AZURE DAY

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

Clemente **Giorio**
Gianni **Rosa Gallina**

MVP Microsoft Most Valuable Professional

deltatre

Microsoft

msc TECHNOLOGY

NET CODE

Azure Meetup TORINO

TORINO.NET

AZURE DAY

Platinum Sponsor

Microsoft

msc TECHNOLOGY

Technical Sponsor

NET CODE

Azure Meetup TORINO

TORINO.NET

stickermule

Packt

JET BRAINS

Generative AI

“Photo of a young office assistant, seated at an office desk, staring into the distance with a thoughtful expression. A lightbulb icon, sym”

Generative AI…for all and everything

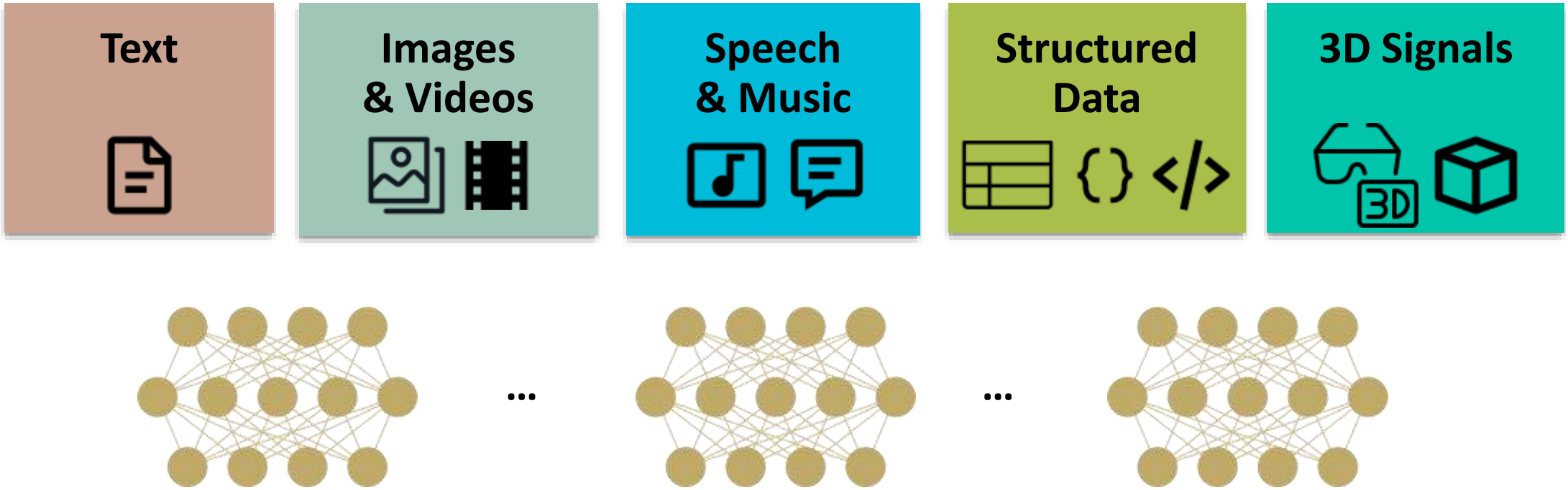
- Arts & Photography
- Design
- Fashion
- Writing
- Sounds & Music
- Gaming
- Architecture
- Marketing
- Customer Support
- Advertising
- Programming
- Scientific Research
- Cinema

…

Generative AI: Magic

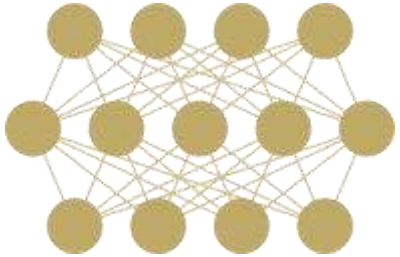
"Photorealistic 8k image showing a wise wizard with an elaborate robe and staff, intently observing a crystal ball. Within the orb, snippets of Python"

Generative AI: Overview



Foundation Models

Generative AI: Overview



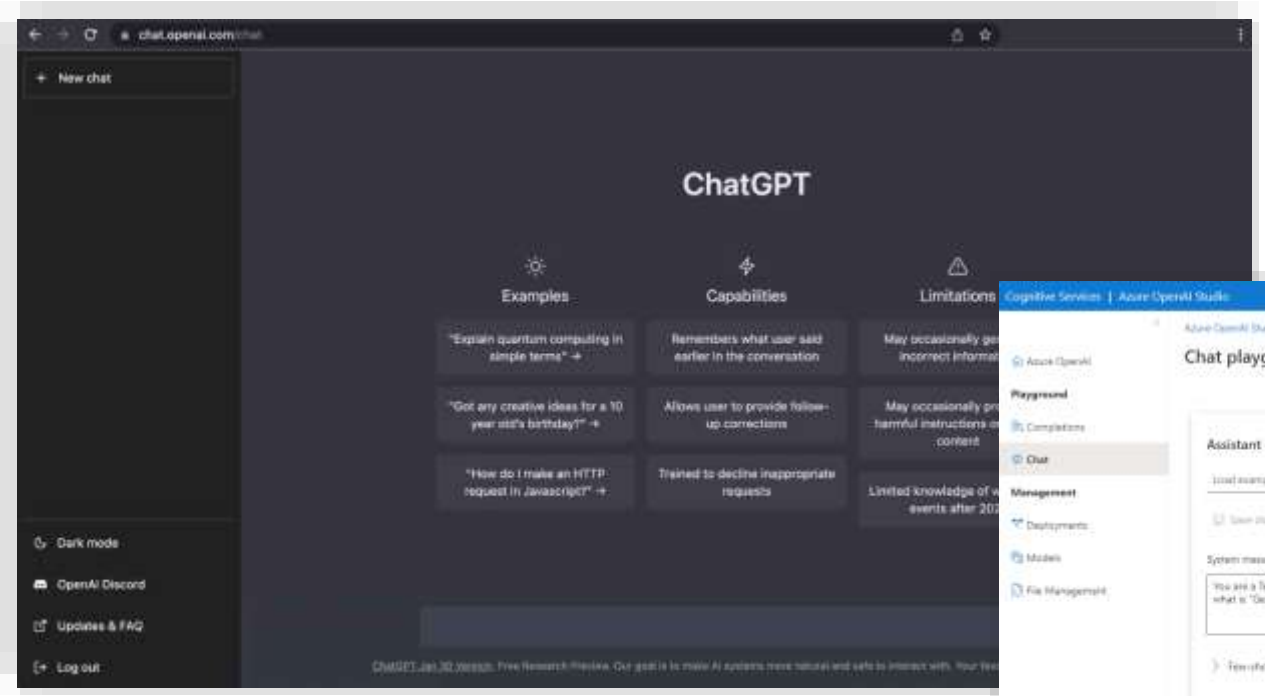
Foundation Models

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

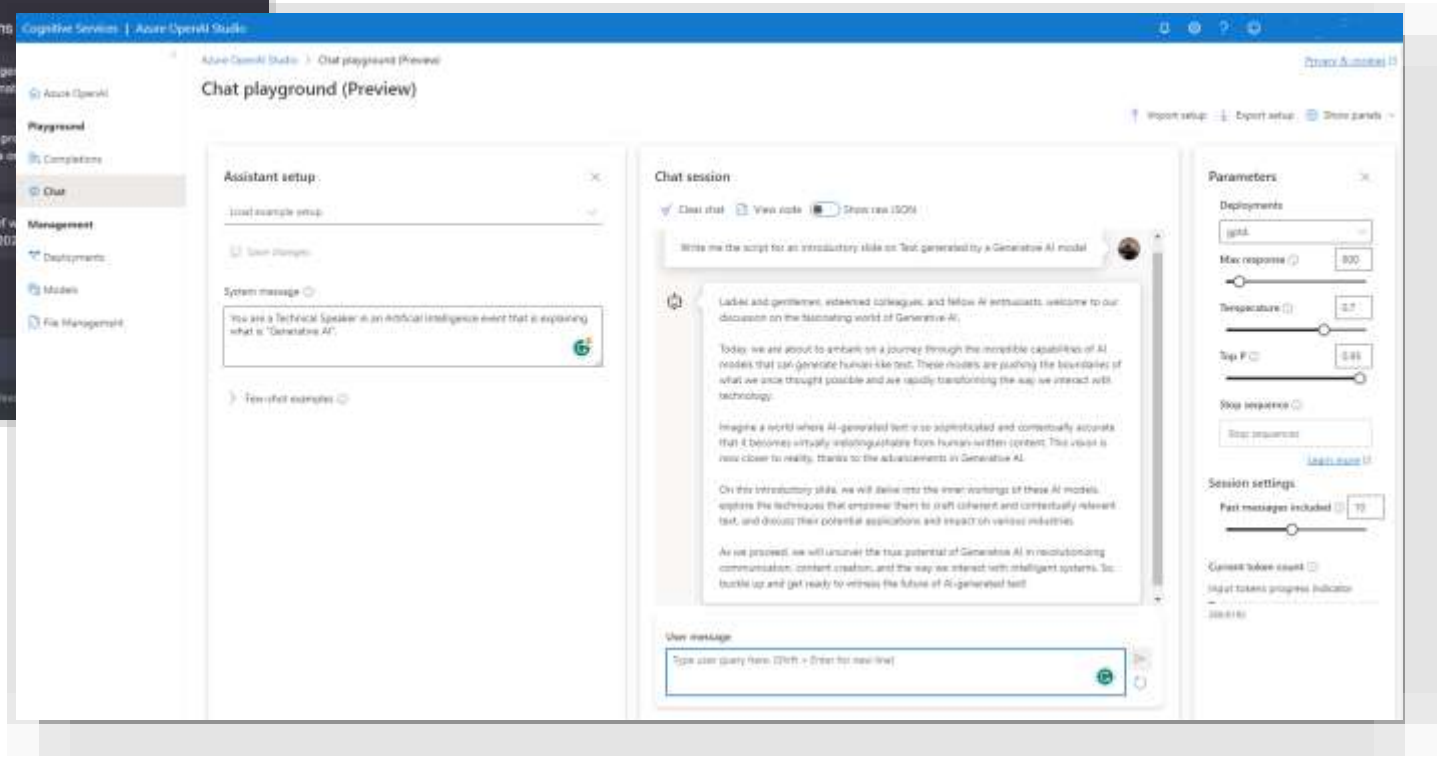
Generative AI: A little bit of history

AI Origins	Neural Networks	Deep Learning boom	Transformers everywhere	Breakthroughs	Generative AI for all & everything
Foundational Research (Logic, Math, Stats, IT)		Audio, Vision, 3D RNNs CNNs GANs RL	Audio Text Images Video 3D		Images Text Video Speech Music 3D Programming …

Generative AI: Text



https://chat.openai.com/



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

OpenAI

Microsoft

ChatGPT

Generative AI: Images

“Gorgeous Mole Antonelliana near the beach”

DALL·E 3

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

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

Generative AI: Images

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

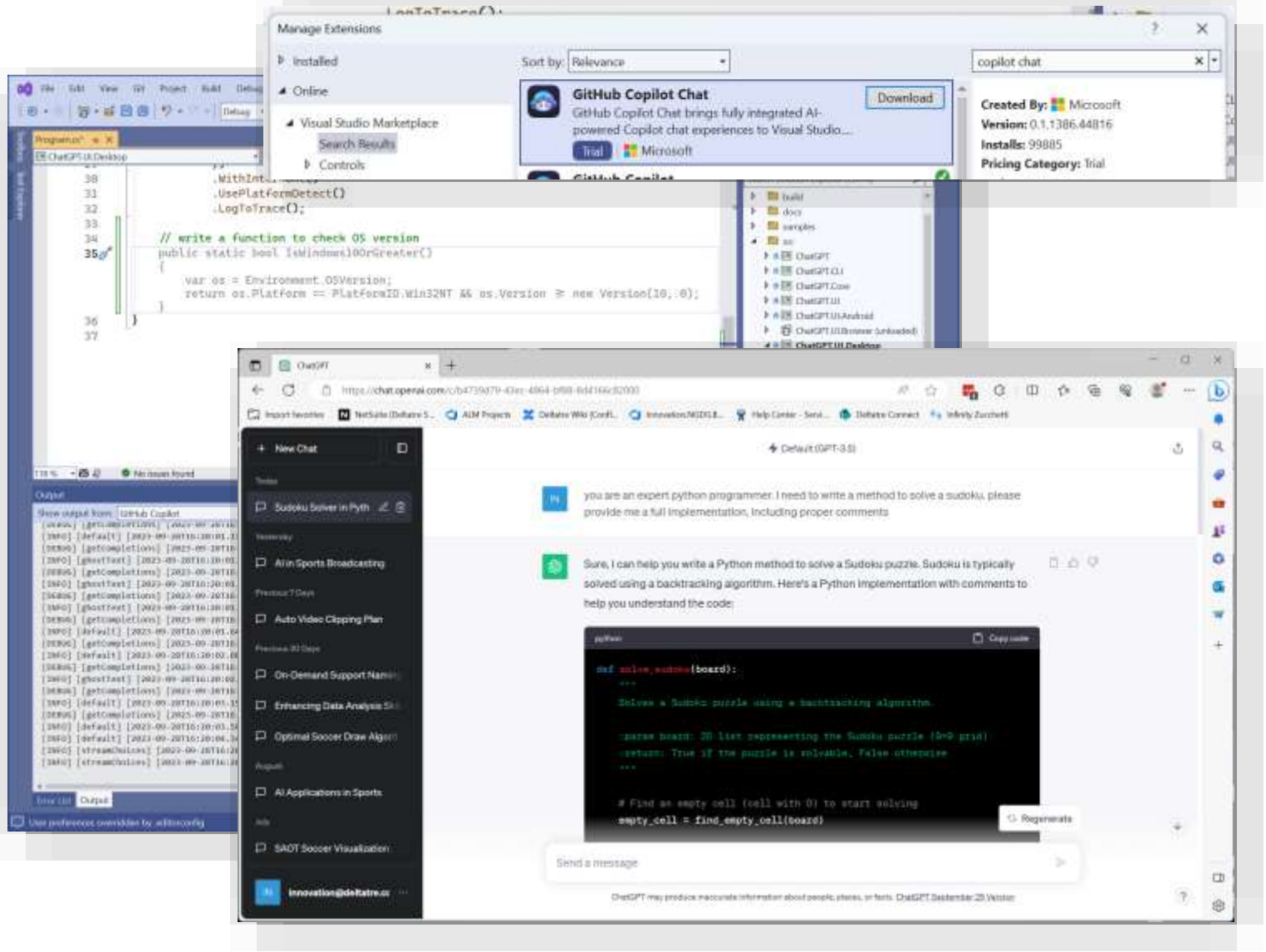
https://midjourney.com/

Generative AI: Developer Assistants

GitHub Copilot

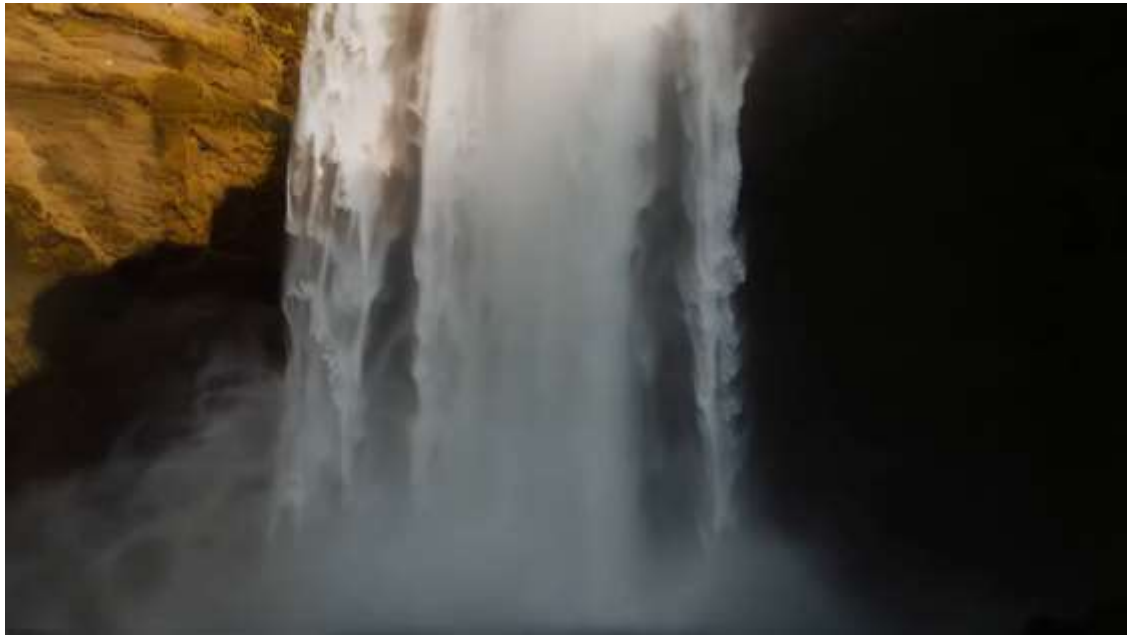
OpenAI ChatGPT

Amazon CodeWhisperer



Generative AI: Video

runway



https://runwayml.com/

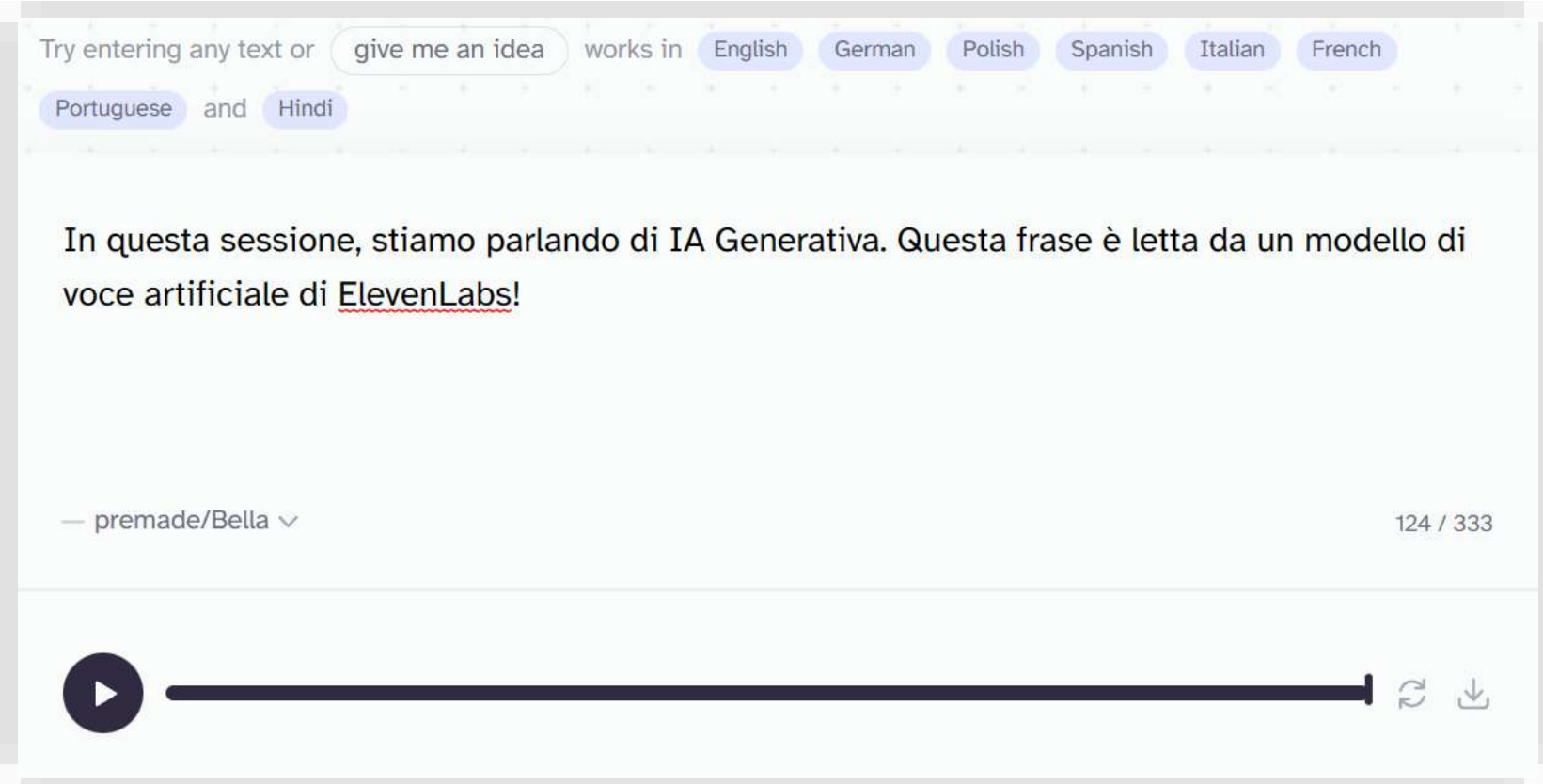
Generative AI: Audio - STT

OpenAI

https://openai.com/research/whisper

Generative AI: Audio - TTS

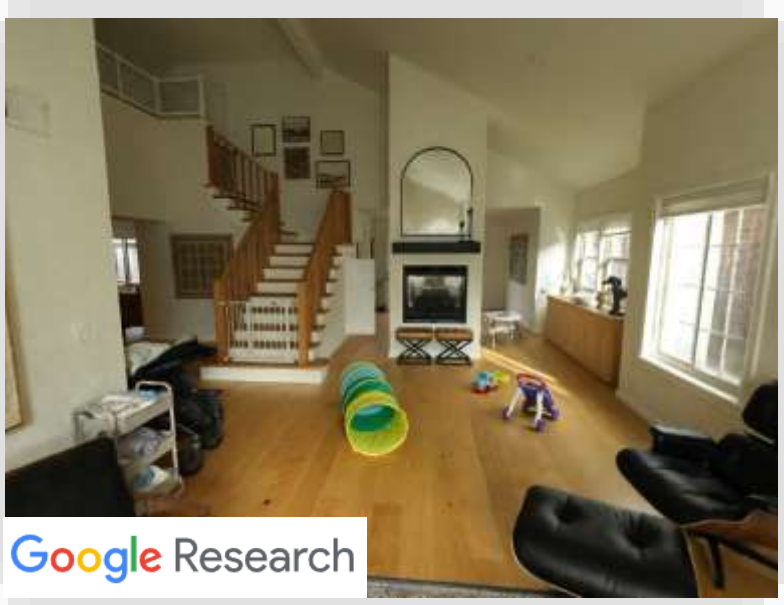
llElevenLabs – Prime Voice AI



https://beta.elevenlabs.io/

Generative AI: 2D to 3D generation

https://developer.nvidia.com/blog/getting-started-with-nvidia-instant-nerfs/



https://jonbarron.info/zipnerf/

https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting

Cinema: Animation

https://www.youtube.com/watch?v=Y1HGgICqZ3c

https://ebsynth.com/

Cinema: Face-Swap and De-Aging

METAPHYSIC

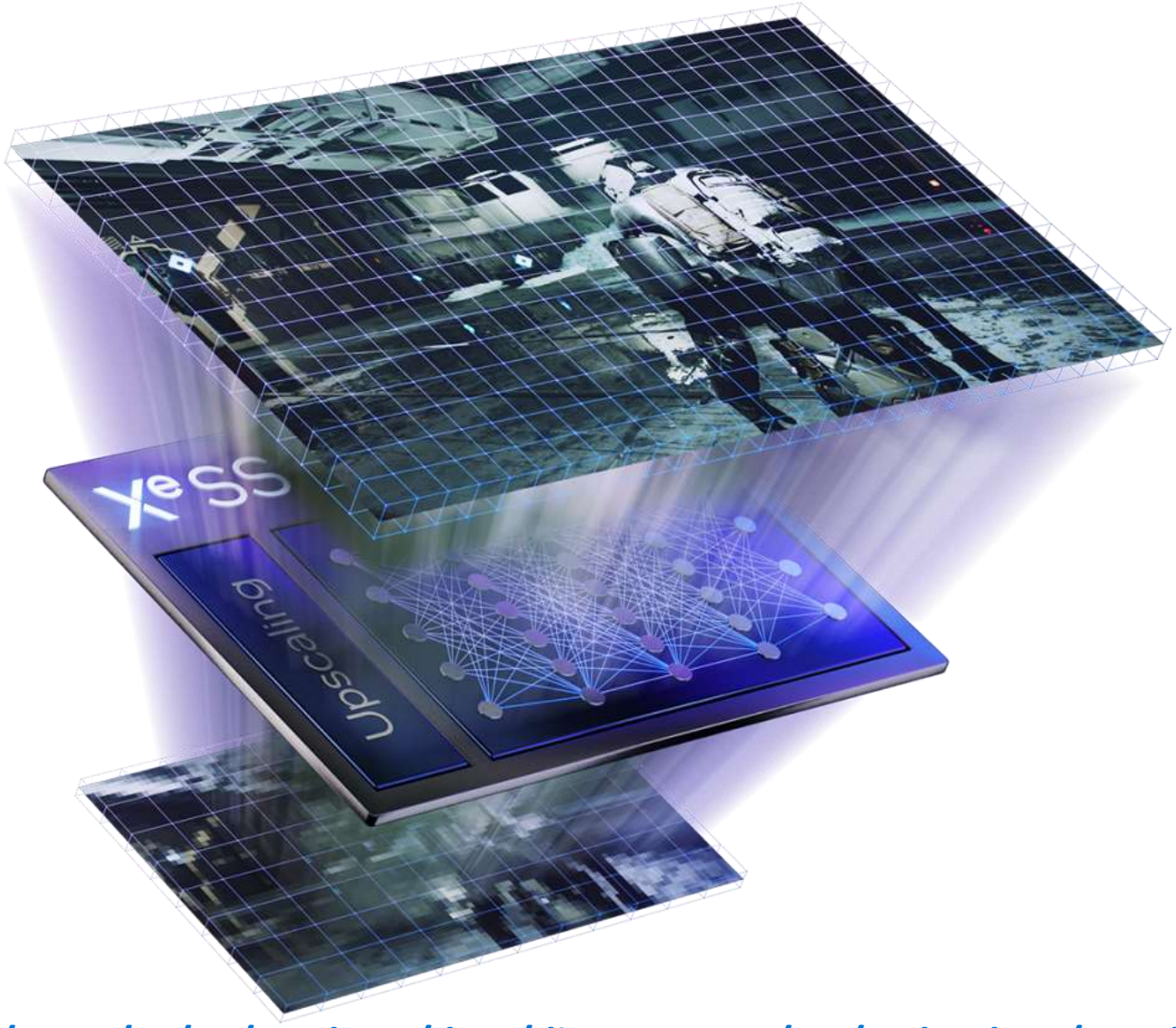
https://metaphysic.ai/

https://www.youtube.com/watch?v=PaI1Vv9MpYY

Games: Super-Resolution

Intel® XeSS

Super Sampling



https://www.intel.com/content/www/us/en/products/docs/discrete-gpus/arc/technology/xess.html

Real Time: High Quality Effects

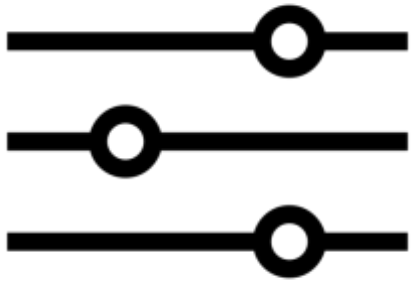
Speaker Focus
Noise removal
Room echo removal
Audio Super-resolution
Acoustic echo cancellation

Virtual Background
Super Resolution
Upscaler
Artifact Reduction
Video Noise Removal

Face Expression Estimation
Eye Contact
Face Tracking
Face Landmark Tracking
Face Mesh
Body Pose Estimation

https://developer.nvidia.com/maxine

Why alternative & O.S. models?



Customization & Flexibility

Embedded/Mobile Devices

Data Policies & Ownership

No Connectivity

Savings & Optimization

Tools and frameworks

OpenVINO™

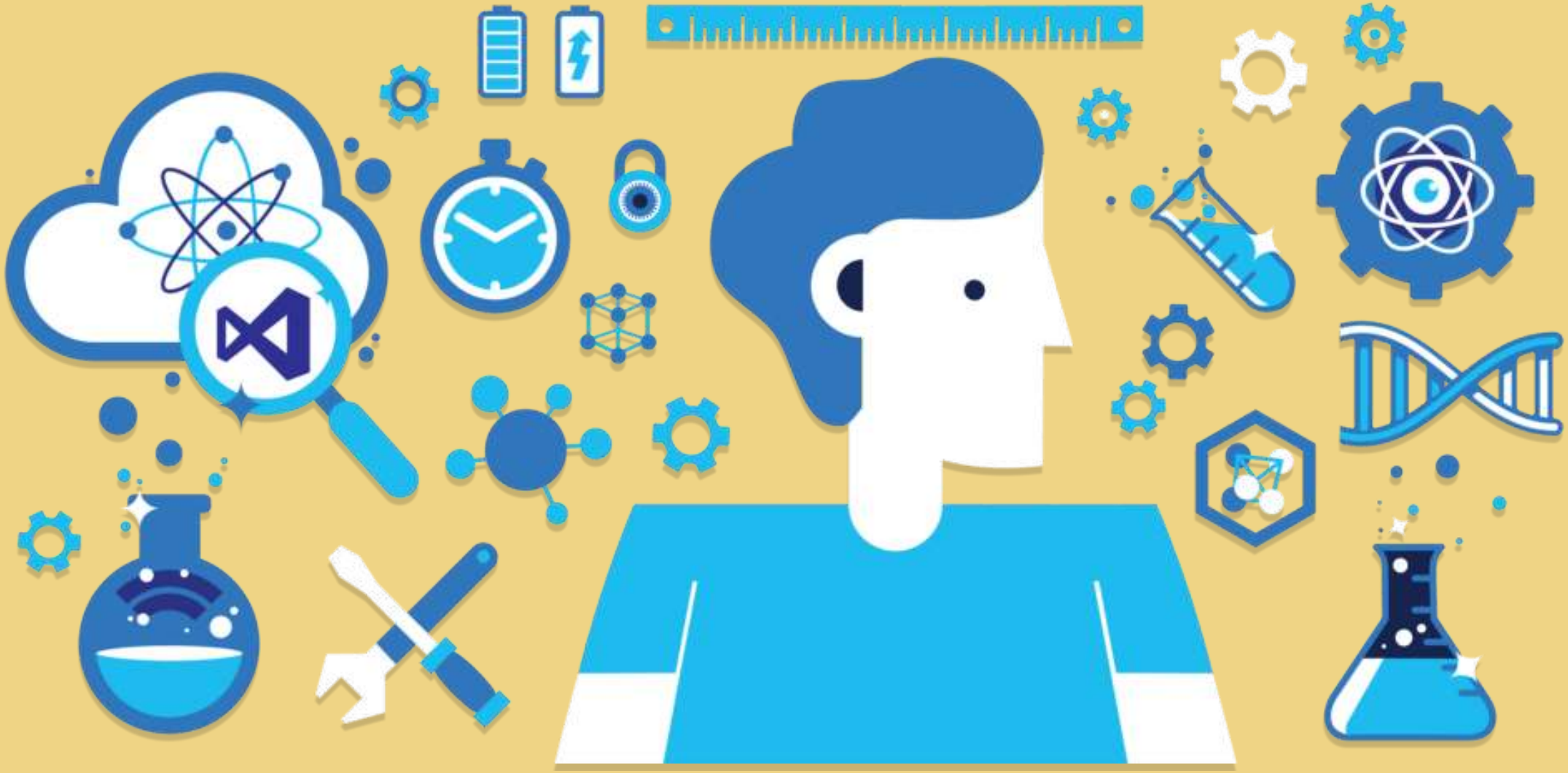
TensorRT

ONNX

TensorFlow

PyTorch

Demo



Azure Machine Learning

Azure AI | Machine Learning Studio

Innovation Lab Experiments
aml-azureday

Deltatre S.p.A. > aml-azureday > Model catalog

All workspaces

Home

Model catalog PREVIEW

Authoring

Notebooks

Automated ML

Designer

Prompt flow PREVIEW

Assets

Data

Jobs

Components

Pipelines

Environments

Models

Endpoints

Manage

Compute

Monitoring PREVIEW

Data Labeling

Linked Services

Model catalog PREVIEW

Search

Announcements

Azure OpenAI language models
Exclusively available on Azure
View models

Discover vision models
Popular OSS vision models hosted by Azure Machine Learning
View models Read blog

Introducing Llama 2
Meta AI
Trained by Meta, hosted by Azure Machine Learning
View models Read blog

Models

tiiuae-falcon-7b (Text generation)	openai-whisper-large (Speech recognition)	databricks-dolly-v2-12b (Text generation)	tiiuae-falcon-40b (Text generation)
Llama-2-70b-chat (Chat completion)	Llama-2-7b (Text generation)	Llama-2-70b (Text generation)	Llama-2-13b (Text generation)
CodeLlama-7b-hf (Text generation)	CodeLlama-7b-Python-hf (Text generation)	CodeLlama-7b-Instruct-hf (Text generation)	CodeLlama-34b-Python-hf (Text generation)
CodeLlama-34b-Instruct-hf (Text generation)	CodeLlama-13b-hf (Text generation)	CodeLlama-13b-Python-hf (Text generation)	CodeLlama-13b-Instruct-hf (Text generation)
Llama-2-7b-chat (Chat completion)	Llama-2-13b-chat (Chat completion)	gpt-4-32k (Chat completions)	gpt-4 (Chat completions)
gpt-35-turbo	tiiuae-falcon-7b-instruct	tiiuae-falcon-7b	tiiuae-falcon-40b

Prev Next

Can't find the model you are looking for? Suggest a model

Can't wait? Try the Import Model notebook Import

Filters

Collections

Curated by AzureML | Azure OpenAI | Meta | Hugging Face

Inference tasks

Text classification | Token classification | Table question answering | Question answering | Zero-shot classification | Translation | Summarization | Conversational | Text generation | Text to text generation | Fill mask | Speech recognition | Chat completions | Embeddings | Image classification | Image segmentation | Object detection

Fine-tuning tasks

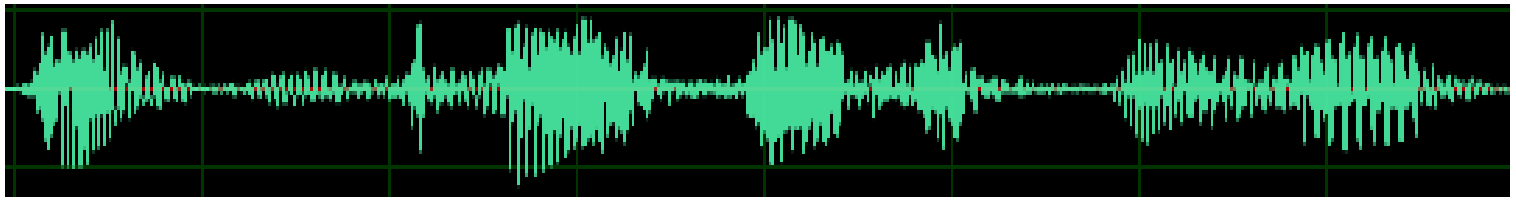
Text classification | Token classification | Question answering | Summarization

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

Audio: Speech-To-Text

Robust Speech Recognition via Large-Scale Weak Supervision

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



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

DEMO
Azure ML Foundation Model

DEMO
Local Audio Transcription (IT/EN) in .NET

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

Text: LLaMA.cpp

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

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

Prompt: "*8k cartoon representation drawn in the style of popular animated series, featuring a llama character with a speech bubble enthusiastically shouting*"

Text: 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

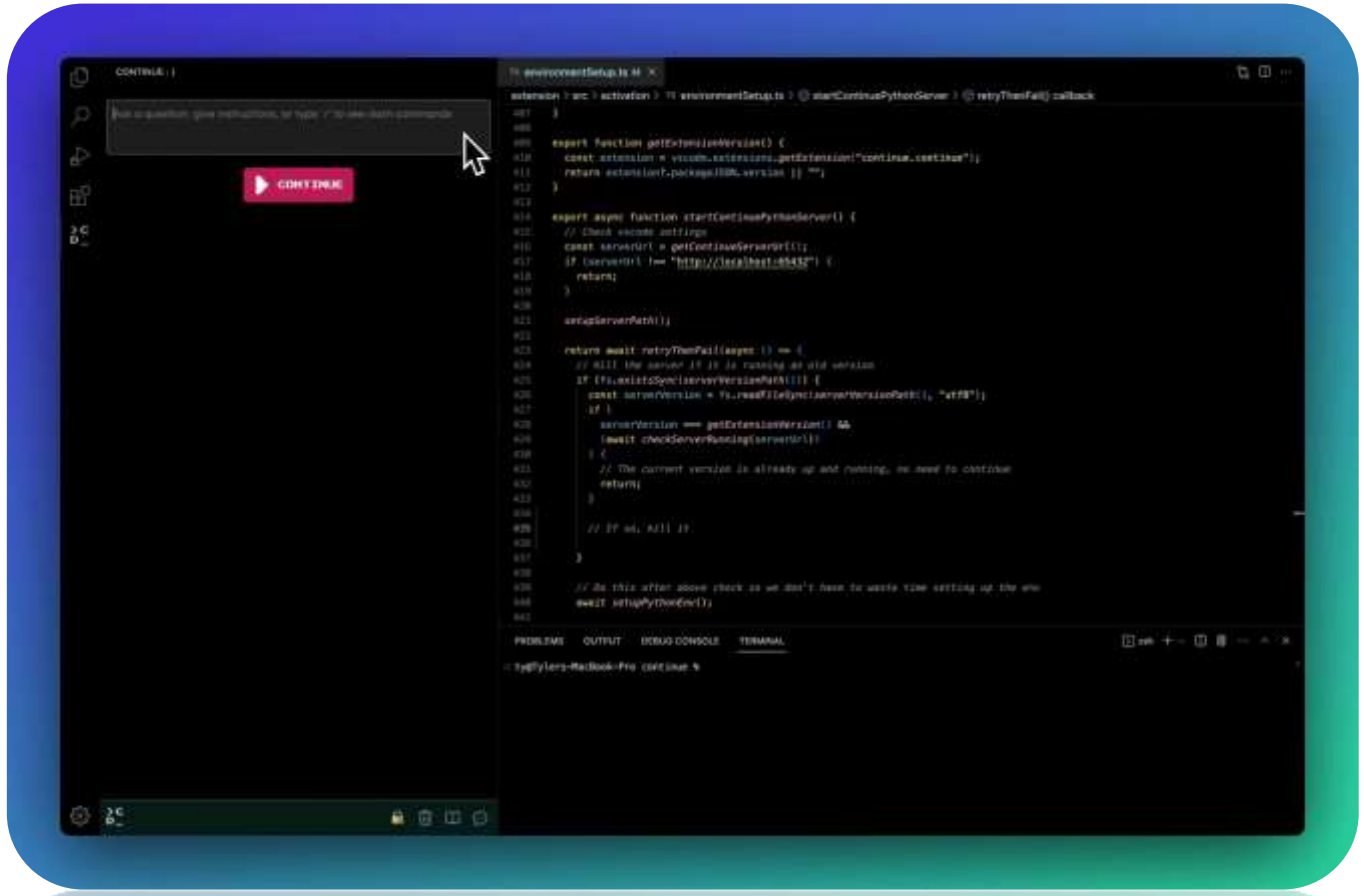
Model	Measure	F16	Q4_0	Q4_1	Q5_0	Q5_1	Q8_0
7B	perplexity	5.9066	6.1565	6.0912	5.9862	5.9481	5.9070
7B	file size	13.0G	3.5G	3.9G	4.3G	4.7G	6.7G
7B	ms/tok @ 4th	127	55	54	76	83	72
7B	ms/tok @ 8th	122	43	45	52	56	67
7B	bits/weight	16.0	4.5	5.0	5.5	6.0	8.5

Supported models:

- LLaMA 🦙
- LLaMA 2 🦙 🦙
- Falcon
- Alpaca
- GPT4All
- Chinese LLaMA / Alpaca and Chinese LLaMA-2 / Alpaca-2
- Vigogne (French)
- Vicuna
- Koala
- OpenBuddy 🐶 (Multilingual)
- Pygmalion 7B / Metharme 7B
- WizardLM
- Baichuan-7B and its derivations (such as baichuan-7b-sft)
- Aquila-7B / AquilaChat-7B
- Starcoder models
- Mistral AI v0.1

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

Text: Developer Assistants



CONTINUE

The open-source autopilot for software development

A **VS Code extension** that brings the power of ChatGPT **(and other LLMs)** to your IDE

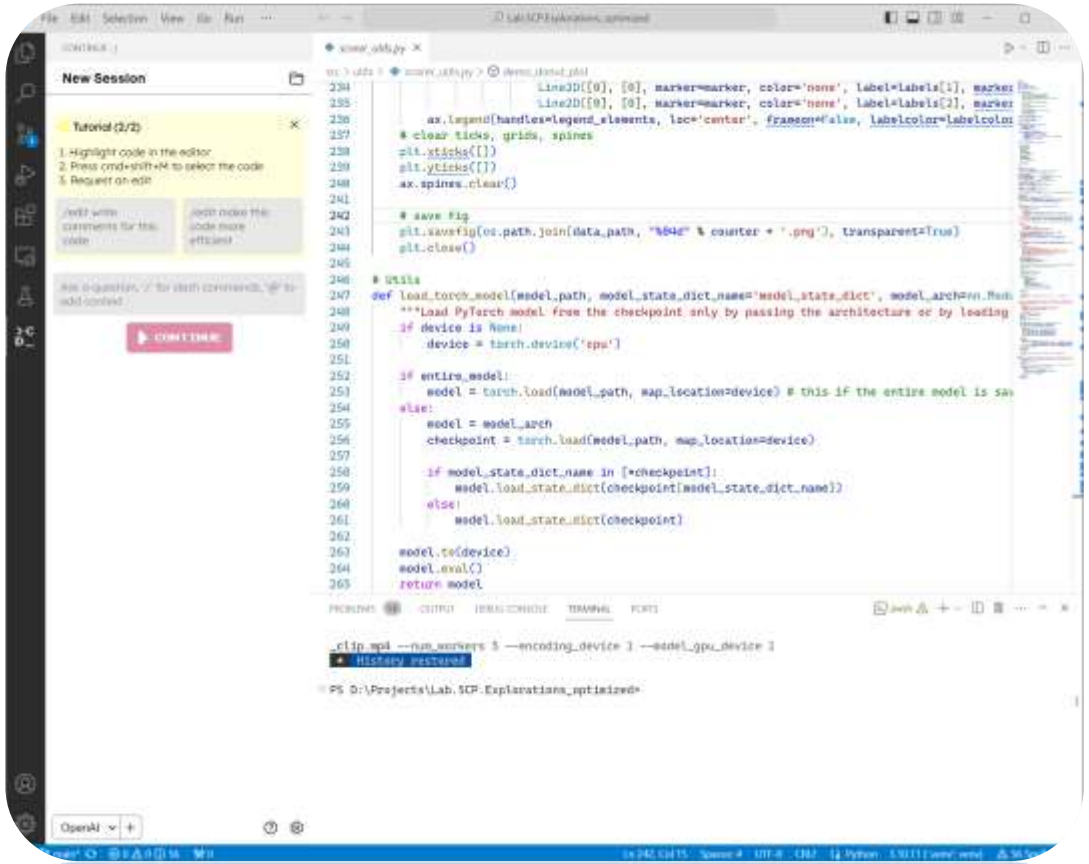
https://continue.dev/

DEMO

LLaMa Sharp

https://github.com/SciSharp/LLamaSharp

Text: Developer Assistants



Answer coding questions
Highlight sections of code and ask Continue for another perspective

Edit in natural language
Highlight a section of code and instruct Continue to refactor it

Generate files from scratch
Open a blank file and let Continue start new Python scripts, React components, C# classes, C++ methods, etc.

Understand errors and exceptions
In case of error or exception, you can send the stack trace into Continue and ask for it to explain the issue to you.

Images: Stable Diffusion

stability.ai

runway

LAION

Hugging Face

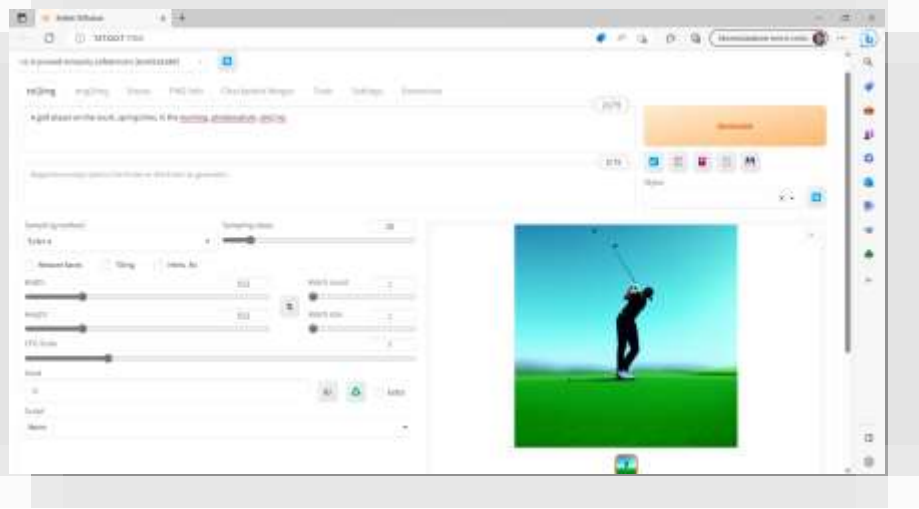
Diffusers

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

Images: Stable Diffusion

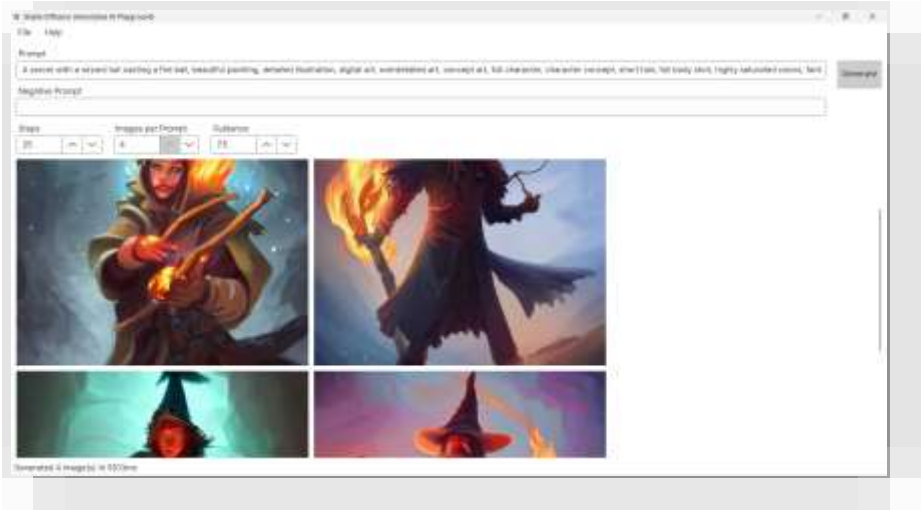
DEMO Web UI Tool

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

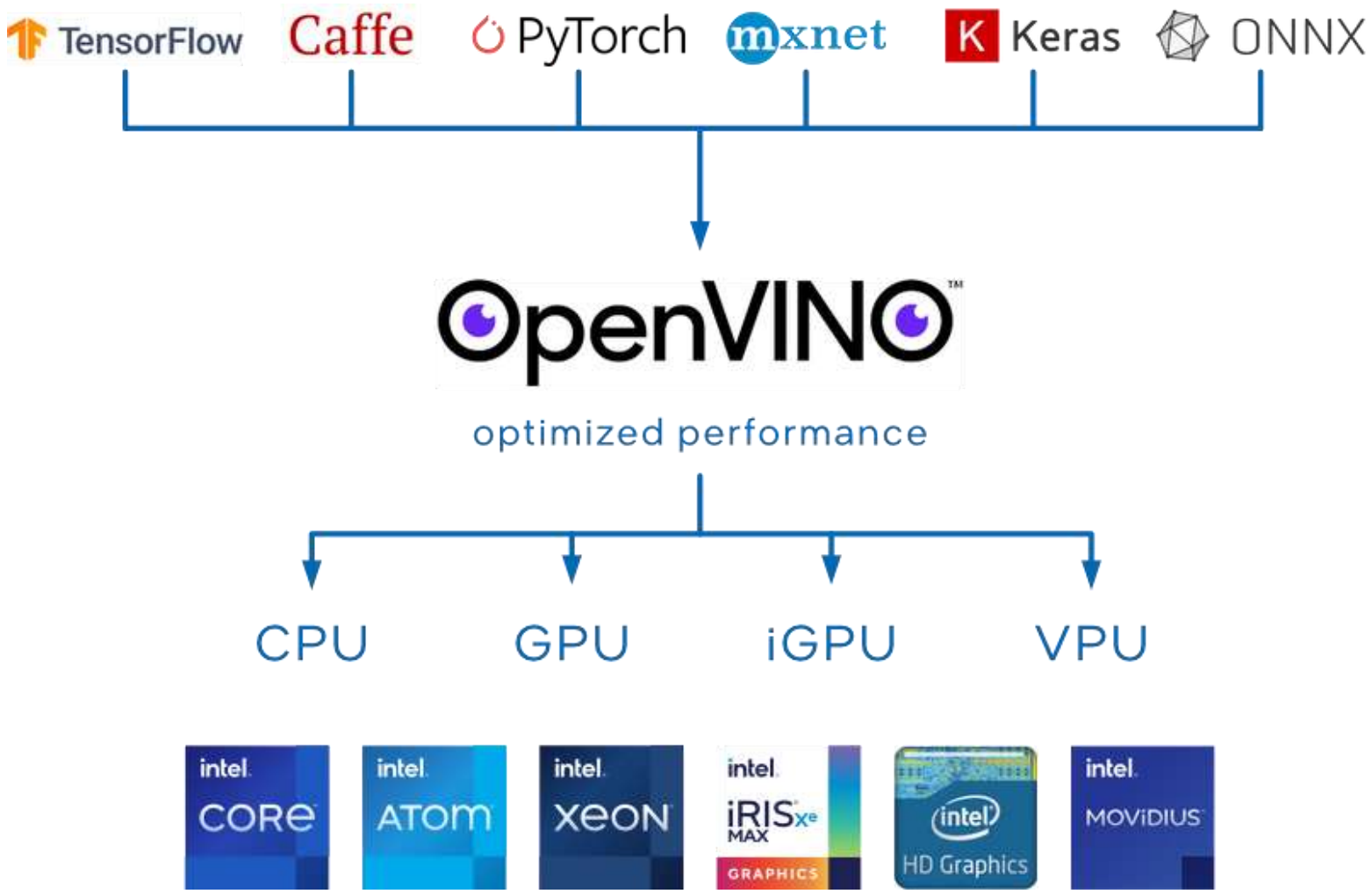


DEMO Generative AI Playground .NET

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



Tools and frameworks: OpenVINO



https://docs.openvino.ai/

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.

https://github.com/openvinotoolkit/openvino_notebooks

Underrated Topics

- **Ethical AI**
- **Model Security**
- **Data Privacy**

Prompt: *”HDR photo of a balance scale with a brain on one side representing AI and machine learning, while the other side showcases a shield and lock indicating security […]”*

Thank You!

ευχαριστώ Salamat Po متشكرم شكراً Grazie

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

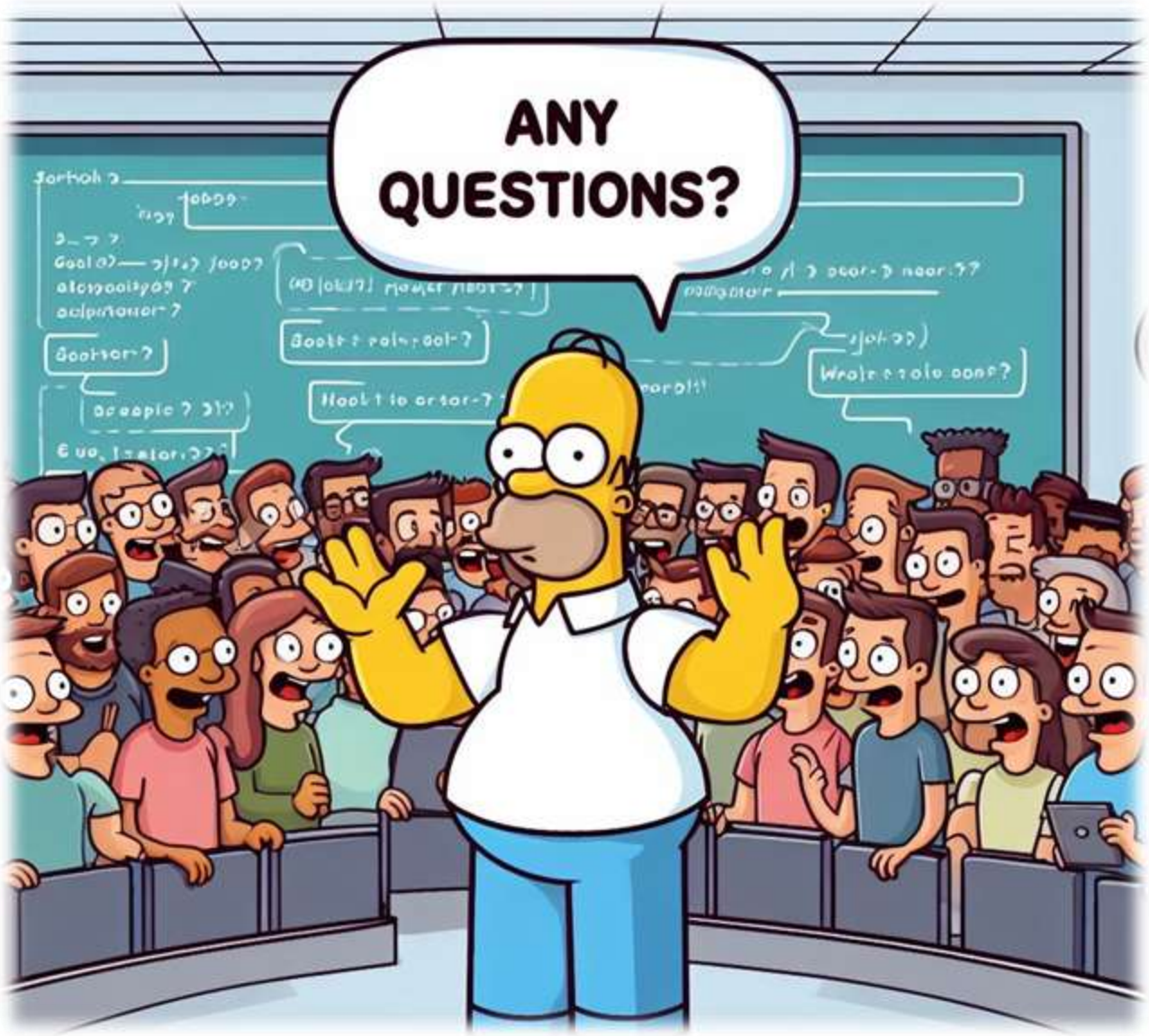
ขอบคุณครับ Obrigado شكريه Terima Kasih Dziękuję

Hvala Köszönöm Tak Dank u wel дякую Tack

Mulţumesc спасибо Danke Cám ơn Gracias

多謝晒 Ďakujem תודה நன்றி Děkuji 감사합니다

QnA



“Simpson character in front of enthusiastic Software Developers. In a speech bubble saying ‘Any questions?’ ”

References (1/2)

- https://github.com/DotNetCodeIT/AzureDay2023Torino
- 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://research.nvidia.com/labs/dir/magic3d/
- https://developer.nvidia.com/blog/getting-started-with-nvidia-instant-nerfs/
- https://jonbarron.info/zipnerf/
- https://github.com/steven2358/awesome-generative-ai
- https://github.com/imaurer/awesome-decentralized-llm
- https://onnx.ai/
- https://docs.openvino.ai/
- https://www.nvidia.com
- https://github.com/openvinotoolkit/openvino_notebooks
- https://www.youtube.com/watch?v=PaI1Vv9MpYY

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/tatsu-lab/stanford_alpaca
- https://github.com/nomic-ai/gpt4all
- 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://github.com/gianni-rg/gen-ai-net-playground
- https://whisper.ggerganov.com/talk/
- https://github.com/voicepaw/so-vits-svc-fork
- https://github.com/facebookresearch/AnimatedDrawings
- https://developer.nvidia.com/maxine
- https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting
- https://ebsynth.com/
- https://metaphysic.ai/
- https://about.fb.com/news/2023/07/llama-2/
- https://github.com/facebookresearch/llama
- https://arxiv.org/pdf/2307.09288.pdf
- https://continue.dev/

About Us

Clemente GIORIO

R&D Senior Software Engineer @ deltatre

@tinux80

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

MVP Microsoft Most Valuable Professional

dotNET{podcast}

intel Software INNOVATOR

[PACKT] PUBLISHING Author

FAB LAB NAPOLI

NVIDIA

NVIDIA Certified Associate - AI in the Data Center

About Us

Ing. Gianni ROSA GALLINA

R&D Technical Lead @ deltatre

@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 Author

SCAN ME

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

AZURE DAY

Platinum Sponsor

Microsoft

msc TECHNOLOGY

Technical Sponsor

NET CODE

Azure Meetup TORINO

TORINO.NET

stickermule

Packt

JET BRAINS