IA GENERATIVA, QUANTO MI COSTI

Francesco Alaimo

info@datasciencefacile.it

THE AI BOOM COULD USE A SHOCKING AMOUNT OF ELECTRICITY (SCI AM)

THE AI BOOM COULD USE A SHOCKING AMOUNT OF ELECTRICITY

- Con l'attuale tendenza, si stima che **NVIDIA** produrrà e venderà 1,5 milioni di server per l'AI entro il 2027, con un consumo stimato di 85,4 TWh/anno
- È più di quanto consumano molti piccoli stati in un anno!

THE AI BOOM COULD USE A SHOCKING AMOUNT OF ELECTRICITY

- Il motore di ricerca di Google, sviluppato con l'aggiunta delle AI, consumerebbe più del corrispondente servizio cui siamo abituati
- Si stima, un consumo pari a quello dell'Irlanda…
- La questione è: l'utente ha davvero bisogno della AI per fare le ricerche?

THE AI BOOM COULD USE A SHOCKING AMOUNT OF ELECTRICITY

- Il consumo è determinato dalle dimensioni dei modelli ma anche dalla necessità di raffreddare i server
- Aumenta se aumentano le richieste e la loro complessità
- Tra i limiti delle IA generative, si dovrebbe aggiungere anche il costo in termini di impatti sull'ambiente

Immagine generata con AI

IL CONSUMO ENERGETICO DELL'AI È FUORI CONTROLLO (WIRED)

IL CONSUMO ENERGETICO DELL'AI È FUORI CONTROLLO

- Le **AI** generative sono, oramai, pervasive
- Dal 2022, data di lancio di **ChatGPT**, applicazioni dei **LLM** sono diventate onnipresenti
- Sviluppare i sistemi di AI generativa richiede una quantità enorme di elettricità e di acqua

IL CONSUMO ENERGETICO DELL'AI È FUORI CONTROLLO

- Secondo le stime dell'agenzia internazionale dell'energia, i data center consumano l'1% dell'elettricità mondiale
- Google, nel 2023, ha aumentato le emissioni di CO_2 del 48% rispetto al 2019, a causa dei data center
- Con lo sviluppo dell'AI questa tendenza potrebbe anche peggiorare

Immagine generata con AI

HOW MUCH ELECTRICITY DOES AI CONSUME? (THE VERGE)

HOW MUCH ELECTRICITY DOES AI CONSUME?

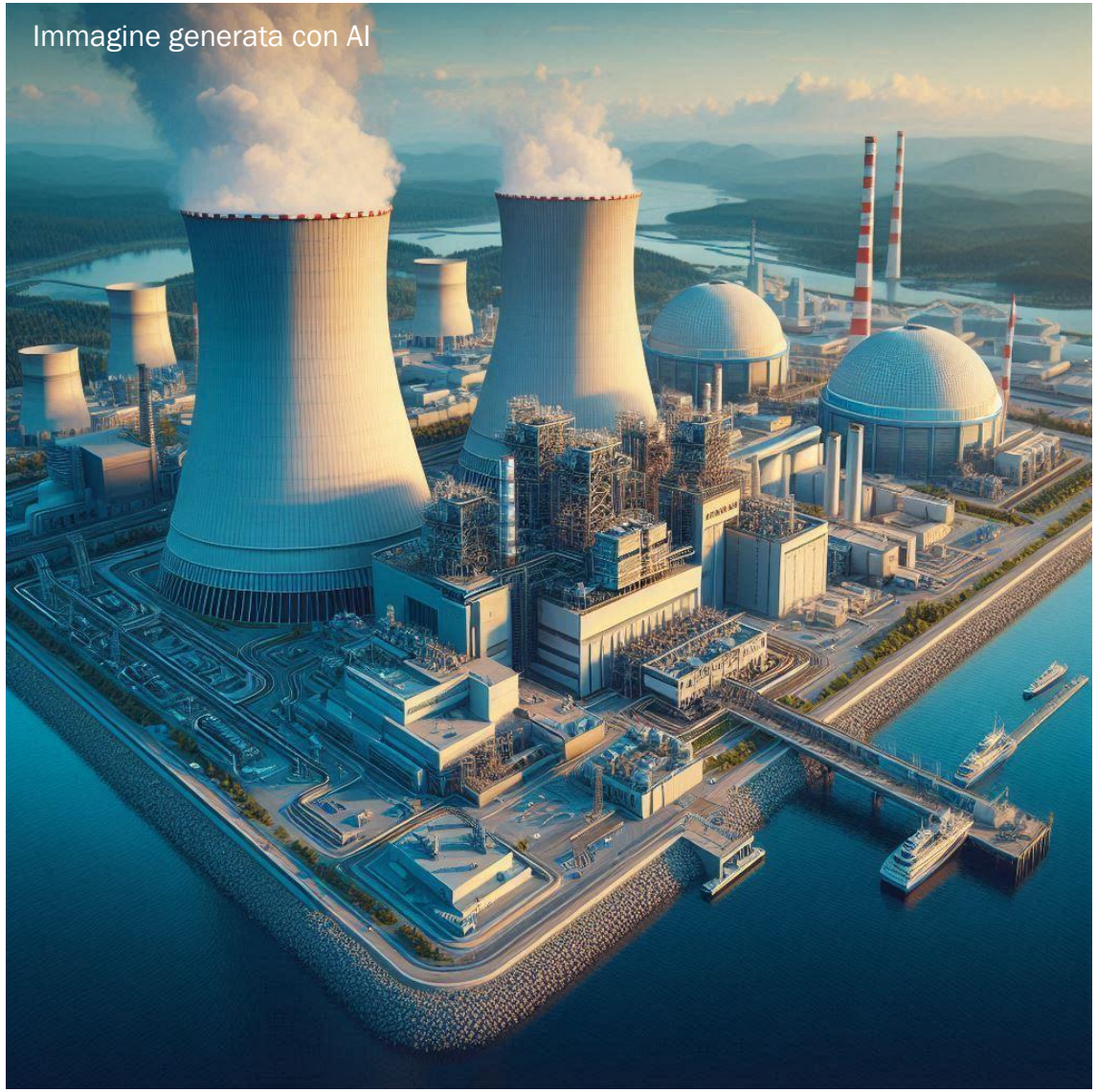
- L’addestramento di un modello è particolarmente energivoro
- Si stima che GPT-3 abbia richiesto circa 1300MWh
- Con Netflix raggiungeremmo gli stessi consumi dopo 1625000 ore di visione (185 anni)

HOW MUCH ELECTRICITY DOES AI CONSUME?

- OpenAI, all'inizio, pubblicava tutti i dettagli sull' addestramento, quale hardware, quanto tempo
- Con gli ultimi modelli queste informazioni non sono più disponibili
- GPT-4 e successivi potrebbero stare funzionando con dei procioni in un fosso...

HOW MUCH ELECTRICITY DOES AI CONSUME?

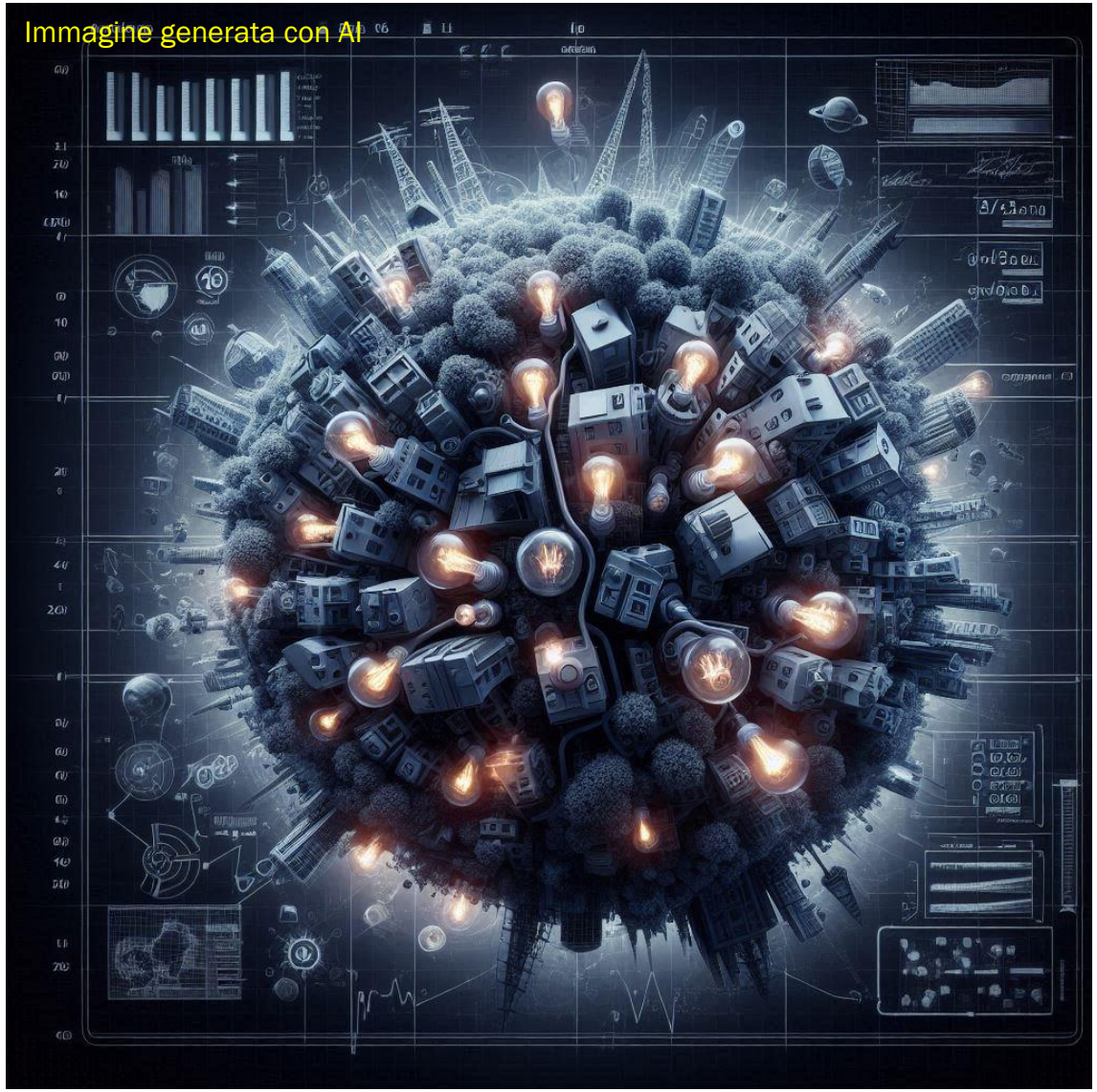
- Alcuni studiosi della Carnegie Mellon University hanno stimato, in uno studio, che la generazione di una immagine richiede la stessa energia che serve per ricaricare uno smartphone (0.012 kWh)



LA SINDROME BIG TECH (Il Manifesto)

LA SINDROME BIG TECH

- A **Three Mile Island**, in Pennsylvania, nel 2028 riaprirà il reattore 1 della centrale nucleare che fu teatro di un grave incidente nucleare nel 1979
- Qualche settimana prima nel film ‘Sindrome cinese’, lo scenario era stato ipotizzato
- Sarà utilizzato, in via esclusiva, da **Microsoft**, per alimentare i suoi datacenter di Intelligenza artificiale



POWER HUNGRY PROCESSING: WATTS DRIVING THE COST OF AI DEPLOYMENT? (HUGGING FACE & CARNEGIE UNIV.)

POWER HUNGRY PROCESSING: WATTS DRIVING THE COST OF AI DEPLOYMENT?

- Tra il 2017 e il 2021, i consumi di Meta, Google, Amazon, Microsoft, per i servizi Cloud, sono raddoppiati
- In questo contesto non è chiaro il contributo sui consumi energetici legati allo sviluppo dell'AI
- Il rapporto tra i consumi legati all'addestramento dei modelli è maggiormente spostato sull'inferenza che si fa sugli stessi

POWER HUNGRY PROCESSING: WATTS DRIVING THE COST OF AI DEPLOYMENT?

- Lo studio si concentra su 88 modelli in 10 attività e 30 set di dati, che spaziano dalle applicazioni del linguaggio naturale alla visione artificiale
- Ogni test viene ripetuto 1000 volte
- Sebbene, in termini energetici, la fase di addestramento sia onerosa, lo studio rivela che la fase di inferenza lo è anche di più

POWER HUNGRY PROCESSING: WATTS DRIVING THE COST OF AI DEPLOYMENT?

Task	Datasets	Task	Datasets
image classification	CIFAR 10 [25] CIFAR 100 [25] ImageNet 1K [45]	**question answering**	SQuAD[44] SQuAD v2 [43] SciQ [23]
image captioning	Visual Genome [24] RedCaps [10] COCO [29]	**summarization**	SAMSum [15] CNN-Daily Mail [20] XSum [35]
image generation	DiffusionDB [54] ImageReward [58] SD Prompts [46]	**text classification**	IMDB [32] Rotten Tomatoes [39] SST 2 [48]
masked language modeling	BookCorpus [59] C4 [42] OSCAR [37]	**text generation**	WikiText [33] BookCorpus [59] OSCAR [37]
object detection	Visual Genome [24] CPPE-5 [9] COCO [29]	**token classification**	ReCoRD [53] WikiANN [38] CoNLL 2003 [50]

POWER HUNGRY PROCESSING: WATTS DRIVING THE COST OF AI DEPLOYMENT?

- Tutti i tests sono stati condotti utilizzando una **NVIDIA A100-SXM4-80GB** GPU su AWS (us-west-2)
- È stato utilizzato il programma **Code Carbon**
- Tenendo conto dell'impronta di CO_2 del sito AWS, sono stati consumati **754,66 KWh** per una produzione di **178,97 Kg** di **CO_2eq**

POWER HUNGRY PROCESSING: WATTS DRIVING THE COST OF AI DEPLOYMENT?

	inference energy (kWh)	
task	**mean**	**std**
text classification	0.002	0.001
extractive QA	0.003	0.001
masked language modeling	0.003	0.001
token classification	0.004	0.002
image classification	0.007	0.001
object detection	0.038	0.02
text generation	0.047	0.03
summarization	0.049	0.01
image captioning	0.063	0.02
image generation	2.907	3.31

POWER HUNGRY PROCESSING: WATTS DRIVING THE COST OF AI DEPLOYMENT?

- **Stable-diffusion-xl-base-1.0** genera **1594 gr** di CO_2 per 1000 inferenze, equivalenti a **6,6 Km** di un veicolo a benzina
- **Distilbert-base-uncased** genera **0,2 gr** per 1000 inferenze
- L'utilizzo dei modelli più energivori, da parte di milioni di utenti globali, ha un impatto sull'ambiente considerevole!

POWER HUNGRY PROCESSING: WATTS DRIVING THE COST OF AI DEPLOYMENT?

- Analizzando i consumi su alcune attività svolte con modelli specifici e con modelli generici, si è visto che i consumi, ogni 1000 inferenze, sono in rapporto 3 su 100 di gr di CO_2eq
- Esempio, una analisi del sentiment con una rete neurale, con uscita 0 o 1, assorbe molto meno che utilizzare un LLM per lo stesso compito, con uscita *positivo* o *negativo*

POWER HUNGRY PROCESSING: WATTS DRIVING THE COST OF AI DEPLOYMENT?

- Per il compito *summarization*, i consumi dei modelli generativi specifici e generici sono vicini anche se aumentano nei modelli con il numero di parametri più alto

POWER HUNGRY PROCESSING: WATTS DRIVING THE COST OF AI DEPLOYMENT?

- Lo studio suggerisce le seguenti azioni di mitigazione:
  - Scelta dell'architettura
  - Utilizzo di tecniche di distillazione
  - Scelta del numero di parametri
  - Scelta dell'hardware, considerando il rapporto prestazioni/consumi
  - Utilizzo di una precisione numerica più bassa (ove possibile)
  - Preferenza di una soluzione **open source** rispetto ad una closed

TRAINING IMAGENET IN 3 HOURS FOR USD 25; AND CIFAR10 FOR USD 0.26 (FAST.AI)

TRAINING IMAGENET IN 3 HOURS FOR USD 25; AND CIFAR10 FOR USD 0.26

- Non sempre la prima soluzione è la migliore
- Si può eseguire una classificazione in meno tempo ed economicamente?
- Alcuni studiosi hanno dimostrato, che è possibile
- **un bravo sperimentatore con un computer lento dovrebbe sempre essere in grado di superare un mediocre sperimentatore, con uno veloce.**

TRAINING IMAGENET IN 3 HOURS FOR USD 25; AND CIFAR10 FOR USD 0.26

- Si può costruire una applicazione di machine learning ottimizzando parti del procedimento, invece che contando solo sull'abbondanza di risorse!
- Il mondo è un luogo con risorse limitate: avere l'accesso a grandi risorse, non deve farcelo dimenticare
- **Nel corso della storia, i vincoli sono stati il motore dell' innovazione e della creatività, ma molti non sembrano apprezzare questa lezione**

Immagine generata con AI

THE CARBON FOOTPRINT OF MACHINE LEARNING TRAINING WILL PLATEAU, THEN SHRINK (GOOGLE, UNIVERSITÀ DI BERKELEY)

THE CARBON FOOTPRINT OF MACHINE LEARNING TRAINING WILL PLATEAU, THEN SHRINK

- Le emissioni di carbonio per l'addestramento dei modelli di ML ha due voci principali:
  - **Operational**: costo energetico per hardware (anche infrastrutturale)
  - **Lifecycle**: costo energetico per la produzione di tutti i componenti coinvolti, dai chip agli edifici dei datacenter

THE CARBON FOOTPRINT OF MACHINE LEARNING TRAINING WILL PLATEAU, THEN SHRINK

- Quattro best practices (4M) possono ridurre i consumi di 100 volte e le emissioni fino a 1000 volte, rispetto alle scelte più comuni
  - MODEL
  - MACHINE
  - MECHANIZATION
  - MAP

THE CARBON FOOTPRINT OF MACHINE LEARNING TRAINING WILL PLATEAU, THEN SHRINK

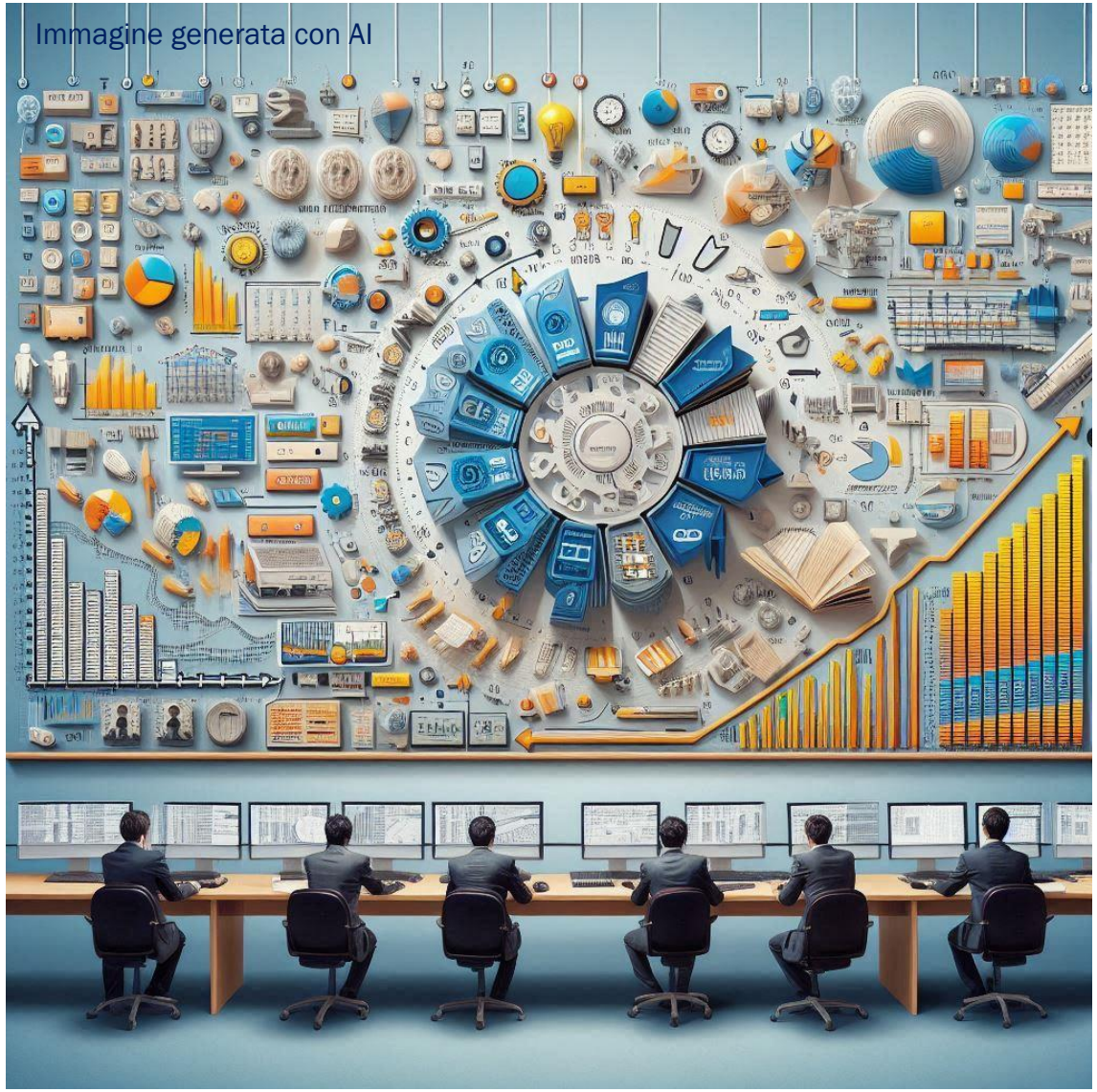
- **Model**: architetture di ML efficienti, l'adozione dei modelli sparsi rispetto a quelli densi, possono ridurre i consumi di un fattore da 5 a 10
- **Machine**: processori ottimizzati per l'addestramento ML, come TPU e GPU recenti, rispetto a processori generici migliora il fattore performance/Watt di un fattore da 2 a 5

THE CARBON FOOTPRINT OF MACHINE LEARNING TRAINING WILL PLATEAU, THEN SHRINK

- **Mechanization**: utilizzare il cloud, invece di data center on premise, migliora l'efficienza energetica, riducendo i costi di un fattore da 1.4 a 2
- **Map**: usare data center che utilizzano l'energia più pulita, riduce l'impronta di carbonio lorda di un fattore da 5 a 105

THE CARBON FOOTPRINT OF MACHINE LEARNING TRAINING WILL PLATEAU, THEN SHRINK

- **GPT3:** ogni token attiva 175B di parametri
- **GLaM**: si basa sulla tecnica MoE (Mixture of Expert) e anche essendo più grande non utilizza più dell'8% dei parametri, in funzione dell'input
  - **GLaM** è stato addestrato utilizzando le raccomandazioni **4M**

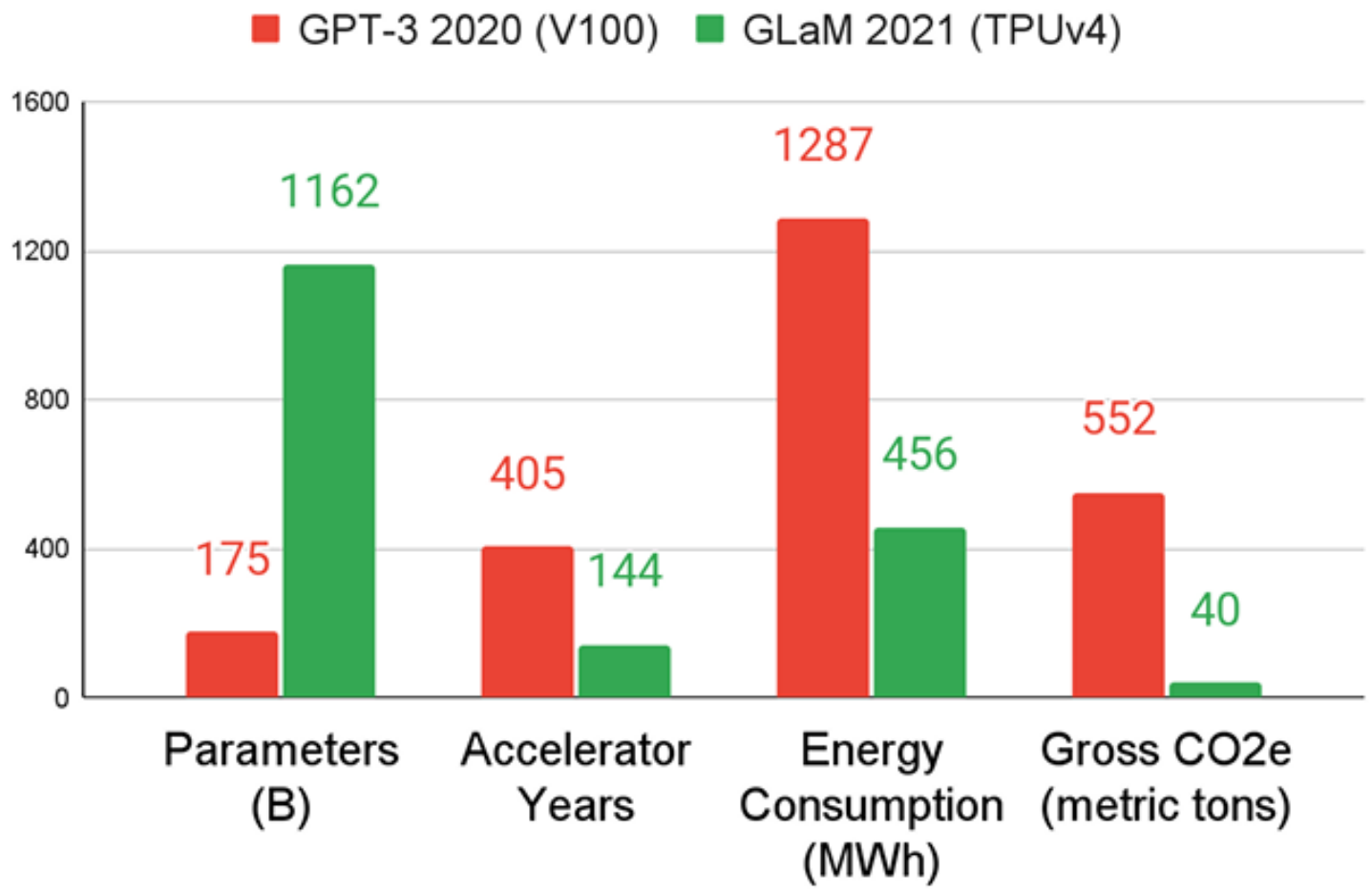


GLAM: EFFICIENT SCALING OF LANGUAGE MODELS WITH MIXTURE-OF-EXPERTS (GOOGLE)

GLAM: EFFICIENT SCALING OF LANGUAGE MODELS WITH MIXTURE-OF-EXPERTS

- È un modello di 1,5 trilioni di parametri di tipo **sparse**
- È stato addestrato con 1/3 dell'energia utilizzata per GPT-3
- Richiede meno potenza elaborativa per l'inferenza a parità di prestazioni

GLAM: EFFICIENT SCALING OF LANGUAGE MODELS WITH MIXTURE-OF-EXPERTS



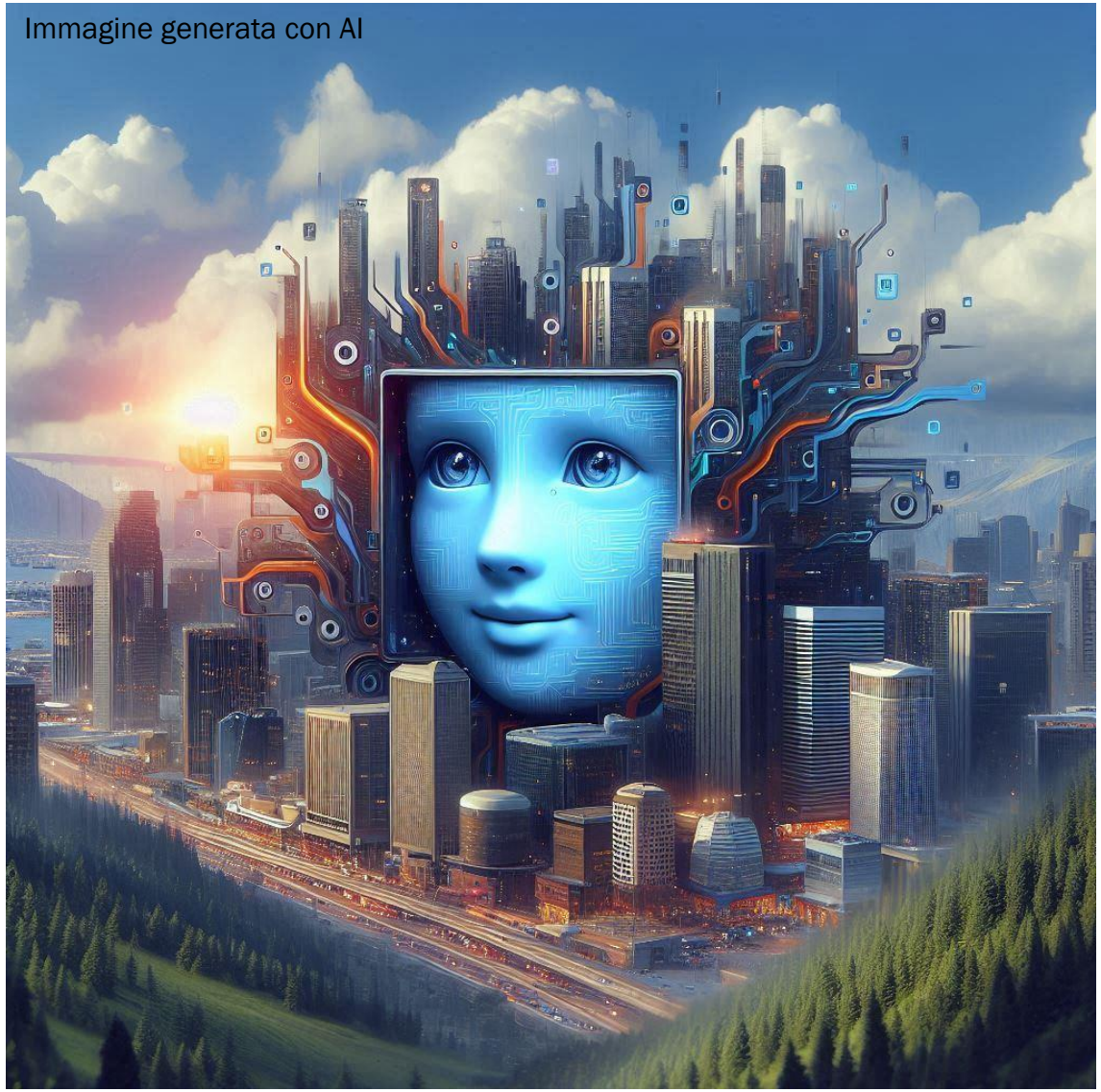


Immagine generata con AI

MIXTURE OF EXPERT EXPLAINED (HUGGING FACE)

MIXTURE OF EXPERT EXPLAINED

- Sono modelli pre-addestrati sparsi, che impiegano un tempo minore per l'addestramento, rispetto ai modelli densi
- A parità di numero di parametri, hanno un'inferenza più veloce rispetto ai modelli densi
- Richiedono un'elevata VRAM perché gli tutti gli esperti devono essere presenti in memoria

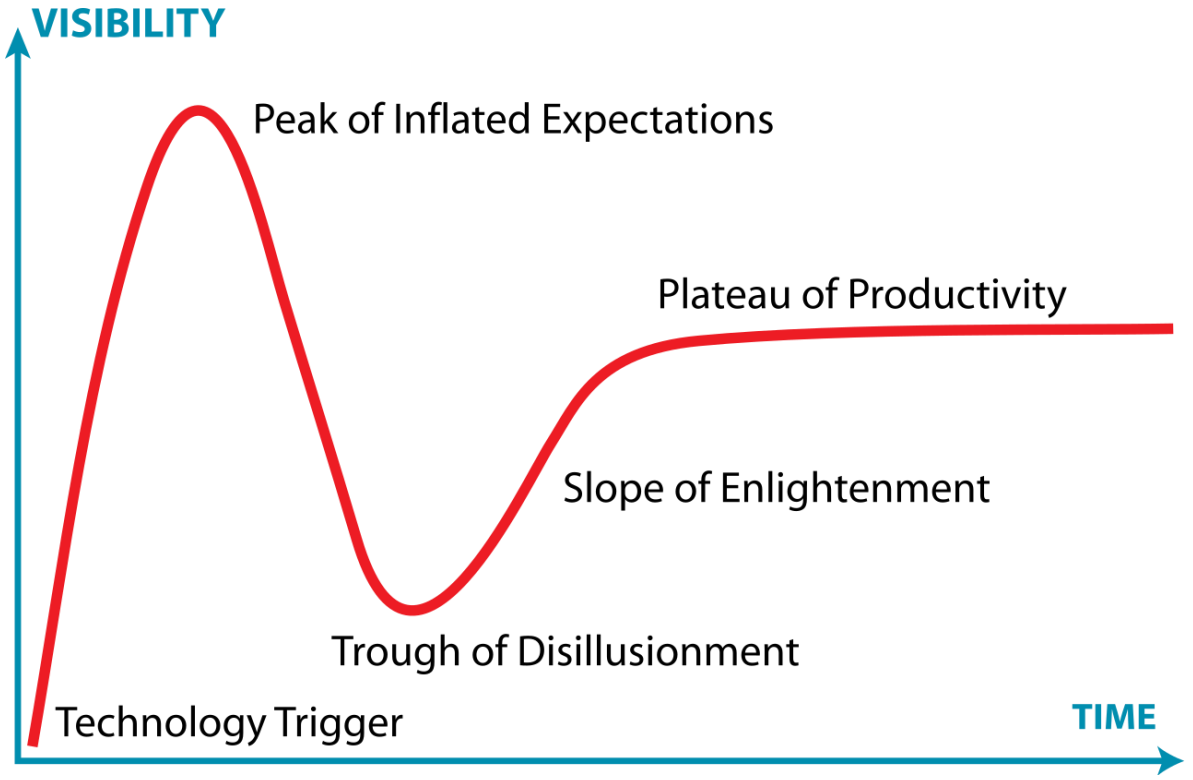
MIXTURE OF EXPERT EXPLAINED

- Un modello **MoE**, dispone di un certo numero di esperti e ogni esperto è una rete neurale
- Gli esperti possono essere organizzati in una gerarchia
- Un rete di *gate* o *router*, determina l'esperto che dovrà vagliare la richiesta

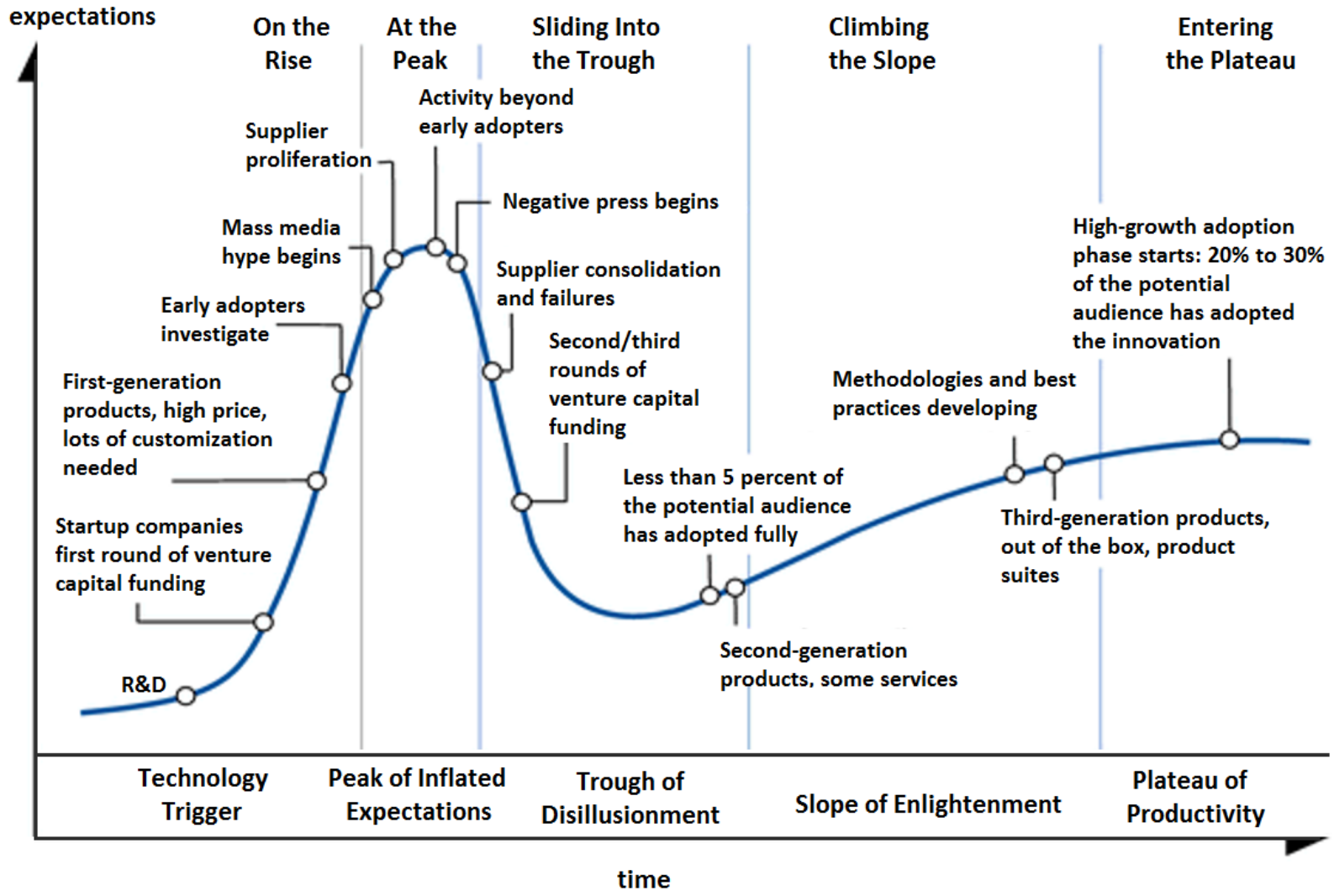
HYPE CYCLE (GARTNER)

HYPE CYCLE (GARTNER)

1. Tecnology trigger
2. Peak of inflated Expectations
3. Trough of disillusionment
4. Slope of Enlightenment
5. Plateau of productivity



HYPE CYCLE (GARTNER)



CONCLUSIONI

- Lo sviluppo delle IA generative, nel futuro, dovrà necessariamente tenere conto dell' impatto sull'ambiente
- Nella fase di plateau dell'**Hype Cycle**, le tecnologie che utilizzeremo dovranno essere le migliori in termini di efficienza energetica e di business case

GRAZIE

FRANCESCO ALAIMO (0009-0004-4842-5216) – ORCID – info@datasciencefacile.it