Generative AI Landscape: From Popular Services to DIY Solutions on Azure

Clemente Giorio
Gianni Rosa Gallina

Platinum Sponsor

Microsoft

avanade

REPLY CLUSTER

PORINI
A DGS COMPANY

Gold Sponsor

LOBRA

Reti
Società Benefit

UNIKEY
Bringing IT knowledge to the people

Silver Sponsor

PHILMARK GROUP

Technical Sponsor

NET CODE

JET BRAINS

Packt

stickermule

Generative AI

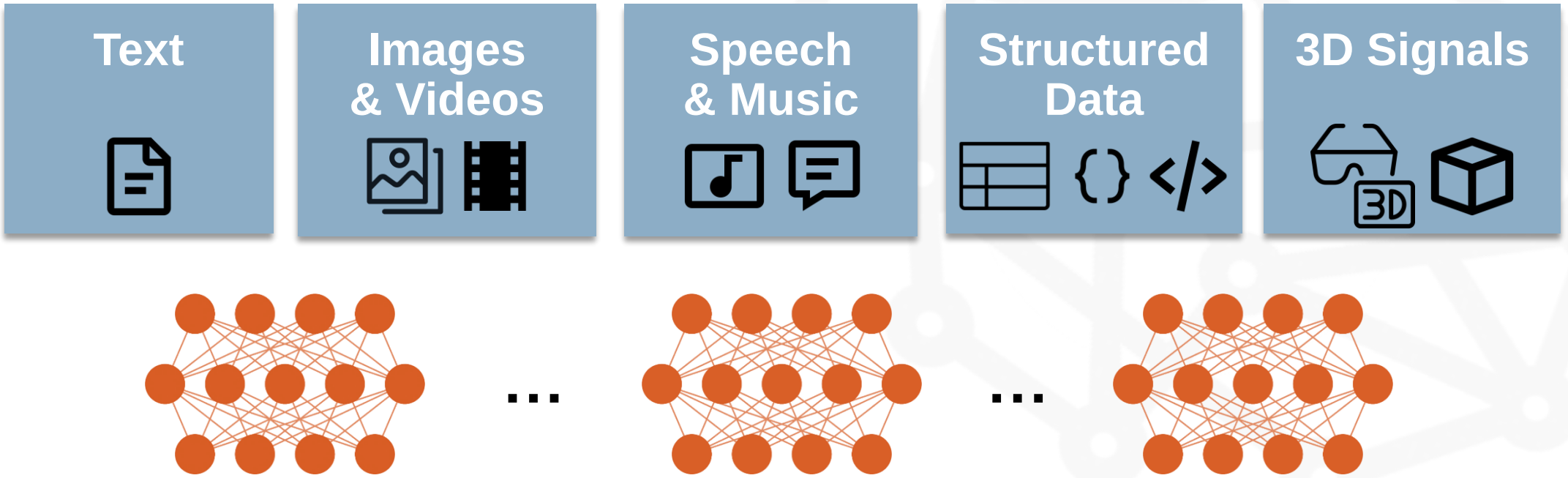
“Photo of a gladiator, seated at an office desk, staring into the distance with a thoughtful expression. A lightbulb icon, symbol of an idea.”

Generative AI
Magic

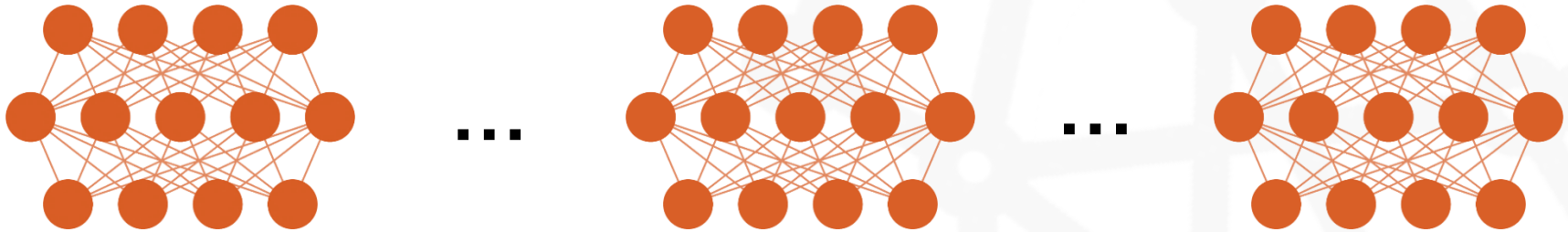
"Any sufficiently advanced technology is indistinguishable from magic"

Arthur C. Clarke

Generative AI
Overview



Generative AI
Overview



Foundation Models

- Question Answering
- Information Extraction
- Image Captioning
- Multimodal Translation
- Text-To-X
- Image-To-X

Generative AI... for all and everything

- Arts & Photography
- Design
- Fashion
- Writing
- Sounds & Music
- Gaming
- Architecture
- Marketing
- Customer Support
- Advertising
- Programming
- Scientific Research
- Cinema

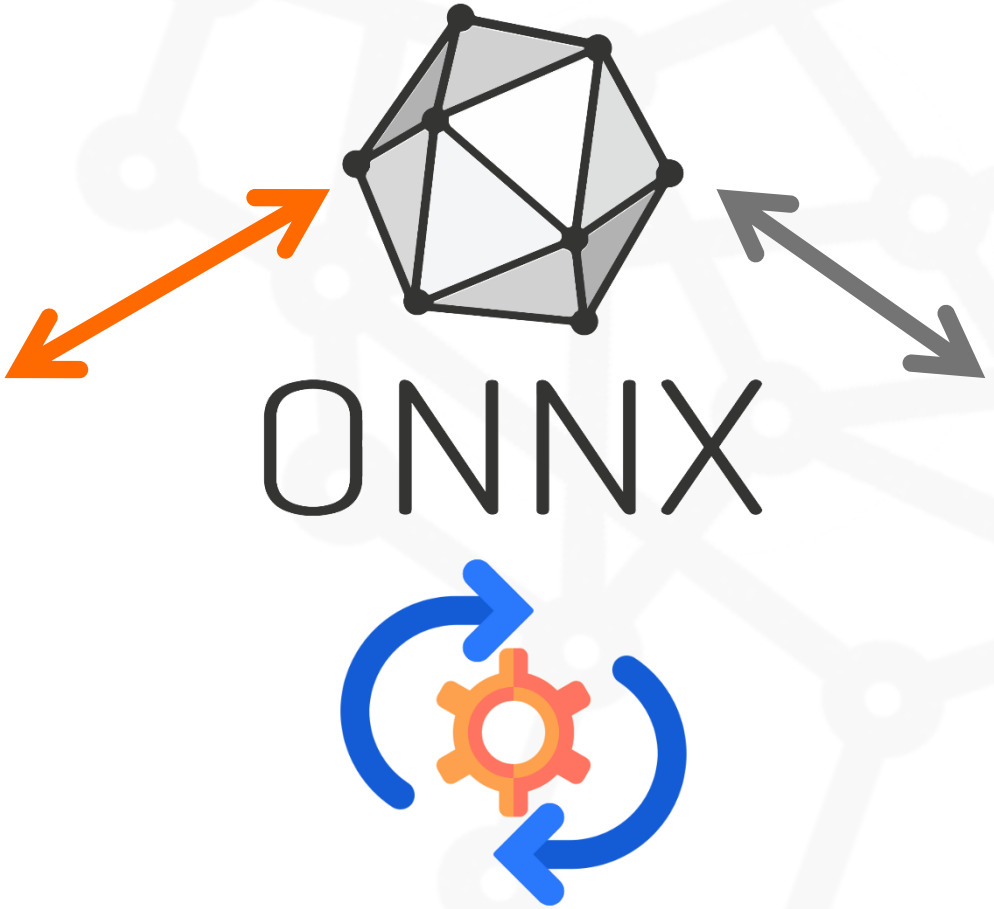
...

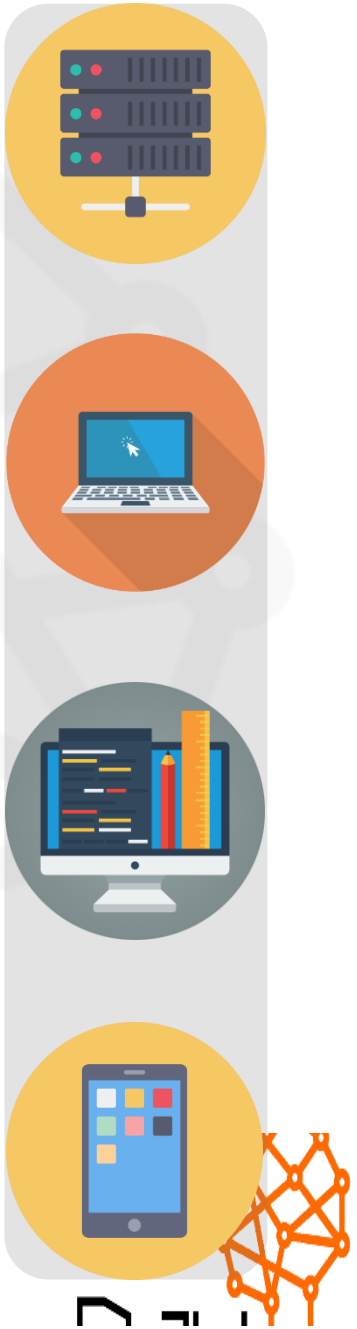
Tools and Frameworks

PyTorch

TensorFlow

OpenVINO™

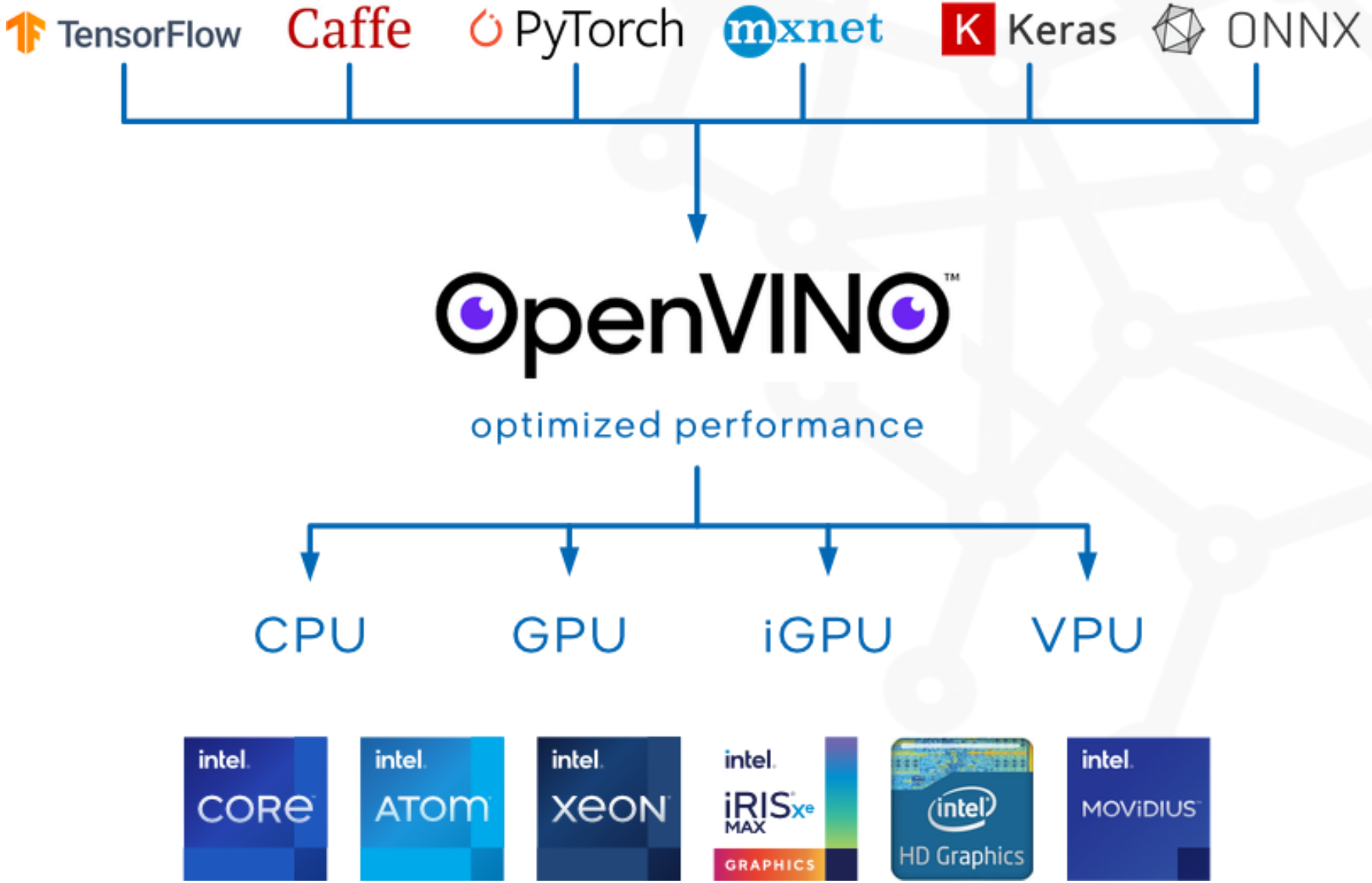




https://onnx.ai/

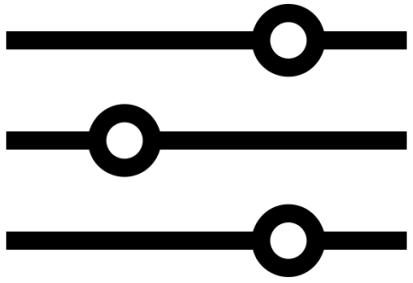
Tools and Frameworks

OpenVINO

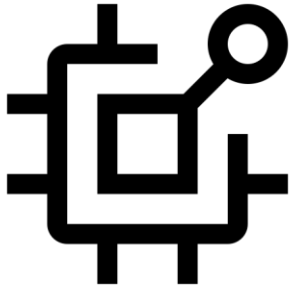


Open Source Models

Reasons



Customization & Flexibility



Embedded/Mobile Devices

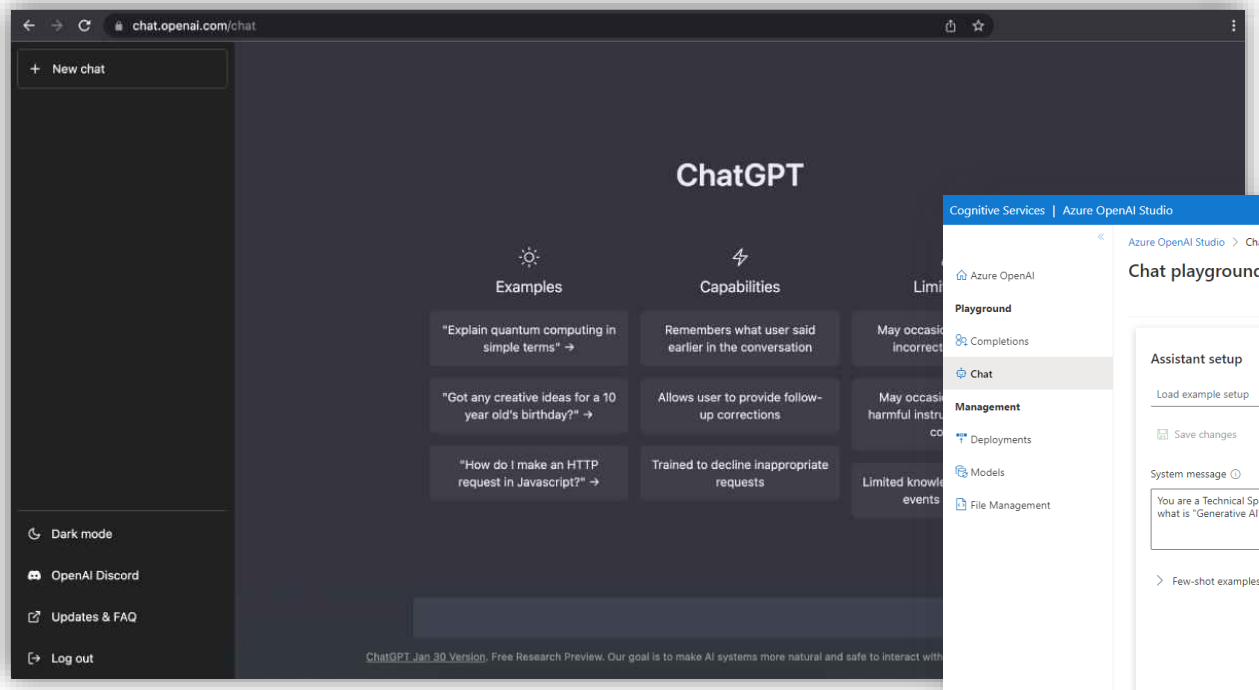
Data Policies & Ownership

No Connectivity

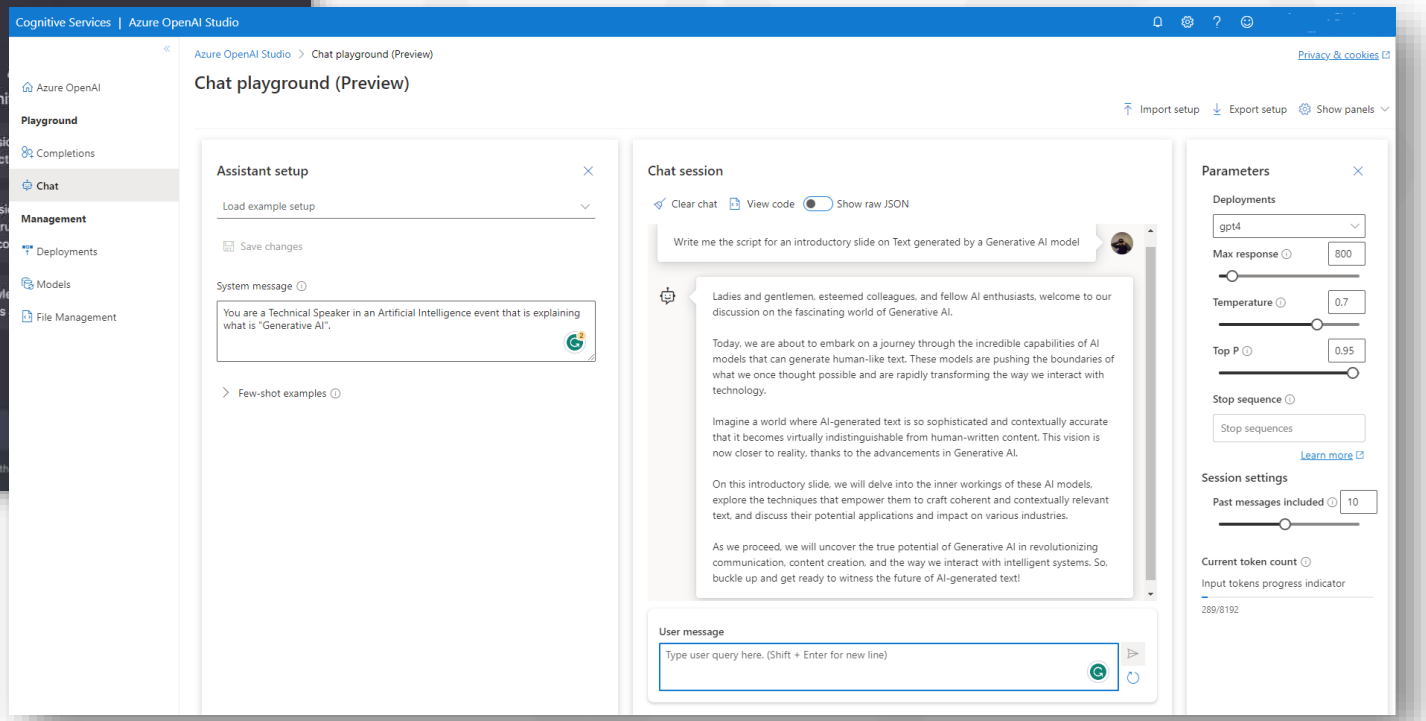
Savings & Optimization

Generative AI

Text



https://chat.openai.com/

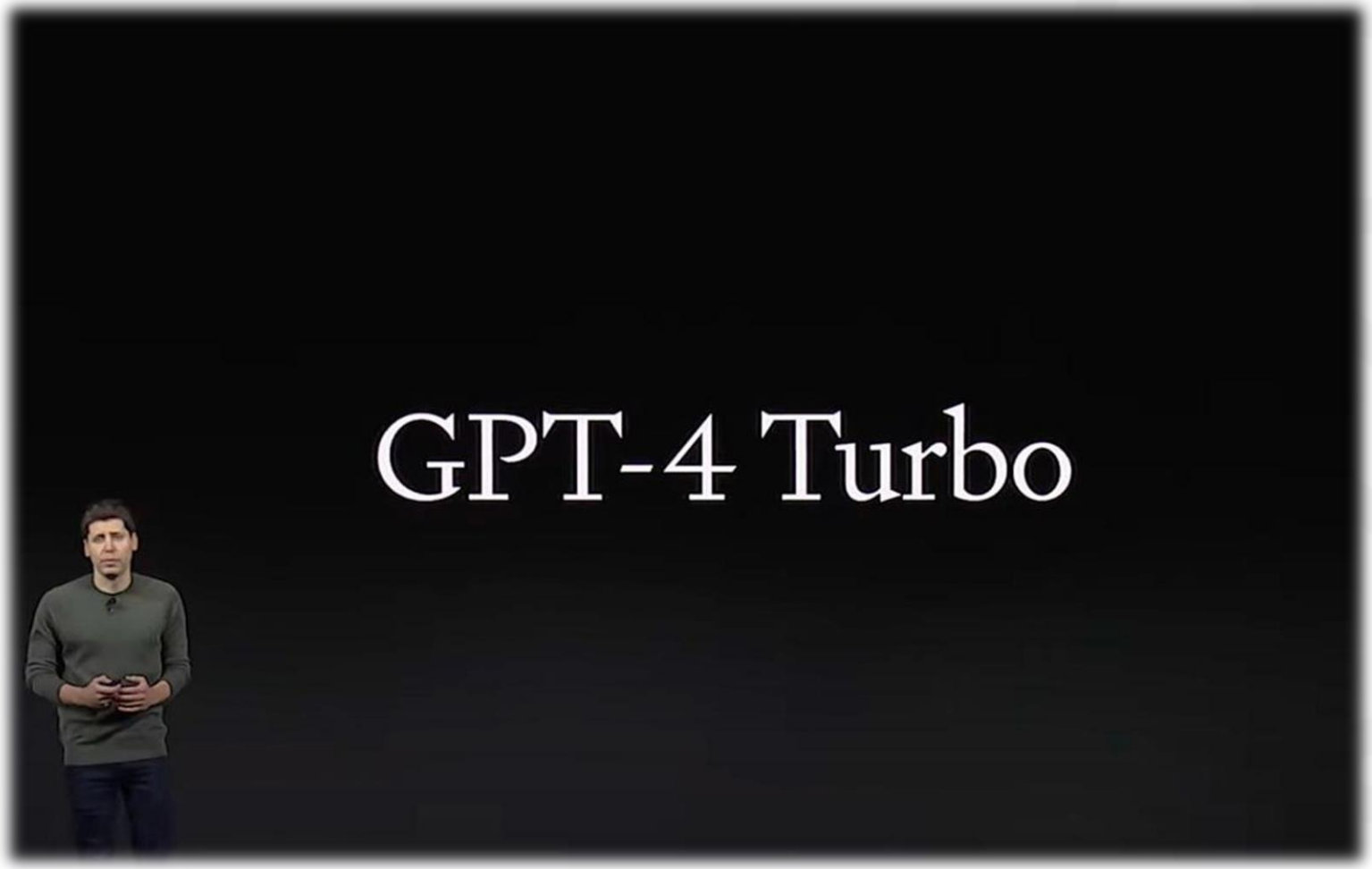


https://azure.microsoft.com/en-us/products/ai-services/openai-service/

OpenAI

Microsoft

Generative AI
OpenAI updates



	GPT-4 Turbo
Cut-Off date	April 2023
Context Window	128K Tokens
Multimodal Abilities	yes

https://openai.com/blog/new-models-and-developer-products-announced-at-devday
Microsoft Ignite 2023: Azure OpenAI Service Announces New Models and Multimodal Advancements

Generative AI
Text-To-Image

"Style painting of a modernized Rome, with the Colosseum"

DALL·E 3

https://openai.com/dall-e-3

https://www.bing.com/images/create/

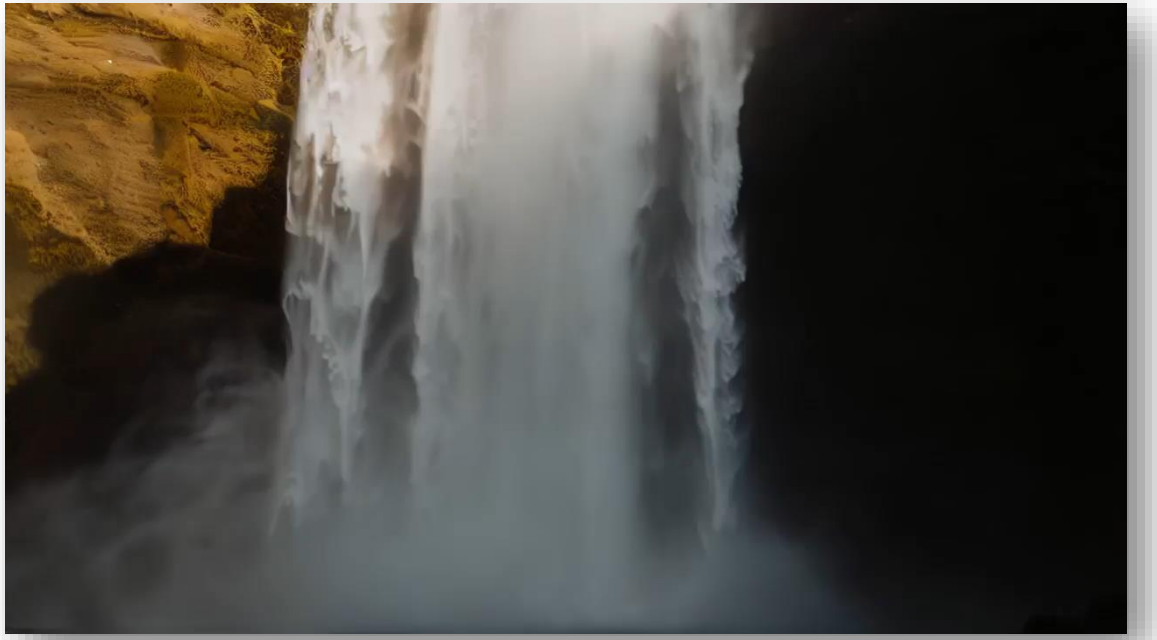
Generative AI
Text-To-Image

“[Inter/Juventus/Milan] club as woman, She wears the [Inter/Juventus/Milan] jersey, ultrarealistic, ultrahd, 4K”

https://midjourney.com/

Generative AI
Text-To-Video

runway



https://runwayml.com/

Images
Stable Diffusion

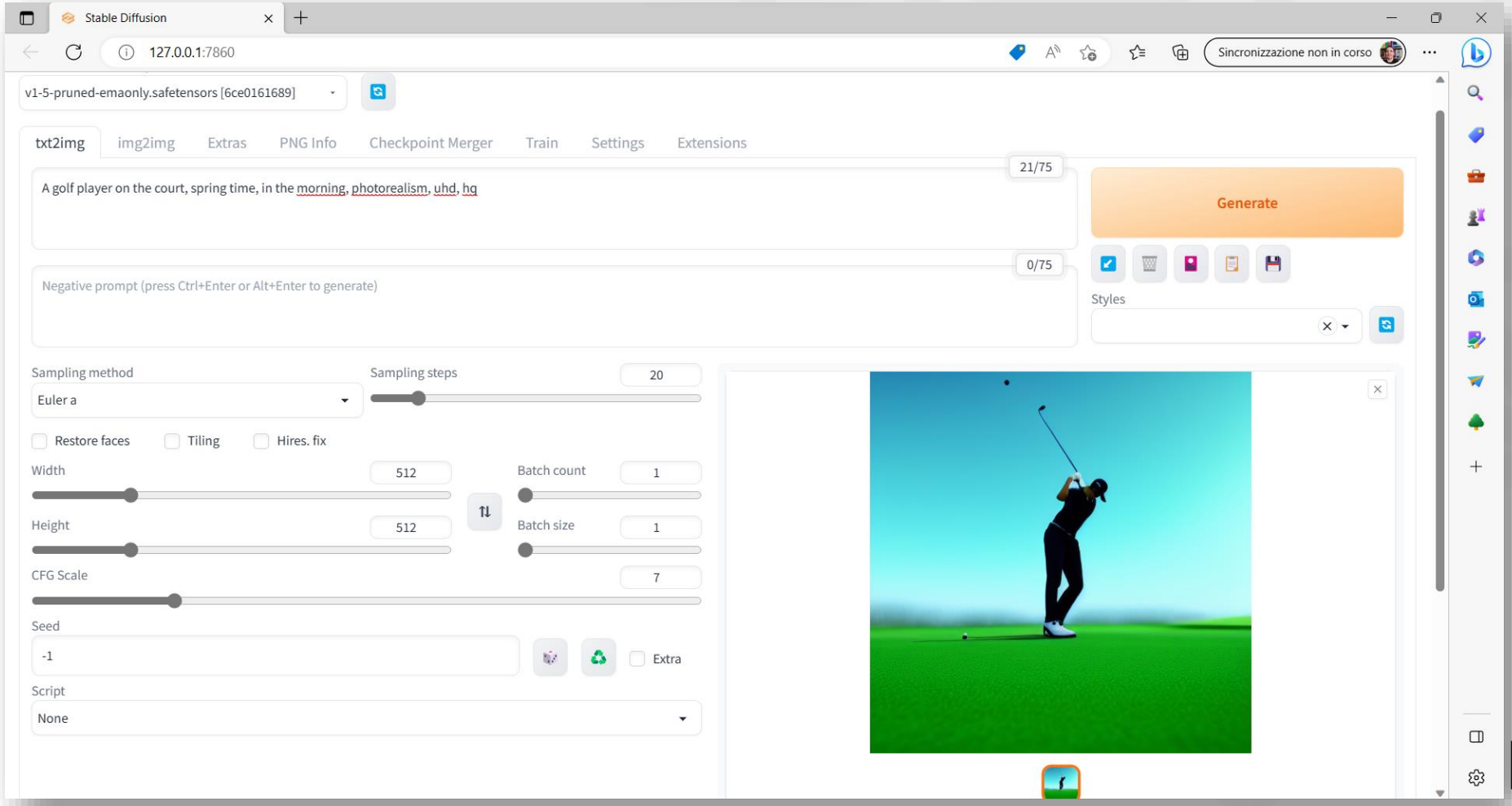



stability.ai runway LAION Hugging Face

https://huggingface.co/blog/stable_diffusion
https://stability.ai/stable-diffusion
https://github.com/huggingface/diffusers
https://runwayml.com/

Diffusers

Images
Stable Diffusion – Web UI Tool

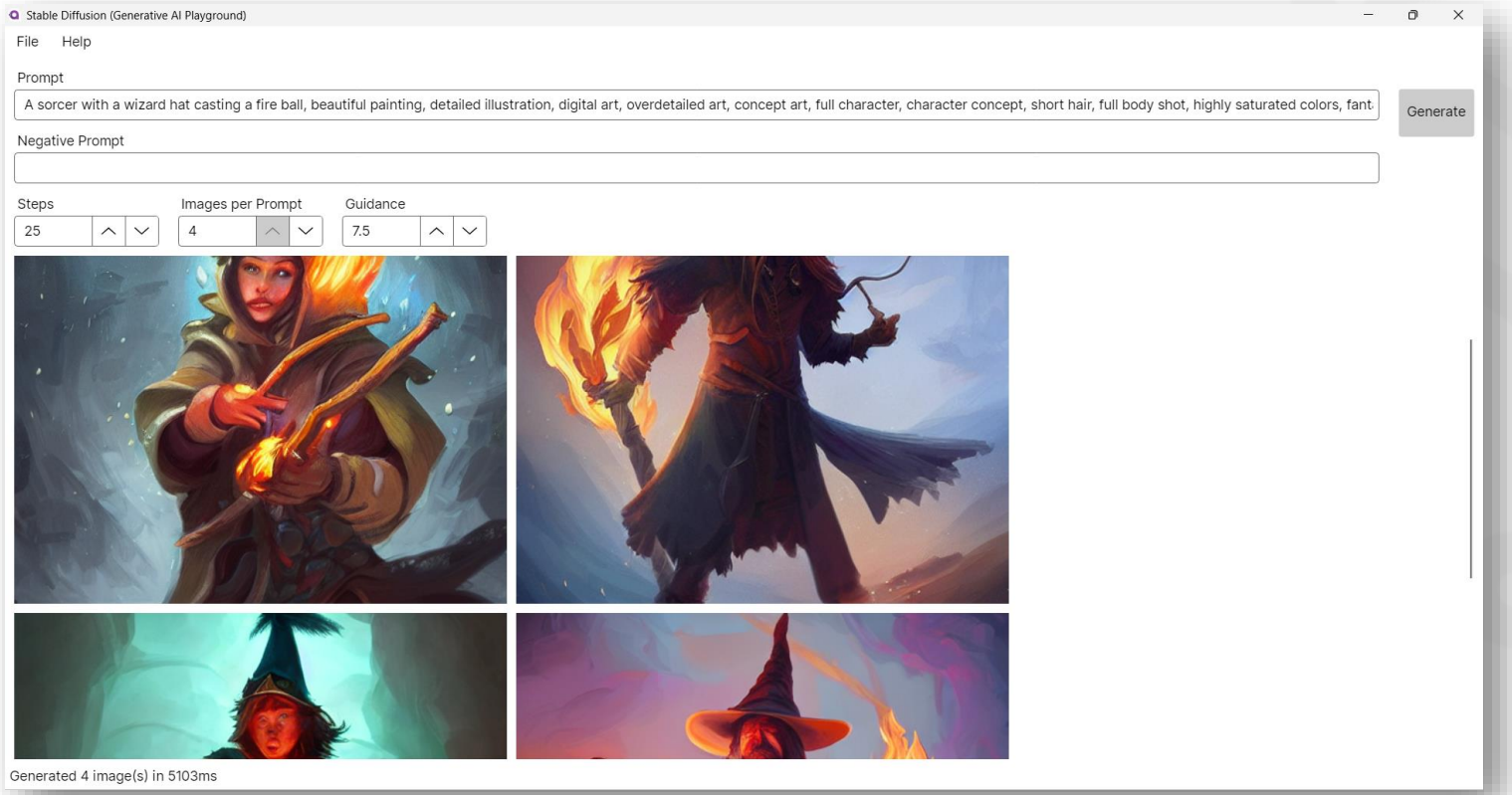


https://github.com/AUTOMATIC1111/stable-diffusion-webui/

Images
Stable Diffusion: ONNX Runtime & .NET

Generative AI Playground .NET

https://github.com/gianni-rg/gen-ai-net-playground
https://github.com/gianni-rg/SharpDiffusion



ONNX

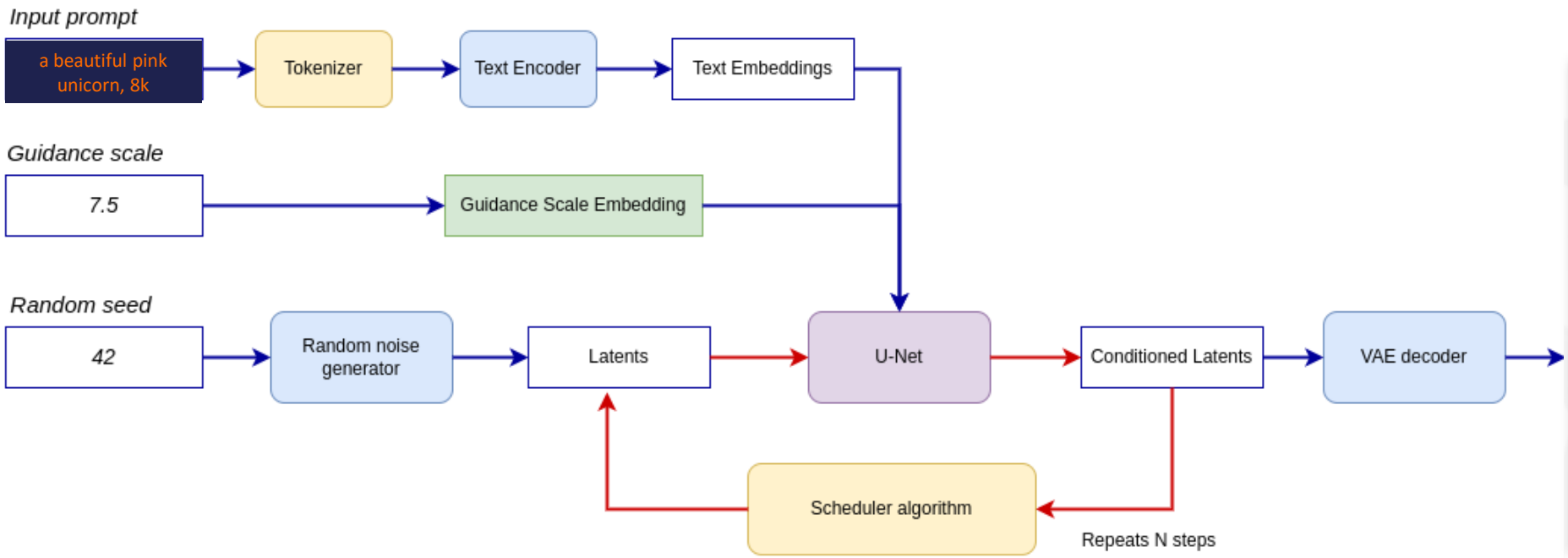
.NET Framework

DEMO

Images
Latent Consistency Models

NEW

Latent Consistency Models: Synthesizing High-Resolution Images with Few-Step Inference
https://arxiv.org/abs/2310.04378



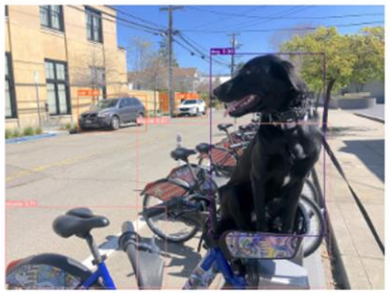
Diffusers

https://huggingface.co/docs/diffusers/main/en/api/pipelines/latent_consistency_models

OpenVINO™ Notebooks

AI Trends - Notebooks

Check out the latest notebooks that show how to optimize and deploy popular models on Intel CPU and GPU.

Notebook	Description	Preview	Complementary Materials
YOLOv8 - Optimization	Optimize YOLOv8 using NNCF PTQ API		Blog - How to get YOLOv8 Over 1000 fps with Intel GPUs?
SAM - Segment Anything Model	Prompt based object segmentation mask generation using Segment Anything and OpenVINO™		Blog - SAM: Segment Anything Model — Versatile by itself and Faster by OpenVINO
ControlNet - Stable-Diffusion	A Text-to-Image Generation with ControlNet Conditioning and OpenVINO™		Blog - Control your Stable Diffusion Model with ControlNet and OpenVINO
			

Text-to-Image Generation with Stable Diffusion v2 and OpenVINO™

Stable Diffusion v2 is the next generation of Stable Diffusion model a Text-to-Image latent diffusion model created by the researchers and engineers from Stability AI and LAION.

General diffusion models are machine learning systems that are trained to denoise random gaussian noise step by step, to get to a sample of interest, such as an image. Diffusion models have shown to achieve state-of-the-art results for generating image data. But one downside of diffusion models is that the reverse denoising process is slow. In addition, these models consume a lot of memory because they operate in pixel space, which becomes unreasonably expensive when generating high-resolution images. Therefore, it is challenging to train these models and also use them for inference. OpenVINO brings capabilities to run model inference on Intel hardware and opens the door to the fantastic world of diffusion models for everyone!

In previous notebooks, we already discussed how to run Text-to-Image generation and Image-to-Image generation using Stable Diffusion v1 and controlling its generation process using ControlNet. Now is turn of Stable Diffusion v2.

Stable Diffusion v2: What's new?

The new stable diffusion model offers a bunch of new features inspired by the other models that have emerged since the introduction of the first iteration. Some of the features that can be found in the new model are:

- The model comes with a new robust encoder, OpenCLIP, created by LAION and aided by Stability AI; this version v2 significantly enhances the produced photos over the V1 versions.
- The model can now generate images in a 768x768 resolution, offering more information to be shown in the generated images

https://github.com/openvinotoolkit/openvino_notebooks

DEMO

Audio
Speech-To-Text: Whisper

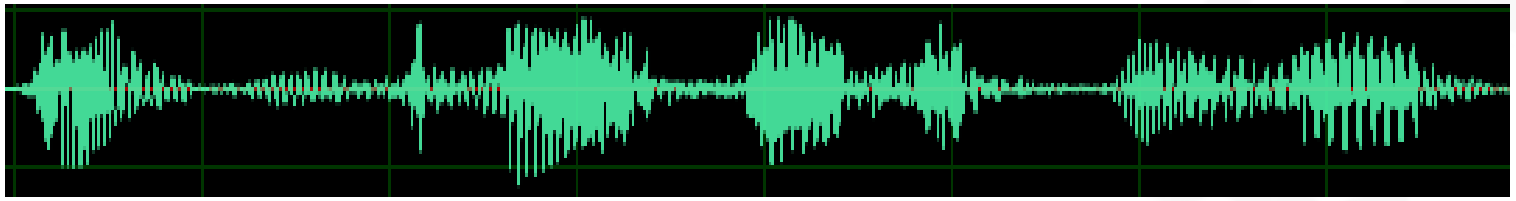
This is the Micro Machine Man presenting the most midget miniature motorcade of Micro Machines. Each one has dramatic details, terrific trim, precision paint jobs, plus incredible Micro Machine Pocket Play Sets. There's a police station, fire station, restaurant, service station, and more. Perfect pocket portables to take any place. And there are many miniature play sets to play with, and each one comes with its own special edition Micro Machine vehicle and fun, fantastic features that miraculously move. Raise the boatlift at the airport marina. Man the gun turret at the army base. Clean your car at the car wash. Raise the toll bridge. And these play sets fit together to form a Micro Machine world. Micro Machine Pocket Play Sets, so tremendously tiny, so perfectly precise, so dazzlingly detailed, you'll want to pocket them all. Micro Machines are Micro Machine Pocket Play Sets sold separately from Galoob. The smaller they are, the better they are.

OpenAI

https://openai.com/research/whisper

Audio

Speech-To-Text: Whisper v3



Robust Speech Recognition via Large-Scale Weak Supervision

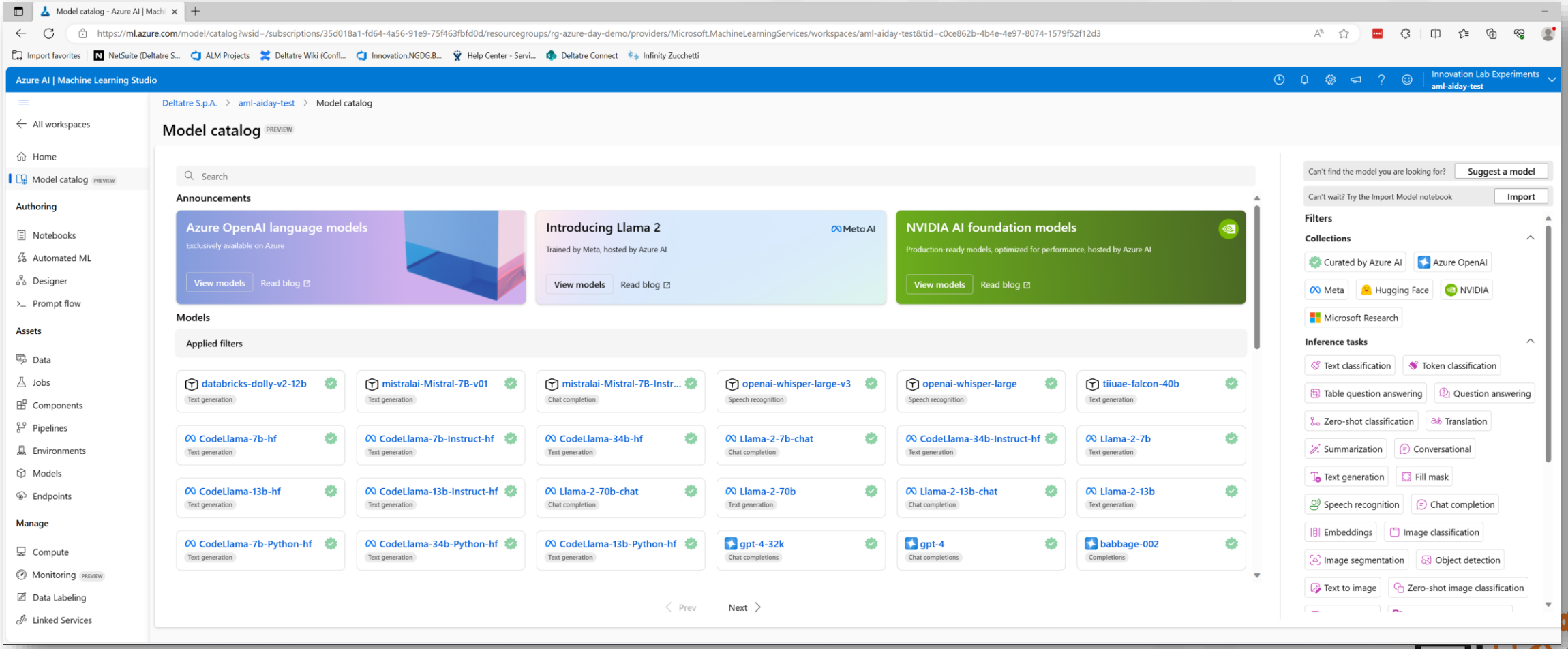
https://arxiv.org/abs/2212.04356
https://github.com/openai/whisper
https://github.com/openai/whisper/discussions/1762

Size	Parameters	English-only model	Multilingual model	Required VRAM	Relative speed
tiny	39 M	`tiny.en`	`tiny`	~1 GB	~32x
base	74 M	`base.en`	`base`	~1 GB	~16x
small	244 M	`small.en`	`small`	~2 GB	~6x
medium	769 M	`medium.en`	`medium`	~5 GB	~2x
large	1550 M	N/A	`large`	~10 GB	1x

Local Audio Transcription in .NET

https://github.com/ggerganov/whisper.cpp
https://github.com/sandrohanea/whisper.net
https://github.com/gianni-rg/gen-ai-net-playground

Azure Machine Learning
Foundation Models



https://learn.microsoft.com/en-us/azure/machine-learning/concept-foundation-models

DEMO

Large Language Models (LLMs)
LLaMA 2

Meta and Microsoft introduce the Next Generation of Llama
https://about.fb.com/news/2023/07/llama-2/
https://github.com/facebookresearch/llama

Llama 2: Open Foundation and Fine-Tuned Chat Models
https://arxiv.org/pdf/2307.09288.pdf

Prompt: "*A cartoon image capturing the lively essence of a popular animated series. The central character is an anthropomorphic llama, standing on two legs, shouting 'NET!!!'*"

License

Model and weights are licensed for both **research AND commercial use**, upholding the principles of openness.

https://ai.meta.com/llama/license/
https://github.com/facebookresearch/llama/blob/main/LICENSE

Large Language Models (LLMs)

LLaMA 2 with LLaMA.cpp

https://github.com/ggerganov/llama.cpp

Model	Original size	Quantized size (4-bit)
7B	13 GB	3.9 GB
13B	24 GB	7.8 GB
30B	60 GB	19.5 GB
65B	120 GB	38.5 GB

Model	Measure	F16	Q4_0	Q4_1	Q5_0	Q5_1	Q8_0
7B	perplexity	5.9066	6.1565	6.0912	5.9862	5.9481	5.9070
7B	file size	13.0G	3.5G	3.9G	4.3G	4.7G	6.7G
7B	ms/tok @ 4th	127	55	54	76	83	72
7B	ms/tok @ 8th	122	43	45	52	56	67
7B	bits/weight	16.0	4.5	5.0	5.5	6.0	8.5

Supported models:

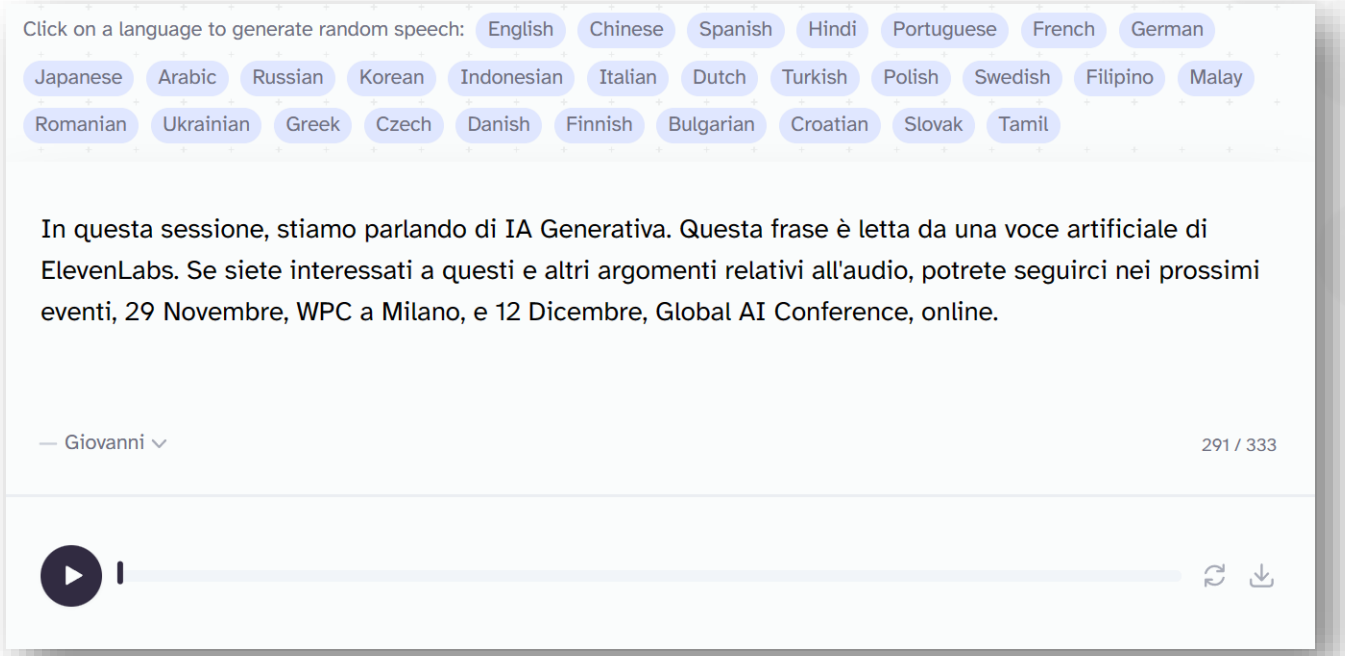
- LLaMA
- LLaMA 2
- Falcon
- Alpaca
- GPT4All
- Chinese LLaMA / Alpaca and Chinese LLaMA-2 / Alpaca-2
- Vigogne (French)
- Vicuna
- Koala
- OpenBuddy (Multilingual)
- Pygmalion/Metharme
- WizardLM
- Baichuan 1 & 2 + derivations
- Aquila 1 & 2
- Starcoder models
- Mistral AI v0.1
- Refact
- Persimmon 8B
- MPT
- Bloom

LLaMa Sharp

https://github.com/SciSharp/LLamaSharp
https://github.com/gianni-rg/gen-ai-net-playground
https://github.com/gianni-rg/LlamaSharpApiServer/

DEMO

Audio
Text-To-Speech



https://elevenlabs.io/

```python
import io
from openai import OpenAI
from pydub import AudioSegment
from pydub.playback import play

client = OpenAI()

def stream_and_play(text):
  response = client.audio.speech.create(
    model="tts-1",
    voice="alloy",
    input=text,
  )

  # Convert the binary response content to a byte stream
  byte_stream = io.BytesIO(response.content)

  # Read the audio data from the byte stream
  audio = AudioSegment.from_file(byte_stream, format="mp3")

  # Play the audio
  play(audio)


if __name__ == "__main__":
  text = input("Enter text: ")
  stream_and_play(text)
```

https://platform.openai.com/docs/guides/text-to-speech

Audio

Text-To-Speech

https://www.wpc.education/


```python
import io
from openai import OpenAI
from pydub import AudioSegment
from pydub.playback import play

client = OpenAI()

def stream_and_play(text):
```

https://globalai.community/events/global-ai-conference-december-2023/

https://elevenlabs.io/

Recap

- Generative AI overview & tools
- Stable Diffusion (Web & .NET)
- Latent Consistency Models
- OpenVINO & Notebooks
- LLMs: LLaMA 2, llama.cpp & .NET
- Whisper
- Azure Machine Learning Catalog

Thank You!

ευχαριστώ Salamat Po متشكرم شكراً Grazie

благодаря ありがとうございます Kiitos Teşekkürler 谢谢

ขอบคุณครับ Obrigado شكريه Terima Kasih Dziękuję

Hvala Köszönöm Tak Dank u wel дякую Tack

Mulţumesc спасибо Danke Cám ơn Gracias

多謝晒 Ďakujem תודה நன்றி Děkuji 감사합니다

AI Day

Microsoft

avanade

REPLY CLUSTER

PORINI
A DGS COMPANY

LOBRA

Reti
Società Benefit

UNIKEY
Bringing IT knowledge to the people

PHILMARK GROUP

JET BRAINS

Packt>

stickermule

NET CODE

References (1/2)

- https://www.bing.com/images/create/
- https://midjourney.com/
- https://azure.microsoft.com/en-us/products/cognitive-services/openai-service
- https://runwayml.com/
- https://research.runwayml.com/gen2
- https://openai.com/dall-e-3
- https://openai.com/research/whisper
- https://www.youtube.com/watch?v=17_xLsqny9E
- https://beta.elevenlabs.io/
- https://github.com/steven2358/awesome-generative-ai
- https://github.com/imaurer/awesome-decentralized-llm
- https://onnx.ai/
- https://docs.openvino.ai/
- https://github.com/openvinotoolkit/openvino_notebooks

References (2/2)

- https://huggingface.co/blog/stable_diffusion
- https://huggingface.co/blog/annotated-diffusion
- https://github.com/huggingface/diffusers
- https://github.com/runwayml/stable-diffusion
- https://github.com/AUTOMATIC1111/stable-diffusion-webui/
- https://github.com/gianni-rg/gen-ai-net-playground
- https://github.com/facebookresearch/llama
- https://arxiv.org/abs/2302.13971
- https://github.com/ggerganov/llama.cpp
- https://github.com/SciSharp/LLamaSharp
- https://github.com/gianni-rg/LlamaSharpApiServer/
- https://github.com/oobabooga/text-generation-webui
- https://github.com/openai/whisper
- https://github.com/ggerganov/whisper.cpp
- https://github.com/sandrohanea/whisper.net
- https://whisper.ggerganov.com/talk/
- https://about.fb.com/news/2023/07/llama-2/
- https://github.com/facebookresearch/llama
- https://arxiv.org/pdf/2307.09288.pdf

About Us

Clemente GIORIO

R&D Senior Principal Engineer EYES ON

X @tinux80

- Augmented/Mixed/Virtual Reality
- Artificial Intelligence, Machine Learning, Deep Learning
- Computer Vision, Multimodal Tracking
- Internet of Things
- Hybrid Clusters

MVP Microsoft Most Valuable Professional

dotNET{podcast}

intel Software INNOVATOR

[PACKT] PUBLISHING Author

FAB LAB NAPOLI AI

NVIDIA Certified Associate - AI in the Data Center

About Us

Ing. Gianni ROSA GALLINA

R&D Technical Lead @ deltatre

𝕏@giannirg

- AI, Machine Learning, Deep Learning on multimedia content
- Virtual/Augmented/Mixed Reality
- Immersive video streaming & 3D graphics for sport events
- Cloud solutions, web backends, serverless, video workflows
- Mobile apps dev (Windows / Android / .NET MAUI / Avalonia)
- End-to-end solutions with Microsoft Azure

MVP Microsoft Most Valuable Professional

Microsoft Specialist
Programming in C#
Programming in HTML5 with JavaScript & CSS3

Microsoft CERTIFIED Solutions Developer
Windows Store Apps Using C#
Web Applications

Microsoft CERTIFIED AZURE AI ENGINEER ASSOCIATE

Microsoft CERTIFIED AZURE DATA SCIENTIST ASSOCIATE

unity Certified Programmer

PLURALSIGHT

SCAN ME

https://gianni.rosagallina.com/en/

Session Feedback

Thank you!

Platinum Sponsor

Microsoft

avanade

REPLY CLUSTER

PORINI
A DGS COMPANY

Gold Sponsor

LOBRA

Reti
Società Benefit

UNIKEY
Bringing IT knowledge to the people

Silver Sponsor

PHILMARK GROUP

Technical Sponsor

NET CODE

JET BRAINS

Packt

stickermule