Come Funziona l' Inferenza degli LLM

Quando si inserisce un *prompt* in un Modello Linguistico di Grande Dimensione (LLM), il modello converte il testo in numeri, li elabora e restituisce una risposta un *token* alla volta. In questo articolo, esamineremo il percorso dell' inferenza degli LLM e vedremo come funziona.

Cosa sono i Modelli Linguistici di Grande Dimensione (LLM)?

Gli LLM sono semplicemente reti neurali costruite sull' architettura *transformer*. A differenza delle architetture precedenti che elaboravano il testo in sequenza, i *transformer* possono analizzare intere sequenze in parallelo, rendendoli più efficienti da addestrare e distribuire.

Il blocco fondamentale di questi modelli è il livello *transformer*, che consiste in due componenti principali:

1. un meccanismo di *self-attention* (auto-attenzione), e
2. una rete neurale *feed-forward*.

Gli LLM impilano dozzine di questi livelli, creando reti profonde capaci di catturare schemi complessi nel linguaggio.

I *transformer* si basano sul *self-attention*, che valuta come ogni parola si relaziona al resto della sequenza, non solo alle parole vicine.

Dimensione del modello = il numero di parametri nella rete. Un modello con 7 miliardi di parametri ha 7 miliardi di numeri in virgola mobile che memorizzano la conoscenza appresa durante l' addestramento. Questi parametri sono organizzati in matrici di peso che eseguono trasformazioni sui dati di input a ogni livello.

Modelli come GPT-4, Claude e Llama sono *transformer* solo *decoder*, il che significa che utilizzano solo la parte *decoder* dell' architettura *transformer* originale. Questo li