UNA SERIE SULL'INTELLIGENZA ARTIFICIALE

Io, Gemini e altri quattro

Parte III

IA generativa speigata in dettaglio, un viagio nel cuore della tecnologia IA

ROCCO MASSIMILIANO MONDO

ABSTRACT

In questo terzo capitolo della serie "Io, Gemini e altri quattro", faccio un salto qualitativo deciso: mi lascio alle spalle il piano della divulgazione generale per addentrarmi nel funzionamento tecnico dell'Intelligenza Artificiale generativa, mantenendo però un registro accessibile a chiunque voglia capire davvero — e non solo usare — questi strumenti.

Il punto di partenza è il processo di inferenza, illustrato attraverso uno schema ciclico che mette al centro la sinergia tra uomo e macchina. L'IA, si chiarisce subito, non lavora in isolamento: è l'utente a definire il problema, a valutare la risposta e a guidarne il progressivo affinamento. Senza questa interazione, il modello non sa né cosa elaborare né come migliorarsi.

*Il cuore dell'articolo è dedicato ai **Large Language Model (LLM)**, descritti come motori di inferenza statistica. Diversamente da un database tradizionale — che memorizza dati espliciti e li recupera tramite query — un LLM distilla, durante la fase di addestramento, miliardi di relazioni matematiche chiamate **pesi**. Non "sa" le cose: le ricostruisce ogni volta, parola dopo parola, calcolando quale **token** sia statisticamente più probabile nel contesto dato.*

1

[Nota]

*(Un **token**, nel contesto del Machine Learning e del Natural Language Processing (NPL) è una unità fondamentale di testo che un modello elabora durante il processo di **Tokenizzazione** in cui si scompongono le frasi intere in parti più piccole o singole parole, comunque unità molto piccole chiamate "**token**")*

L'analogia proposta è quella del lettore umano: nessuno ricorda ogni singola parola di un libro, ma ne comprende il senso e sa ricostruire il ragionamento.

*Si esplora quindi **l'architettura Transformer** e il meccanismo di **Self-Attention**, la vera svolta tecnologica che ha reso gli LLM superiori ai modelli precedenti: anziché leggere una frase in sequenza lineare, il **Transformer** analizza contemporaneamente tutte le parole, cogliendo relazioni di contesto anche a grande distanza — come nel caso classico della parola "pesca", il cui significato cambia radicalmente a seconda delle parole che la circondano.*

*Non mancano le ombre. L'articolo affronta con onestà due limiti strutturali: **le allucinazioni** — ovvero la tendenza del modello a generare contenuti falsi ma plausibili, perché statisticamente "corretti" nel formato — e il **rischio** legato alla*

***sicurezza dei dati**, con l'invito a non condividere informazioni riservate con sistemi LLM pubblici.*

A chiudere, una tavola comparativa tra IA tradizionale e IA generativa sintetizza la distanza concettuale tra i due paradigmi: da macchine che seguono regole rigide a sistemi che manipolano significati. La conclusione è netta: "il limite non è più la potenza di calcolo, ma la qualità dei dati e la capacità umana di porre le domande giuste."

Un articolo denso, visivamente supportato da infografie sull'anatomia degli LLM, che apre la strada ai prossimi approfondimenti sulle applicazioni pratiche — anche in ambito privato e professionale.

PARTE III

IA generativa spiegata in dettaglio: un viaggio nel cuore della tecnologia IA

UNA PREMESSA IMPORTANTE

Nei primi due articoli di questa serie, è stato fatto un lavoro di informazione generale sulle potenzialità della IA e sulla IA generativa in particolare. Abbiamo visto l'impatto che questa nuova tecnologia sta avendo e avrà sulla società, enfatizzando gli aspetti positivi più di quelli negativi, perché credo fermante che stiamo vivendo un salto tecnologico che sarà ricordato negli annali del genere umano. Certo, non siamo ancora arrivati al punto di singolarità, ossia il punto in cui questa tecnologia cambierà per sempre il volto di questa umanità, ma abbiamo intrapreso la via, e sembrerebbe che nessuno dei grandi "Big Tech" voglia più tornare indietro. Quindi il punto è quanta voglia abbiamo di tenerci aggiornati e far parte di questa rivoluzione? Personalmente ne voglio far parte ed intendo documentarmi e attivarmi fattivamente per sfruttare questa tecnologia nel campo dell'ingegneria elettrica.

Con questo nuovo articolo entreremo nel dettaglio tecnico su come funziona realmente l'IA per arrivare più avanti a realizzare un progetto di IA personale, specificamente progettata per applicazioni ingegneristiche e di fisica quantistica applicata.

Da qui in avanti gli articoli saranno più tecnici e un po' più lunghi ma saranno sempre mirati a spiegare in modo semplice e diretto i concetti di base di questa tecnologia e di come applicarli nella vita quotidiana.

[*Nota*]

Questo articolo è realizzato con l'ottica di trasmettere a più persone possibili il maggior numero di concetti sulla IA, mi scuso, anticipatamente, con chi essendo uno specialista del settore, potrebbe trovare troppo esemplificativi gli argomenti trattati.

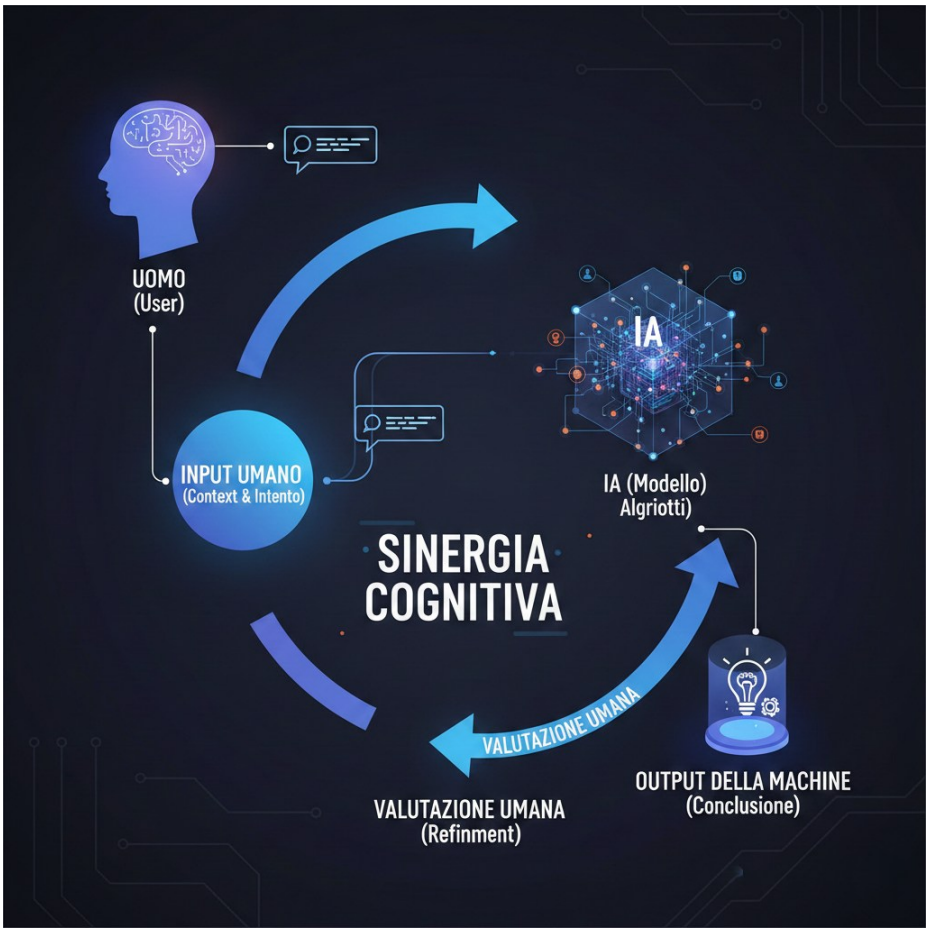
Rocco M. Mondo, Marzo 2026

IL PROCESSO DI INFERENZA

Il flusso di inferenza

Ho impegnato parecchio tempo per pensare alla struttura di questo articolo, cercando un modo semplice ma efficace di poter esporre gli argomenti, anche quelli più complessi, in una forma chiara e comprensibile e dopo tanto tergiversare su come fare ho realizzato che il modo più semplice per affrontare l'argomento e quello di iniziare con un semplice schema che rappresenti l'interazione tra l'uomo e l'IA. Nello schema sotto possiamo distinguere ***attori*** e ***azioni*** del processo, evidenziando anche i vari ruoli che essi hanno nel processo stesso.

Processo di inferenza



La struttura del flusso

1. ***Input Umano (Input):*** L'utente pone un problema o una domanda, portando con sé il bagaglio di esperienza e senso comune (che non appartiene intrinsecamente alla IA).
2. ***Processo di Inferenza (Processing):*** L'IA analizza i pattern nei dati, calcola le probabilità e applica le regole logiche per "estrarre" una conclusione dalle informazioni ricevute.
3. ***Output della Macchina (Output):*** La risposta generata è una sintesi o una previsione basata sulle premesse fornite.
4. ***Valutazione Umana (Refinement):*** L'uomo interpreta il risultato, lo valida o lo corregge, creando un circolo virtuoso di apprendimento e precisione.

Vediamo di commentare il processo di inferenza per poi approfondire alcuni aspetti specifici.

Il grafico mostra subito che il processo di inferenza è rappresentato da un ciclo di lavoro circolare, che si interrompe solo quando l'utente si ritiene soddisfatto della risposta.

Nel caso in cui dato un input, non si ritiene adeguato l'output, il processo continua per affinamento successivo attraverso l'interazione con l'utente. Quindi è evidente che l'utente (l'input umano) ha un ruolo fondamentale nel processo, senza il quale l'IA non sa cosa elaborare e come migliorare la propria risposta.

L'uomo immette il quesito è valuta la risposta, se questa è ritenuta non adeguata va affinata progressivamente, aggiungendo dettagli nuovi al contesto dell'input, ecco che la macchina impara da questa interazione a centrare meglio la successiva risposta, fin quando l'utente non è soddisfatto. Possiamo chiamare questa iterazione processo di ***"sinergia cognitiva".***

IL CUORE DEL PROCESSO: IL LARGE LANGUAGE MODEL (LLM)

Il cuore del processo di inferenza vera e propria risiede nella ***LLM*** ossia nel ***"Large Language Model"***, una sorta di database multidimensionale opportunamente strutturato. Questo modello di struttura dati è la vera rivoluzione **dell'IA Generativa**.

PARTE III

IA generativa spiegata in dettaglio: un viaggio nel cuore della tecnologia IA

Per rendere le cose più semplici, partiamo da concetti che conosciamo, per arrivare a comprendere cosa è una **LLM** e come funziona.

Un **LLM** non è un database tradizionale in cui i dati sono memorizzati in modo esplicito. Se chiedi "Chi è il Presidente?", il programma cerca la stringa di testo esatta in una tabella attraverso una "Query".

Un LLM, invece, **non conserva i testi originali** che ha letto durante l'addestramento. Durante il processo di training, il modello "distilla" le informazioni trasformandole in relazioni statistiche e matematiche. Come conseguenza di ciò, la conoscenza è distribuita in miliardi di numeri chiamati **pesi** (**weights**). Per essere più chiari, è come un essere umano che legge un libro: non ricorda ogni singola parola di ogni pagina (database), ma ne comprende il concetto e sa ricostruire il discorso (LLM).

Ma fisicamente come si presenta un LLM?

È un file (spesso enorme, centinaia di gigabyte) che contiene l'architettura di una **Rete Neurale Artificiale**, nello specifico basata su una struttura chiamata **Transformer**. L'LLM lo trovi sui server delle grandi ***Big Farm come Google*** che mette a disposizione la propria LLM di nome ***Gemini*** per poter fare inferenza.

Fisicamente, viene inserita una domanda di interesse (prompt) attraverso un qualsivoglia browser e il processo che ne segue è il seguente:

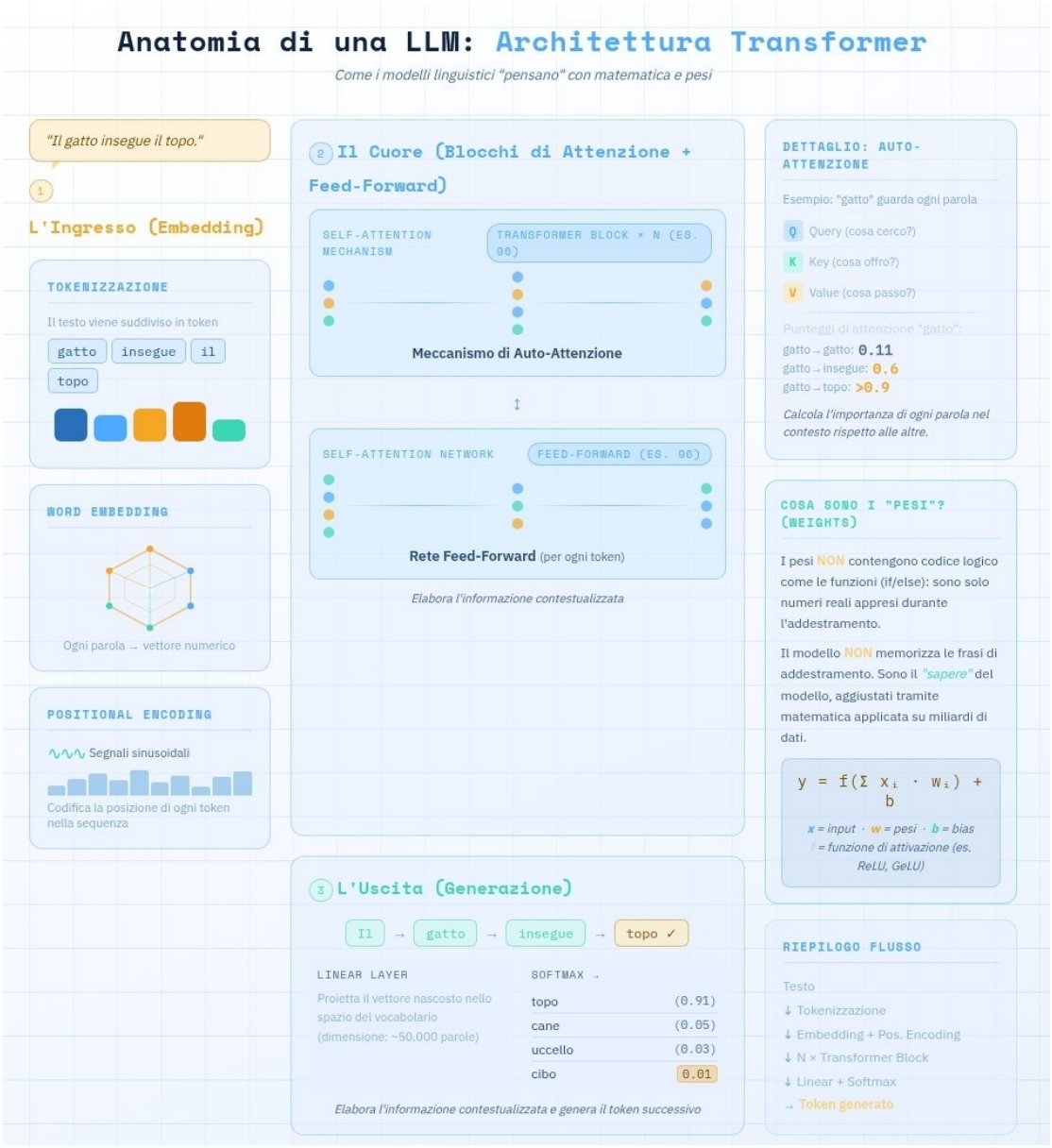
- ***Input****: Il testo viene convertito in numeri (vettori);*
- ***Processo****: Questi numeri passano attraverso i "pesi" della rete neurale. Non c'è una ricerca in un archivio; ci sono miliardi di operazioni matematiche (moltiplicazioni di matrici).*
- ***Output:*** *Il software calcola, parola dopo parola, quale sia la "prossima parola più probabile" per rispondere in modo coerente.*

Quindi **l'LLM** *è un motore di inferenza statistica, non "sa" le cose nel senso tradizionale; le "ricostruisce" matematicamente ogni volta che viene interrogato.*

IL LARGE LANGUAGE MODEL IN DETTAGLIO

Spiegare in dettaglio come è strutturato un LLM non è cosa facile, perché questa tecnologia porta con sé tantissime novità concettuali e di ordine matematico-statistico, quindi l'approccio che userò sarà fondamentalmente visivo e non di eccessivo dettaglio tecnico, per cercare di mantenere il livello adeguato ad un utente avanzato e non per un progettista informatico.

Partiamo da uno schema sviluppato proprio dalla IA:



L'infografia mostra l'anatomia di una LLM con l'architettura Transformer, suddivisa in tre sezioni principali:

1) L'Ingresso (***Embedding***) - Tokenizzazione, Word Embedding e Positional Encoding.

2) Il Cuore - ***Blocchi di Self-Attention*** e ***Feed-Forward Network***.

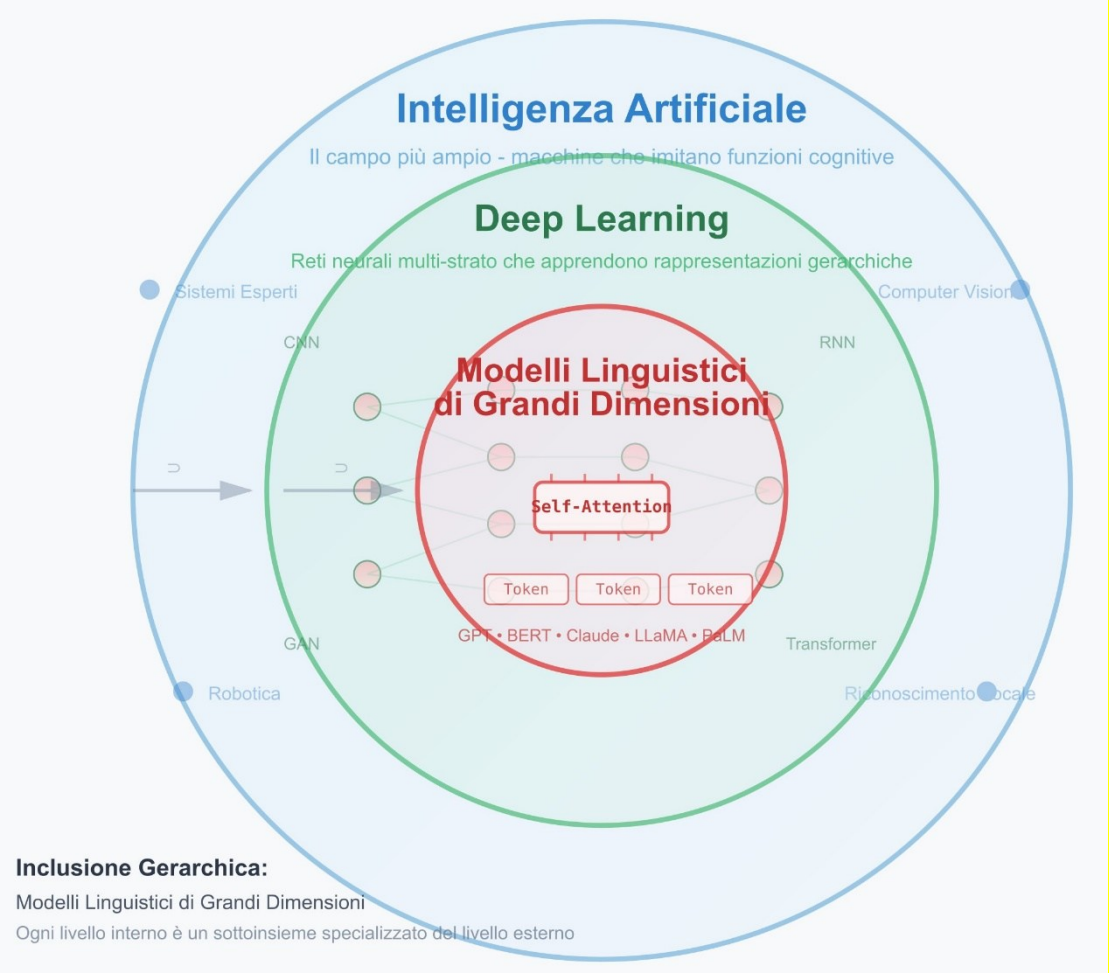
3) L'Uscita (***Generazione***) ***- Linear Layer*** e ***Softmax*** per la generazione del ***token***.

Nell'immagine cé anche un piccolo box che sintetizza il significato dei pesi all'interno della struttura. Una nota è doverosa: "non è necessario comprendere lo schema interno di una LLM, ma è necessario capire l'interazione esterna".

Questa nota "salvagente" vi esula dal capire i meccanismi intrinsechi delle LLM per poter interagire con essa, ma può aiutare a comprendere meglio gli articoli che verranno per poter sfruttare tutti i meccanismi di interazione del mondo IA.

GLI LLM NEL PANORAMA DELL'INTELLIGENZA ARTIFICICALE

La comprensione generale dei Large Language Model (LLM) passa per una contestualizzazione dei modelli all'interno dell'architettura generale dell'Intelligenza Artificiale (IA).



Il mondo della IA è un campo vasto e gli LLM si collocano all'interno del ***Machine Learning*** (apprendimento automatico) e, più precisamente, del ***Deep Learning*** (apprendimento profondo tramite reti neurali).

A differenza dei software tradizionali basati su regole rigide ("se succede A, allora fai B"), questi modelli apprendono dai dati. Immaginiamo di voler insegnare a un computer a riconoscere lo stile di un autore: non gli daremo una lista di regole grammaticali, ma gli "faremo leggere" l'intera bibliografia affinché ne assorba autonomamente i pattern.

PARTE III

IA generativa spiegata in dettaglio: un viaggio nel cuore della tecnologia IA

Come funziona il Deep Learning? (L'esempio del testo)

Il Deep Learning non conta solo le lettere, ma calcola la probabilità statistica di successione tra unità di significato (***token***).

- Esempio inefficace: Contare quante "e" ci sono in una frase.
- Esempio efficace (Predizione): Se un modello legge la sequenza "Il gatto è sul...", statisticamente sa che la parola successiva ha un'alta probabilità di essere "divano" o "tappeto", e una probabilità quasi nulla di essere "democrazia". Analizzando trilioni di frasi, il modello non impara solo le parole, ma la struttura logica del pensiero umano.

LE RETI NEURALI: L'ARCHITETTURA

Per abilitare questo tipo di deep learning, gli LLM sono basati su reti neurali. Le reti neurali artificiali imitano il cervello umano attraverso nodi interconnessi:

- ***Input Layer:*** Riceve il testo (convertito in numeri).
- ***Hidden Layers:*** Elaborano i dati. Immaginali come una serie di filtri: il primo riconosce la grammatica, il secondo il tono (formale/informale), il terzo il concetto astratto.
- ***Output Layer:*** Restituisce la risposta finale.

IL MODELLO TRANSFORMER E LA "SELF-ATTENTION"

La vera rivoluzione è l'architettura Transformer. Prima dei Transformer, i modelli leggevano le frasi da sinistra a destra, "dimenticando" l'inizio della frase quando arrivavano alla fine.

L'esempio del contesto (Ambiguità): Considera la parola "**pesca**".

1. "Ho mangiato una pesca matura."
2. "Oggi vado a pesca al lago." Un modello tradizionale potrebbe confondersi. Un ***Transformer***, grazie ***alla Self-Attention***, analizza contemporaneamente le parole "mangiato" o "lago" per assegnare istantaneamente il significato corretto a "pesca". Riesce a capire che la fine di un paragrafo può cambiare il senso della prima riga.

VANTAGGI E LIMITI: ESEMPI PRATICI

Sicuramente il vantaggio più grande risiede nella potenza del linguaggio naturale.

Un programma classico non capirebbe mai un comando vago come: "Trasforma questo verbale di riunione in una poesia ermetica". Un LLM invece può farlo perché non cerca una corrispondenza esatta di dati, ma manipola i concetti sottostanti.

Il rischio: Le allucinazioni

Il limite principale è che l'LLM non "sa" la verità; calcola solo cosa è probabile dire.
Esempio di allucinazione: Se chiedi a un LLM di scriverti la biografia di un avvocato sconosciuto, il modello potrebbe generare un testo credibile citando lauree prestigiose e casi vinti mai esistiti. Lo fa perché, statisticamente, una biografia di un avvocato "deve" contenere quei dettagli per sembrare corretta, anche se i fatti sono falsi.

La sicurezza dei dati

Immagina un dipendente che incolla il codice sorgente segreto dell'azienda per chiedere all'IA di trovare un errore. Quel codice entra nel database del modello e potrebbe essere usato per rispondere alla domanda di un concorrente mesi dopo. Gli LLM non sono "casseforti", ma "spugne".

Di seguito trovate una tabella comparativa tra ***IA tradizionale*** e le ***LLM generative***:

Caratteristica	**IA Tradizionale (Rule-based / ML classico)**	**Generative AI (LLM / Transformer)**
Logica di base	Basata su regole rigide ("If-Then") e database strut	Basata su probabilità statistica e relazioni semantiche.
Input richiesto	Dati puliti, tabelle, parametri numerici precisi.	Linguaggio naturale, testi disordinati, query vaghe.
Comprensione	Analizza le parole come etichette isolate.	Comprende il **contesto** e le sfumature (ironia, tono).
Output	Classificazioni (Sì/No), previsioni numeriche, filtri.	Generazione di nuovi contenuti (testi, codice, idee).
Esempio	Un filtro antispam che blocca la parola "Gratis".	Un assistente che riassume un libro in stile piratesco.

Leggendo la tabella possiamo concludere quanto segue:

- ***Dalla rigidità alla flessibilità*:** L'IA tradizionale è come un binario ferroviario: efficiente, ma può andare solo dove è stata posata la rotaia. Gli LLM sono come un fuoristrada: possono muoversi su terreni "accidentati" come una conversazione umana piena di errori grammaticali o giri di parole.
- ***Il concetto di "Emergenza":*** Mentre nell'IA tradizionale sappiamo esattamente perché il software ha dato un certo risultato (ha seguito una regola), negli LLM ***emergono*** capacità che non erano state programmate esplicitamente, come la capacità di tradurre lingue che il modello ha visto raramente.
- ***L'importanza del Prompt*:** Nell'informatica classica contava la "query" (il comando tecnico). Oggi conta il **Prompt**, ovvero la capacità dell'uomo di dare istruzioni descrittive.

CONCLUSIONE

Siamo arrivati alla fine di questo articolo, ricco di concetti tecnici ed essenziali per poter abbracciare la rivoluzione della IA. Riassumendo in due parole l'articolo: "siamo passati da macchine che eseguono calcoli a macchine che manipolano significati. Il limite non è più la potenza di calcolo, ma la qualità dei dati e la capacità umana di porre le domande giuste."