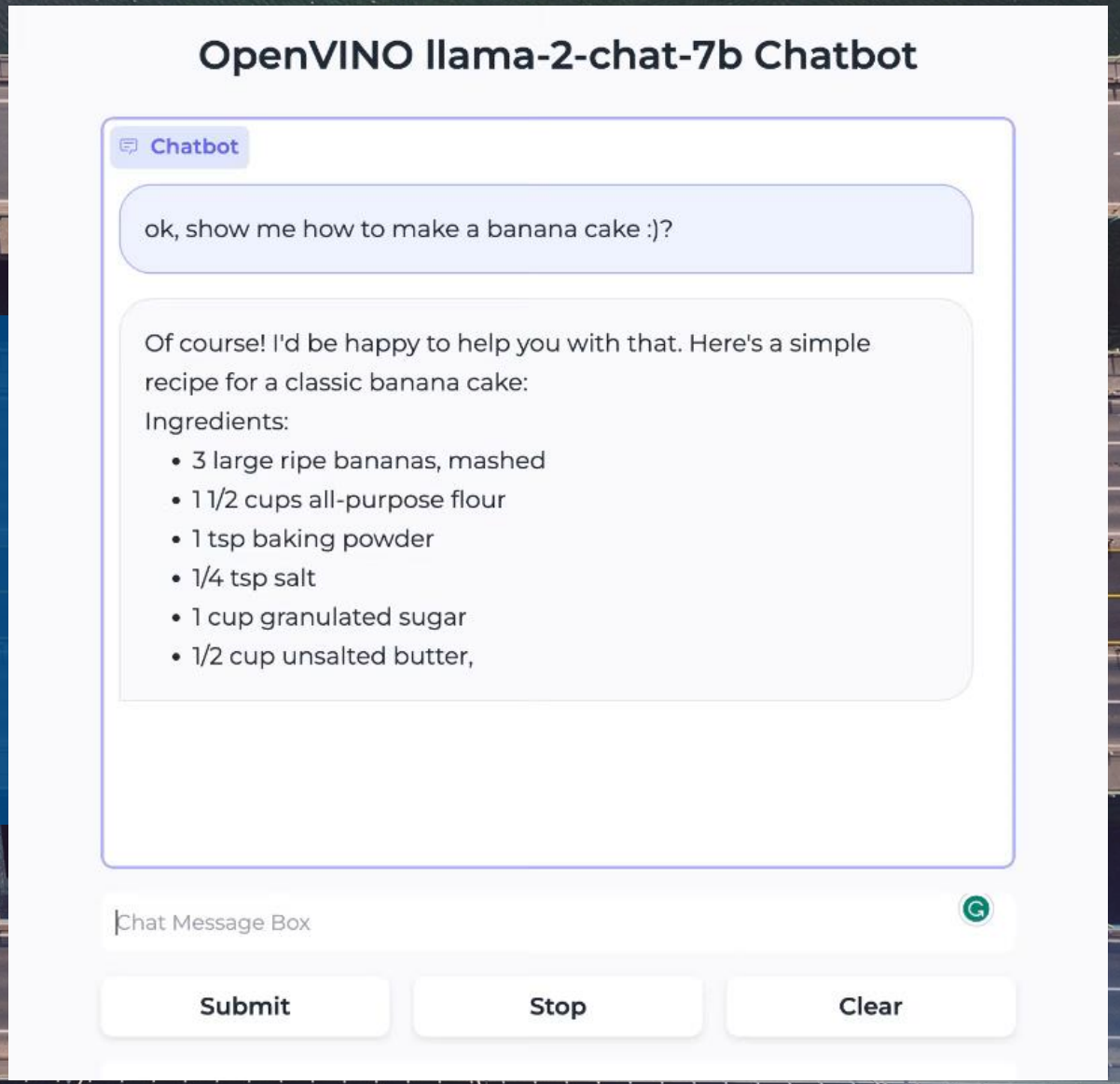
Release 2023.2
Introduced INT8 weight compression on Intel Core CPU and iGPU to accelerate inferencing of LLMs

Release 2023.3 LTS - COMING SOON (JAN '24)

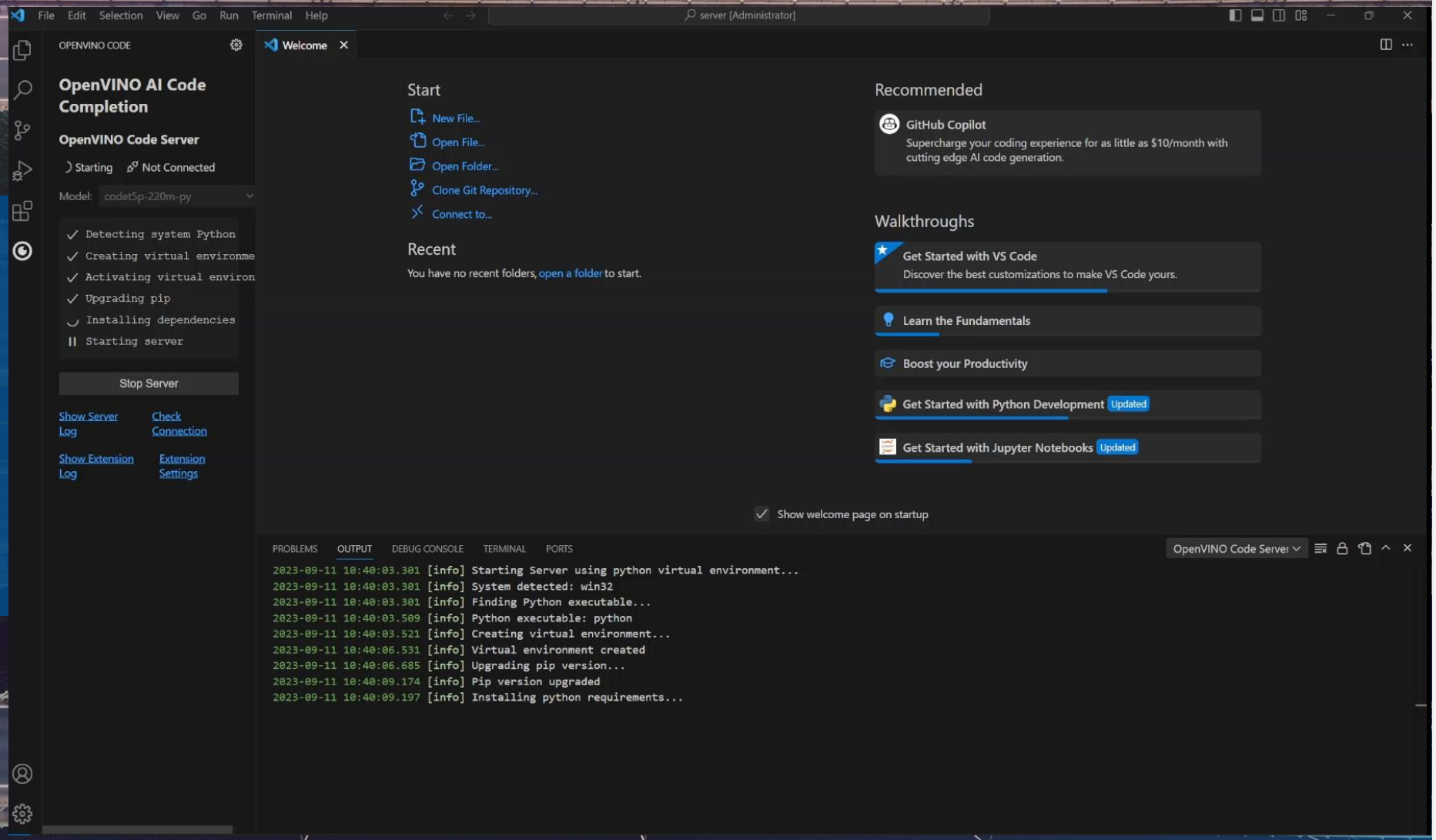
- INT8 weight compression available on Xeon CPU as well as preview for INT4 weight compression
- Introducing OpenVINO Gen AI repository on GitHub* that demonstrates native C and C++ pipeline samples for LLMs. We've started supporting string tensors as inputs and tokenizers natively to reduce overhead and ease of production
- OpenVINO Model Server optimized to support LLMs for seamless deployment

Continue to expand model support and optimize performance across pipeline

OpenVINO™ Demonstrating a Chatbot



OpenVINO™ Demonstrating Code Generation

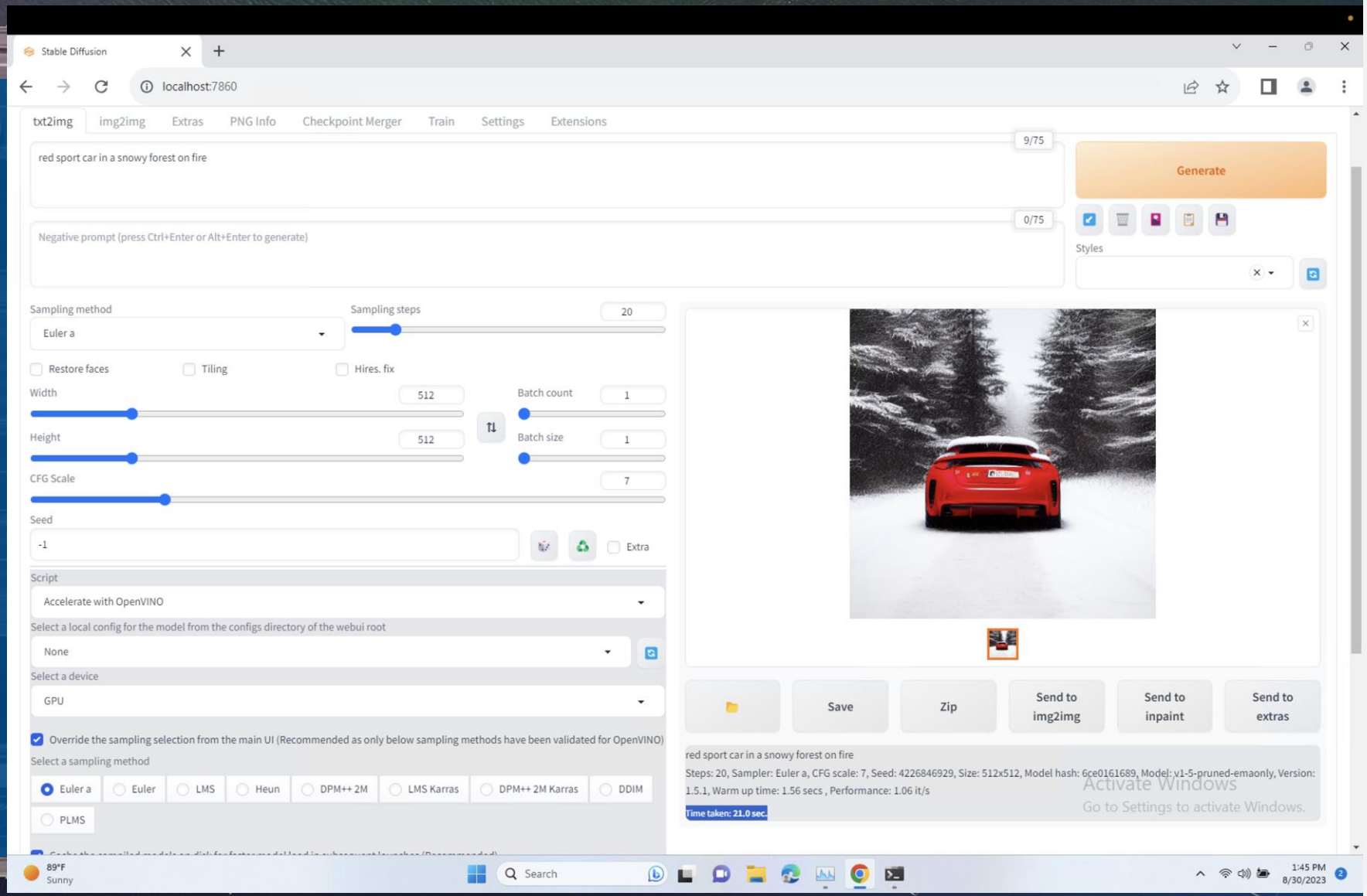


Automatic1111 Stable Diffusion* Web UI

Install OpenVINO Code Completion

```
import torch
import torchvision.models as models
import openvino.frontend.pytorch.torchdynamo.backend
model = models.resnet50(pretrained=True)
input = torch.rand((1,3,224,224))
model = torch.compile(model, backend='openvino')
pred = model(input)
```

OpenVINO 2023.1



Getting Started with Generative AI using OpenVINO™

Notebooks

- Text-to-Image Generation with Stable Diffusion* and OpenVINO™
- Text-to-Image Generation with ControlNet* Conditioning
- Text-to-Image Generation and Infinite Zoom with Stable Diffusion v2 and OpenVINO™
- Video Subtitle Generation with OpenAI* Whisper
- Image generation with Stable Diffusion XL and OpenVINO
- Instruction following using Databricks* Dolly 2.0 and OpenVINO™
- Image generation with DeepFloyd IF and OpenVINO™
- Text-to-Music generation using Riffusion* and OpenVINO

Blogs

- Effortless Image Generation with Optimum-Intel OpenVINO™: Accelerating Stable Diffusion in a Few Lines of Code
- How to run Stable Diffusion on Intel GPUs with OpenVINO
- Chaining multi-modal generative AI models on CPU and GPU

Resources

- Virtual Workshop on Generative AI with OpenVINO
- Gen AI Github* Repository

Optimum Intel

Use OpenVINO as an extension in Hugging Face* transformer models and gain model compression and performance benefits

Choice is Download and Installation Methods

Choose and download free directly from Intel

Intel® Distribution of OpenVINO™ Toolkit

Also available from these sources:

Intel® Developer Cloud | PIP | Docker Hub | Dockerfile | Anaconda Cloud | YUM | APT | Conan | Homebrew | vcpkg

Build from source:

GitHub | Gitee (for China)

Validated Gen AI Models' Specs

Hardware Recommended by Model

What HW is recommended for what model?

Is there a model which works better on intel dGPU vs CPU, iGPU, iVPU (called NPU now?)

Key:

a. Y =Yes it works with some level of optimization in 2023.1 or greater

b. N = Model does not work yet on this device

c. OOM = Device runs out of memory, but int8 compressed version *might* fit, we just haven't tried it yet

d. Infers = We know the model infers but do not have accuracy or performance data yet

Model	Released by	Parameters	CPU	iGPU	dGPU
AquilaChat	Beijing Academy of AI	7B	Y	N	N
Blenderbot	ParlaAI (Meta)	9B	Infers		
BLOOM	BigScience	176B	Infers		
BLOOMZ	YeungNLP	1B	Y		
BLOOMZ	BigScience	560M	Y		
ChatGLM	Tsinghua University	6B	Y	OOM	Y
Code T5	Salesforce	6B	Y	N	N
CodeGen2	Salesforce	1B	Y	Y	Y
CodeGen2	Salesforce	3.7B	Y		
CodeGen2	Salesforce	7B	Y	OOM	Y
DeepFloyd	Stability AI	11B	Infers		
Dolly v2	Databricks	12B	Y	OOM	OOM
Dolly v2	Databricks	2.8B	Y	Y	Y
Falcon	Technology Innovation Institute	7B		OOM	Y
Flan-T5-XXL	Google	11B		N	N
GPT-J	EleutherAI	6B	Y	OOM	Y
GPT-NeoX	EleutherAI	20B	Infers		
Instruct GPT-J	NLPCloud	6B	Infers		
Latent Diffusion LDM	LMU CompVi Lab		Infers		
Llama 2	Meta	7B	Y	OOM	Y
LongChat	LMSys	7B		OOM	Y
MPT Chat	MosaicML	7B	Y	OOM	Y
OpenAssistant StableLM	OpenAssistant	7B	Infers		
OpenChat	OpenChat	13B	Infers		
OpenLLaMA	OpenLM Research	3B	Y	Y	Y
OPT	Meta	2.7B	Y	Y	Y
OrcaMini	Microsoft	3B	Infers		
RedPajama	Together.ai	3B	Y		Y
RedPajama	Together.ai	7B	Y		
Replit Code v1	Replit	3B	Y	Y	Y
Stable Diffusion 1.5	Runway ML	860M	Y	Y	Y
Stable Diffusion 2.1	Stability AI	1.45B	Y	Y	Y
StableLM	Stability AI	3B	Infers		
StableLM	Stability AI	7B	Y	OOM	OOM
Vicuna	LMSys	7B	Infers		
WizardCoder	WizardLM	15B	Infers		
WizardLM	WizardLM	13B	Infers		
XGen Instruct	Salesforce	7B		OOM	Y

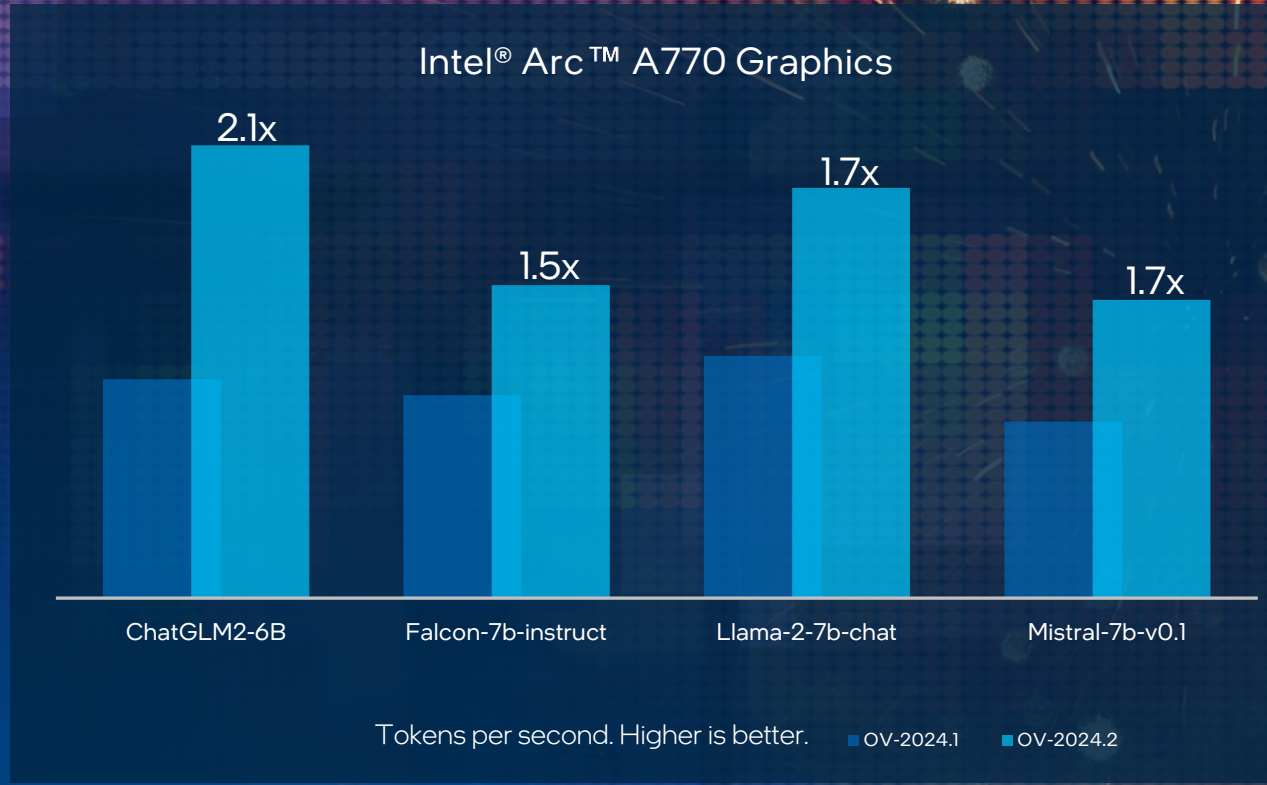
LLMs Functionality Tested & Confirmed

Model	Column2	Column3	Parameters	Column5	Release By	Link
aquilachat-7b	Text Generation	NLP	7B	PyTorch	FlagAI	https://huggingface.co/sammysun0711/aquilachat-7b-hf
Blenderbot-9b	Conversational	NLP	9.4B	PyTorch	ParlaAI (Facebook)	https://huggingface.co/hyunwoongko/blenderbot-9B
Bloom-176B	Text Generation	NLP	176B	PyTorch	Big Science	https://huggingface.co/bigscience/bloom
bloomz-1b4-zh	Text Generation	NLP	1B	PyTorch	YeungNLP	https://huggingface.co/YeungNLP/bloomz-1b4-zh
bloomz-560m	Text Generation	NLP	560m	PyTorch	BigScience	https://huggingface.co/bigscience/bloomz-560m
chatGLM-6b	Text Generation	NLP	6B	PyTorch	THUDM	https://huggingface.co/THUDM/chatglm-6b/tree/main
Clip-VIT_b_16_image_encoder	Image-to-Text	Multimodal	400M	PyTorch	OpenAI	https://github.com/mlfoundations/open_clip/
Clip-VIT_b_16_text_encoder	Text Generation	NLP	400M	PyTorch	OpenAI	https://github.com/mlfoundations/open_clip/
CodeGen2.5	Text Generation	NLP	7B	PyTorch	Salesforce	https://huggingface.co/Salesforce/codegen25-7b-multi
codegen2-1b	Text Generation	NLP	1B	PyTorch	Salesforce	https://huggingface.co/Salesforce/codegen2-1B
codegen2-3_7b	Text Generation	NLP	3.7B	PyTorch	Salesforce	https://huggingface.co/Salesforce/codegen2-3_7B
codegen2-7b	Text Generation	NLP	7B	PyTorch	Salesforce	https://huggingface.co/Salesforce/codegen2-7B
DeepFloyd/IF	Text-to-Image	Multimodal	11B	PyTorch	Stability.AI	https://github.com/deep-floyd/IF/
dolly-v2-12b	Text Generation	NLP	12B	PyTorch	Databricks	https://huggingface.co/databricks/dolly-v2-12b
Dolly-v2-2.8B	Text Generation	NLP	2.8B	PyTorch	Databricks	https://huggingface.co/databricks/dolly-v2-2-8b
Flan-T5-XXL	Text-to-text Generation	NLP	11B	PyTorch	Google	https://huggingface.co/google/flan-t5-xxl
GPT-2	Text Generation	NLP	124M	PyTorch	OpenAI	https://huggingface.co/gpt2
GPT-J-6B	Text Generation	NLP	6B	PyTorch	EleutherAI	https://huggingface.co/EleutherAI/gpt-j-6B
GPT-NeoX-20B	Text Generation	NLP	20B	PyTorch	EleutherAI	https://huggingface.co/EleutherAI/gpt-neox-20b
instruct-gpt-j	Text Generation	NLP	6B	PyTorch	NLPCloud	https://huggingface.co/nlpcloud/instruct-gpt-j-fp16

LLMs Functionality Tested & Confirmed

Model	Use Case	Column3	Parameters	Column5	Release By	Link
Latent Diffusion LDM	Text-to-Image	Multimodal		PyTorch	LMU CompVi Lab	https://huggingface.co/CompVis/ldm-super-resolution-4x-openimages/tree/main
llama-7b	Text Generation	NLP	7B	PyTorch	Facebook	https://huggingface.co/decapoda-research/llama-7b-hf
longchat-7b	Text Generation	NLP	7B	PyTorch	LMSys	https://huggingface.co/lmsys/longchat-7b-16k
mpt-7b-instruct	Text Generation	NLP	7B	PyTorch	MosaicML	https://huggingface.co/mosaicml/mpt-7b-instruct
open-assistant-stablelm-7b	Text Generation	NLP	7B	PyTorch	OpenAssistant/EleutherAI/Stability.AI	https://huggingface.co/OpenAssistant/stablelm-7b-sft-v7-epoch-3
openchat	Text Generation	NLP	13B	PyTorch	OpenChat	https://huggingface.co/openchat/openchat_8192
open-llama-3b	Text Generation	NLP	3B	PyTorch	OpenLLAMA Research	https://huggingface.co/openlm-research/open_llama_3b
open-llama-7b	Text Generation	NLP	7B	PyTorch	OpenLLAMA Research	https://huggingface.co/openlm-research/open_llama_7b
opt-2.7b	Image-to-Text	Multimodal	2.7B	PyTorch	Facebook	https://huggingface.co/facebook/opt-2.7b
orca-mini-3b	Text Generation	NLP	3B	PyTorch	Microsoft	https://huggingface.co/psmathur/orca_mini_3b
red-pajama-incite-chat-3b	Text Generation	NLP	3B	PyTorch	TogetherCompute	https://huggingface.co/togethercomputer/RedPajama-INCITE-Chat-3B-v1
red-pajama-incite-chat-7b	Text Generation	NLP	7B	PyTorch	TogetherCompute	https://huggingface.co/togethercomputer/RedPajama-INCITE-7B-Chat
red-pajama-incite-instruct-3b	Text Generation	NLP	3B	PyTorch	TogetherCompute	https://huggingface.co/togethercomputer/RedPajama-INCITE-Instruct-3B-v1
red-pajama-incite-instruct-7b	Text Generation	NLP	7B	PyTorch	TogetherCompute	https://huggingface.co/togethercomputer/RedPajama-INCITE-7B-Instruct
replit-code-v2-instruct-3b	Text Generation	NLP	3B	PyTorch	Replit	https://huggingface.co/teknium/Replit-v2-CodeInstruct-3B
Stable Diffusion v1.5	Text-to-Image	Multimodal	1.45B	PyTorch	Runway ML	https://huggingface.co/runwayml/stable-diffusion-v1-5
StableDiffusion 2.1	Text-to-Image	Multimodal	1.45B	PyTorch	Stability.AI	https://huggingface.co/stabilityai/stable-diffusion-2-1
stablelm-3b	Text Generation	NLP	3B	PyTorch	Stability.AI	https://huggingface.co/stabilityai/stablelm-tuned-alpha-3b
vicuna	Text Generation	NLP	7B	PyTorch	LMSys	https://github.com/lm-sys/FastChat
wizardcoder-15b	Text Generation	NLP	15B	PyTorch	WizardLM	https://huggingface.co/WizardLM/WizardCoder-15B-V1.0
wizardlm-13b	Text Generation	NLP	13B	PyTorch	WizardLM	https://huggingface.co/WizardLM/WizardLM-13B-V1.1
xgen-7b-instruct	Text Generation	NLP	7B	PyTorch	Salesforce	https://huggingface.co/Salesforce/xgen-7b-8k-inst

Unleash the power of Intel Discrete GPUs with up to 2.1X improvement in LLM performance with OpenVINO 2024.2 as compared to 2024.1



Gains up to 2.1x on Intel® Arc™ A770 Graphics with OpenVINO™2024.2 optimization of 2nd token latency

intel ARC GRAPHICS

OpenVINO™

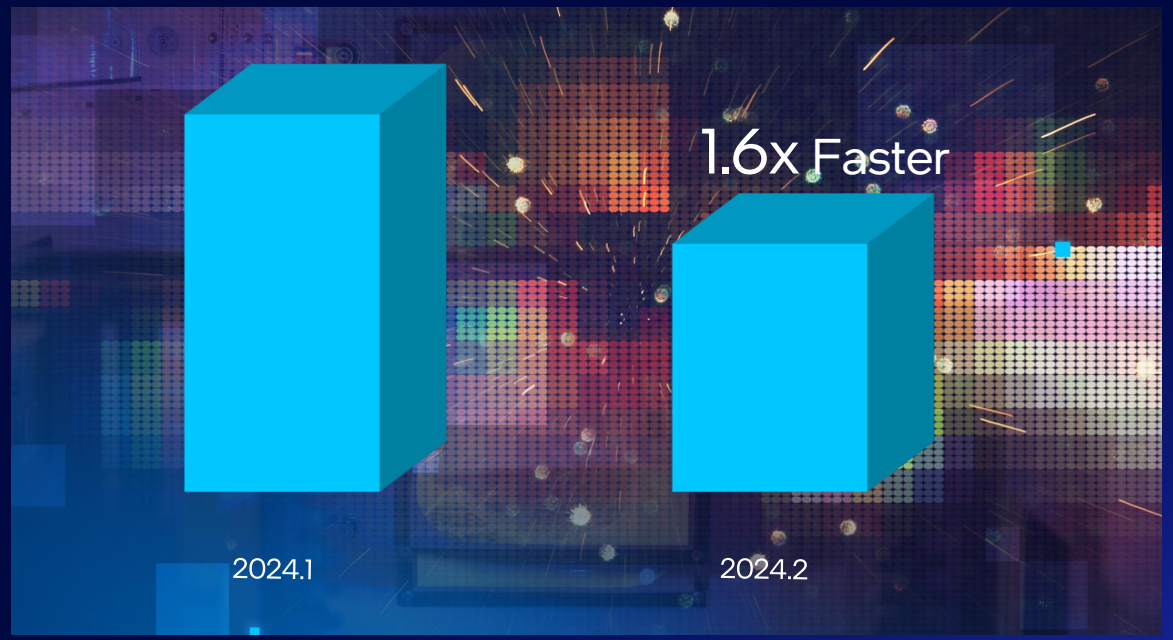
NEW!
Large Language Models Performance Benchmarks for AI PC
not an exhaustive list

ChatGLM2-6B, Llama-2-7b-chat & Mistral-7b-v0.1 - Metric: 2nd token throughput as Tokens Per Second. Input tokens: 1024 | Output token: 128 | Beam search: 1 | Batch size: 1, Precision: INT4
Falcon-7b-instruct - Metric: 2nd token throughput as Tokens Per Second. Input tokens: 32 | Output token: 128 | Beam search: 1 | Batch size: 1, Precision: INT8
For workloads and configurations, see system configuration slides. Results may vary.

Maximize your Stable Diffusion performance with OpenVINO 2024.2

OpenVINO™ release 2024. 2 offers a significant boost in GPU performance over 2024.2

Intel® Discrete GPU Arc™ A770M



Improvement in generation time for responding with a picture /image to a prompt (seconds). Lower is better.

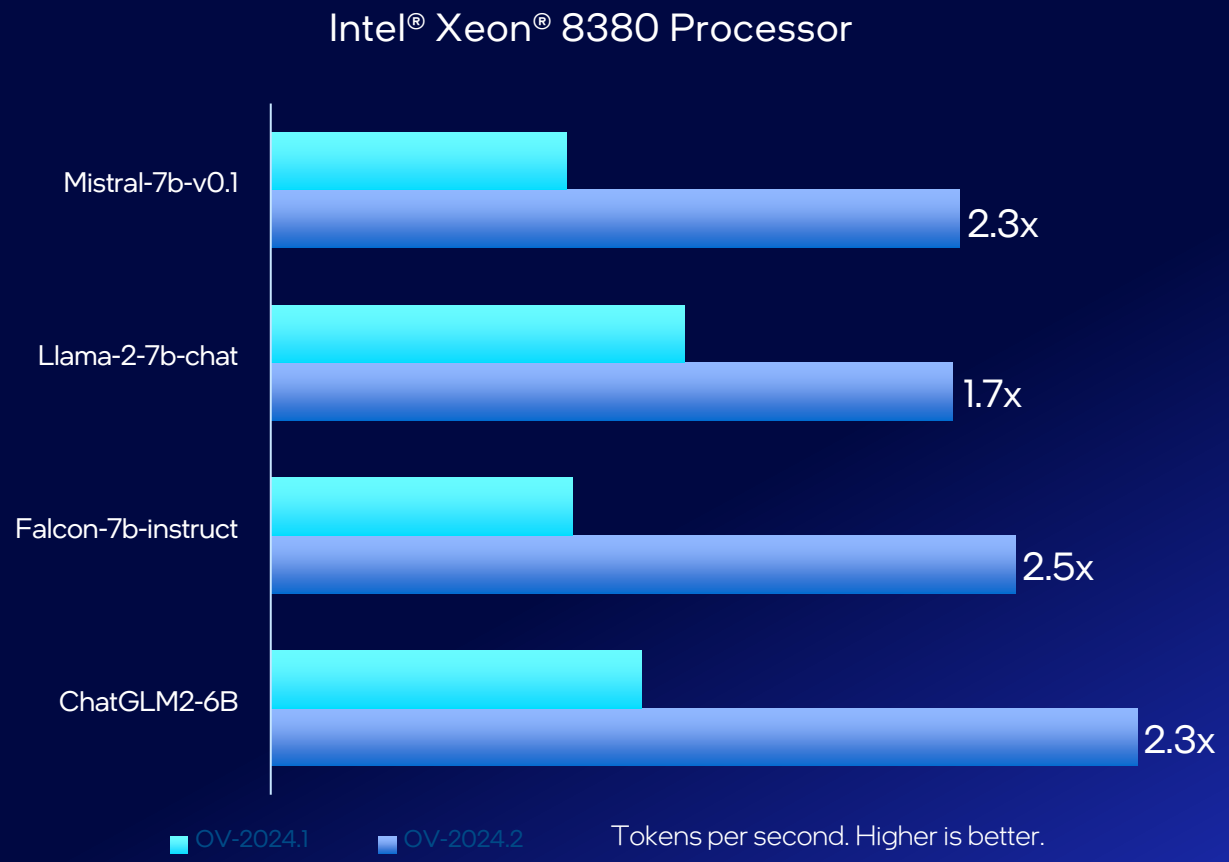
OpenVINO 2024.2 optimizations for dGPUs deliver up to 1.6x performance gains on the Stable Diffusion transformer model.

NEW!
Large Language Models Performance Benchmarks for AI PC
not an exhaustive list

Metric: Generation time for responding with a picture/image to a prompt. Prompt length – 1024, Steps – 20, Precision: FP 16
For workloads and configurations, see system configuration slides. Results may vary.

Optimize LLM Performance for Data Center and Cloud Workloads

OpenVINO™ release 2024. 2 offers up to 2.5x performance gain on 2nd token throughput for Intel® Xeon® Processors.



Servers powered by Intel® Xeon® Processors will benefit from up to 2.5x performance boost for CPUs on 2nd token throughput with OpenVINO 2024.2.

ChatGLM2-6B, Llama-2-7b-chat & Mistral-7b-v0.1 - Metric: 2nd token throughput as Tokens Per Second. Input tokens: 1024 | Output token: 128 | Beam search: 1 | Batch size: 1, Precision: INT8
Falcon-7b-instruct - Metric: 2nd token throughput as Tokens Per Second. Input tokens: 32 | Output token: 128 | Beam search: 1 | Batch size: 1, Precision: INT8
For workloads and configurations, see system configuration slides. Results may vary.

Workloads and System Configurations

CPU Inference Engines:	Intel® Xeon® Platinum 8380
Motherboard	M50CYP2SB1U Coyote Pass
CPU	Intel® Xeon® Gold 8380 CPU @ 2.30GHz
Hyper Threading	on
Turbo Setting	on
Memory	16 x 16 GB DDR4 3200MHz
Operating System	Ubuntu* 22.04.4 LTS
Kernel version	6.5.0-28-generic
BIOS Vendor	Intel Corporation
BIOS Version	SE5C620.86B.01.01.0006.2207150335
BIOS Release	7/15/2022
NUMA nodes	2
Precision	INT8/FP32
Number of concurrent inference requests	80
Test Date	6/3/2024
Power dissipation/socket, TDP in Watt	270
CPU Price on 6/3/2024, Prices may vary	$9,359

GPU Inference Engines:	Intel® Dedicated Graphics Family GPU
GPU	ARC™ 770M
Connection	PCIe G4, 1x16
Batch size	1
Precision	FP16, INT8
Number of concurrent inference requests	Automatic
Memory	16 GB DDR6, 512 GB/s
HPC & AI	FP32, FP16, BF16, INT8, INT4
Form Factor	3/4L Full Height, Passively cooled
X^e cores	32
EUs	512
Device ID	8086-5690
Test Date	6/3/2024
TDP	150 W
Host Machine	Core™ i7-12700H, Serpent Canyon
Motherboard	Intel corp NUC 12SNKi72
CPU	Intel® Core™ i7-12700H CPU @ 2.30GHz
Hyper Threading	on
Turbo Setting	on
Memory	2 x 8 GB DDR4 3200MHz
Operating System	Ubuntu* 22.04.4 LTS
Kernel version	6.5.0-28-generic
BIOS Vendor	Intel Corporation
BIOS Version	SNADL357.0056.2022.1102.1218
BIOS Release	11/2/2022

Notices and Disclaimers

Performance varies by use, configuration, and other factors. Learn more at intel.com/PerformanceIndex.

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® optimizations, for Intel® compilers or other products, may not optimize to the same degree for non-Intel products.

Foundational Generative AI Use Cases

 Large Foundation Models	Models (search HF model hub with below strings): • LLaMa2: meta-llama/Llama-2-7b-chat-hf • LLaMa2 fine tuning: https://huggingface.co/Intel/neural-chat-7b-v2 • MPT: mosaicml/mpt-7b-chat, Intel/neural-chat-7b-v1-1 [intel's fine tune version, better than mpt-7b-chat] • Falcon: tiiuae/falcon-7b • CodeGen: Starcoder: bigcode/starcoderbase-7b Code (llm fine-tuning, inference deployment): • https://github.com/intel/intel-extension-for-transformers/tree/main/workflows/chatbot
 LLM Workflows	LLM BF16 and INT8 inference containers: https://github.com/intel/intel-extension-for-transformers/tree/main/workflows/chatbot/inference PEFT fine-tuning with own dataset for example Alpaca approach: https://github.com/intel/intel-extension-for-transformers/tree/main/workflows/chatbot/fine_tuning INT8 smooth quant workflow: • StarCoder: https://github.com/intel/intel-extension-for-transformers/tree/main/examples/huggingface/pytorch/code-generation/quantization • GPT-j, OPT, MPT, Falcon: https://github.com/intel/intel-extension-for-transformers/tree/main/examples/huggingface/pytorch/text-generation/quantization Model optimization for distillation and sparsity https://github.com/intel/neural-compressor/tree/master/examples/pytorch/nlp/huggingface_models/language-modeling/pruning/eager Synthetic Text data Gen Reference Kit: https://github.com/oneapi-src/text-data-generation

Vertical Generative AI Use Cases

 Financial Services	Business Acceleration	Security and Fraud
	FinGPT-v3 (finetuned with LoRA): https://github.com/AI4Finance-Foundation/FinGPT/tree/master/fingpt/FinGPT-v3 VSE Semantic Search Reference Kit: https://github.com/oneapi-src/vertical-search-engine	Credit Card Fraud Detection **Model:** XGBoost, GNN and HPO **Repo + Readme:** https://github.com/intel/credit-card-fraud-detection **Dataset:** synthetic credit card data: https://github.com/IBM/TabFormer/tree/main/data/credit_card
 Manufacturing	Business Acceleration	
	Visual Quality Inspection: Model: CV models https://github.com/intel/visual-quality-inspection Dataset: MVTec AD	

Vertical Generative AI Use Cases

	Business Acceleration	Personalized Solutions
Retail & Consumers Services	HF Blog: https://huggingface.co/blog/supercharge-customer-service-with-machine-learning	Drone Navigation Ref Kit: https://github.com/oneapi-src/drone-navigation-inspection
Healthcare	Drug Development: alphafold2 https://github.com/IntelAI/models/blob/master/quickstart/aidd/pytorch/alphafold2/inference/README BIOGPT:https://huggingface.co/microsoft/biogpt.md	Precision medicine (analyze genomics, clinical, imaging). Predictive and preventative care Disease Prediction Ref Kit: https://github.com/oneapi-src/disease-prediction PAM Reference Kit https://github.com/oneapi-src/predictive-asset-health-analytics
Network and Security	Security and Fraud	
	Malware Detection (MalConv): • Core model: https://github.com/elastic/ember/tree/master/malconv • https://networkbuilders.intel.com/solutionslibrary/intel-deep-learning-boost-boost-network-security-ai-inference-performance-in-google-cloud-platform-gcp-technology-guide E-mail Phishing Detection: • Core model https://huggingface.co/bert-base-cased. Quantization steps: https://networkbuilders.intel.com/solutionslibrary/intel-deep-learning-boost-intel-dl-boost-improve-inference-performance-of-hugging-face-bert-base-model-in-google-cloud-platform-gcp-technology-guide	