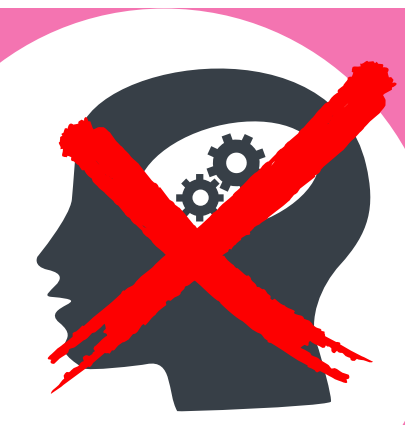
CAPIRE LA AI

Come funziona una LLM ?

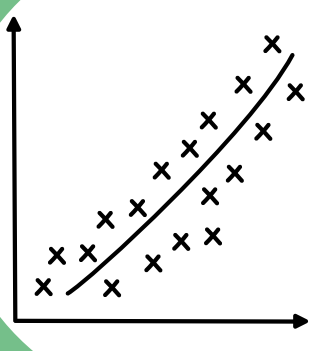


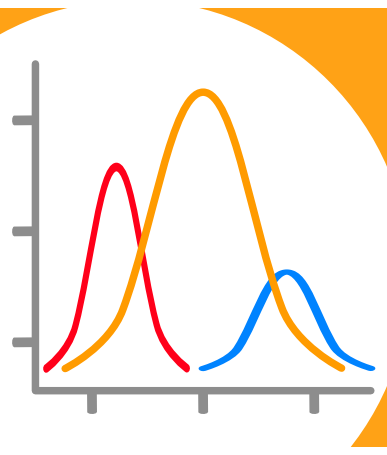
Un LLM non pensa

è un sistema statistico addestrato su enormi quantità di testo per modellare le regolarità del linguaggio, senza accesso diretto al mondo reale. Ogni parola viene codificata come un vettore (“token”) in uno spazio con centinaia di dimensioni; la vicinanza tra vettori riflette la probabilità di apparire in contesti simili. Non c’è semantica innata: è pura mappa statistica...

Ma fa “correlazione ...”

Due parole sono “amiche” se nei dati compaiono insieme più spesso di quanto accadrebbe per puro caso. Non serve sapere cosa significhino: il modello rileva che “pizza” e “mozzarella” si presentano insieme molto più di “pizza” e “batteria dell’auto” e registra quella regolarità.



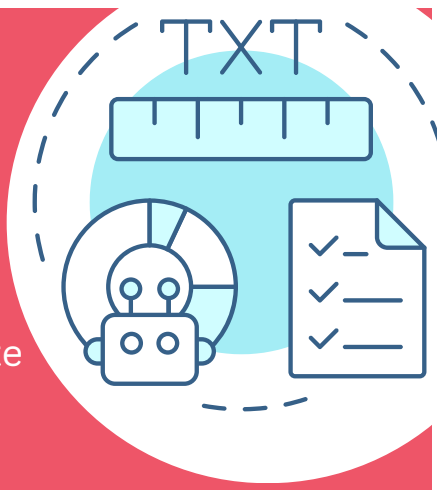


... attua un processo “stocastico” ...

Quando scrive, un LLM non applica logica simbolica o ragionamento causale: genera sequenze di parole campionando dalla distribuzione di probabilità appresa per il contesto dato.
Se il testo è “Il gatto sta...”, la distribuzione assegnerà alta probabilità a “dormendo” e bassa a “pilotando un aereo”.

e poi lo “ottimizza”...

L’abilità dell’LLM non emerge per magia, ma da un processo di minimizzazione di una funzione di perdita (tipicamente la cross-entropy) tra le previsioni del modello e i dati reali. Attraverso il “gradient descent”, miliardi di parametri vengono regolati per ridurre sistematicamente l’errore di previsione sul prossimo “token”. Dopo trilioni di iterazioni, l’output diventa statisticamente indistinguibile dal testo umano. Questo non garantisce verità né comprensione, ma coerenza statistica.



utilizza un “trasformer”...

Il suo cuore è il “self-attention”, un meccanismo che, dato un testo, valuta quanto ogni parola sia rilevante rispetto a tutte le altre del contesto, non solo a quelle vicine. Invece di leggere il testo parola per parola (come facevano le vecchie reti neurali sequenziali), il Transformer considera l’intera sequenza in parallelo, calcolando in un colpo solo relazioni a breve e a lungo raggio... ma resta un simulatore di linguaggio..

... però produce “allucinazioni”...

Il modello può produrre frasi false ma plausibili perché non confronta le sue uscite con lo stato reale del mondo. L’accuratezza è un effetto sistematico, non un vincolo progettuale: sono la conseguenza inevitabile di un sistema che ottimizza per plausibilità linguistica, non per veridicità fattuale.

e dipende dallo “scaling”...

Più parametri, più dati e più calcolo tendono a produrre modelli più capaci. Questo è il principio delle “scaling laws”: allargando la rete e nutrendola di più linguaggio, la mappa statistica diventa più dettagliata. Ma più grande non significa “più intelligente”: significa solo che il completatore di frasi ha un vocabolario statistico più ricco e preciso — e quindi riesce a sembrare ancora più credibile anche quando si inventa tutto.

Dal “Fb” di Walter Quattrociocchi - Università della Sapienza di Roma