BOOM
AiGeneration

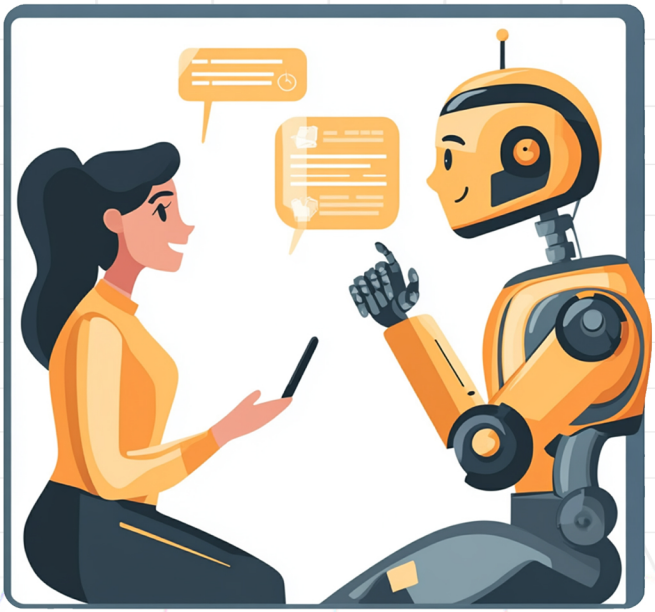
LA GEN AI IN AZIONE

TUTTO QUELLO CHE DEVI SAPERE SUGLI LLM. PT.2

UN CASO D'USO PRATICO

SWIPE FOR MORE

Gli LLM (Large Language Models) stanno rivoluzionando il modo in cui interagiamo con la tecnologia, rendendo più naturale e intuitiva la comunicazione uomo-macchina.



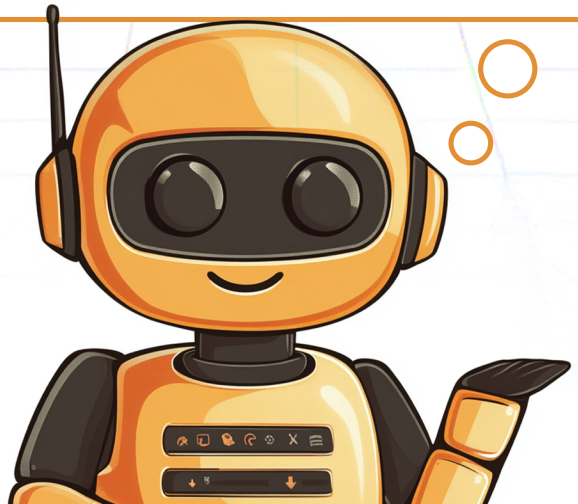
Per comprendere meglio il loro funzionamento, vediamo un caso d'uso pratico:
Chatbot per il Supporto Clienti di una Banca

Un cliente chiede assistenza tramite chatbot:
Domanda: **"Come posso bloccare la mia carta di credito smarrita?"**

1- Analisi della Domanda
(Tokenizzazione e Comprensione)

Il modello scompone la frase in parti più piccole, dette **token**:
"Come", "posso", "bloccare", "la", "mia", "carta", "di", "credito", "smarrita", "?"

Ogni parola viene trasformata in numeri (vettori) per essere compresa dall'IA.



2- Interpretazione del contesto con self-attention

L'LLM applica **multi-head self-attention*** per calcolare l'importanza relativa di ciascun token rispetto al contesto:

Parola	Importanza (%)
bloccare	**40%**
carta	**35%**
credito	**25%**
smarrita	**15%**

- Il modello si concentra su **"bloccare", "carta"** e **"credito"**, che sono essenziali per capire l'intento della richiesta.
- **"smarrita"** ha un peso inferiore, ma contribuisce a determinare l'urgenza dell'azione richiesta.

***Il Multi-Head Self-Attention (MHSA)** è un meccanismo chiave nei modelli Transformer, come GPT-4, che aiuta l'IA a capire il contesto di una frase in modo efficiente.
In breve: divide l'analisi del testo in più **prospettive simultanee** (detti head), permettendo al modello di cogliere diversi significati e relazioni tra le parole.

3- Previsione della risposta più adatta

Dopo aver compreso la domanda, l'LLM deve prevedere la risposta più adatta. Per farlo, il modello analizza tutte le sequenze di testo che ha "visto" durante il suo addestramento e calcola quale risposta sarebbe **la più probabile**.

Ad esempio, nel nostro caso, il modello valuta diverse risposte possibili come:

Risposta Possibile	Probabilità (%)
Contattare il numero di emergenza	**85%**
Bloccare la carta via app	**78%**
Recarsi in filiale	**32%**

L'LLM sceglie la **risposta con la probabilità più alta** e la propone all'utente.

4- Generazione della risposta

Una volta scelta la risposta migliore, il modello inizia a generare il testo, parola per parola. Utilizza un **algoritmo di decoding** che, ad ogni passaggio, sceglie la parola successiva più probabile basandosi sul contesto della conversazione:

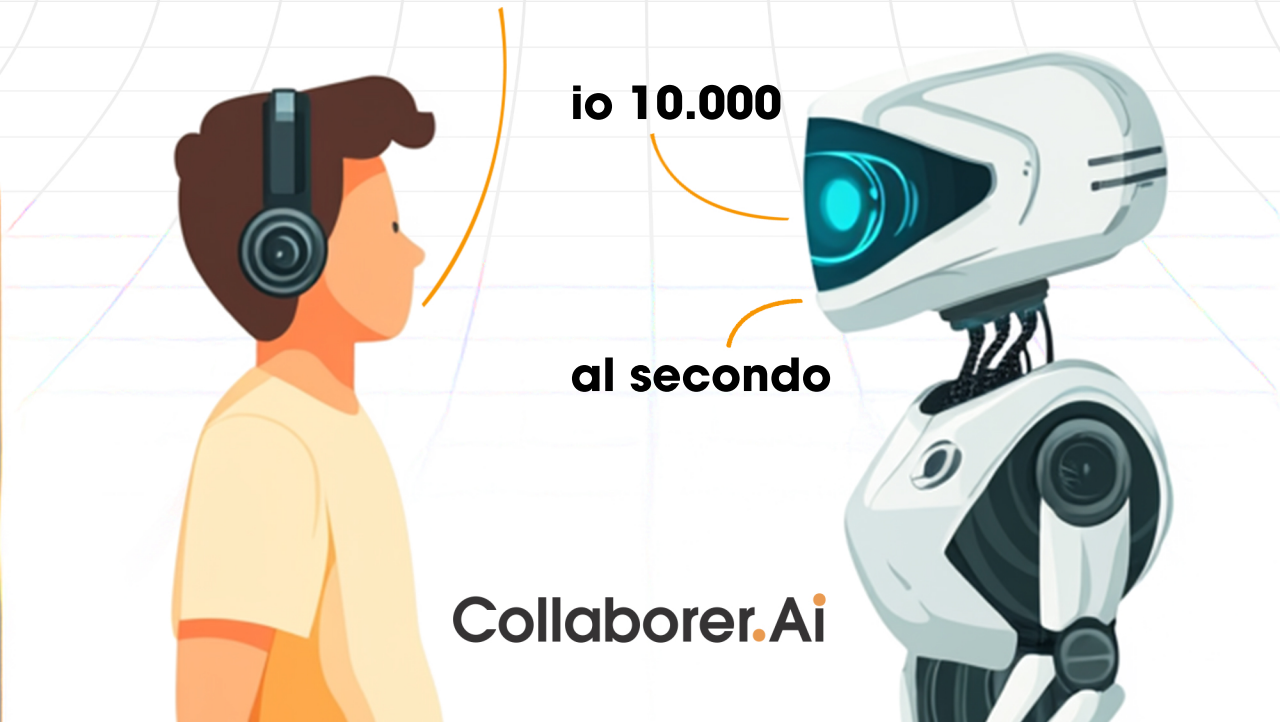
"Per bloccare la tua carta di credito smarrita, puoi accedere all'app mobile e selezionare l'opzione 'Blocca Carta'.
Se preferisci, chiama il numero di emergenza 800-123-456 per assistenza immediata."

BOOM
AiGeneration

Lo sapevi che...

Gli LLM possono generare fino a **10.000 parole al secondo** durante il processo di completamento di frasi o risposte?

Si tratta di una velocità molto più alta rispetto a un operatore umano, che può scrivere in media circa **40 parole al minuto**!



Collaborer.Ai

E tu?

Come pensi che l'uso degli LLM possa migliorare i servizi automatizzati nella tua azienda?

Condividi le tue opinioni nei **commenti**!

Seguici e unisciti a noi per
diventare parte della Community

BOOM
AiGeneration

Collaborer.Ai