I *Large Language Models*: funzionamento, potenzialità e limiti

di *Lucia Passaro*

I *Large Language Models* (LLM) sono oggi tra le tecnologie più diffuse dell'intelligenza artificiale generativa, capaci di produrre testi fluidi, sintetizzare documenti, rispondere a domande e assistere l'utente in una vasta gamma di attività. L'articolo ne ricostruisce il funzionamento, chiarendo come questi sistemi operino attraverso la previsione probabilistica del linguaggio e vengano addestrati su grandi *corpora* testuali. A partire da questa logica, il contributo mette in evidenza potenzialità e limiti: la variabilità delle risposte, il rischio di errori e allucinazioni, e le possibilità di adattamento a domini specialistici mediante strumenti come il *Retrieval-Augmented Generation* (RAG), utili ma non risolutivi rispetto ai problemi di affidabilità. In questa prospettiva, gli LLM possono supportare attività di analisi, ricerca, sintesi e redazione anche in ambito giuridico, ma non sostituiscono la verifica delle fonti, l'interpretazione e la responsabilità della decisione umana.

1. Introduzione / 2. La logica della previsione linguistica / 3. L'addestramento su grandi *corpora* testuali / 4. Generazione del testo e variabilità delle risposte / 5. Errori, *bias* e "allucinazioni" / 6. Adattamento al dominio e recupero di fonti esterne / 7. LLM e decisione giuridica: supporto, non sostituzione / 8. Considerazioni conclusive

1. Introduzione

Negli ultimi anni i *Large Language Models* (LLM) sono usciti dai laboratori di ricerca e sono entrati, con sorprendente rapidità, nella vita quotidiana. Oggi li incontriamo in *chatbot* conversazionali, assistenti alla scrittura, sistemi di traduzione e riassunto, e applicazioni capaci di produrre testi che, almeno in apparenza, somigliano molto a quelli scritti dagli esseri umani. La loro diffusione ha portato l'intelligenza artificiale generativa fuori dal dibattito specialistico: se ne discute nelle scuole, nelle imprese, nelle amministrazioni e, naturalmente, anche nel diritto. Ma che cosa sono davvero questi modelli? E, soprattutto, che cosa fanno quando "rispondono" a una domanda?

In prima approssimazione, un LLM è un sistema addestrato su enormi quantità di testo. Durante questa fase, il modello analizza libri, articoli, siti *web*, documentazione tecnica e molti altri materiali linguistici, non con l'obiettivo di "capirli" come farebbe un lettore umano, ma con quello di apprendere le regolarità che attraversano il linguaggio: ricorrenze, strutture, associazioni, modi tipici di organizzare un discorso. Da qui nasce la capacità di generare testo, riassumere documenti, tradurre, classificare informazioni, rispondere a domande e assistere l'utente in una pluralità di compiti.

È proprio questa plausibilità a spiegare il fascino, e talvolta l'ambiguità, degli LLM. Quando interagiamo con uno di questi sistemi, abbiamo spesso l'impressione di trovarci di fronte a una macchina che "sa", "capisce", "ragiona". In parte è inevitabile: il linguaggio è il veicolo principale attraverso cui gli esseri umani manifestano conoscenza, intenzioni e

comprensione. Se un sistema usa bene il linguaggio, siamo spontaneamente portati ad attribuirgli una forma di intelligenza simile alla nostra. Ma è proprio qui che si impone una distinzione decisiva. Un LLM non "comprende" il testo nel senso umano del termine; non possiede coscienza, esperienza del mondo, intenzionalità, né una nozione intrinseca di vero e falso paragonabile a quella che guida il ragionamento umano. Il suo funzionamento è, in senso stretto, statistico-probabilistico.

Semplificando, un LLM genera una risposta prevedendo, passo dopo passo, quale sequenza di testo sia più probabile in relazione all'*input* ricevuto e ai dati linguistici appresi durante l'addestramento. Sebbene il risultato finale possa apparire sofisticato, alla base vi è un meccanismo di previsione linguistica, e non un accesso diretto a fatti o significati. Tale previsione, tuttavia, non è un semplice "completamento automatico" nel senso banale del termine. I modelli contemporanei si basano in larga parte sull'architettura *Transformer*[1], che consente di trattare il linguaggio in modo fortemente contestuale: le parole e i segmenti di testo vengono interpretati in relazione gli uni agli altri, tenendo conto del contesto in cui compaiono. È questo che permette al modello di mantenere il filo del discorso, adattare il registro, produrre testi coerenti e, spesso, sorprendentemente ben costruiti sul piano formale.

Ma fluidità e affidabilità non coincidono. Un testo ben scritto non è necessariamente un testo corretto; una risposta convincente non è necessariamente una risposta vera. Proprio perché ottimizzato per generare sequenze linguisticamente plausibili, un LLM può produrre affermazioni inesatte, riferimenti inventati o ricostruzioni formalmente impeccabili, ma prive di fondamento. Non si tratta di un difetto occasionale, bensì di una conseguenza del modo in cui questi sistemi operano.

Comprendere gli LLM significa, allora, tenere insieme due prospettive. Da un lato, bisogna riconoscerne la straordinaria efficacia pratica: pochi strumenti digitali recenti hanno mostrato una versatilità comparabile nel trattamento del linguaggio. Dall'altro, occorre evitare il doppio equivoco che, spesso, accompagna il dibattito pubblico: sopravvalutarli come se fossero agenti razionali dotati di comprensione piena del mondo, oppure svalutarli come meri giochi di prestigio linguistico privi di utilità reale. La verità sta in una zona intermedia e molto più interessante: gli LLM sono macchine statistiche estremamente potenti, capaci di operare sul linguaggio con risultati di enorme utilità, ma proprio per questo richiedono di essere comprese, governate e impiegate con prudenza.

Nelle pagine che seguono proveremo a chiarire come funzionano questi modelli, perché sono così efficaci, quali compiti possono supportare e quali limiti non devono essere ignorati. Solo a partire da una comprensione non ingenua del loro funzionamento sarà infatti possibile valutarne l'uso in modo consapevole.

2. La logica della previsione linguistica

Per comprendere che cosa faccia davvero un LLM conviene partire da un dato essenziale: un LLM non "scrive" nel senso umano del termine, ma "prevede". Ricevuto un *prompt* (*i.e.*, una domanda, un'istruzione, un frammento di testo), il modello calcola quale sia la continuazione linguisticamente più probabile e costruisce la risposta un passaggio dopo l'altro. È da questa operazione elementare, ripetuta con grande rapidità e su scala enorme, che nascono testi spesso fluidi, coerenti e sorprendentemente efficaci.

"Prevedere la parola successiva" può evocare un meccanismo rudimentale, simile a un pappagallo stocastico[2]. Ma proprio qui si misura la specificità degli LLM contemporanei. La previsione non avviene infatti su pochi elementi isolati, bensì all'interno di contesti ampi e attraverso rappresentazioni matematiche molto sofisticate delle relazioni tra i diversi segmenti del testo. Il modello non seleziona semplicemente una parola plausibile: ricostruisce, di volta in volta, una continuazione compatibile con il tono, il contenuto, il registro e la struttura dell'*input* ricevuto.

In termini più precisi, il sistema non opera sulle parole come unità naturali, ma su "*token*", cioè frammenti di testo che possono coincidere con parole intere, parti di parola, numeri o segni di punteggiatura. La risposta viene generata *token* dopo *token*. Ogni scelta restringe il campo delle possibilità successive e orienta l'andamento complessivo del testo. Ne deriva una dinamica cumulativa: ciò che il modello ha appena prodotto condiziona ciò che sarà probabile produrre subito dopo.

---

1. L'architettura "*Transformer*" è stata introdotta da A. Vaswani *et al.*, *Attention Is All You Need*, presentato a NeurIPS nel 2017, lavoro che ha segnato un passaggio decisivo nello sviluppo dei moderni modelli linguistici (https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf).

2. L'espressione "*stochastic parrot*" è divenuta celebre dopo il saggio di E.M. Bender *et al.*, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, pubblicato negli atti della Conferenza ACM FaccT del 2021, nel quale gli Autori mettono in guardia contro il rischio di attribuire ai modelli linguistici capacità di comprensione che eccedono il loro effettivo funzionamento statistico (https://dl.acm.org/doi/10.1145/3442188.3445922).

Questa logica spiega insieme la forza e il limite degli LLM. La loro efficacia dipende dal fatto che, grazie all'addestramento su grandi quantità di dati linguistici, hanno appreso regolarità di notevole complessità: associazioni lessicali, strutture sintattiche, forme argomentative, stili discorsivi, modi ricorrenti di organizzare l'informazione. Il limite, però, è altrettanto chiaro: tale competenza riguarda anzitutto la plausibilità del linguaggio, non la veridicità di ciò che viene generato. Un testo può essere ben costruito, convincente e perfino elegante, senza per questo essere corretto. È questo il punto da non perdere di vista. La qualità dell'*output* non dipende da una comprensione del mondo paragonabile a quella umana, ma dalla capacità di stimare quali sequenze linguistiche risultino più appropriate in un certo contesto. Il modello, in altri termini, non "sa" necessariamente ciò che afferma: produce enunciati che assomigliano a quelli che, nei dati di addestramento, tendono a comparire in contesti simili.

La previsione linguistica è, dunque, la chiave interpretativa decisiva degli LLM. Ne spiega l'efficacia, perché spiega la loro capacità di produrre testi fluidi e persuasivi, ma ne chiarisce anche il limite, ricordando che la qualità formale dell'*output* non coincide con la sua attendibilità. Il punto, allora, non è chiedersi se questi modelli "capiscano" davvero, ma imparare a usarli per ciò che sono: potenti strumenti di elaborazione del linguaggio, da integrare nel lavoro umano senza confondere l'assistenza con il giudizio.

3. L'addestramento su grandi *corpora* testuali

La capacità degli LLM di produrre testi plausibili dipende, anzitutto, dall'addestramento su enormi *corpora* testuali. Attraverso questa esposizione su larga scala, il modello non apprende il linguaggio come lo apprende un essere umano, cioè mediante esperienza, intenzione e comprensione del mondo, bensì mediante regolarità statistiche, ricorrenze lessicali, strutture sintattiche e forme discorsive. In breve, impara quali sequenze linguistiche tendono a comparire in determinati contesti e con quali probabilità.

Questo apprendimento non consiste nella semplice memorizzazione di testi, ma in una progressiva ottimizzazione della capacità di previsione. Il modello formula ipotesi sui *token* successivi e corregge ripetutamente i propri parametri interni in funzione dell'errore compiuto: l'addestramento consiste nel ridurre la distanza tra ciò che il modello prevede e ciò che nei dati linguistici di apprendimento compare realmente.

Conta, però, non solo la quantità dei dati, ma anche la loro varietà. Un *corpus* ampio ed eterogeneo consente al modello di sviluppare una competenza linguistica più flessibile, capace di adattarsi a registri e compiti diversi. Al tempo stesso, il modello eredita anche i limiti dei dati da cui apprende: errori, squilibri, semplificazioni, stereotipi.

A questa fase generale di apprendimento, il *pre-training*, si aggiunge spesso una successiva messa a punto, che aiuta a comprendere il comportamento dei modelli conversazionali. Con l'*instruction tuning*, infatti, il sistema viene raffinato su esempi di istruzioni e risposte, così da imparare non solo a continuare un testo in modo plausibile, ma anche a seguire richieste: riassumere, spiegare, tradurre, classificare, riscrivere, adottare un certo tono o rispettare un certo formato. Se il primo addestramento insegna al modello le regolarità del linguaggio, questa fase ulteriore gli insegna, più intuitivamente, a interagire meglio con l'utente.

4. Generazione del testo e variabilità delle risposte

Una volta completato l'addestramento, l'LLM entra nella fase di utilizzo operativo, spesso definita "inferenza". È in questo momento che il modello riceve un *input* e genera una risposta sulla base delle istruzioni ricevute, del contesto disponibile e delle probabilità associate alle possibili continuazioni del testo. Un elemento importante, anche sul piano pratico, è che le risposte degli LLM non sono necessariamente identiche a parità di domanda. Il sistema può formulare la medesima informazione in modi diversi, selezionare esempi differenti, enfatizzare aspetti non coincidenti. In presenza di richieste ambigue, incomplete o formulate in modo leggermente diverso, può anche orientarsi verso soluzioni differenti. Questa variabilità non è un'anomalia, ma una conseguenza diretta della natura probabilistica della generazione.

Il *prompt*, per questa ragione, assume un ruolo decisivo. Più è chiaro, circostanziato e ben contestualizzato, più restringe il campo delle continuazioni plausibili e orienta il modello verso un *output* pertinente. Viceversa, una richiesta vaga amplia lo spazio interpretativo e tende a produrre risposte generiche o disallineate rispetto alle aspettative.

5. Errori, *bias* e "allucinazioni"

Uno degli aspetti più insidiosi dei LLM è che l'errore, spesso, non si presenta come errore. A differenza di altri strumenti informatici, che segnalano con evidenza un malfunzionamento o restituiscono un risultato assente, un LLM può produrre una risposta

falsa, incompleta o infondata in una forma perfettamente scorrevole, grammaticalmente corretta e persino argomentata. È proprio questa dissociazione tra qualità formale e affidabilità sostanziale a renderne l'uso particolarmente delicato. Nel lessico corrente si parla di "allucinazioni" per indicare quei casi in cui il modello genera contenuti che non trovano riscontro nei dati reali o nelle fonti: fatti inesatti, citazioni inventate, riferimenti normativi inesistenti, passaggi logici plausibili, ma privi di base reale. Il termine è in parte metaforico, ma descrive bene un tratto strutturale: il modello è ottimizzato per produrre sequenze linguisticamente credibili, non per verificarne la veridicità.

La possibilità dell'allucinazione deriva infatti dalla logica stessa della generazione. Un LLM non "consulta" una fonte autorevole nel momento in cui risponde, a meno che non gli venga fornita; e non possiede un criterio interno che distingua, in modo affidabile, tra ciò che è vero e ciò che è soltanto verosimile. Quando il *prompt* richiede precisione fattuale, ma mancano elementi sufficienti nel contesto, il modello può colmare i vuoti producendo un testo coerente sul piano linguistico, ma scorretto sul piano informativo.

Un buon *prompt* può ridurre questo rischio, ma non eliminarlo. Una richiesta chiara, circostanziata e ben strutturata restringe lo spazio delle continuazioni plausibili e tende a produrre risposte più pertinenti; una richiesta vaga o ambigua, al contrario, amplia il margine di arbitrarietà. Tuttavia, anche con *prompt* ben formulati, resta fermo il limite di fondo: la plausibilità del linguaggio non coincide con l'attendibilità dei contenuti.

In ambito giuridico, questa caratteristica assume un rilievo particolare. Una risposta può apparire impeccabile nel tono, nella terminologia e perfino nella struttura argomentativa, ma contenere un riferimento normativo errato, una massima inesistente o una ricostruzione troppo semplificata del quadro interpretativo. Il rischio, allora, non è soltanto l'errore in sé, ma la sua persuasività. Proprio perché il testo prodotto "suona bene", l'utente può essere indotto ad abbassare la soglia critica con cui normalmente vaglierebbe una fonte. Gli errori degli LLM non si limitano alle allucinazioni in senso stretto. Possono manifestarsi anche come omissioni rilevanti, eccessi di generalizzazione, fraintendimenti del contesto, appiattimento delle sfumature, riproduzione di *bias* presenti nei dati di addestramento. Gli errori e le allucinazioni non rendono gli LLM inutilizzabili; mostrano, piuttosto, entro quali condizioni debbano essere usati. Il loro valore è reale quando servono a esplorare ipotesi, sintetizzare materiali, produrre bozze, suggerire piste di ricerca o assistere il lavoro. Diventano problematici quando il testo generato viene scambiato per una fonte affidabile in sé, o peggio, per una base sufficiente di decisione.

6. Adattamento al dominio e recupero di fonti esterne

Uno dei modi più comuni per rendere gli LLM più utili in contesti specialistici consiste nell'adattarli a un dominio. Un modello generalista, per quanto potente, resta infatti esposto a un limite evidente: è addestrato su testi eterogenei e non possiede, di per sé, una conoscenza affidabile, aggiornata e controllata di uno specifico settore. In ambiti come il diritto, la medicina o la finanza, questa genericità può tradursi in risposte plausibili, ma troppo vaghe, imprecise o non allineate alle fonti effettivamente rilevanti.

L'adattamento al dominio può avvenire in modi diversi. In alcuni casi si interviene sul modello con ulteriori fasi di addestramento su testi specialistici; in altri, sempre più spesso, si preferisce affiancargli un sistema di recupero di informazioni esterne. È in questo quadro che si colloca il cd. RAG (*Retrieval-Augmented Generation*), oggi tra le soluzioni più diffuse. L'idea è semplice: invece di chiedere al modello di rispondere solo sulla base di ciò che ha appreso durante l'addestramento, gli si consente di appoggiarsi a una base documentale esterna: norme, sentenze, linee guida, manuali, *policy* interne, archivi aziendali o altre fonti pertinenti. Quando riceve una domanda, il sistema cerca anzitutto i documenti più rilevanti e li fornisce al modello come contesto aggiuntivo, così che la risposta venga costruita a partire da materiali più vicini al dominio di interesse.

Da questo punto di vista, la RAG può essere intesa come una forma di specializzazione pratica. Non trasforma il modello in un "esperto" nel senso proprio del termine, ma ne orienta il comportamento entro un perimetro informativo più controllato. In luogo di una risposta costruita soltanto su regolarità linguistiche generali, si ottiene così un *output* maggiormente ancorato a documenti specifici, spesso più aggiornati e pertinenti.

Sarebbe, però, un errore considerare questa tecnica come una soluzione definitiva al problema delle allucinazioni. Il fatto che il modello riceva documenti esterni non garantisce, da solo, né che tali documenti siano selezionati bene, né che siano aggiornati, completi e coerenti, né che il modello li usi correttamente. Anche in un sistema di RAG restano aperti passaggi decisivi: le fonti devono essere curate, mantenute, organizzate, segmentate e, in una certa misura, anche compresse senza perdere contesto o significato.

In altri termini, i "fatti" su cui il sistema si appoggia non sono mai meri dati: sono materiali che devono essere selezionati, mantenuti e resi interrogabili in modo coerente. Una base documentale incompleta o mal governata può produrre risposte fuorvianti quanto un modello privo di supporto esterno. Inoltre, anche quando il recupero è adeguato, il modello deve ancora trasformare quei materiali in una risposta, e proprio in questo passaggio possono emergere omissioni, travisamenti o inferenze scorrette. La RAG non cancella, dunque, il carattere probabilistico della generazione: lo vincola, più realisticamente, entro un contesto informativo meglio controllato. Per questo rappresenta un miglioramento importante, ma non sufficiente: la qualità delle fonti, la loro manutenzione e il controllo umano restano decisivi.

7. LLM e decisione giuridica: supporto, non sostituzione

Proprio per queste ragioni, l'impiego degli LLM nel diritto deve essere ricondotto entro una logica di supporto, non di sostituzione del decisore umano. Il modello può assistere il professionista, ma non può prenderne il posto. La valutazione della rilevanza dei fatti, l'interpretazione delle fonti, il bilanciamento tra principi, la costruzione della motivazione e l'assunzione di responsabilità restano attività intrinsecamente umane, che richiedono competenza tecnica, sensibilità istituzionale e capacità critica.

Nel contesto giuridico, il rischio maggiore non è soltanto l'errore materiale, ma la delega acritica. Quanto più il testo generato appare preciso, ordinato e autorevole, tanto più cresce la tentazione di trattarlo come affidabile in forza della sua sola qualità formale. È invece necessario mantenere una distinzione netta: l'LLM può essere uno strumento di ausilio, ma non una fonte di decisione. Può aiutare a formulare ipotesi, organizzare materiali, velocizzare attività ripetitive, suggerire strutture argomentative e offrire una prima elaborazione del problema; ma la verifica delle fonti, il controllo della correttezza giuridica e la decisione finale devono rimanere affidati all'essere umano.

8. Considerazioni conclusive

I *Large Language Models* rappresentano una delle innovazioni più rilevanti dell'intelligenza artificiale contemporanea. La loro capacità di generare testi complessi e di assistere l'utente in una pluralità di compiti li rende strumenti di indubbia utilità anche nel settore giuridico. Ma la loro efficacia non deve indurre a sopravvalutarne la portata epistemica. Gli LLM non "conoscono" il diritto nel senso in cui lo conosce e lo interpreta il giurista; producono sequenze linguistiche sulla base di regolarità statistiche apprese da grandi *corpora* testuali e, talvolta, orientate da fonti esterne o da adattamenti di dominio.

Da ciò derivano, insieme, le loro potenzialità e i loro limiti. Da un lato, essi possono migliorare efficienza, accessibilità e capacità di trattamento dell'informazione; dall'altro, possono generare errori, allucinazioni, semplificazioni indebite e distorsioni che, in un ambito ad alta sensibilità istituzionale come quello giuridico, impongono particolare cautela. La questione, allora, non è scegliere tra entusiasmo e diffidenza, ma costruire un uso consapevole di questi strumenti: comprenderne il funzionamento, riconoscerne i margini di utilità, delimitarne i rischi.

Il principio metodologico che ne deriva è, in fondo, semplice. Gli LLM possono essere impiegati come strumenti di supporto all'attività giuridica, ma non come sostituti del giudizio, della responsabilità e della decisione umana. Ed è precisamente in questa distinzione tra potenza dello strumento e responsabilità dell'interprete che si gioca la possibilità di un impiego realmente maturo dell'intelligenza artificiale nel diritto.