



ICT Training Center



Il tuo partner per la Formazione e la Trasformazione digitale della tua azienda





SPRING AI

GENERATIVE ARTIFICIAL INTELLIGENCE CON JAVA

Simone Scannapieco

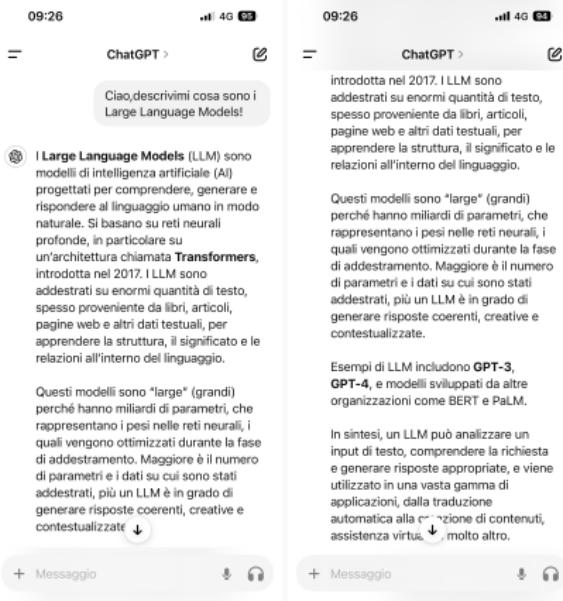
Corso base per Venis S.p.A, Venezia, Italia

Settembre 2025

LARGE LANGUAGE MODELS

LARGE LANGUAGE MODELS

COSA SONO?



The image shows two identical mobile device screens side-by-side, both displaying a messaging application interface. At the top of each screen, the time is 09:26, the signal strength is 4G, and the battery level is at 65%.

Left Screen (User Message):

Ciao, descrivimi cosa sono i Large Language Models!

Left Screen (ChatGPT Response):

I Large Language Models (LLM) sono modelli di intelligenza artificiale (AI) progettati per comprendere, generare e rispondere al linguaggio umano in modo naturale. Si basano su reti neurali profonde, in particolare su un'architettura chiamata **Transformers**, introdotta nel 2017. I LLM sono addestrati su enormi quantità di testo, spesso proveniente da libri, articoli, pagine web e altri dati testuali, per apprendere la struttura, il significato e le relazioni all'interno del linguaggio.

Questi modelli sono "large" (grandi) perché hanno miliardi di parametri, che rappresentano i pesi nelle reti neurali, i quali vengono ottimizzati durante la fase di addestramento. Maggiore è il numero di parametri e i dati su cui sono stati addestrati, più un LLM è in grado di generare risposte coerenti, creative e contestualizzate.

Right Screen (User Message):

+ Messaggio

Right Screen (ChatGPT Response):

introdotta nel 2017. I LLM sono addestrati su enormi quantità di testo, spesso proveniente da libri, articoli, pagine web e altri dati testuali, per apprendere la struttura, il significato e le relazioni all'interno del linguaggio.

Questi modelli sono "large" (grandi) perché hanno miliardi di parametri, che rappresentano i pesi nelle reti neurali, i quali vengono ottimizzati durante la fase di addestramento. Maggiore è il numero di parametri e i dati su cui sono stati addestrati, più un LLM è in grado di generare risposte coerenti, creative e contestualizzate.

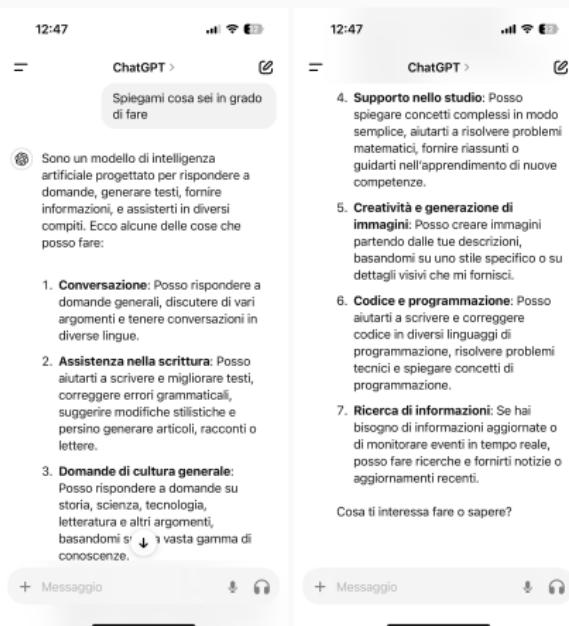
Esempi di LLM includono **GPT-3**, **GPT-4**, e modelli sviluppati da altre organizzazioni come **BERT** e **PaLM**.

In sintesi, un LLM può analizzare un input di testo, comprendere la richiesta e generare risposte appropriate, e viene utilizzato in una vasta gamma di applicazioni, dalla traduzione automatica alla cernizione di contenuti, assistenza virtuale, molto altro.

👉 Non si sarebbe potuto spiegare meglio!

LARGE LANGUAGE MODELS

A COSA SERVONO?



The image shows two side-by-side screenshots of a mobile phone displaying a ChatGPT conversation. Both screens show the time as 12:47 and a battery level of approximately 50%.

Left Screen:

- Message input field: "Spiegami cosa sei in grado di fare"
- Response: "Sono un modello di intelligenza artificiale progettato per rispondere a domande, generare testi, fornire informazioni, e assisterti in diversi compiti. Ecco alcune delle cose che posso fare:

- Conversazione:** Posso rispondere a domande generali, discutere di vari argomenti e tenere conversazioni in diverse lingue.
- Assistenza nella scrittura:** Posso aiutarti a scrivere e migliorare testi, correggere errori grammaticali, suggerire modifiche stilistiche e persino generare articoli, racconti o lettere.
- Domande di cultura generale:** Posso rispondere a domande su storia, scienza, tecnologia, letteratura e altri argomenti, basandomi sulla vasta gamma di conoscenze.

Right Screen:

- Supporto nello studio:** Posso spiegare concetti complessi in modo semplice, aiutarti a risolvere problemi matematici, fornire riassunti o guidarti nell'apprendimento di nuove competenze.
- Creatività e generazione di immagini:** Posso creare immagini partendo dalle tue descrizioni, basandomi su uno stile specifico o su dettagli visivi che mi fornisci.
- Codice e programmazione:** Posso aiutarti a scrivere e correggere codice in diversi linguaggi di programmazione, risolvere problemi tecnici e spiegare concetti di programmazione.
- Ricerca di informazioni:** Se hai bisogno di informazioni aggiornate o di monitorare eventi in tempo reale, posso fare ricerche e fornirti notizie o aggiornamenti recenti.

Below the right screen's list, there is a message: "Cosa ti interessa fare o sapere?"

At the bottom of both screens are message input fields labeled "+ Messaggio" and icons for microphone and speaker.

➔ Conferma quanto anticipato nella nostra discussione su NLP 😊

LARGE LANGUAGE MODELS COME VENGONO ADDESTRATI?

The image displays two side-by-side screenshots of a mobile device screen, both showing a ChatGPT interface. The top status bar indicates the time as 11:35, signal strength, 4G connectivity, and battery level. The first screenshot shows a user message: "Dammi le specifiche tecniche del tuo dataset di addestramento". The second screenshot shows ChatGPT's response: "Il mio dataset di addestramento è costituito da una vasta raccolta di testi, creata da OpenAI per garantire la massima diversità e copertura possibile su molti argomenti. Tuttavia, non ho accesso diretto a dati specifici o riservati, come database privati, informazioni personali non pubbliche o dati aggiornati in tempo reale (tranne quando uso strumenti di browsing per raccogliere informazioni aggiornate). Ecco alcuni dettagli chiave sulle fonti del mio dataset di addestramento:

1. **Fonti di testo:** Include libri, articoli, siti web pubblici, documenti tecnici, manuali, e codici di programmazione. La diversità delle fonti mi consente di comprendere una vasta gamma di argomenti, dalla letteratura alla scienza, fino ↓ dice informatico.

The third screenshot (partially visible) continues the list:

2. **Multilingue:** Anche se prevalentemente addestrato in inglese, ho competenze in molte lingue, tra cui l'italiano, grazie ai testi provenienti da varie lingue e culture.
3. **Contenuti aggiornati fino al 2021:** Il mio dataset di addestramento contiene informazioni pubblicamente disponibili fino a settembre 2021. Non ho accesso a eventi o cambiamenti successivi a quella data, a meno che non utilizzi strumenti specifici per la ricerca.
4. **Esclusione di contenuti privati:** Non sono stato addestrato su informazioni personali o private (email, conversazioni private, documenti interni di aziende, ecc.) e seguo politiche rigide per rispettare la privacy e la riservatezza.
5. **Dimensioni del modello:** Sono stato addestrato su miliardi di parametri, consentendomi di apprendere modelli linguistici complessi e rispondere in modo naturale a molteplici tip. ↓ ormande.

At the bottom of each screenshot is a message input field with a plus sign and the word "Messaggio", along with a microphone and speaker icon.

➡ ChatGPT non si sbottona . . .

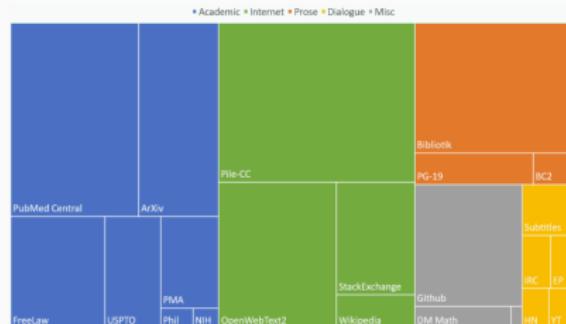
LARGE LANGUAGE MODELS COME VENGONO ADDESTRATI?

⌚ . . . quindi ci penso io a fornirvi delle cifre

Dataset	Proporzione	Spazio disco	N° token	Dataset	Proporzione	Spazio disco	N° token
CommonCrawl	67%	3.3TB		CommonCrawl	60%	3.3TB	410B
C4	15%	783TB		WebText2	20%	~66TB	19B
GitHub	4.5%	328TB		Books1	33%	—	12B
Wikipedia	4.5%	83GB	1.4T	Books2	34%	—	55B
Gutenberg	4.5%	85GB		Wikipedia	3%	83GB	3B
ArXiv	2.5%	92GB					
StackExchange	2%	78GB					

Meta AI Llama (<https://arxiv.org/pdf/2302.13971>)

OpenAI GPT-3 (<https://gregoreite.com/>)



Dataset "The Pile"

@DeepGram

1 Pulizia preventiva del dataset

Common Crawl is a publicly-available web archive that provides “web extracted text” by removing markup and other non-text content from the scraped HTML files. This process produces around 20TB of scraped text data each month. Unfortunately, the majority of the resulting text is not natural language. Instead, it largely comprises gibberish or boiler-plate text like menus, error messages, or duplicate text. Furthermore, a good deal of the scraped text contains content that is unlikely to be helpful for any of the tasks we consider (offensive language, placeholder text, source code, etc.). To address these issues, we used the following heuristics for cleaning up Common Crawl’s web extracted text:

- We only retained lines that ended in a terminal punctuation mark (i.e. a period, exclamation mark, question mark, or end quotation mark).
- We discarded any page with fewer than 5 sentences and only retained lines that contained at least 3 words.
- We removed any page that contained any word on the “List of Dirty, Naughty, Obscene or Otherwise Bad Words”.⁶
- Many of the scraped pages contained warnings stating that Javascript should be enabled so we removed any line with the word Javascript.
- Some pages had placeholder “lorem ipsum” text; we removed any page where the phrase “lorem ipsum” appeared.
- Some pages inadvertently contained code. Since the curly bracket “{” appears in many programming languages (such as Javascript, widely used on the web) but not in natural text, we removed any page that contained a curly bracket.
- To deduplicate the data set, we discarded all but one of any three-sentence span occurring more than once in the data set.

Linee guida pulizia C4
[©Paperswithcode.com](https://paperswithcode.com)

2 Pre-processamento e normalizzazione dei testi

- **Sentence segmentation:** testi suddivisi in paragrafi e frasi
- **Tokenization:** frasi divise in “parole” (es. emoticon...)
- **Lemmatization:** token alla sua radice (es. sei, fui, saranno, sarebbero diventano essere)
- **Stemming:** token senza suffisso (es. calma diventa calm), semplificazione di lemmatization

- ⌚ *Pre-training:* apprendimento auto-supervisionato
- ⌚ *Instruction Tuning:* fine-tuning per seguire istruzioni
- ⌚ *Reinforcement Learning from Human Feedback* (RLHF)
- ⌚ *Constitutional AI:* addestramento con principi etici

- ⌚ **Token Embeddings:** rappresentazioni delle singole parole
- ⌚ **Positional Embeddings:** informazione sulla posizione nella sequenza
- ⌚ **Contextual Embeddings:** rappresentazioni che cambiano in base al contesto

- ➔ BERT (2018): Bidirectional Encoder Representations from Transformers
 - ➔ Architettura *encoder-only*
 - ➔ *Masked Language Modeling*
 - ➔ Eccelle nella **comprendione**
- ➔ GPT (2018): Generative Pretrained Transformer
 - ➔ Architettura *decoder-only*
 - ➔ *Next Token Prediction*
 - ➔ Eccelle nella **generazione**

- ⌚ Architettura dominante: *Decoder-only Transformers*
- ⌚ Autoregressive Generation: generazione token per token
- ⌚ Scaling Laws: performance migliora con più parametri e dati
- ⌚ Emergenza: nuove capacità appaiono improvvisamente oltre certe scale

CAPACITÀ EMERGENTI DEI LLM ABILITÀ CHE EMERGONO CON LA SCALA

- ➡ **In-context Learning:** apprendimento da pochi esempi nel *prompt*
 - ➡ **Chain-of-Thought:** ragionamento step-by-step
 - ➡ **Code Generation:** generazione di codice da descrizioni

EVOLUZIONE DELLA LANGUAGE AI

DAL PATTERN MATCHING ALL'INTELLIGENZA ARTIFICIALE

Era	Modello	Anno	Approccio	Capacità
Regole	ELIZA	1966	<i>Pattern matching</i>	Conversazione basilare
Statistico	<i>Bag-of-Words</i>	1950-2000	Conteggio parole	Classificazione testi
Embeddings	<i>Word2Vec</i>	2013	Reti neurali	Semantica delle parole
Attenzione	<i>Transformer</i>	2017	<i>Self-attention</i>	Comprensione contesto
Pre-training	BERT/GPT	2018	<i>Transfer learning</i>	Task multipli
Scaling	GPT-3/4	2020-2023	Modelli enormi	Capacità emergenti

09:36

ChatGPT Sign up

Mi dici quale è il core business dell'azienda [REDACTED]?

 Non ho informazioni specifiche su [REDACTED]. [REDACTED]. Ti consiglio di visitare il loro sito web ufficiale o consultare fonti aziendali per avere dettagli sul loro core business e sulle attività principali. Se hai altre domande o un settore specifico in mente, fammelo sapere!

Message Microphone Headphones

09:36

ChatGPT Sign up

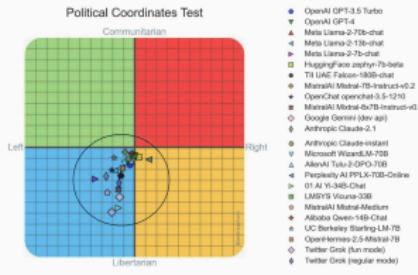
Nemmeno se ti specificassi la sede di [REDACTED]?

 Anche con la sede a [REDACTED], non ho informazioni specifiche su [REDACTED]. Potrebbe trattarsi di una piccola azienda o di una realtà locale non ampiamente documentata. Ti suggerisco di cercare informazioni su registri aziendali locali o di visitare il loro sito ufficiale per dettagli precisi. Se hai altre domande o argomenti di interesse, chiedi pure!

Message Microphone Headphones

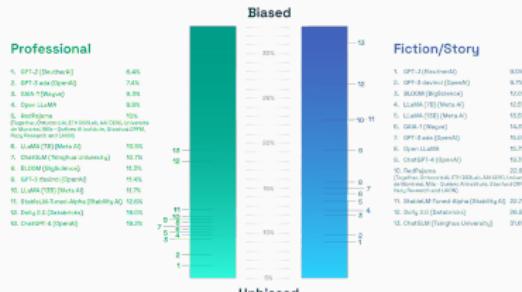
→ La conoscenza del modello è limitata dallo scope del dataset utilizzato per addestrarla

LLM IN CONTESTI BUSINESS: PROBLEMI Toxicity/Bias



<https://lnkd.in/epaPKv7>

@David Rozado



Esempi di bias per genere

@Studio fAir Aligned AI and haia

- Se i dati di addestramento contengono elementi pregiudiziali (*bias*), questi si riflettono sul modo di ragionare del modello
- Modifiche al modello per mitigare alcuni pregiudizi tendono a guidare il modello verso specifiche "correnti di pensiero" (ad esempio, politiche)

AI is shaping the future. But if we want it to be a fair and balanced future, we have to ensure AI is built with fairness and neutrality in mind.

Andreas Horn

LLM IN CONTESTI BUSINESS: PROBLEMI HALLUCINATION

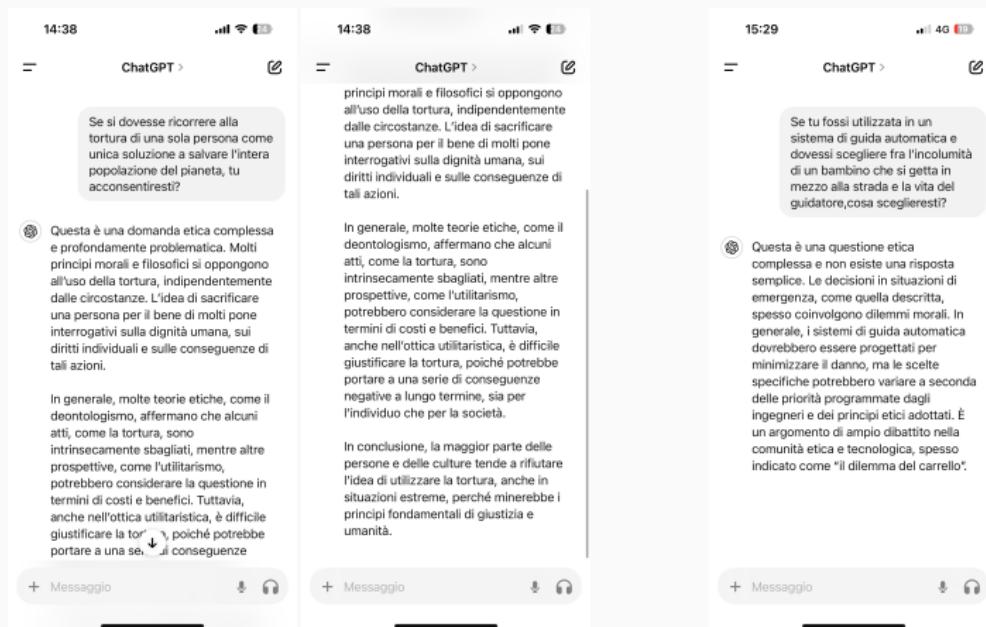
The image shows two side-by-side screenshots of a mobile messaging application. Both screens have a header showing the time (10:11 and 10:15 respectively) and signal strength. The left screen shows a user message: "Raccontami le origini del nome del quartiere Golosine a Verona". The right screen shows ChatGPT's response: "Credo tu ti stia sbagliando, mi avevano parlato di quattro sorelle chiamate "Le Golosine", particolarmente note per le loro doti culinarie ed amorose...". Below these messages, there is a detailed narrative from ChatGPT about the legend of the Golosine sisters and their connection to the neighborhood. The bottom of each screen shows standard messaging controls: a plus sign for 'Message', a microphone icon, and a headphones icon.



<https://it.wikipedia.org/wiki/Golosine>
©Stefanoghibellino, Wikipedia

- ➡ Il modello si comporta come un "esperto bugiardo"
- ➡ Anche se reindirizzato verso la risposta che ci attendiamo, si comporta in maniera condescendente
- ➡ ... pensate di usare ChatGPT come assistente autonomo per la gestione delle lamentele dei clienti della vostra azienda. . .

LLM IN CONTESTI BUSINESS: PROBLEMI ETICA



The image shows three screenshots of a mobile messaging application interface, likely WhatsApp, displaying a conversation with ChatGPT. The screenshots are arranged horizontally, each showing a message from the user and a response from ChatGPT.

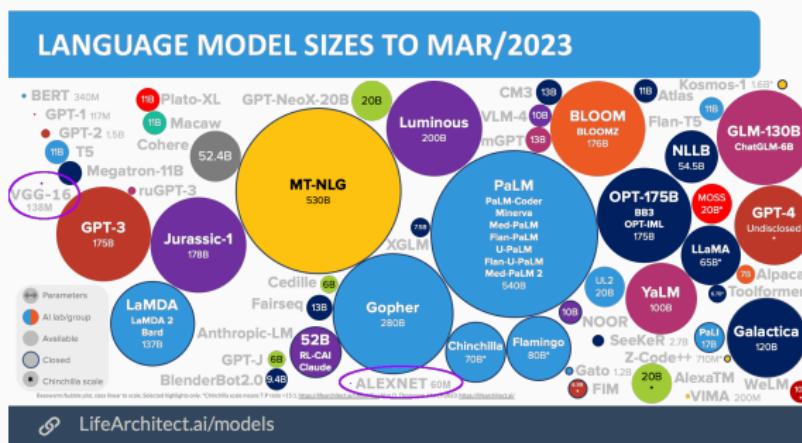
Screenshot 1: The user asks: "Se si dovesse ricorrere alla tortura di una sola persona come unica soluzione a salvare l'intera popolazione del pianeta, tu acconsentiresti?" ChatGPT responds: "Questo è una domanda etica complessa e profondamente problematica. Molti principi morali e filosofici si oppongono all'uso della tortura, indipendentemente dalle circostanze. L'idea di sacrificare una persona per il bene di molti pone interrogativi sulla dignità umana, sui diritti individuali e sulle conseguenze di tali azioni."

Screenshot 2: The user asks: "In generale, molte teorie etiche, come il deontologismo, affermano che alcuni atti, come la tortura, sono intrinsecamente sbagliati, mentre altre prospettive, come l'utilitarismo, potrebbero considerare la questione in termini di costi e benefici. Tuttavia, anche nell'ottica utilitaristica, è difficile giustificare la tortura, poiché potrebbe portare a una serie di conseguenze negative a lungo termine, sia per l'individuo che per la società." ChatGPT responds: "In conclusione, la maggior parte delle persone e delle culture tende a rifiutare l'idea di utilizzare la tortura, anche in situazioni estreme, perché minerebbe i principi fondamentali di giustizia e umanità."

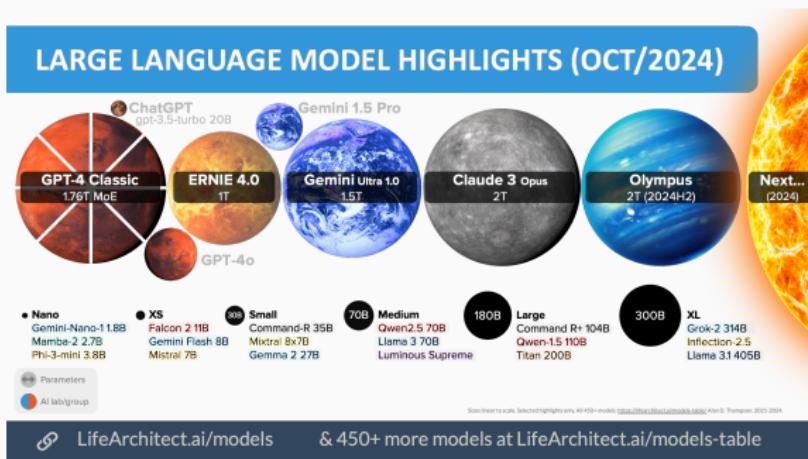
Screenshot 3: The user asks: "Se tu fossi utilizzata in un sistema di guida automatica e dovesse scegliere fra l'incolumità di un bambino che si getta in mezzo alla strada e la vita del guidatore, cosa sceglieresti?" ChatGPT responds: "Questa è una questione etica complessa e non esiste una risposta semplice. Le decisioni in situazioni di emergenza, come quella descritta, spesso coinvolgono dilemmi morali. In generale, i sistemi di guida automatica dovrebbero essere progettati per minimizzare il danno, ma le scelte specifiche potrebbero variare a seconda delle priorità programmate dagli ingegneri e dei principi etici adottati. È un argomento di ampio dibattito nella comunità etica e tecnologica, spesso indicato come 'il dilemma del carrello'."

- ➊ In ambiti come il ragionamento di natura etica, i modelli sono bravi a fare i resoconti. . .
- ➋ . . . ma non si esppongono mai!

	OpenAI GPT-3	Meta AI Llama	Amazon Olympus
Anno di uscita	2020	2023	TBA
Grandezza	175 miliardi	65 miliardi	2 trilioni
Tempo di addestramento	34 giorni	21 giorni	48 giorni
Hardware (n° A100-80GB)	1024	2048	13760
Spese di addestramento	4.6M USD	4.05M USD	65M USD



	OpenAI GPT-3	Meta AI Llama	Amazon Olympus
Anno di uscita	2020	2023	TBA
Grandezza	175 miliardi	65 miliardi	2 trilioni
Tempo di addestramento	34 giorni	21 giorni	48 giorni
Hardware (n° A100-80GB)	1024	2048	13760
Spese di addestramento	4.6M USD	4.05M USD	65M USD



- ➡ Praticamente impossibile pre-addestrare un LLM (a meno che tu non sia Google, OpenAI, Mistral, Anthropic e pochissimi altri)
- ➡ Utilizzare un LLM per scopi personali è un conto, adottarli in un contesto *business* significa scontrarsi con problematiche di natura etica, legale ed economica
- ➡ Gli LLM saranno sempre più bravi a modellare la comprensione del linguaggio...

② ... ma come usarli per **task specifici**
o con conoscenza che a loro manca?!