



# Tabya Conf 2025

Presente e futuro dell'Intelligenza Artificiale



# Rendiamo intelligenti le nostre applicazioni

Alessio Iafrate

Freelance developer

Microsoft MVP



Nicola Paro

Cloud Solution Architect @  
beantech



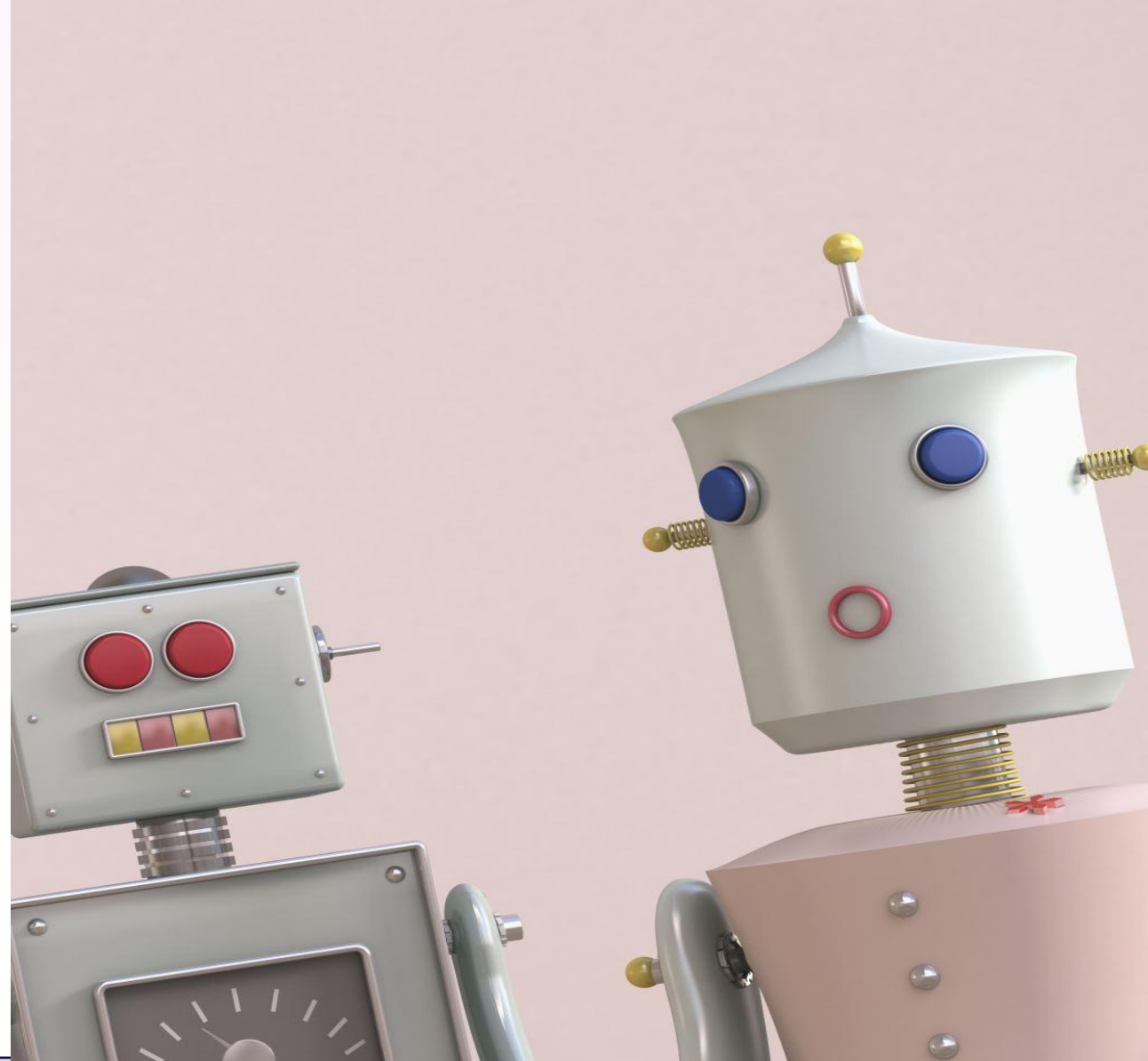
I nostri sponsor

sharpdesign

CodeDesign



In questa  
sessione nessun  
modello di  
ChatGpt o Azure  
OpenAI è stato  
maltrattato



# Voglio rendere intelligenti le mie applicazioni!

Ma...

Non mi serve tutto

Ho sempre bisogno una soluzione multimodale?

Potrei non avere internet

Esistono anche soluzioni offline?

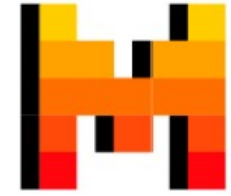
Ho bisogno di fare tante inferenze

Esistono soluzioni più economiche o gratuite?

Non ho un team dedicato

Esistono strumenti che mi garantiscono un risultato certo nelle mie casistiche?

# LLM Locale?



# Ollama



# deepseek

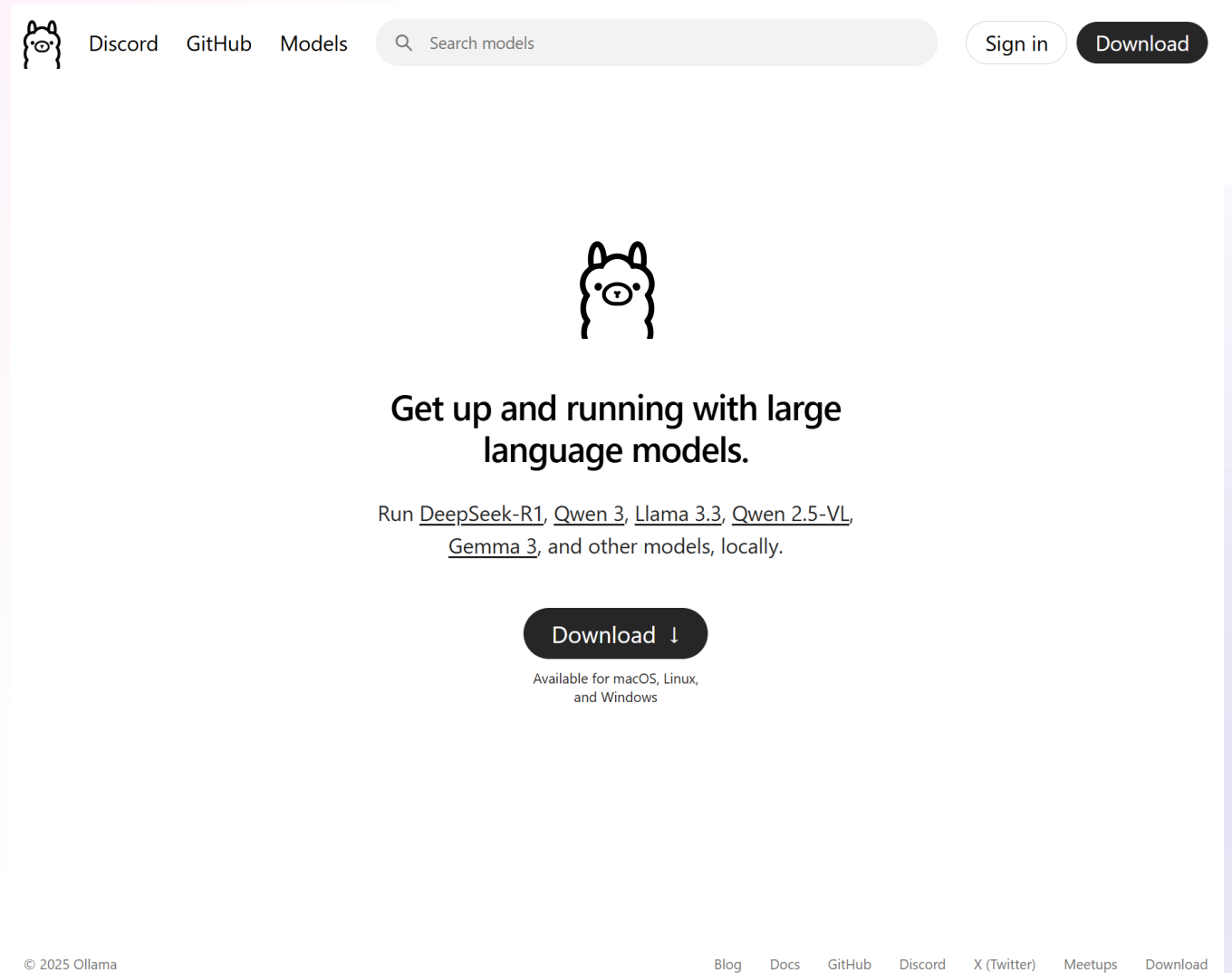


# Ollama

- Sviluppato da un team indipendente con background in AI e developer tools.
- Obiettivo di rendere semplice e accessibile l'esecuzione di **LLM in locale**, per una maggiore **privacy**, **controllo** e **autonomia**.

## Timeline

- 2023 (inizio): Prime versioni interne
- Estate 2023: Lancio pubblico su GitHub con supporto a modelli GGUF e binari ottimizzati.
- 2024: Integrazione API REST, supporto a GPU e ottimizzazioni multi-piattaforma. Aggiunta di modelli più recenti come Mistral e Gemma.



# Ollama

Ollama utilizza  
container di  
modelli LLM (.bin  
o .gguf)

Supporta GPU /  
CPU a seconda del  
sistema

Inferenza locale:  
nessun invio dati a  
server esterni

Compatibile con  
diversi modelli (es:  
LLaMA, Mistral,  
Gemma...)

Supporto multi-  
modello

Integrazione  
semplice: CLI, API  
REST

Supporto a modelli  
di visione e  
modelli tool



# Ollama - Comandi

Usage:

```
ollama [flags]
```


```
ollama [command]
```

Available Commands:

serve	Start ollama
create	Create a model from a Modelfile
show	Show information for a model
run	Run a model
stop	Stop a running model
pull	Pull a model from a registry
push	Push a model to a registry
list	List models
ps	List running models
cp	Copy a model
rm	Remove a model
help	Help about any command



# Ollama – Modelli



Embedding

Vision

Tools

Popular

**gemma3**

The current, most capable model that runs on a single GPU.

vision

1b

4b

12b

27b

4.9M Pulls

21 Tags

Updated 1 month ago

**qwen3**

Qwen3 is the latest generation of large language models in Qwen series, offering a comprehensive suite of dense and mixture-of-experts (MoE) models.

tools

0.6b

1.7b

4b

8b

14b

30b

32b

235b

1.5M Pulls

35 Tags

Updated 2 weeks ago

**deepseek-r1**

DeepSeek's first-generation of reasoning models with comparable performance to OpenAI-o1, including six dense models distilled from DeepSeek-R1 based on Llama and Qwen.

1.5b

7b

8b

14b

32b

70b

671b

**gemma3**

`ollama run gemma3`

4.9M Downloads

Updated 1 month ago

The current, most capable model that runs on a single GPU.

vision

1b

4b

12b

27b

**Models**

View all →

Name	Size	Context	Input
gemma3:latest	3.3GB	-	Text, Image
gemma3:1b	815MB	-	-
gemma3:4b <span>latest</span>	3.3GB	-	-
gemma3:12b	8.1GB	-	-
gemma3:27b	17GB	-	-

**gemma3:1b**

`ollama run gemma3:1b`

4.9M Downloads

Updated 1 month ago

The current, most capable model that runs on a single GPU.

vision

1b

4b

12b

27b

Updated 1 month ago

8648f39daa8f · 815MB

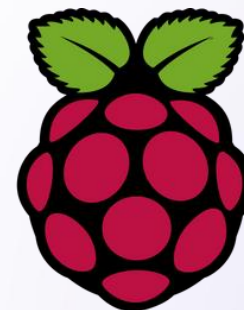
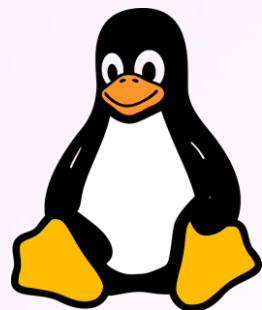
model	arch <b>gemma3</b> · parameters <b>1000M</b> · quantization <b>Q4_K_M</b>	815MB
license	Gemma Terms of Use Last modified: February 21, 2024 By u...	8.4kB
template	{{- range \$i, \$_ := .Messages }} {{- \$last := eq (len (s...	358B
params	{ "stop": [ "<end_of_turn>" ], "temperature": 1, "top_k"...	77B



# Demo

Ollama

# Ollama – Dove gira?



# Azure AI Service





# Azure AI $\neq$ Azure AI Service

# Azure AI



Pre-Built AI



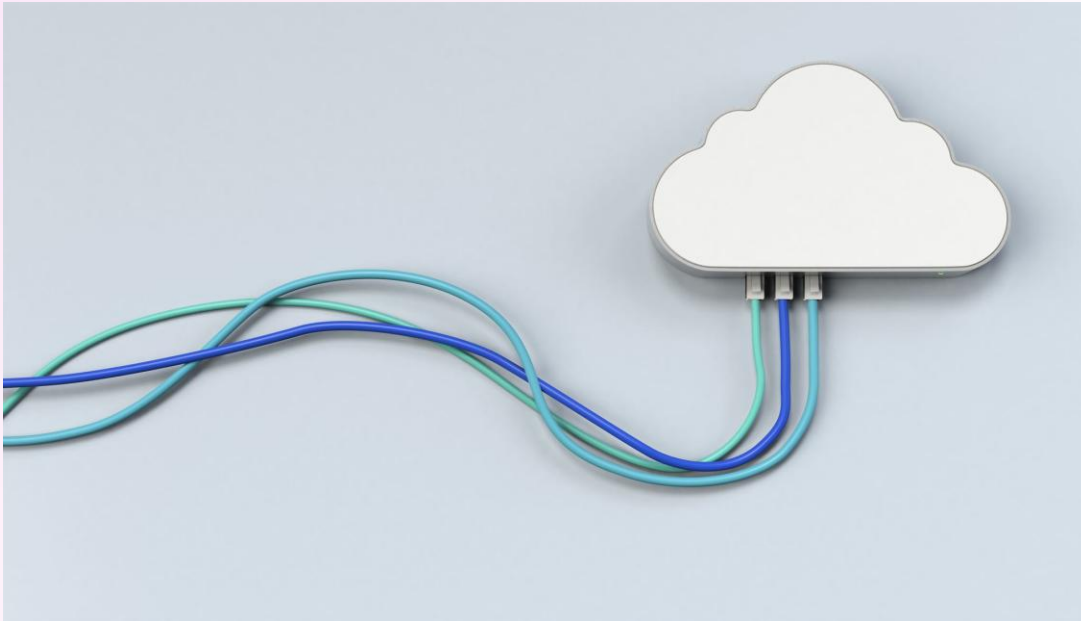
Conversational AI



Custom AI



# Vantaggi



- Azure AI Services offre modelli pre-addestrati per risparmiare tempo.
- L'infrastruttura scalabile consente una gestione efficiente dei carichi di lavoro.
- L'integrazione con altri servizi Azure facilita lo sviluppo rapido.
- Supporta una vasta gamma di casi d'uso, dalle immagini al linguaggio.
- Aggiornamenti continui garantiscono l'accesso alle ultime tecnologie AI.

# Vantaggi Economici



- Riduzione dei costi iniziali grazie a modelli pre-addestrati.
- Minori spese di infrastruttura rispetto a un modello personalizzato.
- Accesso a tecnologie avanzate senza investimento ingente.
- Scalabilità dei costi in base all'utilizzo effettivo.
- Aggiornamenti e supporto continuo inclusi nel servizio.

# Adattabili alle nostre esigenze



Possiamo collegare tra di loro vari servizi per arrivare al risultato che vogliamo ottenere.

Es. le descrizioni ritornate dai servizi sono in inglese, se aggiungo un servizio di traduzione o di text-to-speech posso modificarli prima di presentarli all'utente

# Responsible use of AI with Azure AI services

<https://learn.microsoft.com/en-us/azure/ai-services/responsible-use-of-ai-overview>

# Uso responsabile dell'AI

Microsoft promuove un utilizzo etico e sicuro dell'intelligenza artificiale, basato su principi fondamentali come:

- Equità: evitare pregiudizi nei modelli di IA.
- Affidabilità e sicurezza: garantire che i sistemi siano robusti e sicuri.
- Privacy e sicurezza dei dati: proteggere le informazioni sensibili.
- Inclusività: rendere l'IA accessibile a tutti.
- Trasparenza: fornire spiegazioni chiare sul funzionamento dei modelli.
- Responsabilità: garantire che l'IA sia utilizzata in modo etico.



### Transcribe speech to text

Transcribe call center or meeting conversations. Go global with audio-captioning in more than 100 languages.

[> Learn more](#)



### Transcribe audio with OpenAI Whisper

Transform your call centers using the latest OpenAI Whisper model in Azure AI Speech or Azure OpenAI Service.

[> Read the blog](#)



### Verify and recognize speakers

Confirm a person's identity or recognize who's speaking in a meeting by adding speaker verification and identification to your app.

[> Learn more](#)



### Convert text to speech

Build bots that speak naturally. Differentiate your brand with customized, realistic voices and speaking styles.

[> Learn more](#)



### Build custom voices

Build natural-sounding voices with custom neural voice.

[> Learn more](#)



### Enable multilingual communication

Translate audio or video data from and into an ever-growing list of supported languages. Customize translations to your industry.

[> Learn more](#)



### Speech analytics

Analyze audio or video call recordings to gain deep insights. Summarize key topics and extract or redact personal identification information.

[> Learn more](#)



### Build your avatars

Bring your brand to life using pre-built or custom avatars with natural-sounding voices.

[> Learn more](#)



### Embedd speech

Use embedded speech to power on-device speech to text and text to speech scenarios where cloud connectivity is intermittent or unavailable.

[> Learn more](#)

# Demo AVATAR

<https://speech.microsoft.com/portal/talkingavatar>



# WebRTC

WebRTC (Web Real-Time Communication) is a free and open-source project providing web browsers and mobile applications with real-time communication (RTC) via application programming interfaces (APIs). It allows audio and video communication and streaming to work inside web pages by allowing direct peer-to-peer communication, eliminating the need to install plugins or download native apps.[3]

Supported by Apple, Google, Microsoft, Mozilla, and Opera, WebRTC specifications have been published by the World Wide Web Consortium (W3C) and the Internet Engineering Task Force (IETF).[4][5]

According to the [webrtc.org](http://webrtc.org) website, the purpose of the project is to "enable rich, high-quality RTC applications to be developed for the browser, mobile platforms, and IoT devices, and allow them all to communicate via a common set of protocols".[6]



# Flow



## analizzatore di testo

fornisce l'output sotto forma di sequenza di fonemi



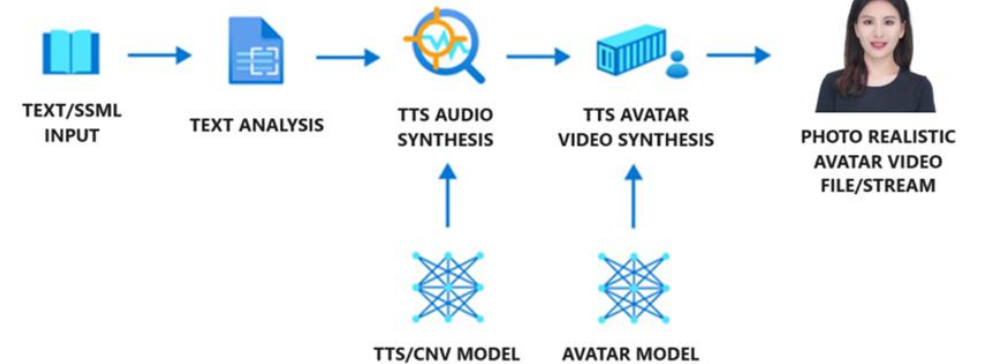
## il sintetizzatore audio TTS

predice le caratteristiche acustiche del testo di input e sintetizza la voce



## sintetizzatore video avatar TTS

prevede l'immagine della sincronizzazione labiale con le caratteristiche acustiche, in modo da generare il video sintetico



Region:

West Europe

Currency:

Euro Zone – Euro (€) EUR

1 USD = 0.8788 EUR

## Text to Speech<sup>8</sup>

### Standard Voice

Neural (real-time and batch): **€13.183** per 1M characters  
Neural HD (real-time and batch)<sup>4</sup>: **€26.365** per 1M characters

### Custom Voice

#### Professional Voice:

Synthesis (real-time and batch): **€21.092** per 1M characters  
Voice model training: **€45.699** per compute hour, up to **€4,387.029** per training  
Endpoint hosting: **€3.55** per model per hour

#### Personal Voice<sup>6</sup>:

Synthesis (real-time and batch): **€21.092** per 1M characters  
Voice creation: Free  
Voice profile storage: **€527.288** per 1,000 voice profiles per month

### Enhanced Add-on feature: Avatar

#### Standard:

Interactive avatar (real-time): **€0.440** per minute  
Avatar video (batch): **€0.879** per minute

#### Custom:

Avatar model training: **€13.183** per compute hour  
Interactive avatar (real-time): **€0.528** per minute  
Avatar video (batch): **€1.758** per minute  
Endpoint hosting: **€0.528** per model per hour

## Speaker Recognition (per transaction billing)

### Speaker Verification<sup>7</sup>

**€4.395** per 1,000 transactions

### Speaker Identification<sup>7</sup>

**€8.789** per 1,000 transactions

Chat with S



### Azure AI Content Understanding

Accelerate multimodal AI solution development.

[Learn more](#)



### Azure AI Content Safety

Monitor text and images to detect offensive or inappropriate content.

[Learn more](#)



### Azure AI Vision

Read text, analyze images, and detect faces with optical character recognition (OCR) and machine learning.

[Learn more](#)



### Azure OpenAI Service

Build your own agent and generative AI applications with cutting-edge language and vision models.

[Learn more](#)



### Azure AI Translator

Translate documents and text in real time across more than 100 languages.

[Learn more](#)



### Azure AI Language

Build conversational interfaces, summarize documents, and analyze text using prebuilt AI-powered features.

[Learn more](#)



### Azure AI Search

Retrieve the most relevant data using keyword, vector, and hybrid search.

[Learn more](#)



### Azure AI Speech

Use industry-leading AI services such as speech-to-text, text-to-speech, speech translation, and speaker recognition.

[Learn more](#)



### Azure AI Document Intelligence

Apply advanced machine learning to extract text, key-value pairs, tables, and structures from documents.

[Learn more](#)

# Demo Document Intelligence Studio

<https://documentintelligence.ai.azure.com/studio>



### Azure AI Content Understanding

Accelerate multimodal AI solution development.

[Learn more](#)



### Azure AI Content Safety

Monitor text and images to detect offensive or inappropriate content.

[Learn more](#)



### Azure AI Vision

Read text, analyze images, and detect faces with optical character recognition (OCR) and machine learning.

[Learn more](#)



### Azure OpenAI Service

Build your own agent and generative AI applications with cutting-edge language and vision models.

[Learn more](#)



### Azure AI Translator

Translate documents and text in real time across more than 100 languages.

[Learn more](#)



### Azure AI Language

Build conversational interfaces, summarize documents, and analyze text using prebuilt AI-powered features.

[Learn more](#)



### Azure AI Search

Retrieve the most relevant data using keyword, vector, and hybrid search.

[Learn more](#)



### Azure AI Speech

Use industry-leading AI services such as speech-to-text, text-to-speech, speech translation, and speaker recognition.

[Learn more](#)



### Azure AI Document Intelligence

Apply advanced machine learning to extract text, key-value pairs, tables, and structures from documents.

[Learn more](#)

# DEMO Vision Portal

<https://portal.vision.cognitive.azure.com/gallery/featured>

# Resources

Demo ai services

<https://github.com/a-iafrate/MauiAzureAIServices>

Demo Ollama

<https://github.com/nicolaparo/OllamaDemo>



# Thanks!



Alessio Iafrate

email: [alessioiafrate@hotmail.com](mailto:alessioiafrate@hotmail.com)

twitter: [@aiafrate](https://twitter.com/aiafrate)

<https://github.com/a-iafrate/>

<https://www.linkedin.com/in/alessio-iafrate/>



Nicola Paro

twitter: [@nicola\\_paro](https://twitter.com/nicola_paro)

<https://github.com/nicolaparo/>

<https://www.linkedin.com/in/nicolaparo>

<https://linktr.ee/nicolaparo>





# Domande?

... e nel frattempo dacci un feedback sulla sessione

