LE BOTTEGHE DELL'INNOVAZIONE SOSTENIBILE - V EDIZIONE

SMARTLAB:

AI generativa su hardware locale

AI Generativa su Hardware locale

Walter Riviera

WONDER

Team

Nicola

Founder

Francesco

CEO-Founder

Walter

Co-Founder

la Repubblica

500 italiani e italiane che contano nell'Intelligenza Artificiale

a cura di redazione Italian Tech

Un giro d'Italia nelle eccellenze in ambito IA: l'università e la ricerca, le startup e le grandi aziende, l'arte e la cultura, le associazioni e la politica. Il primo who's who, in continuo aggiornamento con le vostre segnalazioni

19 MARZO 2024 — 81 MINUTI DI LETTURA

AGGIORNATO ALLE 19:28

la Repubblica

Intel

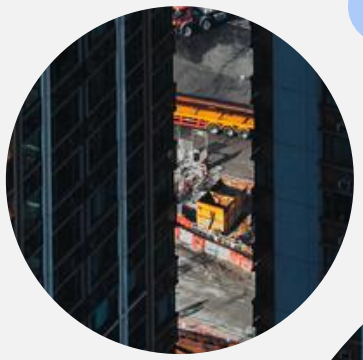
Walter Riviera

Responsabile Tecnico nel Centro di eccellenza IA di Intel in Europa, Medio Oriente e Africa. Ha collaborato con il CERN per l'uso dell'IA nei supercomputer e ha contribuito ai sistemi di addestramento di IA che rispettano la privacy, promuovendo la tecnica Federated Learning.

RTL 102.5 — C — la Repubblica R.it — DEEJAY — RADIO 105 NETWORK

Agenda

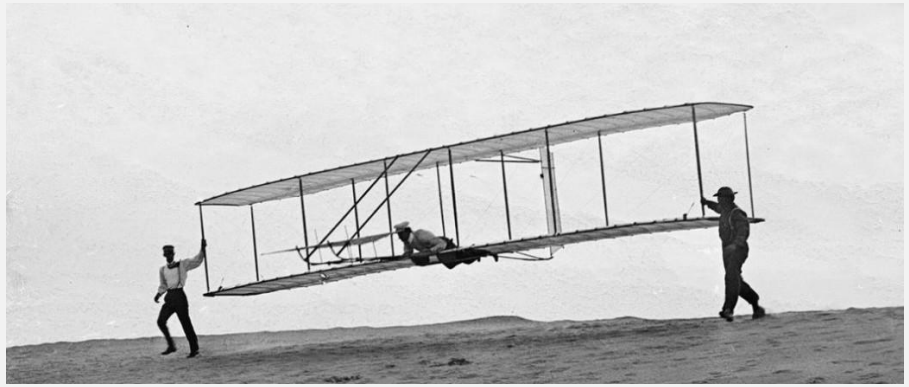
- Panorama Odierno
- Concetti fondamentali AI Generativa
- Sistemi RAG
- Demo-time



Aspettative

Capiremo I meccanismi e daremo definizioni affinchè la conoscenza si regga su basi solide.

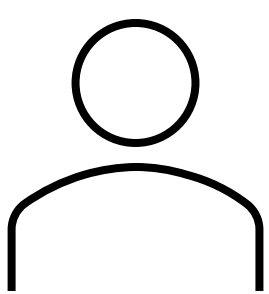
Ma la crescita e la conoscenza di un ambito sono processi non eventi.

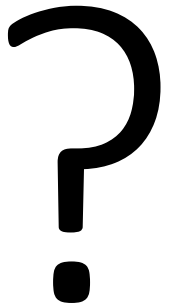


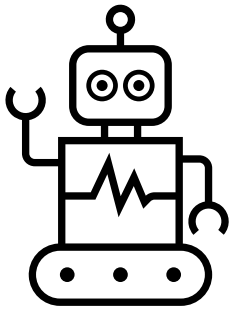
Panorama Odierno

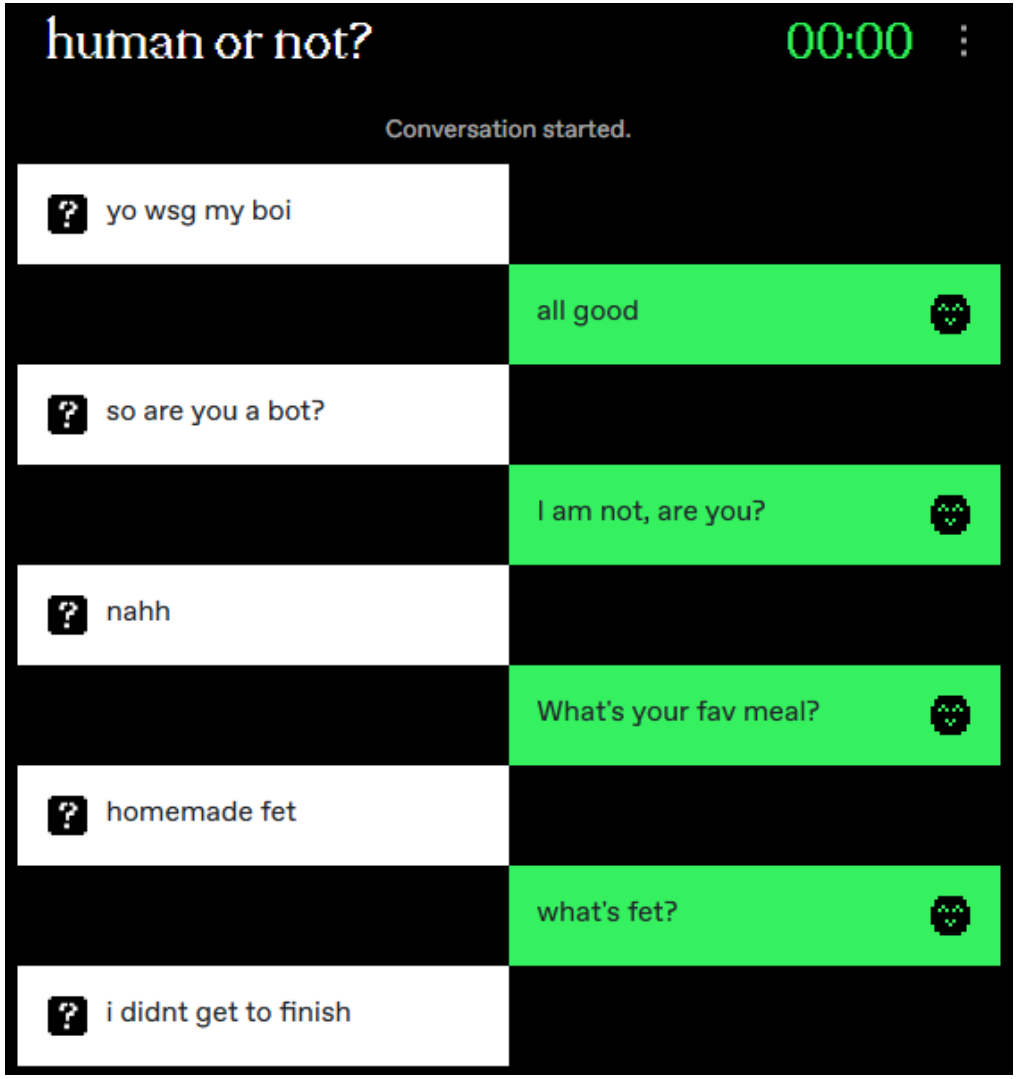
WONDER

La Macchina di Turing









https://app.humanornot.ai/

Il Test di Turing

Si tratta di un esperimento mentale proposto da Turing nel 1950 per determinare se una macchina può essere considerata "intelligente”

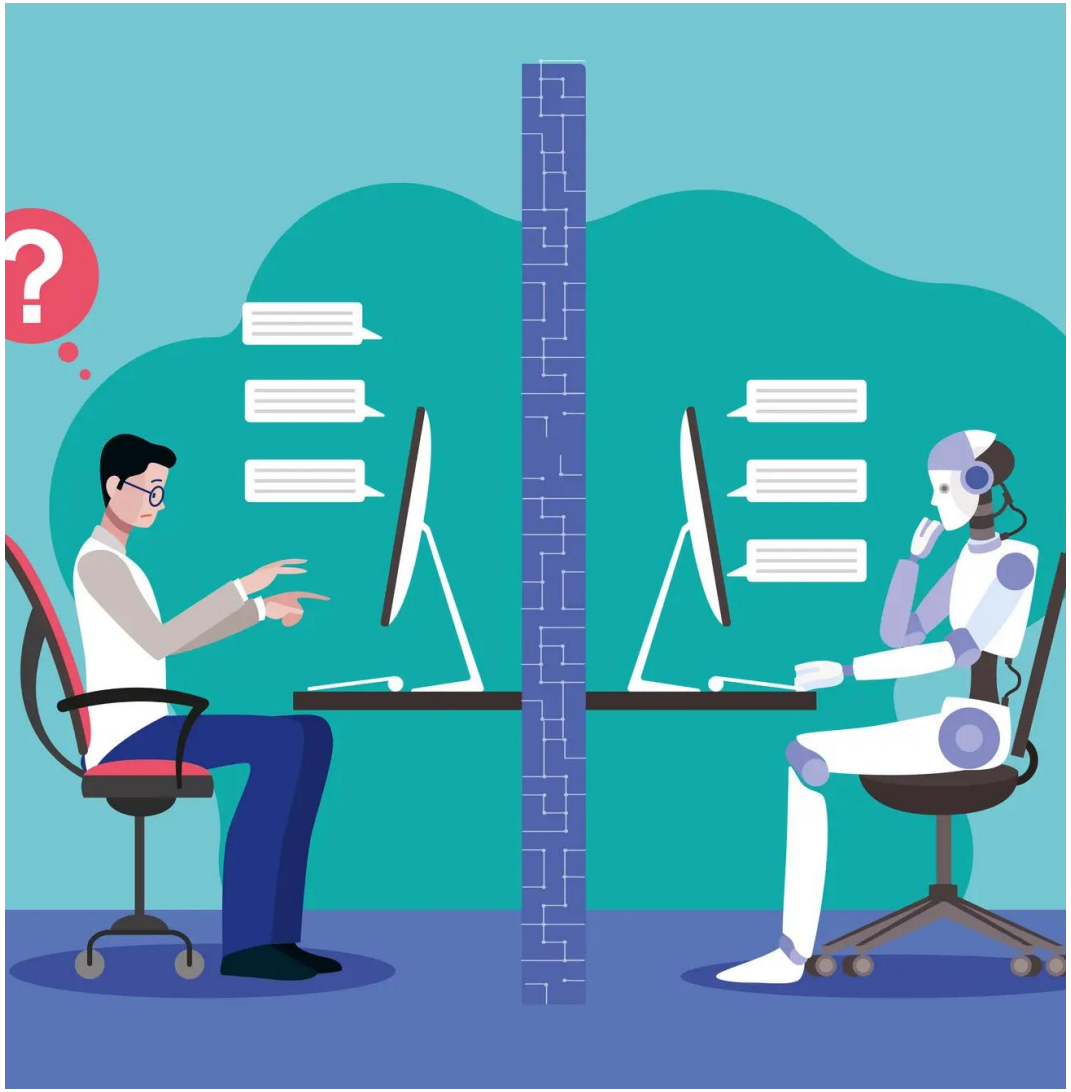
🗨️ Consiste in una conversazione tra un umano e due interlocutori nascosti:

- Un altro umano
- Un’intelligenza artificiale

Se la macchina riesce a **fingere di essere umana** e ingannare l’interrogatore, supera il test.

🔍 *Limiti del test:*

- Valuta solo il comportamento esterno, non la "comprensione"
- Non misura la consapevolezza o l’intelligenza generale
- Alcune IA moderne superano il test, ma sono ancora "deboli”

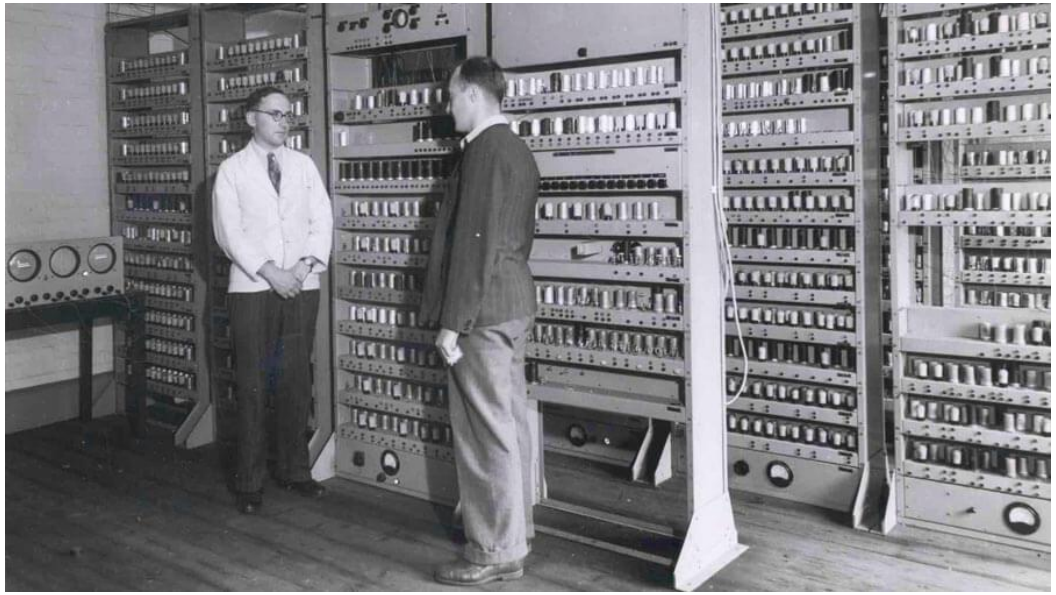


https://app.humanornot.ai/

L’evoluzione dell’AI

Dalle Origini ai Modelli Moderni

- L’IA è nata come disciplina scientifica negli **anni ‘50**, con i primi studi sulla computazione e il ragionamento automatico.
- Ha attraversato fasi di **grande sviluppo** seguite da periodi di **stagnazione e scetticismo**, noti come *"Inverni dell’IA"*.
- Oggi, con i modelli di **Deep Learning** e **LLM (Large Language Models)**, siamo in una nuova fase di forte espansione.



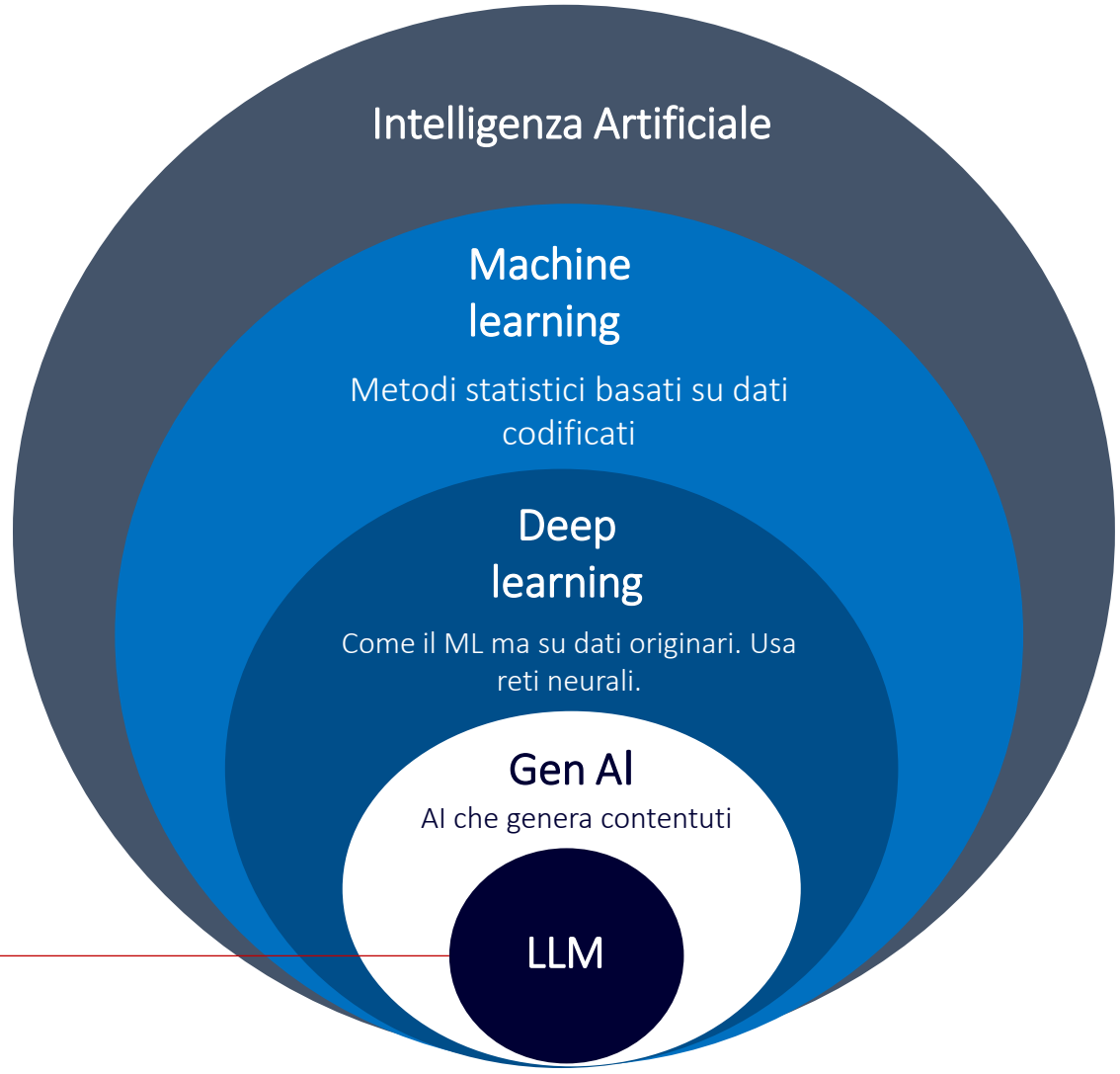
Definizioni

Tokens

Agenti AI

AGI

Large Language Models



Concetti Fondamentali

Tipi d’apprendimento

SEMI-SUPERVISIONATO

Parte dei dati è etichettata, ma non tutto il dataset.

SUPERVISIONATO

Dataset etichettato. Per ogni istanza esiste l’associazione ad un etichetta. Il sistema predice l’etichetta

NON SUPERVISIONATO

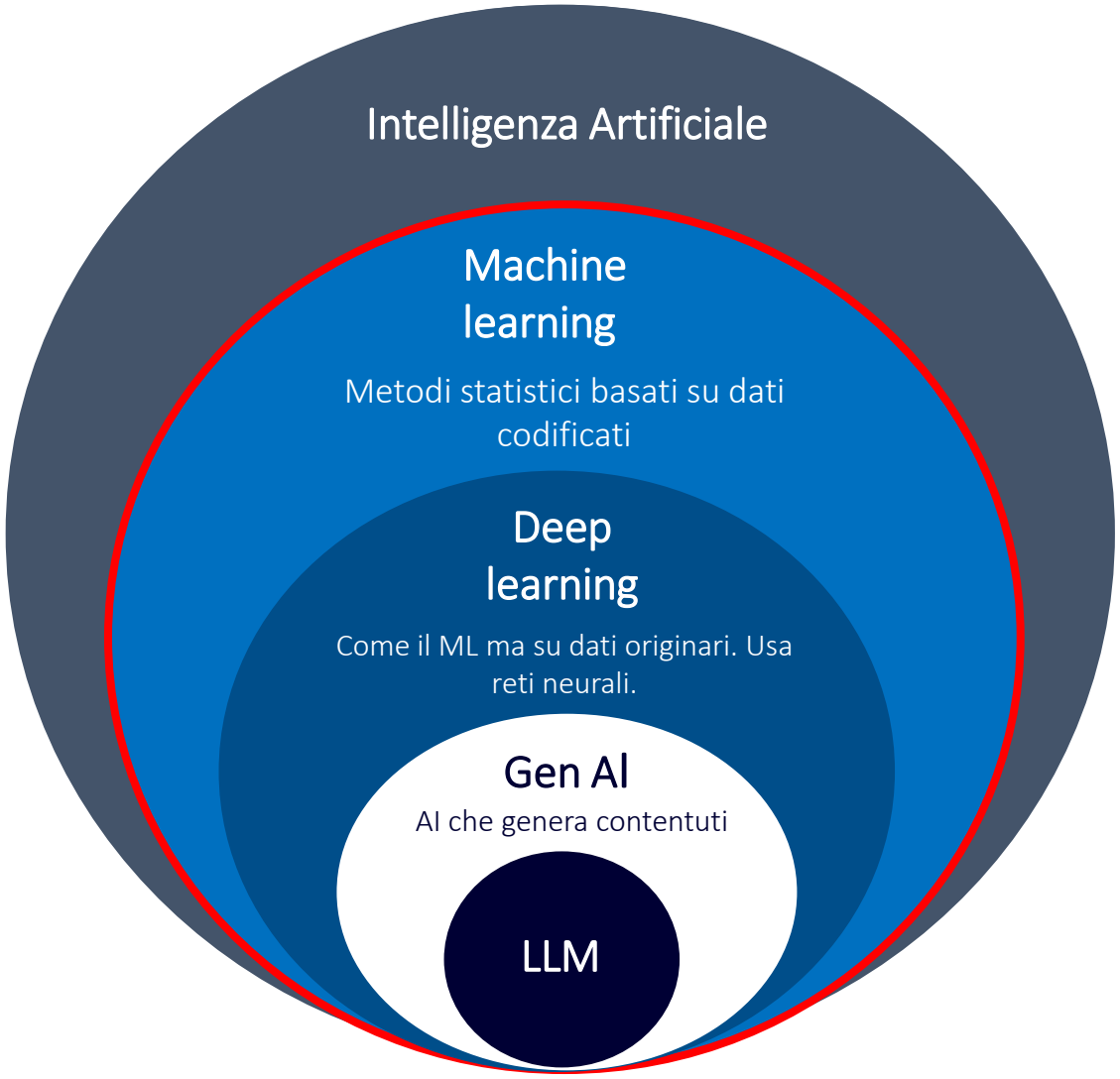
I dati non sono etichettati. E’ il caso del clustering e della generazione di contenuti

PER RINFORZO

Il sistema impara per tentativi partendo da: uno stato, mosse consentite, un reward, vincoli

Machine Learning

?



Machine Learning

$N \times N$

	A	B	C	D	E	F
1	Country	Salesperson	Order Date	OrderID	Units	Order Amoun
2	USA	Fuller	1/01/2011	10392	13	1,440.00
3	UK	Gloucester	2/01/2011	10397	17	716.72
4	UK	Bromley	2/01/2011	10771	18	344.00
5	USA	Finchley	3/01/2011	10393	16	2,556.95
6	USA	Finchley	3/01/2011	10394	10	442.00
7	UK	Gillingham	3/01/2011	10395	9	2,122.92
8	USA	Finchley	6/01/2011	10396	7	1,903.80
9	USA	Callahan	8/01/2011	10399	17	1,765.60
10	USA	Fuller	8/01/2011	10404	7	1,591.25
11	USA	Fuller	9/01/2011	10398	11	2,505.60
12	USA	Coghill	9/01/2011	10403	18	855.01
13	USA	Finchley	10/01/2011	10401	7	3,868.60
14	USA	Callahan	10/01/2011	10402	11	2,713.50
15	UK	Rayleigh	13/01/2011	10406	15	1,830.78
16	USA	Callahan	14/01/2011	10408	10	1,622.40
17	USA	Farnham	14/01/2011	10409	19	319.20
18	USA	Farnham	15/01/2011	10410	16	802.00

FEATURE ENGINEERING

$(f_1, f_2, \ldots, f_K)$

"descrittore"

CLASSIFICATORE

SVM

Random Forest

Naïve Bayes

Decision Trees

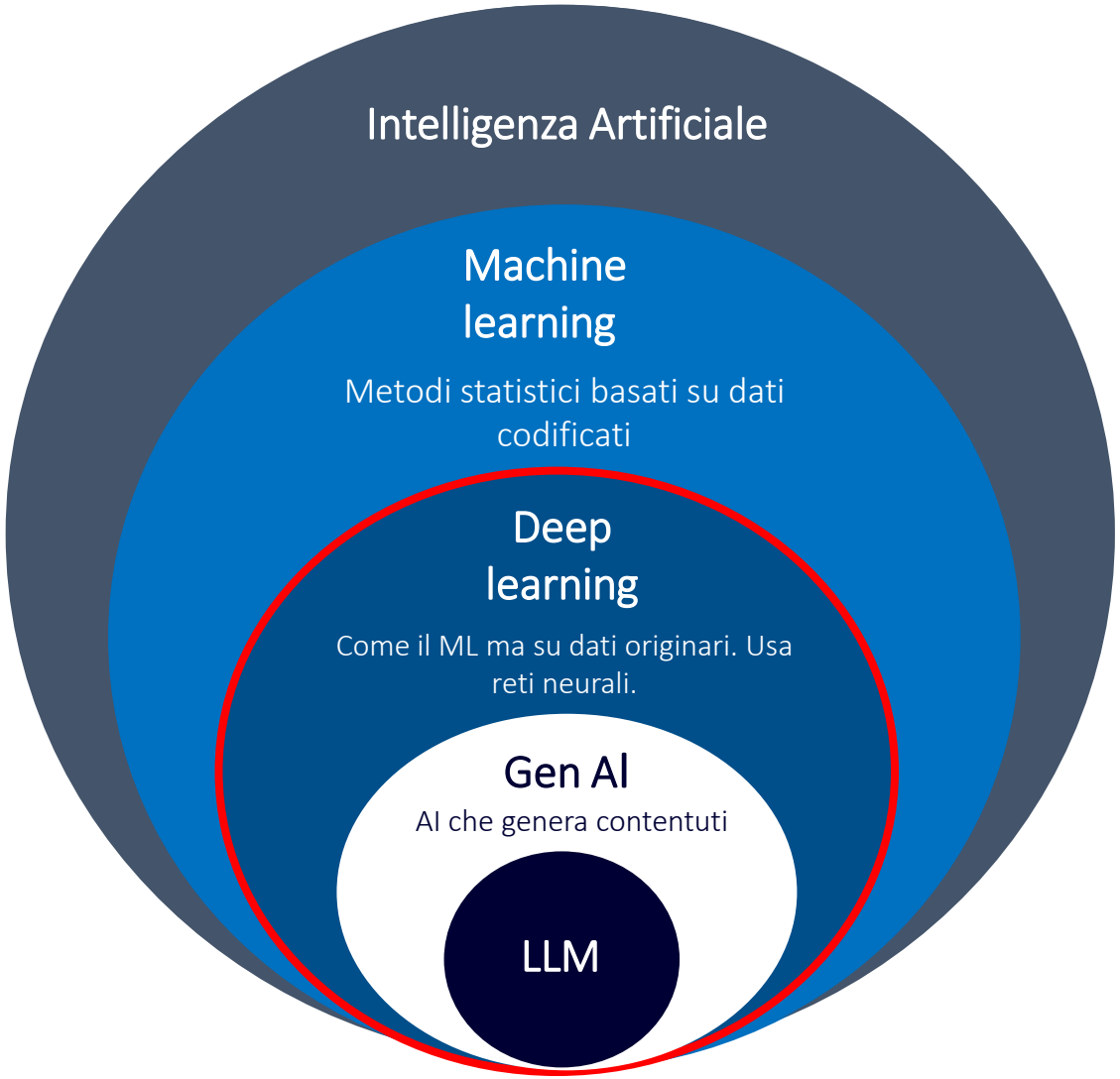
Logistic Regression

Ensemble Methods

Predizione

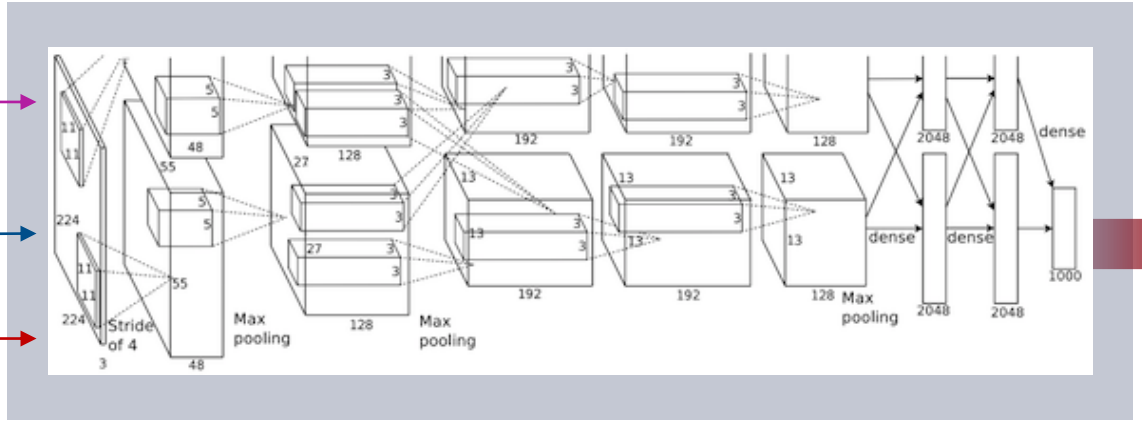
Machine Learning

?



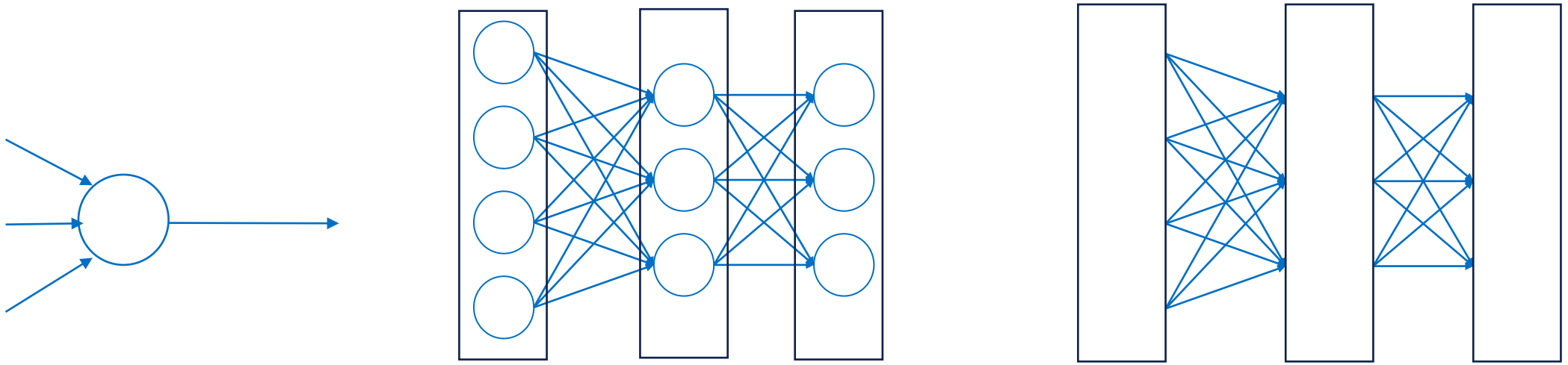
Deep Learning

	A	B	C	D	E	F
1	Country	Salesperson	Order Date	OrderID	Units	Order Amount
2	USA	Fuller	1/01/2011	10392	13	1,440.00
3	UK	Gloucester	2/01/2011	10397	17	716.72
4	UK	Bromley	2/01/2011	10771	18	344.00
5	USA	Finchley	3/01/2011	10393	16	2,556.95
6	USA	Finchley	3/01/2011	10394	10	442.00
7	UK	Gillingham	3/01/2011	10395	9	2,122.92
8	USA	Finchley	6/01/2011	10396	7	1,903.80
9	USA	Callahan	8/01/2011	10399	17	1,765.60
10	USA	Fuller	8/01/2011	10404	7	1,591.25
11	USA	Fuller	9/01/2011	10398	11	2,505.60
12	USA	Coghill	9/01/2011	10403	18	855.01
13	USA	Finchley	10/01/2011	10401	7	3,868.60
14	USA	Callahan	10/01/2011	10402	11	2,713.50
15	UK	Rayleigh	13/01/2011	10406	15	1,830.78
16	USA	Callahan	14/01/2011	10408	10	1,622.40
17	USA	Farnham	14/01/2011	10409	19	319.20
18	USA	Farnham	15/01/2011	10410	16	802.00



Predizione

Neuroni Digitali



Strati di Neuroni

AI Generativa

WONDER

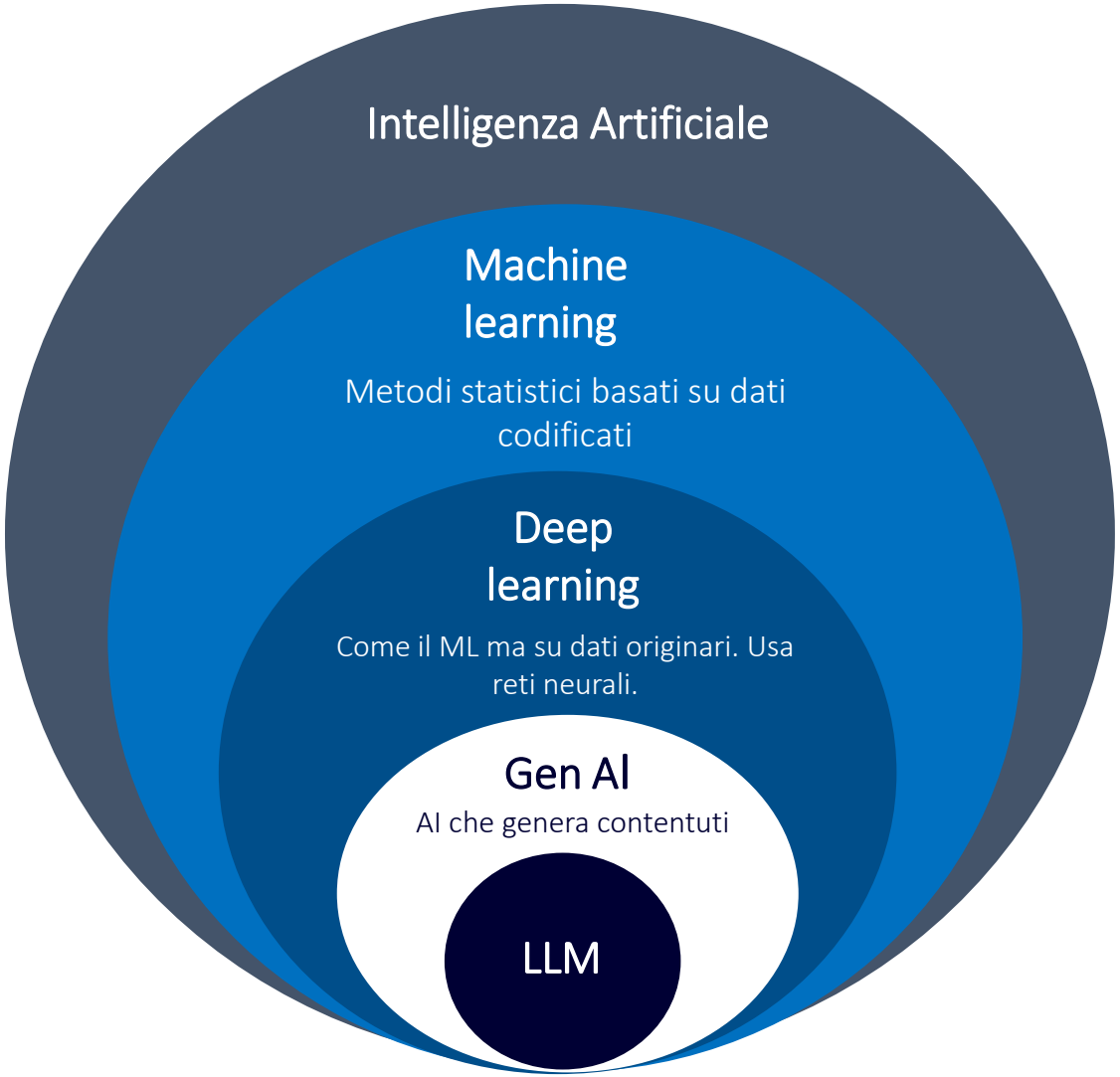
Quante ne sai?

Tokens

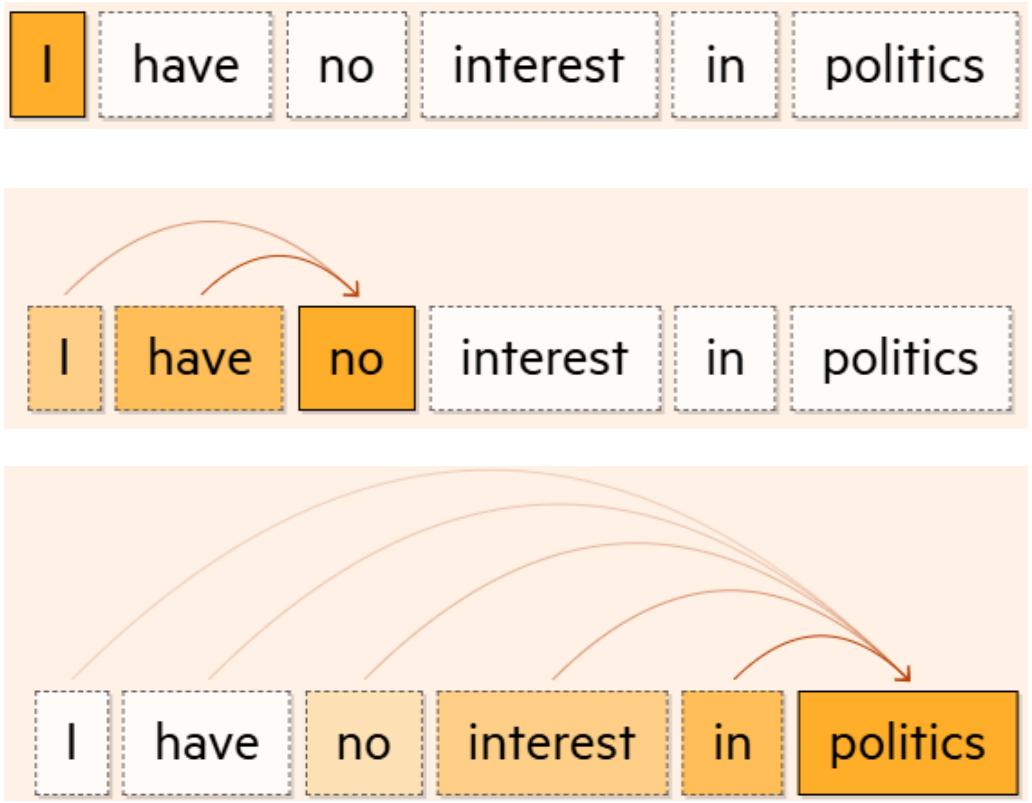
Agenti AI

Artificial General Intelligence (AGI)

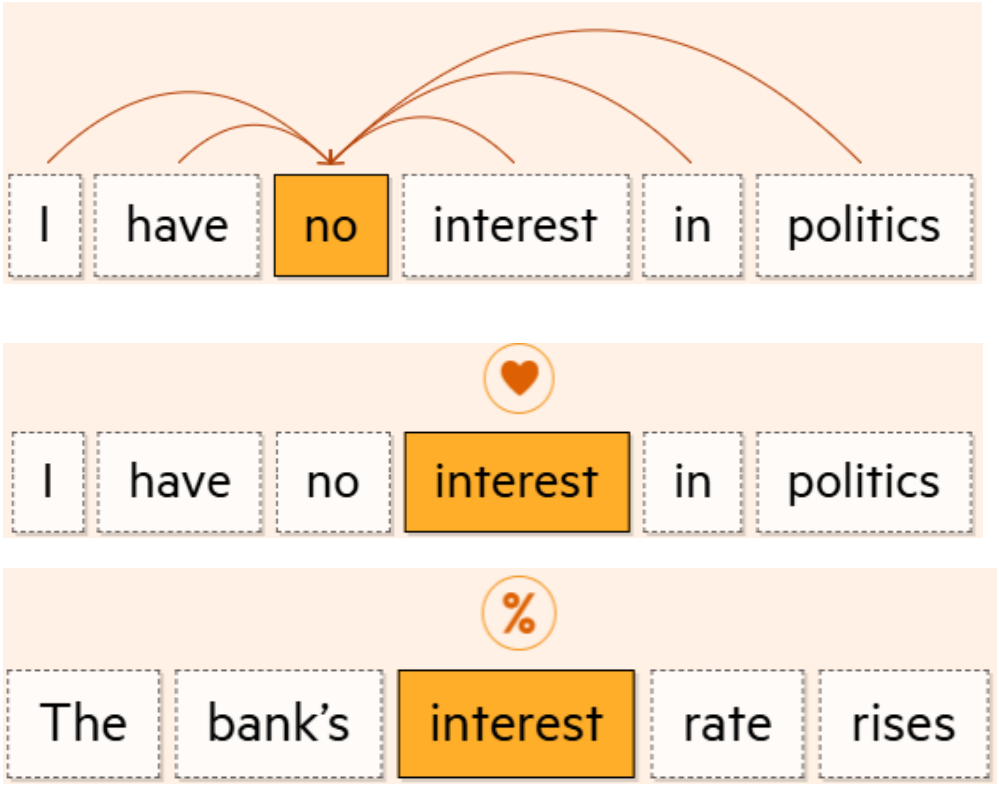
Large Language Model (LLM)



La conoscenza del mondo

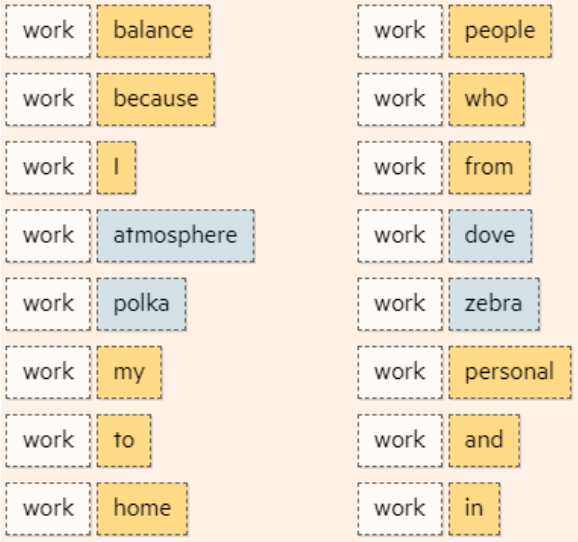


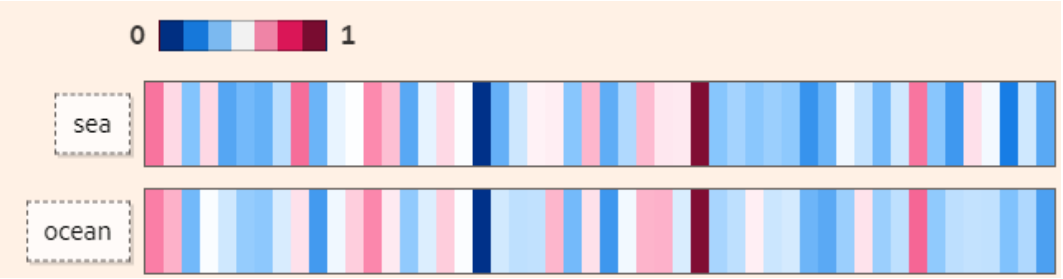
Prima del Transformers

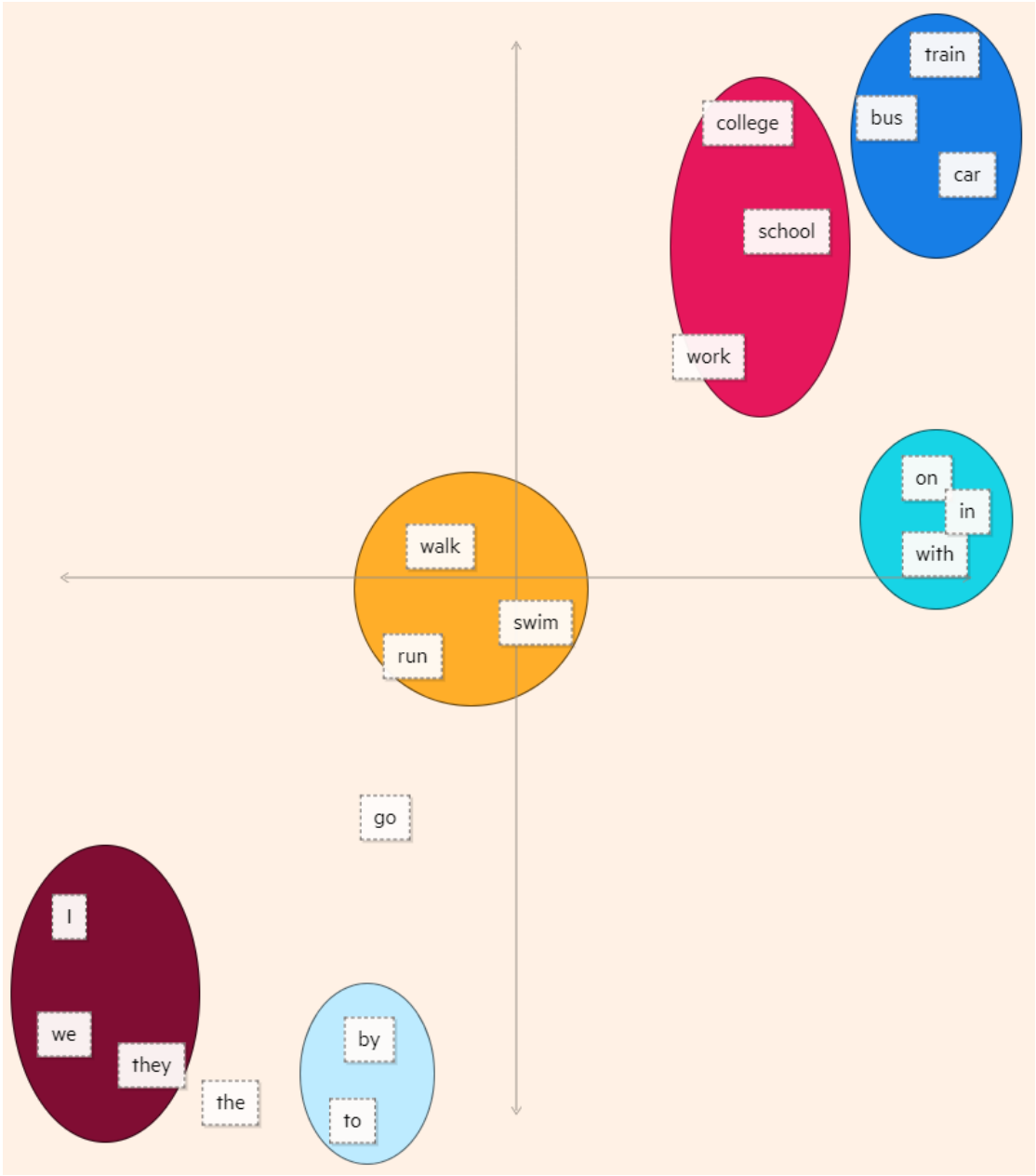


Transformers

La conoscenza del mondo

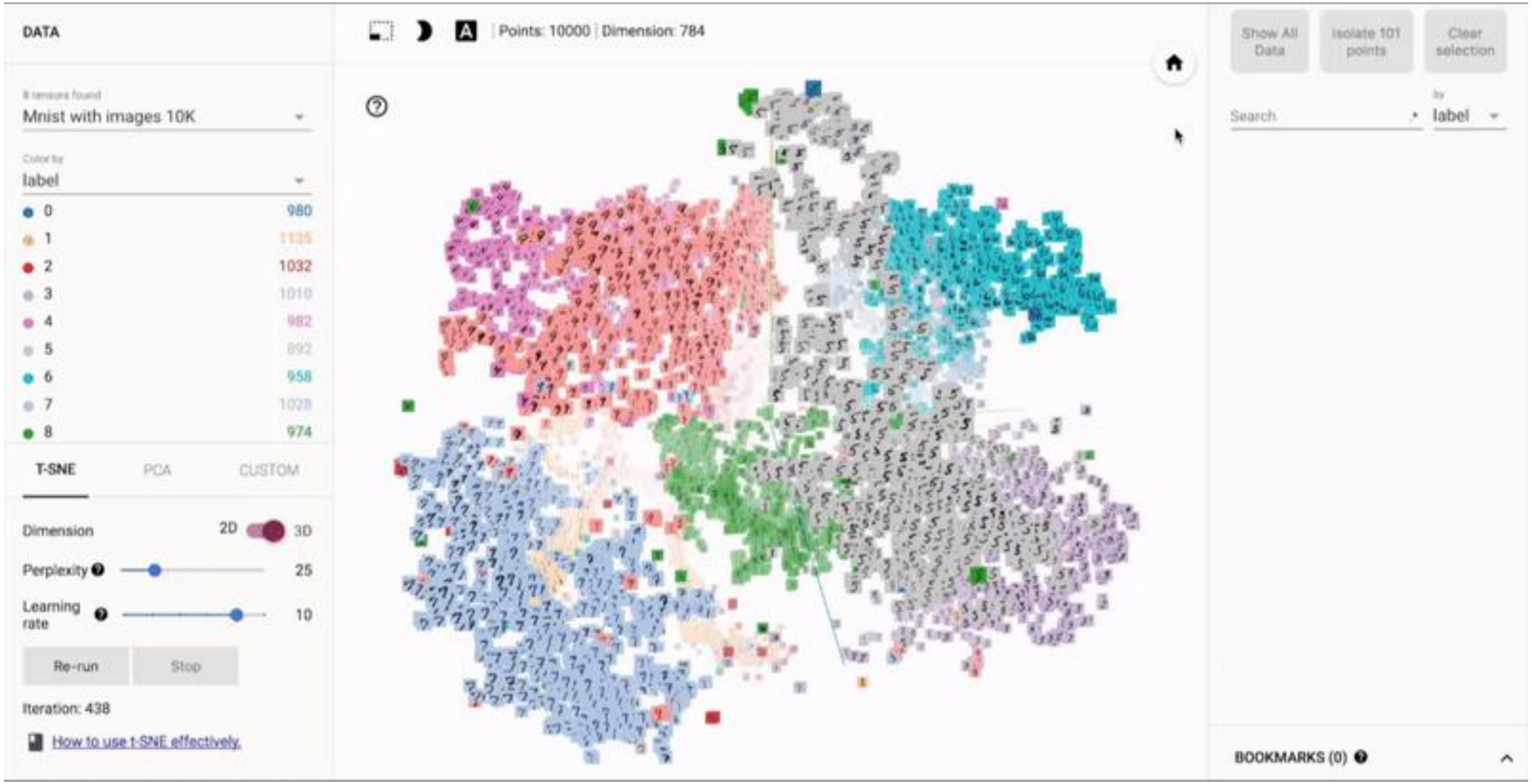






La conoscenza del mondo

https://projector.tensorflow.org/



La conoscenza del mondo

✔ Capacità degli LLM	⚠ Limiti degli LLM
✅ Comprensione del linguaggio naturale	⚠ Allucinazioni (possono inventare dati non veri)
✅ Generazione di contenuti (testi, email, codice, ecc.)	⚠ Memoria limitata (dimenticano interazioni passate)
✅ Traduzione e riscrittura in più stili	⚠ Nessuna comprensione profonda o intenzione
✅ Ragionamento a più passaggi (logica, analogie, schemi)	⚠ Rischio di bias (riflettono pregiudizi dei dati di addestramento)
✅ Apprendimento few-shot / zero-shot	⚠ Mancanza di accesso a dati aggiornati o aziendali senza strumenti RAG

LLM Con ragionamento?

Approccio	Prompt	Comportamento
LLM semplice	Marco ha il doppio degli anni di Luca. Insieme hanno 36 anni. Quanti anni ha ciascuno?	Risposta secca (possibile errore o hallucination)
LLM con ragionamento	Marco ha il doppio degli anni di Luca. Insieme hanno 36 anni. Quanti anni ha ciascuno?	Spiega il processo → più trasparente e affidabile
LLM base + prompt forzato	Ti darò un problema. Prima di rispondere, analizza ogni passaggio. Spiega cosa fai in ciascun passaggio. Rispondi solo alla fine. Ecco il problema: Marco ha il doppio degli anni di Luca. Insieme hanno 36 anni. Quanti anni ha ciascuno?	Può simulare il ragionamento se istruito chiaramente

Modelli Multimodali

Testo

Immagini

Audio

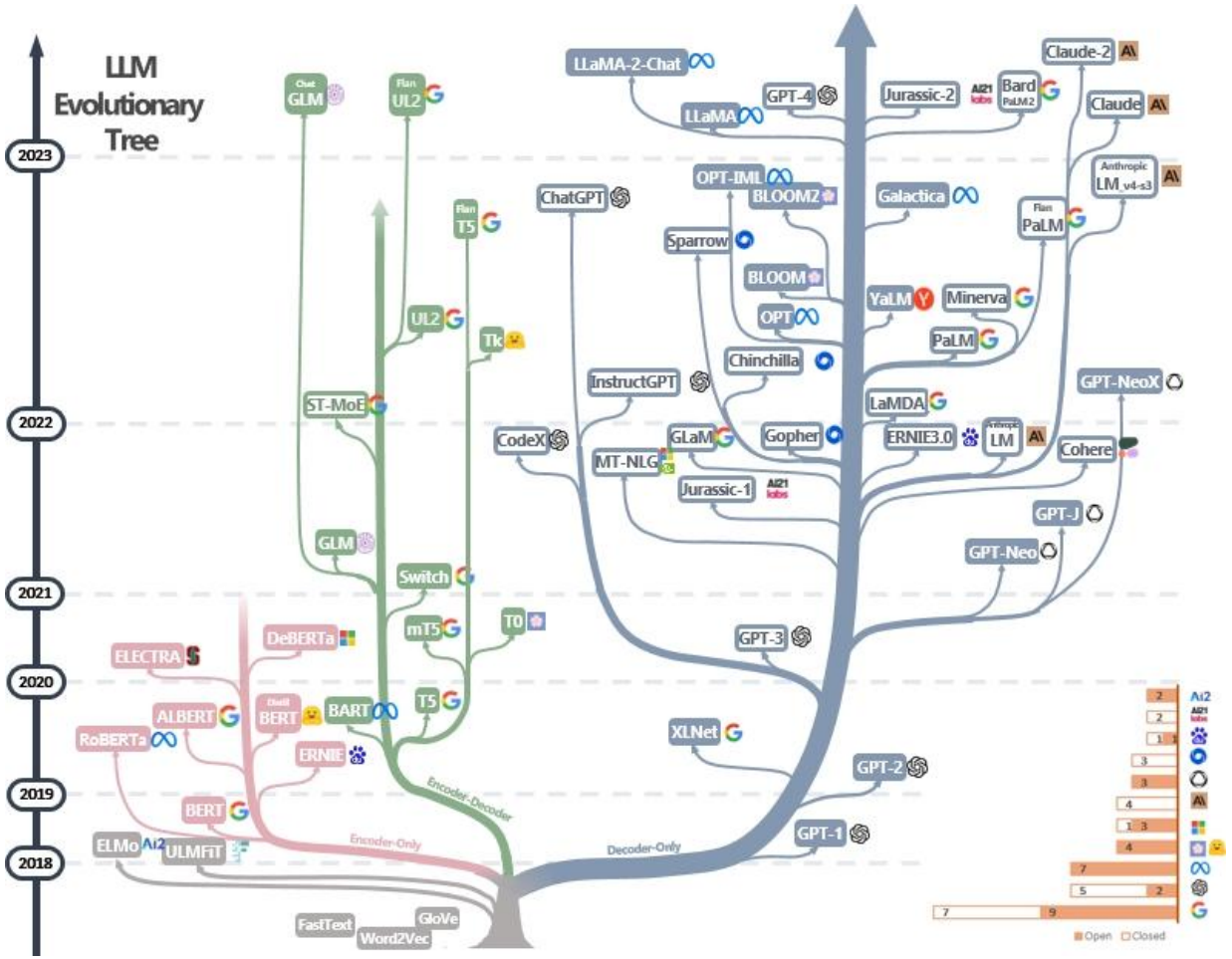
Elemento	LLM	Modello Multimodale
Input	Solo testo	Testo, immagini, audio, video
Output	Solo testo	Multiformato
Allenamento	Dataset testuali	Dataset multi-paired (testo + immagine, ecc.)
Capacità	Linguistica	Semantica trasversale tra modalità

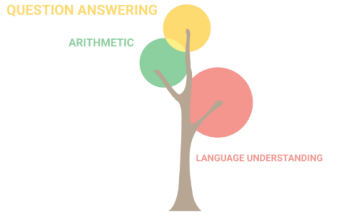
LLM to Video

LLM to Video

https://lumalabs.ai/ray

La conoscenza del mondo



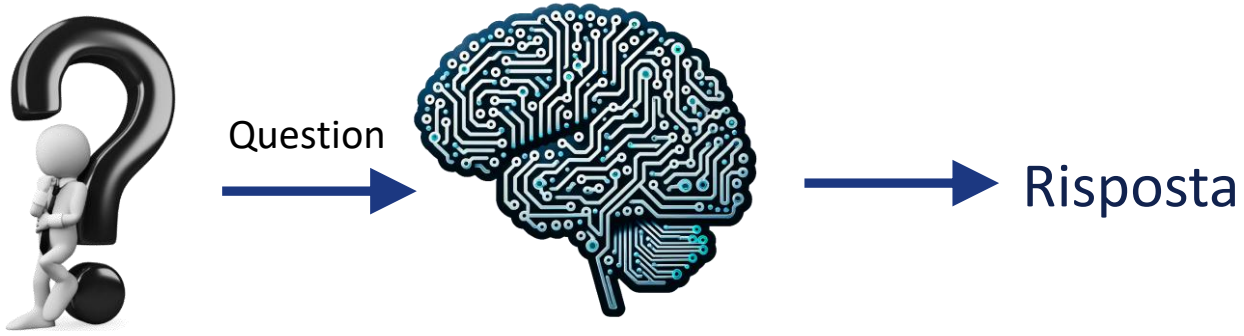


8 billion parameters

Sistemi RAG

WONDER

LLM uso originario



BLAME GAME —

Air Canada must honor refund policy invented by airline's chatbot

Air Canada appears to have quietly killed its costly chatbot support.

ASHLEY BELANGER - 2/16/2024, 6:12 PM

https://arstechnica.com/tech-policy/2024/02/air-canada-must-honor-refund-policy-invented-by-airlines-chatbot/

SI, MA!

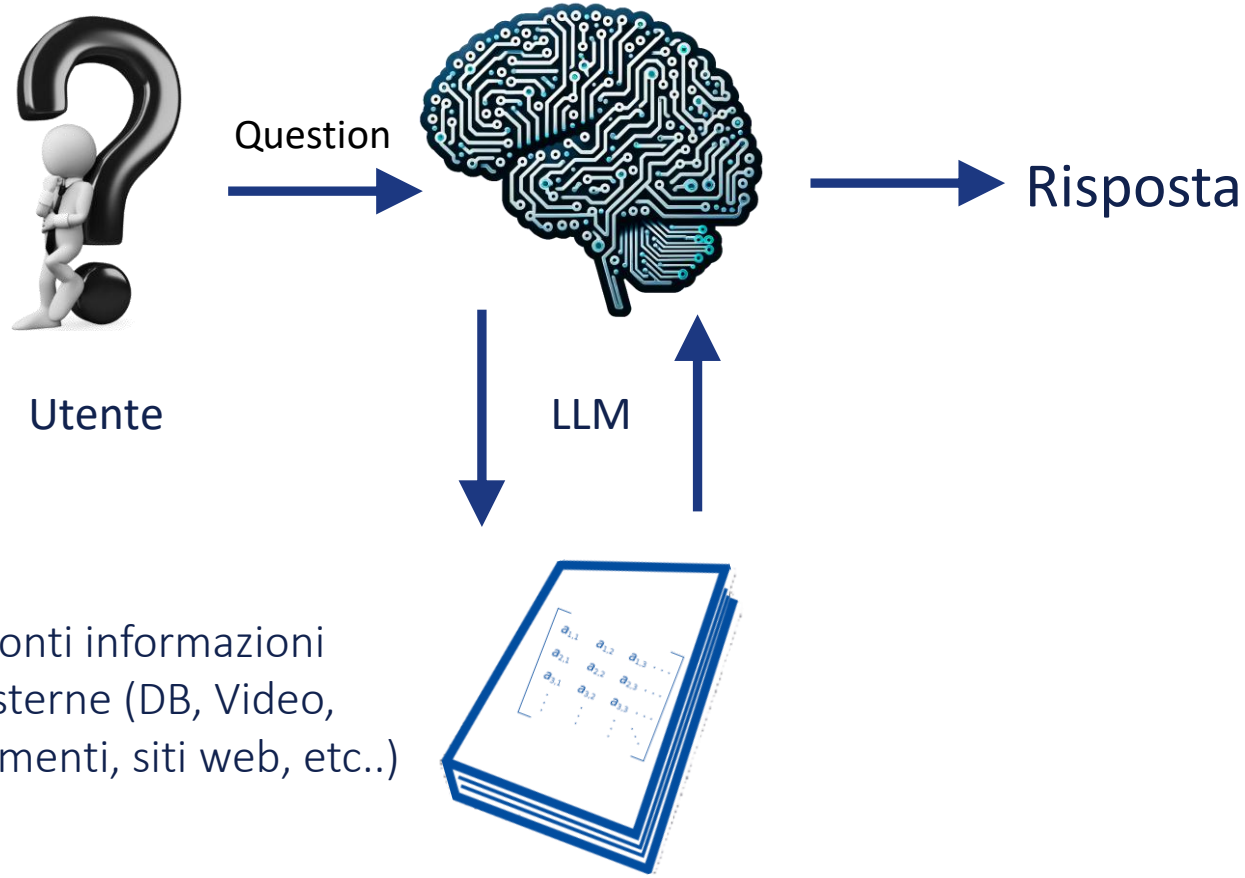
- E' Statico!
- Puo' generare risposte inaccurate
- Se non addestrato su un topic?
- Costoso da "ospitare":

LLM = molti parametri

↓

Tanta memoria, tanta capacita' di calcolo!

RAG – Retrieval Augmented generation



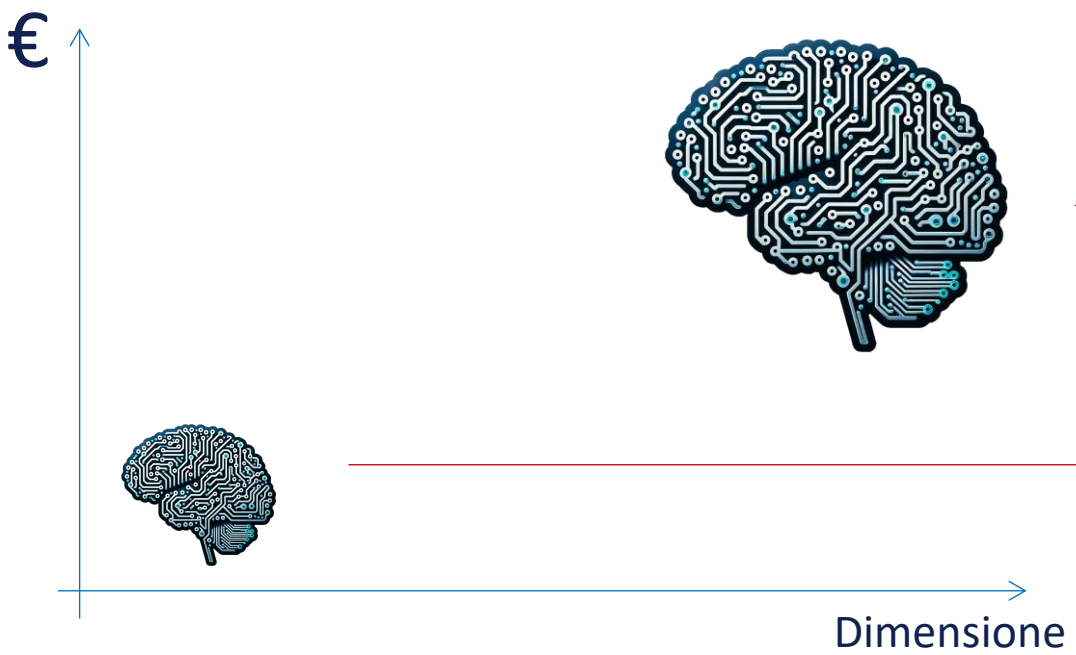
SI, PERCHE

- Conoscenza dinamica
- Si attiene alle fonti
- Riduce allucinazioni
- Piu' economico:

Modelli piu' piccolo (SLM)

Meno memoria, Meno computazione

Perchè saperlo?



Modello più «potente»
= più computazione,
Più memoria

Modello più «ridotto»
= meno computazione,
meno memoria

Quanti token mi servono?

- Regola 80-20: 100 token ≈ 75 parole.
- Moltiplica le parole per 1,33 per ottenere una stima dei token o viceversa.
- Conta entrambe le direzioni (prompt + output) nel budget.
- I prompt sono le domande. Partendo da una banca dati, si può stimare il caso pessimo

Perchè saperlo?

Pricing models

https://aws.amazon.com/bedrock/pricing/

On Demand and Batch +

Latency Optimized (Public Preview) −

Latency-optimized inference for foundation models in Amazon Bedrock delivers faster response times for models and helps improve responsiveness for your generative AI applications. You can use the use latency-optimized inference for Amazon Nova Pro, Anthropic's Claude 3.5 Haiku model, and Meta's Llama 3.1 405B and 70B models. As verified by Anthropic, with latency-optimized inference on Amazon Bedrock, Claude 3.5 Haiku runs faster on AWS than anywhere else. Additionally, with latency-optimized inference in Bedrock, Llama 3.1 405B and 70B runs faster on AWS than any other major cloud provider. Learn more here.

Provisioned Throughput +

Custom Model Import +

Marketplace models +

Perchè saperlo?

Google models

Google Cloud

Gemini 2.5

Model	Type	Price (/1M tokens) <= 200K input tokens	Price (/1M tokens) > 200K input tokens
Gemini 2.5 Pro	Input (text, image, video, audio)	$1.25	$2.5
	Text output (response and reasoning)	$10	$15
Gemini 2.5 Flash	Input (text, image, video)	$0.15	$0.15
	Audio Input	$1	$1
	Text output (no thinking)	$0.60	$0.60
	Text output (thinking- response and reasoning)	$3.50	$3.50

* If a query input context is longer than 200K tokens, all tokens (input and output) are charged at long context rates.

https://cloud.google.com/vertex-ai/generative-ai/pricing

Perchè saperlo?

Microsoft

o3

o3 is a powerful reasoning model from the o-series of reasoning models, pushing the frontier across coding, math, science, and visual perception. It excels in complex queries requiring multi-faceted analysis and performs strongly in visual tasks like analyzing images, charts, and graphics. The model features a 200K token context window and has a knowledge cutoff of June 2024.

Model	Pricing (1M Tokens)	Pricing with Batch API (1M Tokens)
o3 2025-04-16	Input: **$10** Cached Input: **$2.50** Output: **$40**	N/A

o4-mini

o4-mini is a compact, efficient, and cost-effective reasoning model from OpenAI's o-series. It excels in math, coding, and visual tasks. The model features a 200K token context window and has a knowledge cutoff of June 2024.

Model	Pricing (1M Tokens)	Pricing with Batch API (1M Tokens)
o4-mini 2025-04-16	Input: **$1.10** Cached Input: **$0.28** Output: **$4.40**	N/A

https://azure.microsoft.com/en-us/pricing/details/cognitive-services/openai-service/#pricing

Demo Time

GRAZIE

C S
M T
Innovation HUB

CSMT INNOVATION HUB
VIA BRANZE 45 BRESCIA
WWW.CSMT.IT
INFO@CSMT.IT
030 6595108

Per informazioni e approfondimenti, contattare:

LICIA ZAGNI
Head of Sales & Marketing
CSMT Innovation Hub
T. +39 030 6595110 | M. +39 328 4505280

VIOLA NICOLARDI
Technology Transfer Engineering / Digital & AI
CSMT Innovation Hub
T. +39 030 6595125 | M. +39 335 6044069