INTRODUZIONE
ai
LARGE LANGUAGE MODEL

Ormai tutti conoscono le potenzialità di CHATGPT4 nel conversare con un uomo, tradurre in lingue diverse un testo, elaborare relazioni, riassumere un articolo, annotare delle immagini, simulare un esperto professionista

Queste capacità sono date dall'introduzione di una nuova classe di algortimi i generative pre-trained transformer (GPT) che sono in grado di elaborare in modo efficiente testi anche lunghi, per estrarne il *significato*

I LARGE LANGUAGE MODEL (LLM) SONO SOSTANZIALMENTE RETI NEURALI ALLENATE E SPECIALIZZATE PER ELABORARE IL LINGUAGGIO NATURALE, USANDO COME TRAINING SET DELLE BASI GIGANTESCHE DI TESTI (detti CORPUS)

(c) Fusco Francesco NS Innovation Hub

OBIETTIVI DELLA PRESENTAZIONE

Far conoscere lo stato dell'arte, le tecnologie e gli algoritmi che stanno alla base dei LLM e che rendono questi strumenti emulatori delle capacità linguistiche-cognitive dell'uomo

Ma anche evidenziare i limiti ed i rischi insiti in questi strumenti, che anche se utili dal punto di vista pratico non sono intelligenti come sembra, ma piuttosto assomigliano a dei pappagalli stocastici.

(c) Fusco Francesco NS Innovation Hub

LA RETE NEURALE CHE PRESTA ATTENZIONE

La facoltà **dell'attenzione umana** è un'abilità cruciale nella comprensione del linguaggio naturale. Ogni volta che leggiamo o ascoltiamo un testo, la nostra mente seleziona le parole chiave e le mutue relazioni semantiche per ottenere una comprensione del significato complessivo.

Ricordiamo che per **significato** si intende Il contenuto espressivo di qualsiasi mezzo di comunicazione (parole o frasi, gesti, segni grafici ecc.). In particolare in linguistica, ciò che si vuol dire pronunciando una frase o una parola, il messaggio cioè che con queste si trasmette a chi riceve.

Il vocabolario è un utile strumento per apprendere il significato delle parole, perché spiega le parole con altre parole.

L'architettura dei GPT, una delle più recenti e avanzate tecniche di elaborazione del linguaggio naturale, offre una soluzione matematica convincente .

(c) Fusco Francesco NS Innovation Hub

I vettori di embedding

- E’ difficile usare le parole come input in una rete neurale, per poterle elaborare queste devono essere rappresentate in forma numerica (token) numerosi come gli entri del vocabolario
- Un vettore di embedding rappresenta una parola in un computer. Questo vettore ha una dimensione fissata per ogni parola
- I vettori di embedding vengono creati attraverso una fase di apprendimento dai vettori di one-hot contecniche di auto-encoder, così che le parole semanticamente simili vengano posizionate in punti vicini di u n iperspazio.
- Dimesioni tipiche dei vettori di embedding sono 512 o 1024.

(c) Fusco Francesco NS Innovation Hub

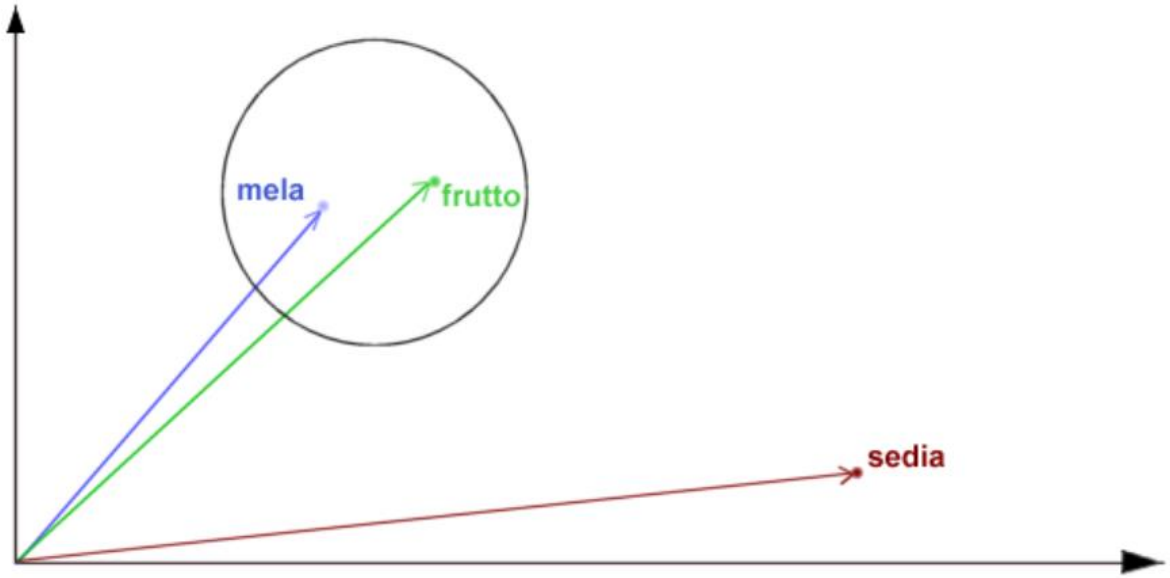
Il clustering è lo strumento per apprendere conoscenza

Le tecniche di Word Embeddings si basano sulla creazione di rappresentazioni vettoriali dense di parole da un corpus di testo. Queste rappresentazioni sono utilizzate per identificare la somiglianza semantica tra le parole e per compiere operazioni matematiche con le parole vettorializzate.

L'obiettivo delle tecniche di Word Embeddings è di creare vettori di parole in cui parole simili sono rappresentate da vettori simili. Questo significa che le parole che hanno significati simili o che compaiono spesso insieme nel contesto saranno rappresentate da vettori che sono molto vicini tra loro nello spazio vettoriale.

(c) Fusco Francesco NS Innovation Hub

La distanza tra i vettori embedding



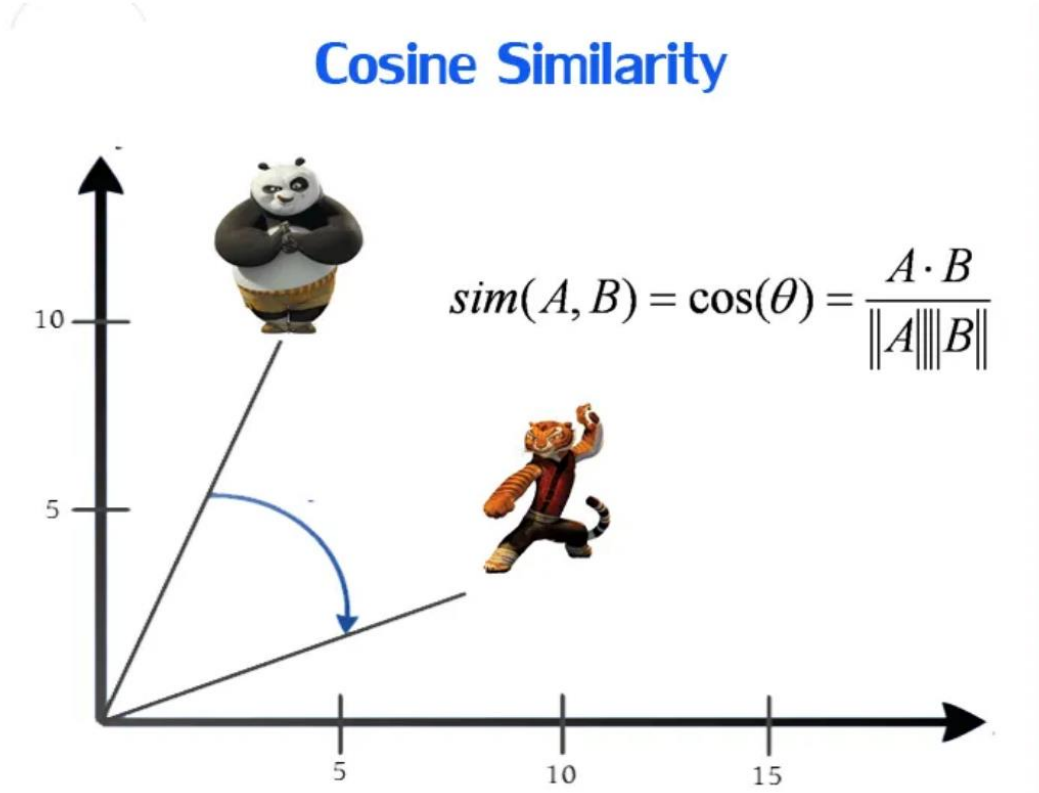
Nello schema a destra la rappresentazione vettoriale di **mela e frutto** è molto più simile (distanza euclidea minore) rispetto a quella con **sedia**.

il pre-training basato sulla co-occorrenza utilizza tecniche di conteggio delle parole per calcolare la probabilità della co-occorrenza di due parole in un corpus di testo. In questo modo, vengono appresi i vettori di embedding per ogni parola del corpus in base alla loro prossimità semantica.

I vettori di embedding vengono utilizzati per calcolare i pesi di attenzione attraverso il cosidetto "query-key-value attention mechanism" da assegnare ad ogni parola dell'input.
Questo peso o grado di attenzione che viene mostrata ad ogni parola dell'input è calcolato in riferimento ad un'altra parola (query) che il modello individua in base ai suoi obiettivi.

(c) Fusco Francesco NS Innovation Hub

Il prodotto scalare di due vettori



Se rappresentiamo i vettori di embedding in uno spazio vettoriale (nell'esempio bidimensionale) la similitudine di due vettori parole è dato dall'avere l'angolo θ compreso quasi nullo, ovvero il coseno prossimo a 1. Quindi è sufficiente calcolare il prodotto scalare normalizzato tra due vettori di embedding per sapere quanto sono simili i token che rappresentano

Facciamo un esempio

Il primo passo è quello di calcolare il **prodotto scalare** tra i vettori di query e chiavi, ottenendo così i punteggi di attenzione per ogni token dell'input rispetto alla query:
“`

attenzione(“La”) = embedding(“La”) x embedding(“frutto”)
attenzione(“mela”) = embedding(“mela”) x embedding(“frutto”)
attenzione(“è”) = embedding(“è”) x embedding(“frutto”)
attenzione(“un”) = embedding(“un”) x embedding(“frutto”)
attenzione(“frutto”) = embedding(“frutto”) x embedding(“frutto”)
attenzione(“molto”) = embedding(“molto”) x embedding(“frutto”)
attenzione(“sano”) = embedding(“sano”) x embedding(“frutto”)

attenzione(“La”) = 0.14
attenzione(“mela”) = 0.26
attenzione(“è”) = 0.1
attenzione(“un”) = 0.12
attenzione(“frutto”) = 0.3
attenzione(“molto”) = 0.06
attenzione(“sano”) = 0.12

pesi_attenzione(“La”) = 0.013
pesi_attenzione(“mela”) = **0.147**
pesi_attenzione(“è”) = 0.104
pesi_attenzione(“un”) = 0.116
pesi_attenzione(“frutto”) = **0.174**
pesi_attenzione(“molto”) = 0.089
pesi_attenzione(“sano”) = **0.156**



LA FUNZIONE MATEMATICA DI ATTENZIONE

Il meccanismo che ha costituito la rivoluzione dei GPT, come classe specifica degli LLM, è l'attenzione che permette di focalizzarsi su differenti parti della sequenza di input mentre genera l'output
Un modello di linguaggio dotato di attenzione genera frasi che hanno senso per un umano, perché ogni prossima parola tiene conto del contesto

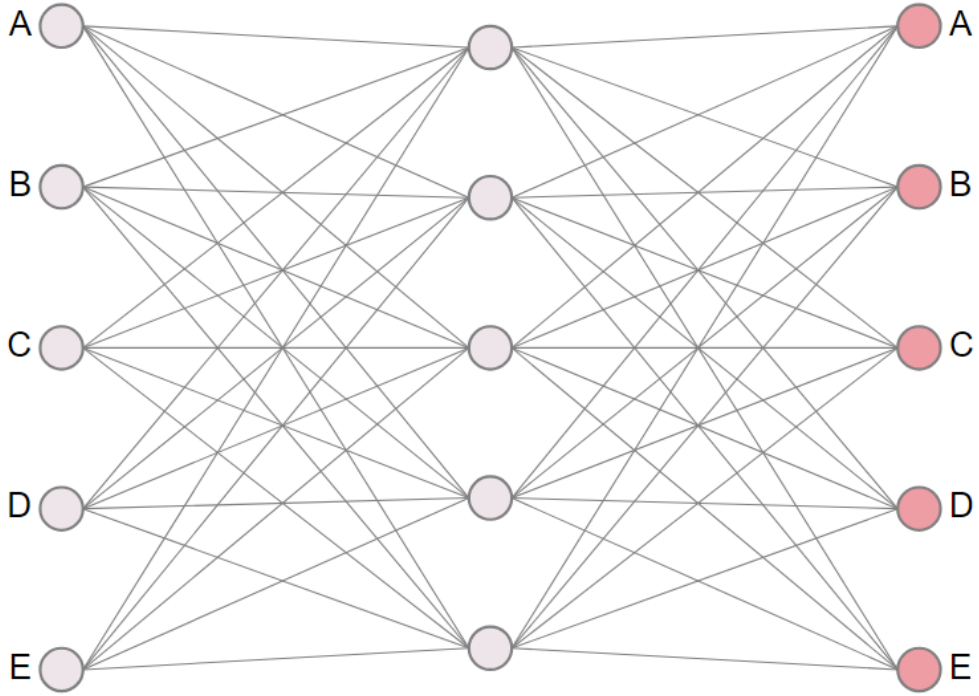
In termini matematici l'attenzione è una funzione complessa che utilizza le matrici che abbiamo generato nella fase di encoding

$$\mathbf{ATTENTION\ (Q,K,V) = softmax(QK^T/radq(d\text{-}k))*V}$$

Ma forse è meglio vederlo all'opera con un esempio di generazione di testo

Vediamo come funziona il word embedding

Un modello neurale semplice a 3 livelli di apprendimento del word embedding è quello dell'autoapprendimento, ovvero quando un vettore rappresentativo di una parola deve imparare se stesso Supponiamo di avere cinque parole a, b ,c, d, e e vogliamo che la rete neurale sotto raffigurata impari tali parole, ovvero a|a, b|b, c|c, d|d, e|e



Ad ogni parola in input assegniamo un vettore casuale, dopo di che lasciamo che la rete supervisionata apprenda, cambiando i pesi, in modo da ridurre l'errore (la distanza) tra una parola di input e la corrispondente parola di output

Un esempio interattivo è questo
https://ronxin.github.io/wevi/

La potenza degli LLM

Grazie al fatto che non sono cognitivi ma processano pattern di dati molto bene gli LLM hanno applicazioni anche in altri campi
Per esempio, osservando GPT-3 i ricercatori dei OpenAI che lo hanno creato, si sono resi conto che faceva più di quanto si aspettassero, per esempio riuscendo a scrivere porzioni di codice e, in qualche modo, imparando dalle proprie attività. **Niente di stupefacente né di senziente**.

C'è quindi un rapporto stretto tra LLM e contenuti di diverso tipo: testo, immagini, audio, video, ma non solo. **Dall-e 2** è un modello LLM e viene usato per scoprire nuove proteine, ChatGPT è un modello LLM e viene usato anche per scrivere software.

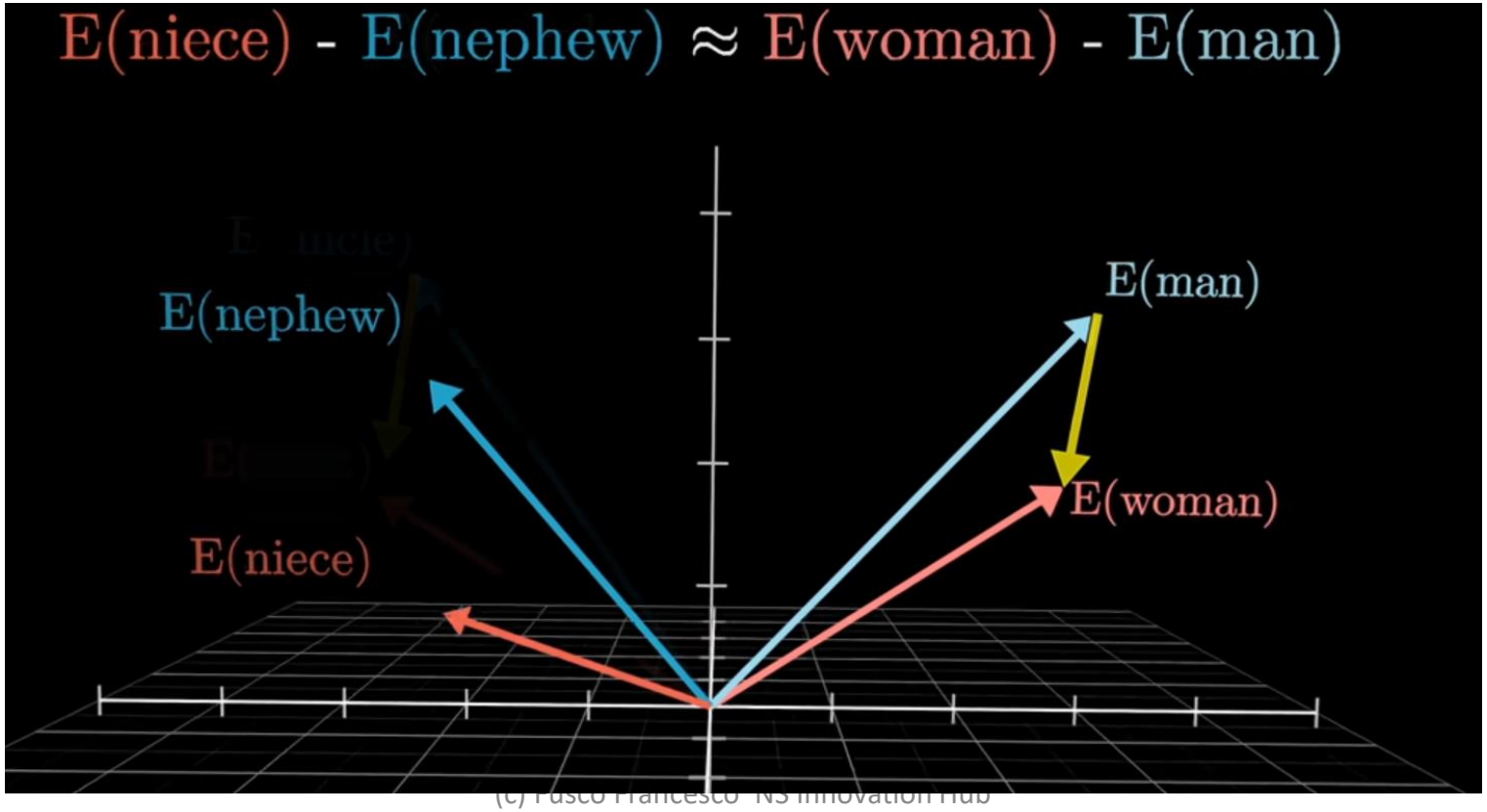
QUESTA POTENZA INASPETTATA E' LA VERA CHIAVE CHE RENDE GLI LLM UN VALIDO STRUMENTO DI IA NEI VARI CAMPI DI APPLICAZIONE TRA CUI LA MEDICINA

Alcuni problemi degli LLM

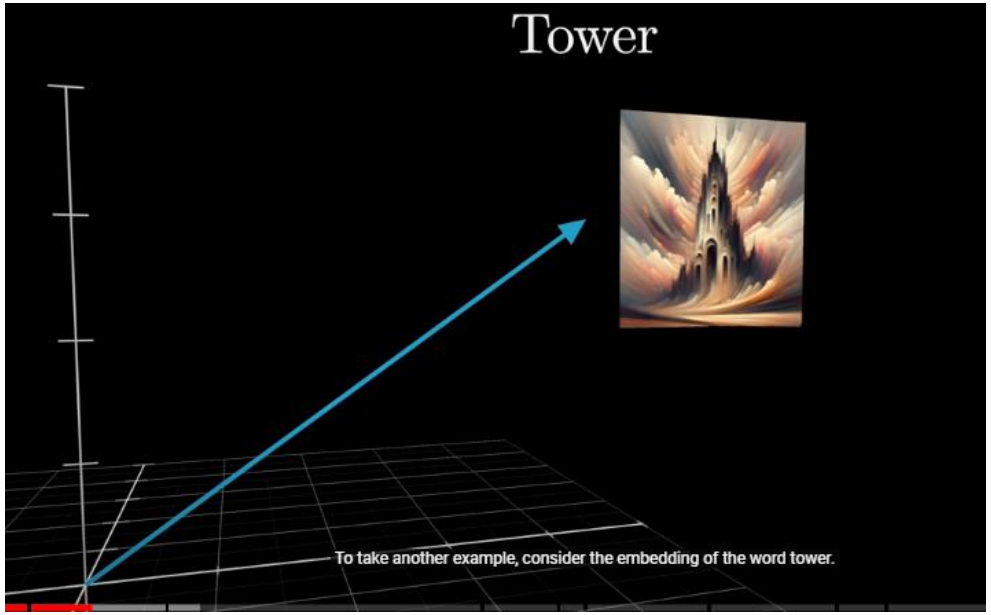
- Possono soffrire di allucinazioni
- Possono imparare da informazioni false
- Hanno una accuratezza che cambia in funzione del campo di applicazione e del compito assegnato
- Sono molto costosi da mantenere
- Non hanno una propria etica
- Non si evolvono da soli

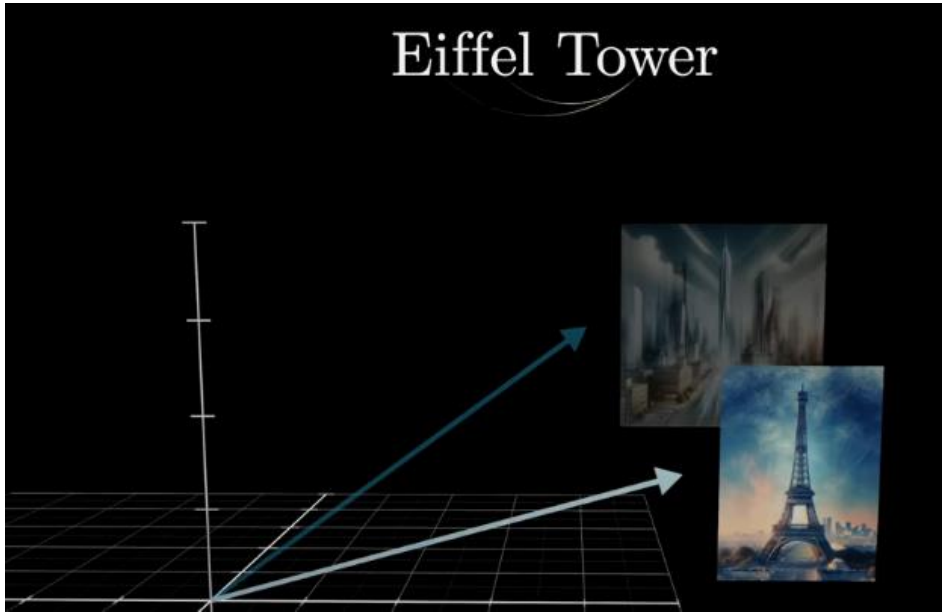
RECAP SULL'ATTENZIONE

- Iperspazio della codifica delle parole preserva il significato e le relazioni semantiche



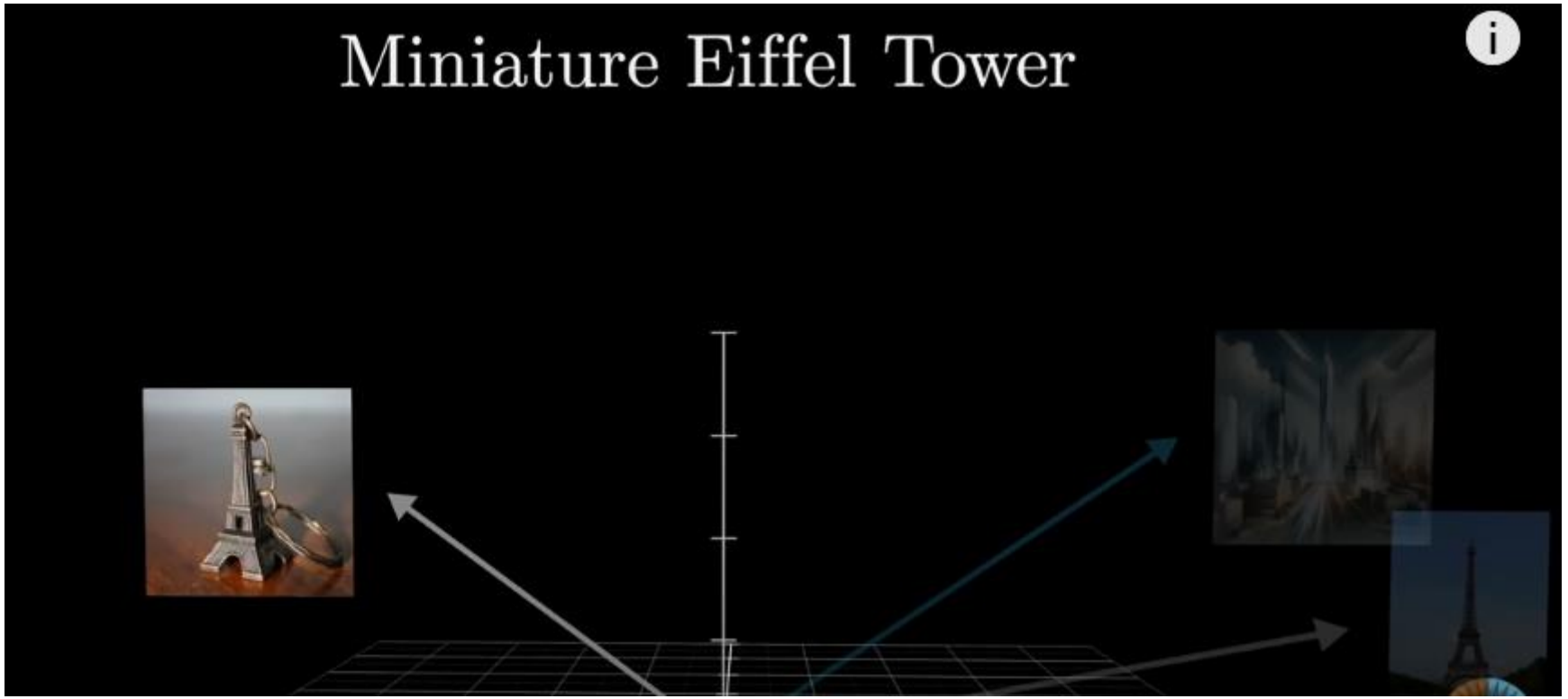
RECAP SULL'ATTENZIONE





L'embedding di una parola viene modificata leggermente in funzione del contesto (parola che precede)

ATTENZIONE SINGOLA



L’embedding di una parola viene modificata anche pesantemente in funzione del contesto (più parole che precedono)

Le barriere da superare verso un AGI

- **Mancanza di apprendimento rapido da memoria episodica**
- **Dimenticanza catastrofica per nuovi apprendimenti**
- **Assenza di curiosità per apprendimento continuo**
- **Mancanza di intenzionalità, data dal bisogno**
- **Assenza di immaginazione per auto-simulazioni tra alternative**
- **Assenza di un modello interno per analisi controfattuale**
- ***Assenza di mentalizzazione***

COME MIGLIORARE LE PERFORMANCE DEGLI LLM

Apparentemente strumenti come GPT4 Vision offrono un'accuratezza paragonabile a quella dei medici in alcuni test a scelta multipla (81,6% vs 77,8%), ma quando gli si chiede il motivo della propria scelta corretta, questi strumenti sbagliano nel 60% dei casi, dimostrando una mancanza di reale comprensione del fenomeno in esame.

Questo intervento vuole sottolineare come migliorare le prestazioni per superare alcuni limiti degli LLM, come l'approccio essenzialmente statistico, il tasso variabile di accuratezza nelle risposte, la possibilità di allucinazioni, l'assenza di senso comune e la carenza di conoscenza non linguistica del mondo esterno.

I due approcci

Due approcci operativi migliorano sufficientemente le prestazioni:

1. ***PROMPT TUNING***

2. ***REINFORCED LEARNING BY HUMAN FACTOR***

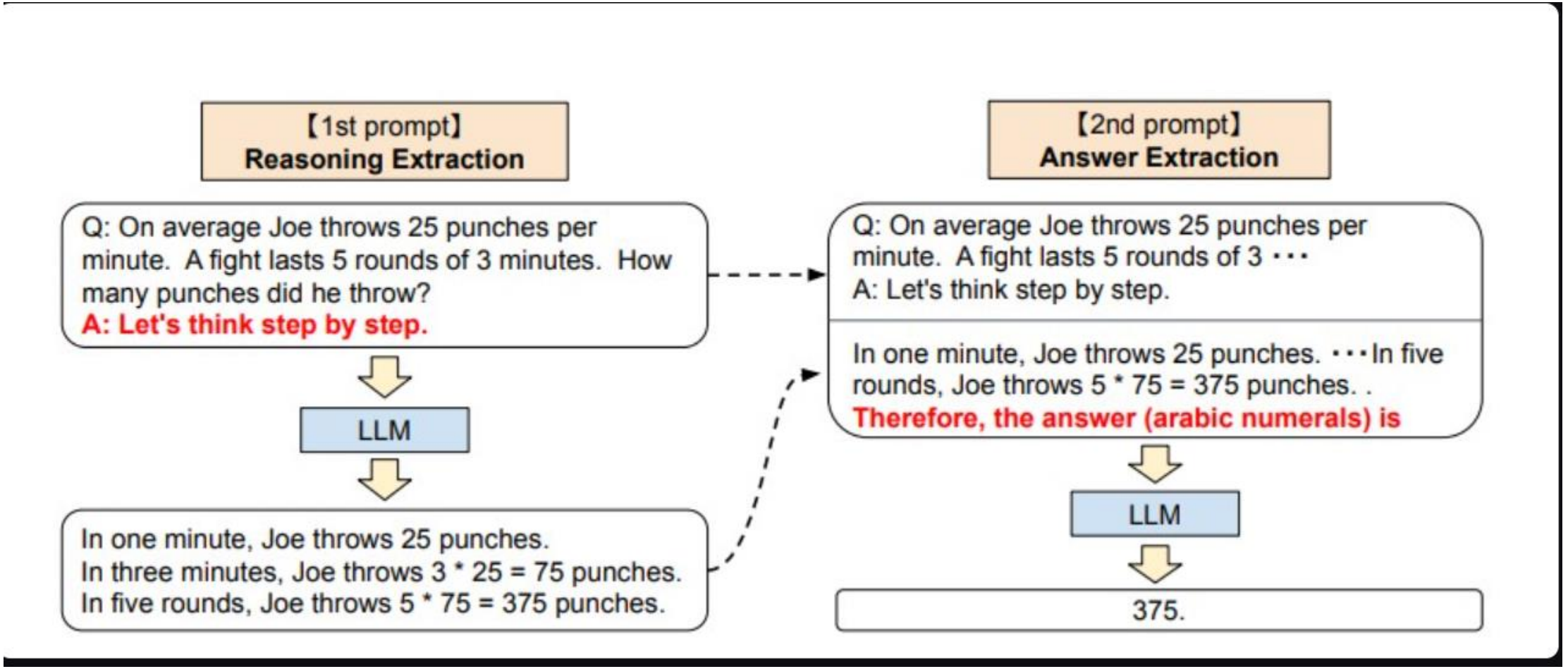
PROMPT ENGINEERING

- ***Il prompt engineering è la pratica di scrivere prompt ben strutturati che possono essere meglio interpretati da un modello di IA generativa. Un prompt strutturato indica al LLM quale attività eseguire e che tipo di output generare.***
- ***Può contenere istruzioni, contesto, dati di input e indicatori di output. Utilizzando il prompt engineering, possiamo utilizzare gli LLM per svolgere varie attività, dalla semplice risposta a domande alla complessa generazione di testi creativi.***
- ***Si basa su una proprietà emergente, l'apprendimento nel contesto, che consente agli LLM di apprendere da suggerimenti dati nel prompt, lasciando al meccanismo dell'attenzione di generare output migliori.***

CHAIN OF THOUGHT

- ***Il CoT è una tecnica attraverso la quale costringiamo gli LLM a produrre una sequenza di passaggi intermedi che portano alla risposta desiderata. Migliora le capacità di ragionamento degli LLM. È vantaggioso perché consente al modello di concentrarsi sulla risoluzione di un passaggio alla volta, anziché dover considerare l'intero problema tutto in una volta.***
- ***Può essere particolarmente utile per problemi complessi che sarebbero difficili o impossibili da risolvere in un unico passaggio.***
- ***Fornisce una finestra interpretabile sul comportamento del modello.***
- ***Possiamo vedere come il modello è arrivato alla sua risposta seguendo la sequenza di passaggi necessari e inglobare questo output intermedio nella query successiva.***

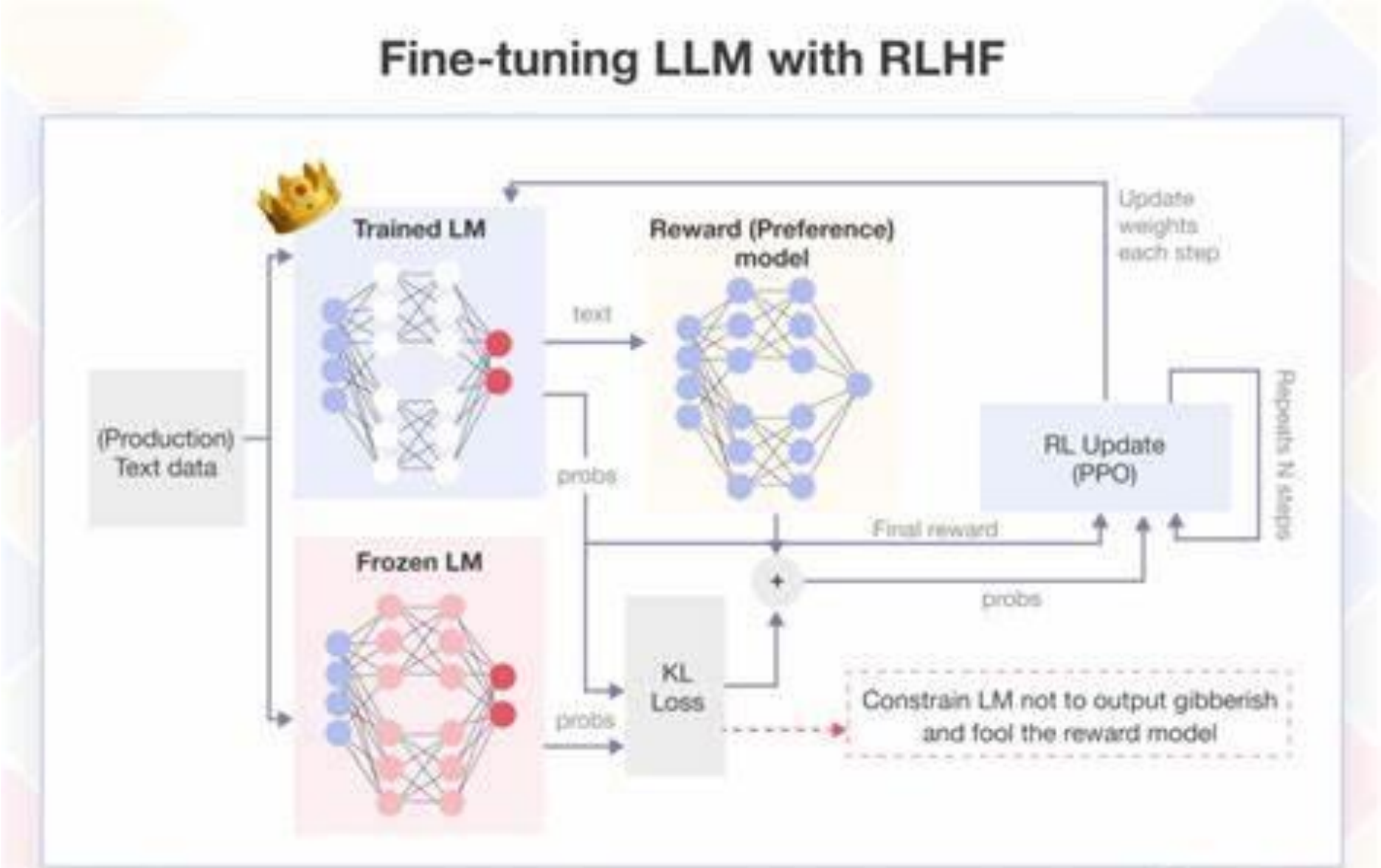
ESEMPIO DI CHAIN OF THOUGHT



Tree-of-Thought (ToT) prompting

- ***Nei problemi complessi come quelli medicali dove occorre considerare una varietà di parametri, estratti attraversi diversi esami diagnostici il cui risultato influenza l'ipotesi iniziale e costringe i medici a valutare ipotesi alternative uno strumento che si sta rilevando fondamentale è il ToT Prompting***
- ***La tecnica ToT si ispira all'approccio della mente umana per risolvere compiti complessi attraverso tentativi ed errori. In questo processo, la mente esplora lo spazio delle soluzioni attraverso un processo di pensiero ad albero, consentendo di tornare indietro quando necessario.***
- ***Per implementare il ToT, aumentiamo un LLM con moduli aggiuntivi tra cui un agente suggeritore, un modulo di controllo, un modulo di memoria e un controller ToT.***

Reinforced learning by human factor



GRAZIE