14 giugno 2023 23:06

CLIAPI 5: Intelligenze Artificiali Generative Tematiche: come usarle

di Alessandro Pedone

In questo quinto articolo della serie “Capire L’Intelligenza Artificiale Per Investirci” (CLIAPI) desidero mostrare una delle evoluzioni più significative dell’uso dell’intelligenza artificiale che cambierà il nostro modo di utilizzare i siti internet. Al tempo stesso, desidero illustrare come utilizzare i modelli di intelligenza artificiale generativa (come Chat GPT) in modo proficuo, riducendo il più possibile la probabilità di acquisire informazioni errate e/o perdere tempo.

Per fare questo sarebbe molto utile capire alcuni concetti di funzionamento di questi strumenti, ma per non appesantire troppo la lettura, relegherò le questioni più tecniche alla fine dell’articolo. Partiamo subito affrontando la questione del futuro ragionevolmente prevedibile di questi strumenti e poi di come utilizzarli bene.

Il futuro del Web nell’era delle intelligenze artificiali generative

L’errore tipico che commettono quasi tutti quelli che non comprendono come funziona la tecnologia (in particolare quella informatica) è di giudicarla in base a quello che è oggi, ignorando che nel giro di pochi trimestri sarà una cosa completamente diversa.
Io ho avuto il privilegio di vivere l’avvento di internet da un osservatorio privilegiato (quello di editore e direttore responsabile di riviste nazionali dedicate agli sviluppatori di software) mentre si stava sviluppando. Alla fine degli anni ‘90, ho visto il convergere delle tecnologie legate all’IP (Internet Protocol) con le telecomunicazioni ed ho compreso che in futuro tutti avremmo avuto un dispositivo mobile perennemente connesso alla rete e che questo avrebbe cambiato drasticamente le nostre vite quotidiane.
La grande massa delle persone che vedevano le prime pagine del World Wide Web (come si chiamava allora) su un browser Netscape non riusciva a vedere quello che sarebbe diventata la rete oggi perché non riusciva a comprendere gli sviluppi: la trasformazione delle pagine testuali prima in immagini e video, poi in vere e proprie applicazioni software.

Accadrà la stessa cosa con l’intelligenza artificiale generativa.
L’uso che possiamo fare adesso di questi strumenti è estremamente primitivo (per questo il prossimo paragrafo si concentrerà nello spiegare come utilizzarle bene).

Man mano che la tecnologia si perfezionerà, l'intelligenza artificiale generativa diventerà la principale interfaccia attraverso la quale utilizzeremo tutte le funzioni che oggi passano per le pagine web. L'interfaccia attuale (fatta di menù, zone del video dove cliccare per attivare funzioni, caselle di input nelle quali scrivere del testo e più in generale tutto quello che utilizziamo con il mouse e la tastiera) progressivamente verrà sempre più confinata in ambiti specialistici (chi sviluppa software, chi scrive testi di professione, varie applicazioni molto specifiche, ecc.).
Così come oggi, la gran parte dei servizi internet non viene fruita attraverso i computer tradizionali, ma attraverso i dispositivi mobili (smartphone e tablet); in futuro, quando queste tecnologie che padroneggiano il linguaggio si perfezioneranno, semplicemente diremo alle macchine quello che vogliamo e queste lo comprenderanno con un livello di precisione sufficientemente affidabile da basarci la relazione uomo-macchina.
Per la quasi totalità dei compiti che oggi svolgiamo con i dispositivi mobili (ma anche per molti che oggi eseguiamo ancora con i computer tradizionali) non saranno più necessari né tastiera, né mouse, né menù, né bottoni da cliccare, ecc.

Questo è uno scenario che richiede qualche anno prima di realizzarsi completamente.
Parallelamente a questa rivoluzione informativa, vivremo la trasformazione dei siti web che forniscono informazioni. A breve, insieme alle intelligenze artificiali generative di tipo generalista (come Chat GPT, Bing, Google Bard, Claude di Antropic e molte altre) si svilupperanno intelligenze artificiali specializzate in specifici settori come lo sport, i viaggi, la salute, l'economia, la finanza, i più svariati settori legati al divertimento, ecc.

I siti web si trasformeranno radicalmente. L'interfaccia principale con la quale si cercheranno informazioni non sarà più la struttura della classica pagina web da ricercare sul motore di ricerca e poi raggiungere attraverso i vari link. Il modo principale con il quale si cercheranno informazioni sarà attraverso una intelligenza artificiale generativa selezionata in base alla reputazione dell'organizzazione che l'ha creata.

In Italia, il primo esempio di intelligenza artificiale specializzata nel settore della finanza personale è Kappa (http://www.k-finance.it), ma sono sicuro che presto ne nasceranno centinaia, se non migliaia, nei settori più disparati. Esattamente come - all'inizio di internet - i siti web nascevano come funghi, così nasceranno intelligenze artificiali generative specializzate in tutti i campi del sapere umano.

Come utilizzare le intelligenze artificiali tematiche
In questa prima fase, dal momento che la tecnologia è ancora molto primitiva, il rischio principale che si corre è quello di usare male questi strumenti.
E' normale che chi li usa inizialmente non capisca bene cosa ha senso e cosa non ha senso chiedere per ottenere una risposta utile.
Il concetto principale che è necessario stamparsi bene in testa è che questi strumenti **non sono affatto intelligenti**, a dispetto del nome. Sono essenzialmente delle **macchine statistiche.**
Tutto quello che fanno è **mettere in ordine le parole** (in realtà pezzi di parole, detti token) secondo una probabilità che è definita dall'interazione fra il modello ed il contesto. Per contesto si intendono le parole inserite in input nel modello.
E' utile sapere che il contesto non sono solo le parole scritte dall'utente. Ogni specifica interfaccia di intelligenza artificiale generativa aggiunge alle parole scritte dall'utente una serie di parole per migliorare la qualità della risposta. Inoltre, le parole prodotte dal modello stesso diventano parte del contesto per definire le probabilità che generano le parole successive.
Questo fa sì che se si pone la stessa domanda in momenti diversi a queste macchine si ottengono risposte diverse.
Un secondo concetto importante da capire è che queste macchine non estraggono informazioni nel senso in cui siamo abituati a pensare oggi. Non c'è una ricerca ed estrazione di informazioni da un enorme database, come pensano quasi tutti.
Queste macchine generano un testo originale. E' come se avessero un archivio infinito di parole sparpagliate, come se fossero mattoncini del lego, ed ogni volta ricostruissero tutto da zero.
Nel prossimo paragrafo, per chi vuole approfondire, proveremo a spiegare meglio il concetto degli embeddings per chiarire questo punto. Per chi non vuole approfondire, è sufficientemente capire che il testo prodotto da queste macchine è in un certo senso "inventato" su base probabilistica. E' assolutamente normale che **contenga errori.**
La cosa più problematica è che questi errori sono tutt'altro che evidenti, perché sono inseriti all'interno di frasi scritte molto bene e contenenti tante informazioni corrette. Quindi è molto facile prendere cantonate!

Compresi questi aspetti, al momento - mentre attendiamo che queste macchine si evolvano sempre di più, riducendo al minimo gli errori - il segreto è in primo luogo quello di sapere cosa NON chiedere.
Non ha alcun senso chiedere di fare calcoli, elaborazioni, comparazioni, previsioni, ecc.
Al momento, tutto quello che sanno fare queste macchine è mettere in ordine le parole sulla base di una probabilità che è data dalle parole iniziali e dalle probabilità inserite nel modello durante la fase di apprendimento.
Al contrario, se chiediamo di spiegare concetti, fornire definizioni, fare elenchi di informazioni per dare spunti o idee, sintetizzare o riformulare testi, tradurre, ecc. le risposte saranno con grande probabilità utili.
Al momento, dobbiamo pensare a Kappa (come a tutti questi LLM, Large Language Model), come a una enorme enciclopedia con cui possiamo “parlare”. Ad esempio, se chiediamo, come è stato fatto realmente, a Kappa: “Che cos’è la duration?” otteniamo una risposta che con elevata probabilità sarà utile perché rispondere correttamente a questa domanda richiede proprio l’abilità di mettere in ordine le parole in modo corretto.
Se invece chiediamo a Kappa, sempre come è stato fatto realmente: “Quanto occorre per avere una rendita di 2000 Euro al mese?”. La risposta che si potrà ottenere sarà nel migliore dei casi inutile, nel peggiore dei casi errati. Questo non soltanto perché la domanda non contiene tutte le informazioni necessarie per poter fare il calcolo (ad esempio se si richiede una rendita vitalizia, rendita certa, reversibile, età del percettore, ecc.) ma proprio perché queste macchine - al momento - **non sanno fare i calcoli.**
Arriverà - ne sono sicuro - il giorno in cui si integreranno gli LLM con pezzi di software che fanno calcoli, ricerche in database, ecc, ma al momento, se chiediamo cose che richiedono la capacità di fare calcoli o estrarre informazioni specifiche da basi di dati, otterremo delle risposte inutili o errate.
La cosa più grave è che queste macchine provano sempre a dare delle risposte e, talvolta, l’utente potrà credere che dietro ad esse ci siano dei calcoli. In realtà si tratta solo di un tentativo di mettere insieme le parole (ed i numeri trattati come parole) in modo il più statisticamente probabile.
Ad esempio è possibile chiedere a Kappa un elenco di ETF, ma bisogna essere consapevoli che Kappa non estrae queste informazioni da un database di ETF strutturato. Ricrea quel testo sulla base delle probabilità. Quando si generano informazioni così specifiche attraverso le probabilità, commettere errori è quasi una certezza. Andare a ricostruire uno specifico codice ISIN componendo i caratteri su base probabilistica è qualcosa di molto difficile. Spesso ci prende, ma è quasi un miracolo della tecnologia che lo faccia così spesso.
Informazioni come i costi degli strumenti, le statistiche relative alla composizione ecc. sono tutte informazioni che Kappa tenta di ricostruire, ma non bisogna assolutamente contare sull’affidabilità di queste informazioni.
Per questo Kappa ha integrato la possibilità di fare verifica con l’intelligenza umana.

Ad oggi, una serie di risposte che fornisce sono necessariamente imprecise.
Come abbiamo scritto nel paragrafo precedente, le cose evolveranno rapidamente.
Siamo sicuri che in futuro, attraverso l’interfaccia di Kappa, si riuscirà ad avere informazioni complesse che richiedono calcoli, estrazione di informazioni da database, comparazioni, grafici, ecc.
Al momento però è bene utilizzare Kappa per il tanto di buono che può dare e non chiedergli quello per il quale non è progettato.
Riassumendo in tre punti.
1 - Ci sono alcune domande che non ha senso fare a Kappa e sono tutte quelle che presuppongono fare dei calcoli o usare algoritmi tradizionali.
2 - Ci sono alcune domande che sono perfette per Kappa e sono tutte quelle che richiedono di mettere in ordine le parole. Si può pensare Kappa come un’enorme enciclopedia specializzata sui temi finanziari alla quale si possono direttamente fare delle domande.
3 - Ci sono, infine, delle domande che si pongono a metà fra queste due categorie e che può essere utile porre a Kappa, ma le risposte vanno prese con molta cautela. Si tratta di domande che non prevedono calcoli ma non sono neppure conoscenza pura e semplice: suggerimenti, spunti, consigli.

ui siamo nella “via di mezzo”. Il nostro consiglio è di fare questo genere di domande a Kappa, ma chiedere sempre la verifica umana.
Abbiamo riunito alcune risposte reali che Kappa ha dato nei passati giorni in questo articolo: Kappa risponde (a domande utili) per avere un’idea delle cose che ha senso chiedere. Nei prossimi giorni amplieremo questo elenco e ne faremo uno strumento sempre più utile.

Ma come funzionano?
Questo paragrafo è leggermente più tecnico e può essere saltato da chi non ha tempo o voglia di approfondire, lo

scopo è quello di approfondire il concetto che ho esposto sopra: “tutto quello che sanno fare queste macchine è mettere in ordine le parole sulla base di una probabilità che è data dalle parole iniziali e dalle probabilità inserite nel modello durante la fase di apprendimento”. Ma come fanno? E come fanno le intelligenze artificiali generative tematiche a far dare risposte diverse se utilizzano come base un motore generalista?
Dobbiamo dare per scontato che chi legge conosce il concetto di rete neurale, altrimenti l’articolo viene insopportabilmente lungo. Se non conoscete questo concetto vi rimando al primo articolo della serie nel quale l’ho spiegato in termini comprensibili ad una persona che non sa niente di programmazione né di informatica.
Se hai almeno un’idea approssimativa di cosa sia una rete neurale puoi leggere il resto. I Large Language Model (LLM) sono costituiti da vari livelli di reti neurali con miliardi di parametri addestrati con enormi quantità di testo.
In questi modelli, le parole (in realtà i modelli usano pezzi di parole, detti token, ogni volta che scrivo parole, tecnicamente dovrei scrivere token, ma complicherebbe inutilmente la spiegazione) sono rappresentate da un insieme molto grande di numeri, circa 300 per ciascuna parola. Ciascuno di questi numeri esprime quanto è probabile che quella parola sia vicina ad un’altra sotto un certo aspetto. Il concetto di vicinanza infatti, non è da esprimersi solo in senso fisico. E’ più ampio e può riguardare una vicinanza in termini di svariate categorie utili per costruire una frase in modo sensato. Queste caratteristiche riguardano l’aspetto sintattico, non semantico. Riguardano, cioè, il modo in cui la parola può essere disposta correttamente all’interno della frase e come, statisticamente, si lega ad altre parole con la quale in genere è associata.
Questo insieme di numeri prende il nome di **embedding**. Le parole inserite in ingresso in un LLM sono tradotte in embedding e date in pasto alla rete neurale. Sulla base di tutti questi numeri, il modello restituisce la parola successiva più probabile, considerate tutte le relazioni che legano i vari embedding e sulla base dei pesi dei miliardi di nodi che compongono il modello.
Compreso questo concetto degli embedding, ci domandiamo adesso come sia possibile fare in modo che un LLM generalista diventi più “specializzato” e si trasformi in una intelligenza artificiale generativa tematica.
Questo avviene lavorando su tre livelli:
1) cambiando i pesi nei nodi più superficiali dell’LLM che si utilizza con un procedimento che prende il nome di “fine-tuning”;
2) creando un database di embedding specifico sulla base dei contenuti che si vuole personalizzare;
3) lavorando sulla “personalità” ovvero modificando il contesto al quale è associato l’input dell’utente.

Il primo modo è il più costoso in termini di risorse. Andando a cambiare i pesi del modello in sostanza si crea un modello proprietario. Si cambiano le probabilità con le quali il modello compone i testi dati gli stessi embedding in input.
Il secondo, che si può usare in combinazione con il primo, non lavora sul modello ma sull’input che si da al modello stesso. E’ come se si creasse un gergo che utilizza le stesse parole (token) utilizzate dal modello, ma assegnando ad esse caratteristiche ricavate dai dati con i quali il modello è personalizzato.
Il terzo piano sul quale si può lavorare è quello di aggiungere alla domanda che pone l’utente un contesto che orienta le risposte del modello.
L’ideale, ovviamente, è lavorare contemporaneamente sui tre livelli.
In questo modo una intelligenza artificiale generativa tematica fornirà risposte decisamente diverse, ed in modo consistente, tutte le volte che le domande fanno riferimento agli embedding “proprietari”.
Questo non significa che il modello fa una ricerca nel database proprietario ed estrae, con un copia&incolla, le informazioni. Si tratta sempre di generare le risposte su base probabilistica, ma grazie alla combinazione dei tre fattori sopra elencati, le probabilità che dia risposte in linea con quelle desiderate è molto più elevata rispetto ai modelli generalisti.
Compreso il funzionamento di queste macchine, appare evidente come sia molto più utile utilizzare intelligenze artificiali generative tematiche rispetto a quelle generaliste, se le domande che vogliamo porre al sistema riguardano argomenti specifici.