

Che cos'è un LLM? Come i modelli linguistici generano testo – e perché hanno allucinazioni

Autore: Zsolt Lengyel, iTStudy Hungary Training and Research Centre

Revisore: Mária Hartyányi, iTStudy Hungary Training and Research Centre

Obiettivo di apprendimento

Chatbot come ChatGPT, Claude o Gemini rispondono con un linguaggio fluente e sicuro – anche quando sbagliano. Lo scopo di questo materiale è spiegare, con esempi concreti, che cos'è un modello linguistico di grandi dimensioni (LLM), come produce risposte testuali e perché può generare informazioni inesatte, inventate o fuorvianti. Al termine del microcontenuto formativo, i docenti saranno in grado di definire il concetto di “token”, spiegare il principio della previsione del token successivo, riconoscere le forme tipiche e i segnali di allarme delle allucinazioni e utilizzare strategie e strumenti pratici per verificare i risultati prodotti dall'IA.

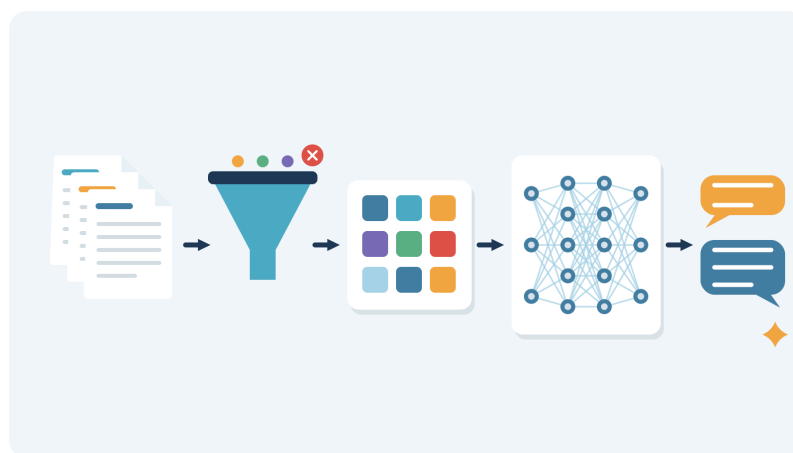
Introduzione

Il 30 novembre 2022 è stata rilasciata la prima versione pubblica di ChatGPT: nel giro di cinque giorni si era registrato un milione di persone e da allora i modelli linguistici di grandi dimensioni sono diventati strumenti di lavoro quotidiani. Le loro risposte sono ben strutturate e sicure di sé. Proprio per questo i loro errori sono pericolosi: una risposta sbagliata ha lo stesso aspetto di una risposta giusta.

Ai docenti non serve conoscere la matematica alla base dei modelli linguistici di grandi dimensioni, ma un'idea di base di come funziona un LLM spiega molte cose: perché è così bravo con le lingue, perché a volte inventa e perché verificarne i risultati è una responsabilità professionale.

Come nasce un modello linguistico di grandi dimensioni

Un modello linguistico di grandi dimensioni (LLM) è un sistema di intelligenza artificiale addestrato su enormi quantità di testi, capace di elaborare e generare linguaggio. Viene definito “grande” perché contiene un numero molto elevato di parametri, cioè valori numerici che vengono modificati durante l'addestramento e che permettono al modello di rappresentare ciò che ha appreso dai dati. La creazione di un LLM prevede quattro fasi principali.



Dalla selezione dei testi alla suddivisione in token, dall'addestramento alla creazione di un assistente conversazionale: le principali fasi di sviluppo di un LLM.

Raccolta e filtraggio dei dati

Gli sviluppatori scaricano gran parte del web pubblico ed eliminano spam, duplicati, pagine di scarsa qualità e dati personali: restano decine di terabyte di testo. Il modello può “sapere” solo ciò che era contenuto in questi dati filtrati – e solo fino a una certa data (la *data limite delle conoscenze*, o knowledge cutoff).

Tokenizzazione

Il modello non vede lettere o parole, ma numeri. I **token** sono frammenti di parola, parole o segni di punteggiatura, ciascuno con un numero identificativo. Il vocabolario di un modello tipico contiene circa centomila token. Una parola italiana lunga viene spesso suddivisa in più token – uno dei motivi per cui gli LLM sono sorprendentemente scarsi nel contare quante volte compare una lettera in una parola.

Pre-addestramento

Il modello riceve una sequenza di token e cerca di prevedere quello successivo, poi verifica se ha indovinato e regola leggermente i propri parametri. Questo processo si ripete migliaia di miliardi di volte. Il risultato è un “completatore di testo” molto sofisticato, ma non ancora un assistente.

Post-addestramento

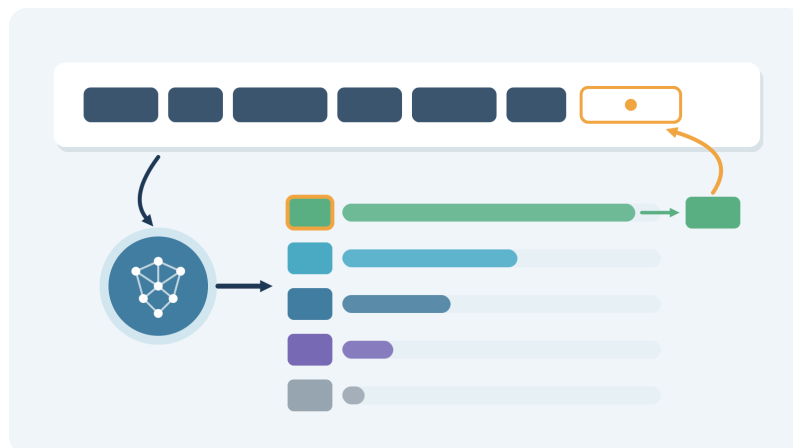
In seguito, il modello viene addestrato su esempi di domande e risposte redatti con cura, in modo che impari a rispondere in modo utile, cortese e in un formato adeguato – e a capire quando rifiutarsi. Infine, persone o modelli valutatori mettono in ordine risposte alternative, così il modello impara quali risposte sono preferite.

Generare una risposta: un token alla volta

Quando inviamo un messaggio, il modello calcola per ogni possibile token successivo quanto è probabile che segua il testo scritto fin lì. Ne sceglie uno, lo aggiunge al testo e ripete il processo – token dopo token. Quella che sullo schermo vediamo come “digitazione” è l'arrivo di questi token. Prendiamo l'inizio di una frase da officina:

“Prima di tagliare il tubo d'acciaio con la smerigliatrice angolare, indossa sempre ...”

Continuazioni come *gli occhiali di protezione*, *i guanti* o *le cuffie antirumore* sono molto probabili, perché frasi simili comparivano molte volte nei dati di addestramento; *le pantofole* sono estremamente improbabili. Il modello non consulta una regola – prosegue il testo in modo statisticamente plausibile.



Previsione del token successivo: il modello assegna un punteggio alle possibili continuazioni e ne aggiunge una al testo

È importante ricordare tre conseguenze di questo funzionamento:

Le risposte variano

Il modello non sceglie sempre il token più probabile in assoluto, quindi la stessa domanda può produrre risposte diverse.

Il modello ha bisogno di token per “pensare”

Quando gli è stato chiesto di fornire solo il risultato, un modello ha risposto che tre macchine, che da sole impiegano 12, 8 e 6 ore, completerebbero insieme un lavoro in 3,43 ore. Quando gli è stato permesso di procedere passo dopo passo, è arrivato al risultato corretto di 2 ore e 40 minuti. I modelli “di ragionamento” generano una simile catena di passaggi prima della risposta.

Il modello non ricorda come un essere umano

Ogni risposta viene generata a partire dal testo che il modello riceve in quel momento (vedi M1-T1-MLC5); le informazioni aggiornate arrivano solo da strumenti come la ricerca sul web, l'esecuzione di codice o i file allegati.

La **tokenizzazione** è il processo con cui il testo viene suddiviso in piccole unità, a ciascuna delle quali viene poi assegnato un numero identificativo univoco. Per esempio, in GPT-4 questo “dizionario” – l'elenco di tutti i token possibili – è composto da 100.277 elementi. Così, quando arriva al modello, il nostro messaggio è una sequenza di numeri – per esempio [15339, 1917, 264, 3918].

La generazione della risposta funziona allo stesso modo, ma in ordine inverso: il modello restituisce i token uno dopo l'altro, e poi un semplice programma li riconverte in testo leggibile. Sullo schermo la vediamo come una digitazione continua, ma in realtà i token arrivano uno dopo l'altro.

Proviamo!

La tokenizzazione si può sperimentare su un sito interattivo chiamato Tiktokenizer (<https://tiktokenizer.vercel.app/>). Inserendo un testo, vedremo come viene

suddiviso in token dal modello GPT-4o: quali frammenti di parola diventano un token autonomo e quali no.

Perché l'IA ha le allucinazioni

Si sente spesso dire, in modo fuorviante, che “l'IA mente”. In effetti capita di ricevere una risposta sbagliata a una domanda, ma ciò deriva dal modo in cui il modello è stato addestrato: nei suoi esempi di addestramento alle domande seguono risposte sicure, per cui il modello impara a produrre un testo che ha la forma di una risposta anche quando non dispone di conoscenze affidabili.



Le allucinazioni umane nascono dall'immaginazione. (Fonte: ChatGPT)

È forse una fortuna che in letteratura si sia affermato il termine **allucinazione**, anche se il processo sottostante è radicalmente diverso dall'allucinazione umana: **l'IA non possiede la capacità umana di immaginare; non “vede” né “immagina” cose che non ci sono, ma genera informazioni plausibili sulla base di schemi appresi – informazioni che a volte possono essere errate.**

Nel contesto dei modelli linguistici di grandi dimensioni, per allucinazione si intende un contenuto generato dall'IA che appare fluente, coerente e plausibile, ma che è errato nei fatti o non è supportato da prove affidabili. Si verifica perché il modello genera testo prevedendo sequenze di parole statisticamente probabili, anziché verificare le proprie affermazioni rispetto alla realtà.

I ricercatori di OpenAI hanno inoltre dimostrato che i metodi di addestramento e di valutazione più diffusi premiano il tirare a indovinare più dell'ammettere l'incertezza (Kalai et al., 2025). I modelli più recenti dicono “non lo so” più spesso, ma il problema non è scomparso. Una risposta frutto di allucinazione è statisticamente coerente con le risposte reali – stesso tono, stesso formato – semplicemente non è vera.

Esempi tipici

- *Fatti e persone inventati.* Se gli si chiede di una persona poco conosciuta, può capitare che il chatbot fornisca una biografia dettagliata, con professione, opere e date, messa insieme in modo del tutto casuale.
- *Riferimenti e normative inventati.* Se gli si chiede la base giuridica di una norma di sicurezza, un modello può citare un decreto con numero, data e articolo in un formato impeccabile – che però non esiste o disciplina tutt'altro. Nel 2023 un tribunale di New York ha sanzionato degli avvocati che avevano presentato una memoria contenente sentenze inventate da ChatGPT (Mata v. Avianca).
- *Valori tecnici plausibili ma sbagliati.* Una coppia di serraggio per un bullone, la sezione di un cavo, una temperatura di cottura o una tolleranza possono sembrare corrette perché rientrano nell'intervallo abituale – ma differire dal valore indicato dal produttore o dalla norma. Anche i calcoli possono sbagliare facilmente!
- *Il rischio della sostituzione silenziosa.* Il modello può scambiare termini tecnici rari per errori di battitura e sostituirli con parole simili e più comuni. Per esempio, può sostituire automaticamente il nome poco noto di una pianta, di un animale o di un componente meccanico con una parola familiare dall'aspetto o dal suono simile. La risposta che ne risulta può sembrare comunque fluente e convincente, pur andando in una direzione completamente diversa e sbagliata.
- *Omissioni e pregiudizi.* Alla domanda su come tagliare un tubo metallico, un'IA può suggerire correttamente una smerigliatrice angolare con disco da taglio – senza dire nulla sulla protezione di occhi, mani e udito. I modelli possono inoltre riprodurre stereotipi presenti nei dati di addestramento.



Una risposta frutto di allucinazione può sembrare ufficiale e completa – finché qualcuno non ne verifica la fonte

Le allucinazioni sono più probabili quando l'argomento è raro, locale o recente (normative nazionali, procedure aziendali, modifiche del mese scorso), quando sono richiesti numeri, nomi o citazioni precisi e quando la domanda contiene un presupposto falso.

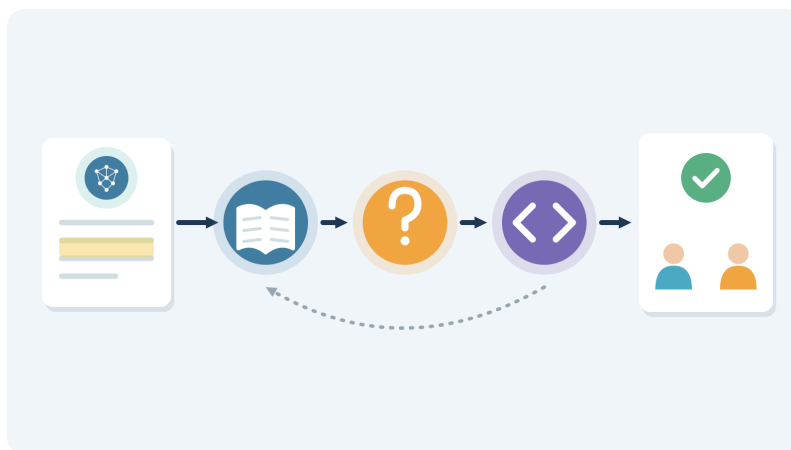
Poni questa domanda a ChatGPT:

Raggruppa i principali ambiti e tipi di compito in base a quanto tendono a essere affidabili le risposte di ChatGPT. Crea una tabella che mostri l'intervallo di affidabilità stimato, il principale rischio di errore e il livello di verifica necessario. Considera le percentuali come risultati misurati solo se supportate da ricerche affidabili; in caso contrario, indicale chiaramente come stime.

La tabella fornita in risposta offre indicazioni importanti per un uso consapevole dell'IA. Suggerisce, per esempio, che la probabilità di allucinazioni è relativamente bassa per le attività di programmazione (5-15%), ma può arrivare fino al 60% quando si tratta di informazioni aggiornate senza accesso a fonti web.

Individuare gli errori e verificare i risultati

Fluidità e sicurezza non sono segni di correttezza. La domanda giusta non è “sembra giusto?”, ma “come faccio a sapere che è giusto?”. Le seguenti strategie possono aiutare:



Verificare la fonte, porre domande di approfondimento e usare gli strumenti – e solo dopo portare il contenuto in classe

Verificare le fonti

Chiedere le fonti e poi aprirle: il documento esiste e dice davvero ciò che la risposta afferma? Confrontare con riferimenti autorevoli: il manuale del produttore, la norma tecnica, il libro di testo, il programma di studio.

Porre domande di approfondimento

“Su quali parti della tua risposta non sei sicuro?”, “Da dove viene questo valore?”, “Che cosa cambierebbe se...?” Porre di nuovo la stessa domanda in una nuova chat o ad un altro modello: risposte incoerenti sono un segnale di allarme.

Far usare gli strumenti all'IA

Per i fatti recenti, chiedere una ricerca sul web; per i calcoli matematici e tecnici, chiedere di “usare il codice”. In questo caso l'IA scrive un programma per risolvere il problema e lo esegue per fornire la risposta, il che riduce notevolmente la possibilità di errori.

Ancorare le risposte ai propri documenti

Allegare il capitolo o la scheda tecnica pertinente e dare questa istruzione: “Rispondi solo sulla base del documento allegato; se la risposta non è presente, dillo.”

Verificare prima dell'uso in classe

Usare le proprie competenze professionali – o quelle di un collega – per controllare ogni fatto, numero e istruzione di sicurezza su cui gli studenti faranno affidamento.

Cinque controlli prima dell'uso in classe

1. Quali affermazioni sono fatti, numeri, nomi o riferimenti – e li ho verificati uno per uno?
2. C'è qualcosa di critico per la sicurezza? Corrisponde ai dati del produttore e alle norme vigenti?
3. Che cosa manca (dispositivi di protezione, prerequisiti, eccezioni)?
4. Livello, terminologia e contesto sono adatti ai miei studenti?
5. Firmerei questo materiale con il mio nome?

Una questione pedagogica ed etica

L'allucinazione non è un semplice inconveniente tecnico. Quando un docente distribuisce una scheda di lavoro generata dall'IA, i suoi errori portano con sé l'autorevolezza del docente, e gli studenti – che stanno ancora costruendo le proprie conoscenze professionali – sono i meno capaci di individuarli. In laboratorio, un valore sbagliato o un passaggio di sicurezza mancante possono causare danni alle attrezzature o infortuni. Gli errori memorizzati dagli studenti sono difficili da correggere.

La responsabilità resta quindi al docente: l'IA può preparare una bozza, ma è il docente a validarla. È altrettanto importante insegnare agli studenti a fare lo stesso – chiedersi da dove provengono le informazioni, confrontarle con fonti affidabili e considerare l'IA un punto di partenza, non l'autorità finale. Questo fa parte dell'alfabetizzazione all'IA e all'informazione descritta in DigComp 3.0.

Un po' di storia

Fin dall'invenzione della scrittura, l'umanità ha cercato continuamente di conservare, tramandare e rendere accessibile alle generazioni future il sapere scritto accumulato.

La comparsa della scrittura, intorno al 3200 a.C. in Mesopotamia, ha permesso al sapere di sopravvivere alla memoria umana. Le tavolette d'argilla cuneiformi erano durevoli, ma potevano essere lette solo da una ristretta cerchia di scribi istruiti. I rotoli di papiro e di pergamena hanno reso il sapere scritto più trasportabile e hanno permesso la nascita delle grandi biblioteche dell'antichità, ma l'accesso è rimasto in gran parte riservato a pochi privilegiati.

Nel Medioevo il rotolo è stato gradualmente sostituito dal codice, la forma rilegata e suddivisa in pagine del libro. I libri copiati a mano nei monasteri erano estremamente costosi e richiedevano moltissimo tempo, per cui i libri e il sapere che contenevano rimanevano rari. La stampa a caratteri mobili ha portato un cambiamento radicale a metà del XV secolo: i testi potevano essere riprodotti in grandi quantità e a costi sempre più bassi. La diffusione dei libri, seguita dalle biblioteche pubbliche, dall'istruzione di massa e dalla crescente alfabetizzazione, ha ampliato costantemente il numero di persone che potevano accedere direttamente al sapere accumulato.

Alla fine del XX secolo, i computer e internet hanno segnato un altro grande salto. La digitalizzazione e le reti globali hanno reso disponibili quantità enormi di informazioni in modo quasi istantaneo. Molte barriere fisiche sono scomparse: oggi un solo smartphone può dare accesso a risorse che un tempo richiedevano biblioteche, viaggi o privilegi istituzionali. Ma questo ha creato una nuova sfida. Il problema non è più soprattutto la mancanza di informazioni, ma orientarsi nella loro abbondanza, comprenderle e distinguere le informazioni affidabili dai contenuti inaffidabili.

I modelli linguistici di grandi dimensioni rappresentano un ulteriore passo nella democratizzazione dell'accesso al sapere. Internet ha dato alle persone l'accesso alle informazioni; gli LLM vanno oltre, offrendo un'interfaccia in linguaggio naturale per lavorare con quelle informazioni. Gli utenti non hanno più bisogno di sapere esattamente quale libro o banca dati consultare, quale termine tecnico cercare o addirittura come formulare la query di ricerca perfetta. Possono porre domande nel linguaggio di tutti i giorni, chiedere approfondimenti, spiegazioni più semplici o più dettagliate, esempi, sintesi, confronti o traduzioni. Lo stesso argomento può essere spiegato a un principiante o a uno specialista, con diversi livelli di dettaglio e in lingue diverse.

Uno dei principali vantaggi degli LLM è proprio la possibilità di ridurre le difficoltà legate non solo alla ricerca delle informazioni, ma anche alla loro comprensione. Le tecnologie precedenti hanno progressivamente reso i libri e le informazioni accessibili a un numero sempre maggiore di persone. I modelli linguistici aggiungono una nuova possibilità: adattare le spiegazioni alle domande, alla lingua, al contesto e al livello di conoscenza di chi le richiede.



Conservare il sapere umano: dalle tavolette d'argilla ai modelli linguistici (Fonte: ChatGPT)

Questo segna anche una differenza importante rispetto alle forme di sapere precedenti. Un libro conserva un testo fisso; un LLM genera una nuova risposta a ogni domanda sulla base degli schemi appresi durante l'addestramento. Ciò lo rende straordinariamente flessibile, ma non infallibile. Può produrre risposte inesatte o persino del tutto sbagliate, e l'origine di un'affermazione non è sempre riconducibile a una fonte specifica. **Man mano che l'accesso al sapere diventa più facile, la capacità di valutare e verificare le risposte che riceviamo diventa sempre più importante.**

Dalla tavoletta d'argilla, passando per il libro stampato e internet, fino all'LLM, possiamo quindi osservare un lungo processo in cui **il sapere è diventato accessibile a sempre più persone, con sempre meno barriere. Il contributo distintivo degli LLM è che la prossima fase della democratizzazione non riguarda più soltanto l'accesso alle informazioni, ma sempre più l'accesso alla comprensione.**

Indicazioni per i docenti

- Un LLM genera testo un token alla volta, scegliendo continuazioni plausibili sulla base degli schemi appresi durante l'addestramento.
- Una risposta chiara e formulata con sicurezza non è necessariamente corretta: il modello può generare informazioni senza verificarle.
- Presta particolare attenzione a numeri, riferimenti, normative, termini poco comuni, argomenti locali o recenti e istruzioni di sicurezza.
- Controlla le fonti, fare domande di approfondimento, utilizza strumenti come la ricerca sul web o il codice e fornire all'IA documenti attendibili.
- Verifica sempre i materiali prima di utilizzarli in classe e insegna agli studenti a controllare criticamente le risposte dell'IA.

Bibliografia

- Kalai, A. T., Nachum, O., Vempala, S. S. & Zhang, E. (2025): Why Language Models Hallucinate. arXiv:2509.04664.

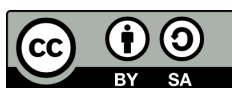
- Huang, L. et al. (2024): A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Transactions on Information Systems. DOI: 10.1145/3703155.
- Mata v. Avianca, Inc., 678 F. Supp. 3d 443 (S.D.N.Y. 2023).
- Cosgrove, J. & Cachia, R. (2025): DigComp 3.0 – European Digital Competence Framework. 5^a ed., Publications Office of the European Union.
- Lengyel, Zs. (2026): Introduction to the concepts of artificial intelligence; Concrete applications for content generation. Presentazioni del corso di formazione VITA.

Link

- Tiktokenizer – come il testo viene suddiviso in token: tiktokenizer.vercel.app
- Visualizzazione 3D interattiva di un LLM: bbycroft.net/llm
- A. Karpathy: Deep Dive into LLMs like ChatGPT (video): youtube.com/watch?v=7xTGNNLPyMI

Copyright © 2026 Consorzio VITA

Quest'opera è concessa in licenza secondo i termini della Creative Commons
Attribuzione-Condividi Allo Stesso Modo 4.0 Internazionale (CC BY-SA 4.0).



Finanziato dall'Unione europea. Le opinioni espresse sono esclusivamente quelle degli autori e non riflettono necessariamente quelle dell'Unione europea o dell'Agenzia nazionale.