

LARGE LANGUAGE MODELS

Comprendere il funzionamento degli LLM per integrarli criticamente nella didattica

LUCA ZORLONI

THE BIG INTERVIEW 17.02.2026

In Estonia ChatGPT arriva in classe. La ministra dell'Istruzione Kallas ci spiega perché, anziché vietare l'AI, è meglio insegnare agli studenti a usarla


Dopo i docenti, il progetto AI Leap in collaborazione con OpenAI arriva a 15mila studenti. Ecco come funziona il piano estone per applicare l'intelligenza artificiale all'apprendimento

<https://www.wired.it/article/estonia-chatgpt-scuola-ministra-kallas-istruzione-intervista/>

La ragione per cui l'Estonia ha lanciato il programma AI Leap risiede nel nostro timore principale. Ossia continuare a **insegnare e apprendere, soprattutto a insegnare, esattamente come abbiamo fatto negli ultimi vent'anni**, senza prestare attenzione al fatto che gli studenti hanno già ChatGPT sul telefono o sul computer.

Il modo in cui svolgono i compiti, scrivono i loro elaborati, studiano... Fingiamo semplicemente che tutto ciò non esista e ci limitiamo a punire gli studenti per aver copiato. Il problema è che dobbiamo riprogettare il modo in cui insegniamo, perché la tecnologia esiste e gli studenti la utilizzano.

Dobbiamo insegnare in modo diverso affinché questa tecnologia li aiuti dal punto di vista cognitivo a sviluppare il pensiero anziché diminuire la loro capacità di ragionamento. Purtroppo, e dico purtroppo, **dobbiamo ricorrere alla tecnologia americana perché non disponiamo di alternative europee**".



Con l'avvento di ChatGPT o di altri modelli linguistici, la **capacità di pensare in modo più profondo e ampio può essere facilmente delegata** al computer. E ciò significa che gli studenti stessi non svilupperanno né cresceranno cognitivamente in termini di capacità di pensiero o analisi.

Gli esseri umani diventeranno, diciamo chiaramente, più stupidi perché non saranno in grado di pensare. O avranno dati ma non saranno capaci di analizzarli o di trarne alcuna comprensione significativa. Tutta questa capacità cognitiva di pensare. Ecco, questo è il rischio maggiore di questa tecnologia. E "significativo" comporta che gli insegnanti comprendano che dobbiamo insegnare agli studenti a pensare con l'uso della tecnologia. Quindi si introduce la tecnologia nel processo di apprendimento per costringere gli studenti a imparare e a pensare con la tecnologia, non a delegare il pensiero alla tecnologia.

SUPERARE IL "MARKETING" DELL'IA

Troppo spesso l'Intelligenza Artificiale viene presentata come un'entità magica o una minaccia per le materie umanistiche. In questa lezione, la spogliamo della retorica commerciale per analizzarla come uno **strumento computazionale basato sulla statistica**.

L'obiettivo di queste lezioni è riflettere sull'integrazione dei **Large Language Models (LLM)** all'interno dell'agenda di ricerca e della didattica letteraria. Più che parlare genericamente di "Intelligenza Artificiale" — termine spesso abusato dal marketing per vendere licenze — è opportuno concentrarsi sulla ricerca scientifica legata a modelli specifici

L'approccio proposto non è puramente strumentale (un "tutorial" sull'uso di ChatGPT), ma **metodologico**: come queste tecnologie possono diventare parte integrante della nostra ricerca? Possono aiutarci a raffinare la nostra metodologia o a conoscere meglio il nostro oggetto di studio? come i modelli generativi di grandi dimensioni.

COSA SIGNIFICANO LLM E GPT

LLM (Large Language Model)

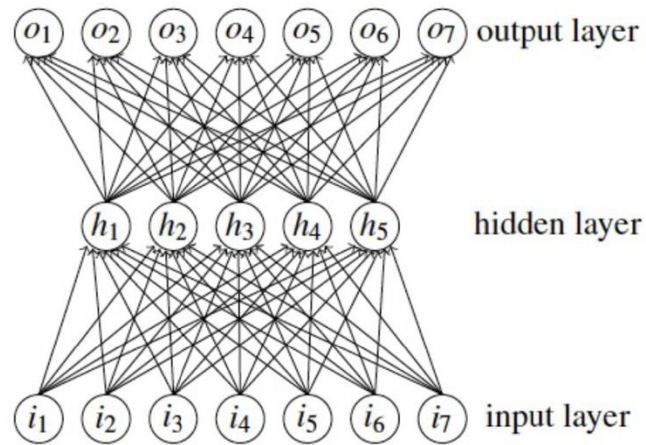
- **LARGE (Grande):** si riferisce alla vastissima mole di dati utilizzata per l'addestramento e al numero enorme di parametri (la "dimensione") del modello.
- **LANGUAGE (Linguistico):** denota il focus specifico sull'elaborazione del linguaggio naturale (NLP - *Natural Language Processing*).
- **MODELS (Modelli):** indica che si tratta di una struttura che formula previsioni o risposte basandosi su pattern (modelli ricorrenti) estratti dai dati linguistici.

GPT (Generative Pre-trained Transformer)

- **GENERATIVE (Generativo):** si riferisce alla capacità del modello di creare nuovo testo, non limitandosi a recuperare informazioni esistenti.
- **PRE-TRAINED (Pre-addestrato):** indica che il modello è già stato istruito su un immenso corpus di dati prima di essere messo a disposizione dell'utente.
- **TRANSFORMERS (Trasformatore):** è il nome specifico dell'architettura di rete neurale utilizzata (che approfondiremo nella sezione successiva).

ChatGPT è una APplicazione, un'interfaccia attraverso cui usiamo il modello GPT nelle sue varie versioni

LE RETI NEURALI E IL DEEP LEARNING



Il cuore degli LLM è la **Rete Neurale**, un modello computazionale ispirato vagamente al funzionamento dei neuroni biologici:

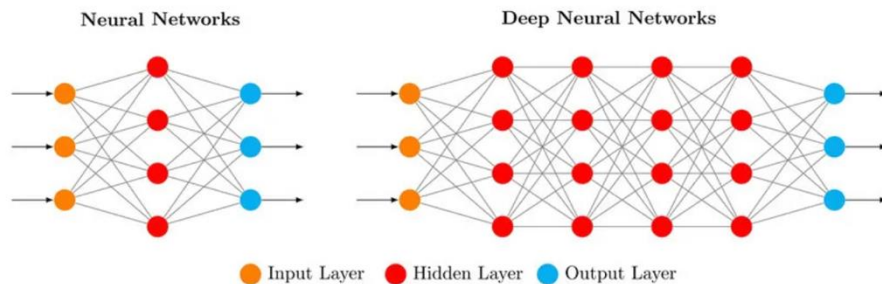
Strati (Layers) e DNN: Una rete è composta da uno strato di *Input*, vari strati *Nascosti* (Hidden Layers) e uno di *Output*.

- I modelli con più di uno strato nascosto sono definiti **Reti Neurali Profonde** (Deep Neural Networks - DNN).
- Queste reti hanno spesso oltre **100 strati nascosti** tra i nodi di input e quelli di output.

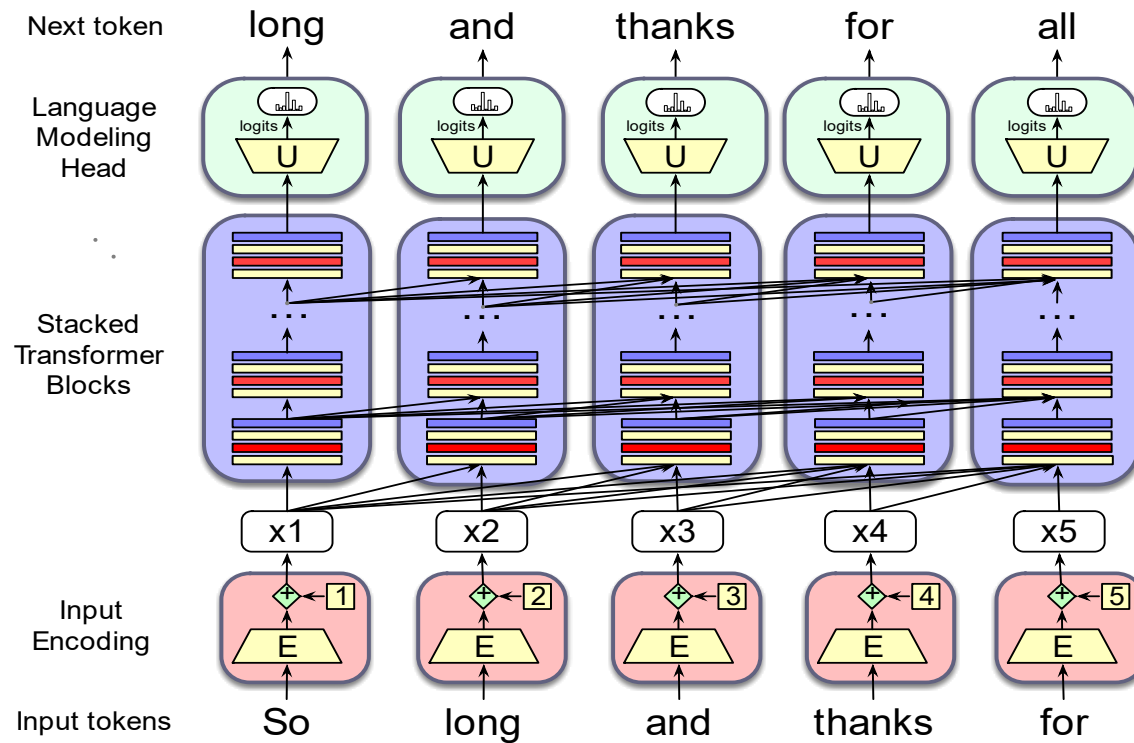
Connettività: All'interno di queste strutture, ogni neurone è connesso a ogni singolo neurone dello strato precedente e di quello successivo. Questa fitta rete di connessioni permette al modello di apprendere relazioni estremamente complesse tra i dati.

Vettori (Embedding): La macchina non legge "lettere", ma trasforma ogni parola (o *token*) in un vettore numerico, ovvero una coordinata in uno spazio a centinaia di dimensioni. In questo spazio, parole semanticamente simili (es. "rosa" e "fiore") si trovano vicine.

Apprendimento: Durante l'addestramento, la rete aggiusta i "pesi" (la forza) delle connessioni tra i neuroni per prevedere con massima precisione la parola successiva in una sequenza.

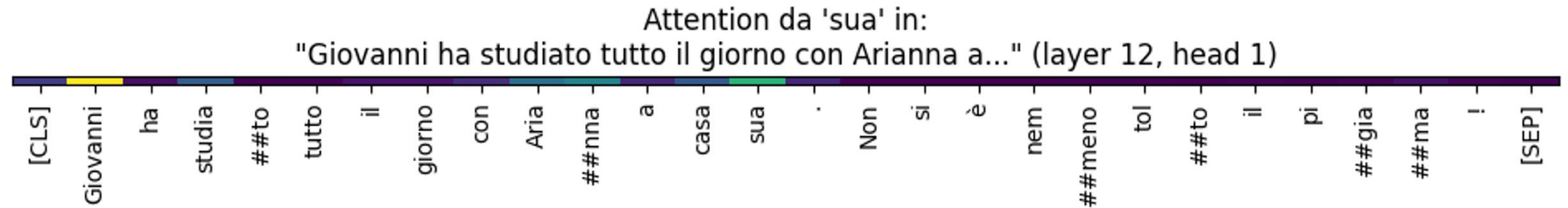


LA RIVOLUZIONE DEI TRANSFORMER



Architettura di un **transformer** che funziona sequenzialmente («**autoregressive**») per fare **generazione di testo**.

Addestrato imparando a predire la parola che segue guardando solo quello che precede.



La vera rivoluzione che ha permesso la nascita di GPT è il **Transformer**. A differenza dei modelli precedenti, introduce il meccanismo di **Attention**:

Self-Attention: Il modello analizza ogni parola di una frase in relazione a tutte le altre simultaneamente. Assegna dei "pesi" per capire quali parole sono cruciali per il senso di un'altra.

- *Esempio:* In "Giovanni ha studiato con Arianna a casa sua", l'attenzione permette di legare "sua" a "Giovanni" o "Arianna" pesando il contesto circostante.

Multi-head Attention: Il modello non ha un solo sguardo, ma "molte teste" che guardano il testo a diversi livelli (morfologico, sintattico, semantico) contemporaneamente.

DALL'ANALISI ALLA GENERAZIONE: IL CICLO DI PREVISIONE

Ma come si trasforma questa analisi statistica in una risposta testuale? Il processo di generazione è un ciclo continuo basato sulla probabilità:

Proiezione sul Vocabolario: Una volta che l'attenzione ha "arricchito" il vettore dell'ultima parola con tutto il contesto precedente, questo vettore viene proiettato sull'intero vocabolario del modello (una lista che può contenere decine di migliaia di termini).

Distribuzione di Probabilità: Il modello assegna un punteggio a ogni termine del vocabolario. Se l'input è "*Arma virumque...*", la parola "*cano*" riceverà una probabilità altissima, mentre "*pizza*" una probabilità vicina allo zero.

Campionamento e Autoregressione: Il modello sceglie una parola tra quelle più probabili (influenzato dalla *Temperatura*). Una volta scelta, la nuova parola non è solo l'output finale, ma viene **riaggiunta all'input**. Il modello ricomincia il ciclo analizzando ora "*Arma virumque cano*" per prevedere il termine successivo.

Questo processo "autoregressivo" significa che la macchina genera il testo una parola alla volta, dove ogni nuova parola contribuisce a ridefinire il contesto per quella successiva.

IN PRATICA

LLM: un modello computazionale capace di generare sequenze di parole.

Obiettivo: Calcolare la probabilità di una parola dato un contesto.

Esempio: "Oggi il cielo è molto nuvoloso, credo che presto inizierà a ____"

- a) piovere (Alta probabilità)
- b) ristorante (Bassa probabilità)

in una lingua, certe sequenze sono più probabili di altre. Se dico 'mettersi a...', la probabilità che segua 'piovere' o 'correre' è molto più alta di 'citofono'. Gli LLM fanno questo, ma su una scala immensa."

=> siamo partiti dall'ipotesi distribuzionale

L'ADDESTRAMENTO E LA CONFIGURAZIONE

Per arrivare a generare testi coerenti, il modello attraversa diverse fasi di "istruzione" e richiede parametri di controllo durante l'uso.

Pre-training: Apprendimento autonomo su masse di dati (senza etichette).

Fine-tuning: Specializzazione su compiti specifici.

Temperatura: (T): È il parametro che controlla l'aleatorietà della generazione.

- **Bassa Temperatura (vicina a 0):** Il modello è "conservativo", sceglie quasi sempre la parola più probabile. Ottimo per analisi grammaticali.
- **Alta Temperatura (vicina a 1):** Il modello diventa "creativo", dando spazio a parole meno probabili. Utile per riscritture stilistiche.

PRE-TRAINING

Pre-training (Pre-addestramento): È la fase "generalista". Il modello viene alimentato con una mole colossale di dati (web, libri, articoli, Wikipedia) per imparare come funziona la lingua, la grammatica e i pattern del mondo. In questa fase il modello impara a prevedere la parola successiva in contesti generici.

Ricordate che questo modello è chiamato «apprendimento autonomo» (self supervision) perché **non richiede etichette manuali**: la sequenza naturale delle parole fornisce la supervisione.

L'addestramento consiste nel **ridurre l'errore nella previsione** della parola corretta.

FINE-TUNING

Fine-tuning (Affinamento): È la fase "specialistica". Una volta che il modello ha una conoscenza generale, viene addestrato su dataset più piccoli e curati per compiti specifici.

1. Selezione del modello base: Si parte da un modello già esistente (es. Llama o GPT-4).

2. Preparazione del Dataset: Si raccoglie una lista di esempi (Prompt + Risposta corretta). Ad esempio: "Analizza questa frase di Tucidide" -> "Analisi filologica dettagliata".

4. Training Session: Si fa girare il modello su questi dati specifici. La macchina confronta la sua risposta con quella "corretta" del dataset e corregge i propri parametri interni per minimizzare l'errore.

5. Validazione: Si testa se il modello è diventato effettivamente più bravo sul greco antico senza aver "dimenticato" come si parla in generale.

PROMPT

Un prompt è un **testo di input** che serve a guidare il modello generativo verso un output che si avvicina maggiormente ai **desideri** dell'utente.

Il prompt è fondamentale perché fornisce al modello il **contesto** che servirà a generare la risposta.

L'input dato a un LLM non è solo una domanda, ma un modo per **«educare»** il modello sulla situazione specifica. Se il prompt contiene:

- **Fatti verificati**, il modello li userà per la risposta.
- **Esempi di output** desiderato, il modello seguirà il pattern.
- **Passaggi logici**, il modello migliorerà la coerenza del ragionamento.
- **Domande ben formulate**, si ridurranno bias ed errori.

RHLF: REINFORCED LEARNING FROM HUMAN FEEDBACK

Come si fa? Si forniscono coppie di "Esempio-Risposta" di alta qualità. Ad esempio, per creare un esperto di greco antico, lo si "nutre" con migliaia di esempi di analisi.

Cosa succede tecnicamente? Durante il fine-tuning, i "pesi" della rete neurale (le connessioni che abbiamo visto nella sezione 2.1) vengono leggermente modificati per adattarsi ai pattern del nuovo dataset, rendendo il modello molto più preciso in quel settore specifico.

RLHF (Allineamento Umano): Un'ulteriore fase di fine-tuning in cui esseri umani valutano le risposte del modello, aiutandolo a diventare più preciso, sicuro e utile.

Si chiede al modello di **generare più risposte** per 1 prompt

Gli annotatori umani leggono le risposte e le **ordinano per qualità/preferenza**

Questi giudizi vengono usati per addestrare un **modello di ricompensa**

Il modello di ricompensa viene usato per **aggiornare** il LLM attraverso un metodo chiamato **reinforcement learning**

TEMPERATURA

Temperatura (T): È il parametro che controlla l'aleatorietà della generazione durante l'uso (inferenza).

Bassa Temperatura (vicina a 0): Il modello è "conservativo", sceglie quasi sempre la parola più probabile. Ottimo per analisi grammaticali e compiti di precisione.

Alta Temperatura (vicina a 1): Il modello diventa "creativo", dando spazio a parole meno probabili. Utile per riscritture stilistiche o brainstorming.

PUNTI DI FORZA: LE RISPOSTE DI UN LLM

Fluidità Sintattica e Coerenza Formale: Sono in grado di produrre testi grammaticalmente perfetti in molte lingue, inclusi il latino e il greco (sebbene con accuratezza variabile). Eccellono nel generare testi che "suonano" naturali e ben strutturati.

Manipolazione Stilistica e del Registro: Questa è forse l'area di maggiore interesse per le materie letterarie. L'LLM può riscrivere un testo cambiando il registro (da formale a colloquiale) o imitando lo stile di un autore specifico (*pastiche*).

- *Esempio:* Riscrivere un'invettiva di Catullo nello stile di un moderno saggio accademico o viceversa.

Sintesi e Astrazione Informativa: Gli LLM sono ottimi sintetizzatori. Possono leggere testi lunghi e complessi per estrarne i concetti chiave, creare abstract o riassunti tecnici rispettando vincoli precisi di lunghezza.

PUNTI DI FORZA: LE RISPOSTE DI UN LLM

Traduzione e Adattamento: Oltre alla traduzione interlinguistica, eccellono nella "traduzione intralinguistica" (es. parafrasi di testi poetici o semplificazione di passi filosofici complessi per scopi didattici).

Annotazione e Classificazione su Larga Scala: Possono essere istruiti per identificare velocemente in un corpus vasto:

- **Entità Nominate (NER):** Nomi di dei, eroi, luoghi geografici o personaggi storici.
- **Figure Retoriche:** Metafore, similitudini, iperbati (con le dovute verifiche umane).

Generazione di Codice per le Digital Humanities: Sono strumenti preziosi per i ricercatori che hanno bisogno di scrivere piccoli script (in Python o R) per estrarre dati da database digitali o analizzare frequenze lessicali in grandi biblioteche digitali.

I LIMITI: ALLUCINAZIONI E BIAS

Dove l'LLM fallisce: Il Problema delle Allucinazioni

Le **allucinazioni** si verificano quando il modello genera informazioni false o inesistenti presentandole con estrema sicurezza.

Primato della Probabilità sulla Verità: Il modello non consulta un database di fatti, ma calcola la parola "statisticamente più plausibile". Se una menzogna suona grammaticalmente e stilisticamente corretta, il modello la genererà.

Mancanza di "Groundedness" (Ancoraggio): La macchina non ha esperienza del mondo reale. Per lei, "Seneca" è un insieme di coordinate numeriche, non un uomo storico.

Allucinazioni Bibliografiche: Il modello può inventare titoli di saggi o citazioni critiche mescolando nomi di studiosi reali con argomenti plausibili. Questo è il miglior esercizio di **verifica delle fonti** per uno studente: chiedere all'IA una bibliografia e stanare i titoli inventati.

Bias: Il modello riflette i pregiudizi presenti nei testi di addestramento.

Scatola Nera: È difficile spiegare *perché* il modello ha dato una certa risposta.

PROMPT ENGINEERING

Tecnicamente, il prompt definisce lo stato iniziale della memoria di lavoro del modello (la *Context Window*). Poiché l'LLM è un sistema probabilistico, il prompt funge da "filtro": esso attiva specifici percorsi nella rete neurale, escludendone altri.

Se scrivo "*In latino...*", il modello sposta immediatamente i pesi dell'attenzione verso il vocabolario e le strutture latine, "spegnendo" (statisticamente parlando) le probabilità legate ad altre lingue.

LE COMPONENTI DI UN PROMPT EFFICACE

Un prompt professionale per la ricerca letteraria dovrebbe contenere quattro elementi:

Istruzione (Task): L'azione specifica da compiere (es. *"Analizza"*, *"Traduci"*, *"Riscrivi"*).

Contesto (Context): Informazioni di sfondo (es. *"Sei un esperto di metrica oraziana"* o *"Il testo appartiene all'età flaviana"*).

Dati di Input: Il testo su cui lavorare (es. il carme di Catullo).

Indicatori di Output: Il formato desiderato (es. *"Fornisci una tabella"* o *"Usa un registro aulico"*)

STRATEGIE DI APPRENDIMENTO NEL CONTESTO

Zero-Shot: Chiedere una risposta "a freddo", senza esempi. Utile per compiti semplici.

Few-Shot: Fornire 2 o 3 esempi di coppie "testo-analisi". È la tecnica d'oro per la filologia: istruisce il modello sullo stile critico richiesto.

Chain of Thought (CoT): Chiedere al modello di "pensare ad alta voce" (es. *"Prima scansiona i piedi del verso, poi identifica la cesura"*). Obbliga la macchina a generare token logici intermedi, riducendo le allucinazioni.