ChatGPT e oltre: Viaggio nel mondo dell'intelligenza artificiale generativa e delle sue potenzialità

Pietro Segreto

Abstract

Il 30 novembre 2022 OpenAI lancia ChatGPT. Il chatbot registra nel giro di pochi mesi un numero altissimo di interazioni convertendosi in un successo mondiale e riuscendo nell'impresa di diffondere in modo pervasivo il concetto di intelligenza artificiale generativa nel dibattito pubblico. Ma che cosa si intende effettivamente per intelligenza artificiale generativa e che architettura tecnologica sottende uno strumento come ChatGPT?

Il presente contributo prende le mosse da questi interrogativi per esplorare le potenzialità offerte dall'uso di modelli di IA generativa dando uno sguardo ai temi più caldi della ricerca scientifica recente sul tema. L'obiettivo ultimo è offrire una panoramica scientificamente aggiornata che permetta di orientarsi in un campo nuovo e fluido come quello attuale, in cui le innovazioni tecnologiche si affastellano a un ritmo così alto da rendere difficile anche agli addetti ai lavori restare aggiornati.

La prima parte dell'articolo inquadra il tema, chiarendo il concetto di intelligenza artificiale generativa e concentrandosi nello specifico sui Large Language Models e la loro architettura. Lungi dal voler entrare nelle specifiche statistiche e ingegneristiche, l'intento è aiutare ad acquisire un'intuizione sul suo funzionamento, per poter riflettere in modo informato su potenzialità e rischi.

Successivamente, l'articolo analizza i temi più interessanti dell'attuale ricerca scientifica sull'IA generativa. Nello specifico, parleremo di *Retrieval Augmented Generation* (RAG), che migliora la generazione di testo attraverso l'accesso a informazioni aggiornate e contestuali, e *Agentic AI*, che introduce agenti intelligenti capaci di prendere decisioni autonome.

Chiude il contributo una riflessione sul significato dell'intelligenza artificiale generativa e le sue potenzialità per la gestione e la lettura dei dati di un ente o un'azienda, cercando, alla luce di quanto esposto prima, di indicare possibili vie di sviluppo e integrazione di questa tecnologia.

Parole chiave: Intelligenza artificiale generativa, ChatGPT, Large Language Models, Retrieval Augmented Generation, Agentic AI.

On 30 November 2022, OpenAI launched ChatGPT. The chatbot registers a very high number of interactions within a few months, becoming a worldwide success and succeeding in the feat of pervasively spreading the concept of generative artificial intelligence in the public debate. But what is actually meant by generative artificial intelligence and what technological architecture underlies a tool like ChatGPT? This contribution takes these questions as a starting point in order to explore the potential offered by the use of generative AI models by taking a look at the hottest topics of recent scientific research on the subject. The ultimate goal is to offer a scientifically up-to-date overview that allows one to find one's way in such a new and fluid field, in which technological innovations are piling up at such a high pace that even insiders find it difficult to stay up-to-date. The first part of the article frames the topic, clarifying the concept of generative artificial intelligence and focusing specifically on Large Language Models and their architecture. Far from wishing to go into statistical and engineering specifics, the intention is to help gain an insight into how it works, in order to be able to reflect in an informed manner on its potential and risks. Next, the article analyses the most interesting topics of current scientific research on generative AI. Specifically, we discuss Retrieval Augmented Generation (RAG), which improves text generation through access to up-to-date and contextual information, and Agentic AI, which introduces intelligent agents capable of making autonomous decisions. The contribution closes with a reflection on the significance of generative artificial intelligence and its potential for managing and reading the data of an entity or a company, seeking, in the light of the foregoing, to indicate possible avenues for the development and integration of this technology.

Keywords: Generative Artificial Intelligence, ChatGPT, Large Language Models, Retrieval Augmented Generation, Agentic AI.

Introduzione

L'arrivo dell'intelligenza artificiale generativa, segnata in modo emblematico dal lancio di ChatGPT nel novembre 2022[1], ha innescato una corsa all'innovazione tecnologica che ha toccato industria e servizi di molti Paesi. ChatGPT, con la sua capacità di generare testi coerenti e convincenti su una vasta gamma di argomenti, ha reso tangibile il potenziale dei Large Language Models (LLM), modelli di intelligenza artificiale in grado di processare e imitare testi in linguaggio naturale[2]. L'impatto è stato immediato, pur sollevando una serie di interrogativi non solo sulla loro capacità di svolgere compiti complessi, ma anche sulle limitazioni intrinseche e sulle implicazioni etiche della loro adozione su larga scala. Questi strumenti, infatti, promettono di rivoluzionare il mondo del lavoro in molti settori produttivi e culturali; ma, pongono importanti questioni riguardanti la qualità dell'informazione, la creatività umana e il futuro del lavoro stesso in un mondo sempre più automatizzato.

L'intelligenza artificiale generativa è un campo in rapida evoluzione che si concentra sulla creazione automatica di contenuti, sfruttando tecniche avanzate di deep learning e grandi quantità di dati. È un sottodominio dell'intelligenza artificiale che include sistemi in grado di generare testi, immagini, video e persino musica, spesso con una qualità che può competere con quella dei creatori umani. I modelli sono addestrati su enormi dataset (insiemi di dati) che consentono loro di 'apprendere' le strutture e le sfumature dei dati forniti come input, costruendo una rappresentazione interna che gli permette di riprodurre quanto appreso nei loro output. Per comprendere appieno l'impatto e il funzionamento dell'IA generativa, è essenziale esplorarne i fondamenti tecnici e teorici. Nel corso dell'articolo vedremo nel dettaglio il funzionamento dei Large Language Models (LLM), il cui scopo è la generazione di

1 Samantha Lock, *What is AI chatbot phenomenon ChatGPT and could it replace humans?* «The Guardian», 5 dicembre 2022, <https://www.theguardian.com/technology/2022/dec/05/what-is-ai-chatbot-phenomenon-chatgpt-and-could-it-replace-humans> (Ultima consultazione: 23 settembre 2024).

2 *What are large language models (LLMs)?* «IBM», 27 September 2023, <https://www.ibm.com/topics/large-language-models> (Ultima consultazione: 27 settembre 2024).

testo[3]. I LLM si basano su architetture complesse, i transformer: introdotti per la prima volta nel 2017[4], hanno superato i modelli precedenti grazie alla loro capacità di gestire in modo efficace e scalabile contesti linguistici lunghi e complessi, utilizzando meccanismi di attenzione che permettono di focalizzarsi su parti specifiche dell'input durante l'elaborazione. Ciò ha reso possibile la generazione di testi più coerenti e attenti al contesto rispetto ai metodi tradizionali. L'addestramento dei LLM avviene su enormi dataset che possono comprendere miliardi di parole, e richiede una potenza computazionale significativa. Una volta addestrati, questi strumenti possono essere utilizzati in una vasta gamma di applicazioni, dalla scrittura creativa alla traduzione automatica, fino alla generazione automatica di codice. Tuttavia, nonostante i loro successi, i LLM presentano anche limiti, come la tendenza a produrre output non sempre accurati o pertinenti, e la riproduzione dei bias nei dati di addestramento.

Oltre alle architetture di base, l'attuale ricerca scientifica ha portato a una serie di sviluppi innovativi che ne stanno espandendo le capacità e le applicazioni. Il *Retrieval Augmented Generation* (RAG) è uno degli approcci più interessanti e sul quale si concentra molta attenzione da parte di aziende e ricercatori[5]. La forza del RAG, come vedremo più avanti, sta nella possibilità di ancorare le risposte dell'IA a una knowledge base strutturata e personalizzata con i propri dati, permettendo la generazione di output meno generici, più orientati al dominio di conoscenza dell'utilizzatore e quindi più affidabili. Questo approccio permette di migliorare la precisione e la rilevanza dei contenuti generati, soprattutto in contesti dove l'accuratezza dell'informazione è cruciale. Un altro ambito di ricerca è quello dei sistemi agentici (*Agentic AI*), che si basa sull'idea di costruire un team di agenti autonomi potenziati dall'IA generativa che sappiano lavorare insieme per la risoluzione di un problema. Costruire simili sistemi non è semplice, ma le potenzialità di sviluppo sono alte[6].

---

3 Esistono altri tipi di IA generativa per la produzione di output diversi, immagini, audio, video, ma non ne parleremo in questa sede.

4 Ashish Vaswani [et al.], *Attention Is All You Need.* 2017, <https://arxiv.org/pdf/1706.03762> (Ultima consultazione: 17 settembre 2024).

5 Patrick Lewis [et al.], *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. 2020, <https://arxiv.org/abs/2005.11401v4> (Ultima consultazione: 17 settembre 2024).

6 Andrew Ng, *The Dawning Age of Agents*. «The Batch», 6 marzo 2024, <https://www.deeplearning.ai/the-batch/the-dawning-age-of-agents/> (Ultima consultazione: 17

Machine learning e deep learning

Per iniziare diamo alcune definizioni. Quando parliamo di intelligenza artificiale ci riferiamo a un campo di studi multidisciplinare il cui obiettivo è costruire una macchina 'intelligente', che mostri quindi tratti associati all'intelligenza umana (ragionamento, creatività, emotività, ecc.). L'IA non è una novità, ma ci accompagna dalla metà del secolo scorso e negli ultimi anni ne abbiamo spesso fatto uso senza nemmeno rendercene conto[7].

Il campo di ricerca che ha avuto più successo è il machine learning, che ha beneficiato del grande sviluppo di componenti hardware e software nell'industria informatica fondamentali per costruire sistemi potenti e performanti in poco tempo e a costi contenuti, oltre all'esistenza di una vasta comunità di ricercatori e sviluppatori informatici che ha saputo fare dell'open source un proprio punto di forza, sia per la condivisione delle ricerche sia per la costruzione di ampi dataset fondamentali per l'addestramento delle macchine. L'obiettivo del machine learning è la definizione di algoritmi che permettano a una macchina di svolgere un problema individuando il modello matematico più opportuno alla risoluzione dello stesso e ottimizzando in modo automatico le risposte (output) tramite una fase di addestramento e valutazione degli output[8]. Un tipico esempio è il caso della regressione lineare. Si tratta di un potente e diffuso algoritmo utilizzato in contesti di machine learning supervisionato per predire il trend di un fenomeno. Nel caso del machine learning supervisionato, per la risoluzione del problema si raccolgono dati strutturati (o etichettati, dall'inglese *labeled*) in variabili indipendenti (x) e dipendenti (y). Immaginando di rappresentare i valori di x e y sull'asse delle ascisse e delle ordinate, l'obiettivo dell'algoritmo è quello di individuare la retta che meglio approssima i valori e che meglio rappresenta il trend del fenomeno che si vuole predire[9].

---

settembre 2024).

7 Per una breve storia dell'AI, si consiglia la lettura di Bruce G. Buchanan, *A (Very) Brief History of Artificial Intelligence*, «AI Magazine», 26 (2006), 4, pp. 53-60.

8 *What is Machine Learning (ML)?* «IBM», <https://www.ibm.com/topics/machine-learning> (Ultima consultazione: 17 settembre 2024).

9 Facciamo un esempio: immaginiamo di voler costruire un sistema che aiuti a predire il costo di un'automobile a partire dalle sue caratteristiche. Il primo step è quello di preparare un set di dati che raccolga il costo di diversi modelli automobilistici sulla base di alcuni parametri (casa di produzione, dimensioni dell'auto, potenza del

Una volta che in fase di valutazione il modello risponde sufficientemente bene, allora può essere utilizzato in contesti reali con un certo grado di affidabilità.

Un sottodominio del machine learning è noto con il nome di deep learning[10], la cui specificità risiede nell'utilizzo delle reti neurali. Si tratta di una tecnologia che risale anche alla metà del secolo scorso, ma che negli ultimi anni ha mostrato le sue potenzialità grazie ai passi avanti in termini di software e hardware necessari per la costruzione di questi sistemi. L'unità di base di una rete neurale è il neurone artificiale, che prende ispirazione dal funzionamento dei neuroni nel cervello umano. Un neurone artificiale opera la stessa operazione che abbiamo visto nell'esempio precedente descritto in nota: riceve degli input (x) e produce un output (ŷ, predizione di y) sulla base di una funzione f (chiamata, funzione di attivazione). Più neuroni uniti tra loro formano uno strato, più strati formano una rete neurale. È un tipo di tecnologia che funziona bene nella classificazione e raggruppamento dei dati ad alta velocità ed è efficace anche in contesti di apprendimento non supervisionato, in cui i dati di addestramento non sono strutturati in modo da avere per ogni input un output sulla base del quale modellare la risposta[11].

Non abbiamo fatto che veloci accenni a un campo molto più vasto, ma per quanto sintetici abbiamo ritenuto fossero utili per introdurre il paragrafo successivo in cui entreremo più nelle specifiche dei transformer, un'architettura informatica che ha segnato un grande passo in avanti nel campo del machine learning, aprendo la strada ai recenti

---

motore, accessori, ecc.) e a partire dal set si addestra il modello applicando una specifica funzione matematica (ŷ = f(x)) in modo che, dati questi parametri (valore x della funzione, input), si ottenga una predizione del costo (valore ŷ della funzione, output). Il compito della macchina, in fase di addestramento, è ridurre la discrepanza tra la predizione (valore ŷ della funzione) e il costo effettivo dell'auto (valore y della funzione, indicato nel set di dati).

10 Jim Holdsworth – Mark Scapicchio, *What is deep learning?* «IBM», 2024, <https://www.ibm.com/topics/deep-learning (Ultima consultazione: 17 settembre 2024)>.

11 Tornando al caso del costo dell'automobile, se volessimo ricorrere a un algoritmo di machine learning non supervisionato non avremmo bisogno di un dataset con valori di x (i parametri che nel nostro esempio definiscono il costo dell'auto) associati a y (il costo vero dell'auto), ma sarebbe sufficiente ricorrere a un set di dati con solo i parametri x. Ovviamente, cambia il tipo di problema: un simile approccio non serve per predire l'evoluzione di un fenomeno, ma per individuare pattern di correlazione non osservati dall'analisi umana.

exploit di ciò che viene definito intelligenza artificiale generativa. Sotto l'aggettivo 'generativa' si raccolgono sistemi diversi il cui compito è la generazione di contenuti (testo, audio, immagini, video), ma non in tutti i casi i transformer sono coinvolti. Eppure, in più di un'occasione l'IA basata su architettura transformer ha dimostrato di riuscire lì dove altre fallivano. Cerchiamo allora di capire perché.

LLM e Transformer

I Large Language Models (LLM) sono un punto di svolta nel campo dell'intelligenza artificiale, in particolare nell'ambito dell'elaborazione del linguaggio naturale. La loro architettura, basata sui transformer, gli consente di acquisire una profonda comprensione statistica delle strutture linguistiche. Durante l'addestramento, che avviene su larghissimi dataset testuali, i LLM apprendono come le parole e le frasi si relazionano tra loro in vari contesti, riconoscendo pattern e dipendenze che vanno oltre la semplice sequenzialità delle parole. Tale capacità permette loro di generare testo grammaticalmente corretto, spesso coerente e contestualmente appropriato.

Ciò che rende difficile scrivere per una macchina è innanzitutto il fatto che deve gestire dati in sequenza: il linguaggio naturale è un fenomeno temporale e sequenziale, in cui uniamo unità morfologiche e sintattiche in un certo ordine fondamentale per la comprensione della parola e della frase. Ma noi esseri umani non ci fermiamo qui, costruiamo anche discorsi sulla base di queste unità. E per poter costruire un discorso coerente è necessario tenere in memoria il contesto rappresentato dall'insieme di quanto detto nelle frasi che lo compongono. A complicare il tutto, aggiungiamo il fatto che il linguaggio naturale è plastico, ogni grammatica, con le sue regole, non è che un tentativo di scattare una fotografia che non è rappresentativa del tutto. Eppure, ciò non ci impedisce di capirci l'un l'altro.

Se agli albori della ricerca nel *Natural Language Processing* (il campo dell'intelligenza artificiale il cui compito è usare il machine learning per permettere alle macchine di 'comprendere' il linguaggio naturale) si è prevista la possibilità di istruire una macchina definendo un insieme di regole prestabilito che permettesse a un chat di interagire in modo convincente, presto ci si è resi conto che una simile strada non teneva in considerazione proprio la plasticità della lingua. La prima tecnologia in

grado di affrontare meglio la sequenzialità del linguaggio e a mostrare una capacità di memoria più elevata di altre soluzioni passate è quella delle reti neurali ricorrenti (*Recurrent Neural Networks*, RNN), utilizzate per problemi di tipo ordinale o temporale, che riescono a prelevare informazioni dagli input precedenti per influenzare input e output attuali. Ma nemmeno le RNN riescono a essere efficaci in contesti testuali che superino la dimensione di qualche frase[12].

Il transformer, che ha fatto la sua comparsa con il contributo di Vaswani [et al.], *Attention is all you need* nel 2017, ha superato questi limiti grazie al meccanismo di *self-attention*, che consente al modello di calcolare l'importanza relativa di ogni parola rispetto alle altre all'interno della sequenza. In pratica, si assegna un peso diverso a ciascuna parola, a seconda della sua rilevanza nel contesto generale. Ciò è utile per catturare le dipendenze a lungo raggio, ad esempio, come un pronome alla fine di una frase si riferisca a un nome menzionato all'inizio. Un simile approccio ha reso possibile l'elaborazione di sequenze più lunghe e complesse in modo efficiente, aprendo la strada a sistemi più potenti e scalabili.

Il transformer descritto nel contributo *Attention is all you need* riprende l'architettura encoder-decoder già applicata in passato per soluzioni di intelligenza artificiale il cui scopo fosse la traduzione automatica di un testo. L'encoder elabora l'intera sequenza di input (la frase in una lingua di partenza) e ne genera una rappresentazione interna o 'codifica'. Successivamente, il decoder utilizza la rappresentazione per generare la sequenza di output (la traduzione della frase nella lingua di destinazione).

Vediamo però i principali componenti dei transformer.

1. I sistemi di machine learning sono potenti macchine statistiche. In quanto tali, lavorano con numeri. Il primo passaggio, quindi, è quello di convertire le parole di una sequenza in numeri che la macchina possa processare. Questo è quanto avviene nel processo di tokenizzazione, in cui a ogni parola viene associato un numero identificativo (token) che ne rappresenta la posizione all'interno del dizionario che contiene tutte le parole che il modello può utilizzare. Esistono

12 Per un approfondimento non specialistico sul tema si consiglia la lettura di Melanie Mitchell, *L'intelligenza artificiale. Una guida per esseri umani pensanti*, Torino: Einaudi, 2022, pp. 177-234.

approcci diversi, dal momento che un token può corrispondere a una parola intera o a una sua parte (ad esempio, a un morfema)[13].
2. I token passano attraverso l'unità di *embedding,* in cui a ognuno di loro viene attribuito un insieme di valori numerici (vettori) che ne permettono la rappresentazione all'interno di uno spazio geometrico che ne approssima il significato in relazione agli altri token. Semplificando, il significato di una parola viene rappresentato come l'area di una figura geometrica; due concetti simili o strettamente correlati dal punto di vista semantico, saranno più vicini anche nello spazio geometrico che li rappresenta[14].
3. Dal momento che il transformer lavora in parallelo (i singoli token non vengono trasmessi come input uno dopo l'altro, ma tutti insieme), per consentire al modello di comprendere l'ordine delle parole nella sequenza originaria, si aggiungono ai valori di embedding altri valori numerici che permettono di preservare l'informazione. Questo processo si chiama *positional encoding*, ed è fondamentale per mantenere il contesto e la coerenza nel testo generato.
4. La somma dei vettori generati dall'*embedding* e dal *positional encoding* passa all'unità di *self-attention*, che analizza le relazioni tra i token, prestando attenzione ad aspetti diversi dell'input. Il processo si ripete in parallelo perché ci sono più unità che lavorano insieme a più teste (*multi-head attention*). Ogni testa analizza un aspetto diverso del linguaggio: ciò non avviene esplicitando che cosa analizzare, ma è il modello stesso che, sulla base di quanto imparato in fase di addestramento, assegna a ognuna di loro un ruolo. Le unità di *self-attention* sono presenti sia nell'encoder e che nel decoder.

Ricapitoliamo adesso il flusso di lavoro che porta alla generazione del testo. Si inizia da una sequenza di input che viene convertita in token, al quale si aggiungono valori numerici tramite i processi di *embedding* e *positional encoding* per preservare il significato della sequenza a livello semantico e sintattico. I dati passano all'unità di *self-attention*

13 Per una comprensione immediata dei concetti di token e di tokenizzazione, si consiglia di accedere alla seguente piattaforma di OpenAI (https://platform.openai.com/tokenizer) in cui, selezionando il modello che si preferisce e inserendo una frase, il sistema mostra la suddivisione in token.

14 Il modo più semplice di afferrare questi concetti è visualizzarli. L'Embedder Projector (https://projector.tensorflow.org/) visualizza su tre dimensioni gli embedding di alcuni dataset, così è possibile esplorarne i nodi e le relazioni.

dell'encoder che elabora una rappresentazione dell'input; l'output dell'encoder viene trasmesso al decoder per influenzare la generazione del testo. Il compito del decoder è predire la sequenza di output che meglio risponde alla sequenza di input. Il decoder viene inizializzato da un token di inizio frase (*Start of Sequence*, SOS), che supera i livelli di *embedding* e *positional encoding*, per passare alle unità di *self-attention* del decoder. Qui si calcola la probabilità che ogni singola parola del vocabolario usato dal sistema sia valida come output. La parola con la probabilità più alta diventa il nuovo input del decoder, e si continua così fino alla generazione dell'intera sequenza di output.

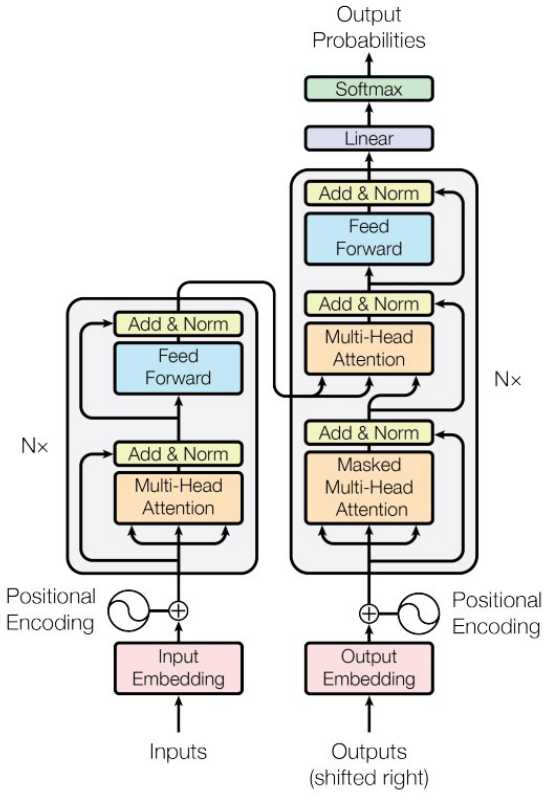

Figura 1 - Rappresentazione dell'architettura Transformer (l'immagine è tratta dall'articolo di Vaswani [et al.], *Attention is all you need*, cit., p. 3).

Esistono varianti dell'architettura transformer che utilizzano solo la componente encoder o solo la componente decoder. Un esempio di modello solo encoder è BERT[15] (*Bidirectional Encoder Representations from Transformers),* sviluppato da Google. BERT è progettato per comprendere il contesto bidirezionale in un testo, il che significa che analizza simultaneamente il contesto delle parole sia da sinistra a destra che da destra a sinistra, rendendolo efficace per compiti di comprensione del linguaggio come la classificazione del testo e l'analisi del sentiment.

Al contrario, GPT[16] (*Generative Pre-trained Transformer*) di OpenAI è un esempio di modello solo decoder. GPT è ottimizzato per la generazione di testo, partendo da un prompt e continuando a generare parole in una sequenza. Il focus sul decoder rende GPT efficace per compiti come il completamento del testo, dove la generazione di output coerente e continuo è essenziale.

Il processo di addestramento dei Transformer è intensivo dal punto di vista computazionale e richiede enormi quantità di dati e risorse. Inoltre, ci sono tecniche diverse a seconda del tipo di architettura che si sceglie.

Ad esempio, BERT è addestrato utilizzando una tecnica chiamata *Masked Language Modeling* (MLM), dove alcune parole in una frase sono mascherate e il modello deve predirle basandosi sul contesto rimanente. È una tecnica che consente di imparare il significato delle parole nel loro contesto bidirezionale, migliorando la capacità di comprendere il linguaggio naturale.

GPT, invece, utilizza un approccio di *Causal Language Modeling* (CLM), dove si addestra a predire la parola successiva in una sequenza basandosi su tutte le parole precedenti. È un approccio usato per compiti di generazione di testo, poiché il modello impara a generare sequenze di parole in modo fluido e coerente.

15 Per maggiori informazioni si può consultare la scheda disponibile su HuggingFace, <https://huggingface.co/docs/transformers/model_doc/bert> (Ultima consultazione: 30 settembre 2024).

16 Per maggiori informazioni si può consultare la scheda disponibile su HuggingFace, <https://huggingface.co/docs/transformers/model_doc/openai-gpt> (Ultima consultazione: 30 settembre 2024).

I prompt

L'utente può interagire con i LLM attraverso i prompt, che corrispondono alle istruzioni fornite, attraverso le quali l'IA interpreta ciò che l'utente desidera. I prompt non solo attivano il modello, ma hanno anche la capacità di orientare e influenzare la qualità del contenuto generato. Un prompt ben formulato può indirizzare alla generazione di risultati non solo accurati, ma anche rilevanti e contestualmente appropriati, mentre un prompt vago o mal strutturato può portare a risposte poco utili o addirittura fuorvianti.

La scrittura efficace dei prompt, conosciuta come *prompt engineering*, è una disciplina che sta acquisendo sempre più rilevanza. Questa pratica consiste nel formulare input che possano massimizzare le prestazioni del modello, ottimizzando la chiarezza e la precisione delle risposte generate. La progettazione del prompt richiede una comprensione approfondita di come il sistema interpreta l'input[17] e delle dinamiche che influenzano la generazione di output. In pratica, il *prompt engineering* non significa solo dare comandi, ma strutturare quei comandi in modo da guidare il modello verso risultati ottimali, influenzando quindi la distribuzione statistica dell'output. Ad esempio, un prompt potrebbe includere istruzioni dettagliate su come rispondere, specificando lo stile di scrittura, il livello di dettaglio e persino il tono desiderato. Questo approccio può essere utile in scenari dove è necessario produrre contenuti altamente specifici o in contesti professionali dove la precisione è fondamentale.

Ci sono diversi approcci al *prompt engineering*[18].

17 Ad oggi, gli input che possiamo fornire a un LLM non sono solo testuali, ma anche visivi (immagini e foto) e sonori (registrazioni audio). I LLM sono diventati sistemi multimodali. A titolo di esempio (con una certa cautela perché si tratta comunque di demo) si consiglia la visione delle seguenti presentazioni per avere un'idea delle loro potenzialità: OpenAI, *Live demo of GPT4-o voice variation*. «YouTube», 2024, <https://www.youtube.com/watch?v=D9byh4MAsUQ&ab_channel=OpenAI> (Ultima consultazione: 27 settembre 2024); Google, *Project Astra: Our vision for the future of AI assistants*. «YouTube», 2024, <https://www.youtube.com/watch?v=nXVvvRhiGjI&ab_channel=Google> (Ultima consultazione: 27 settembre 2024).

18 Per una review esaustiva si consiglia di fare riferimento a Sander Schulhoff [et al.], *The Prompt Report: A Systematic Survey of Prompting Techniques*. 2024, <https://arxiv.org/html/2406.06608v1#Ch2.S4> (Ultima consultazione: 17 settembre 2024).

- *Zero-shot learning*: è il primo livello di prompting, in cui si richiede di produrre un output solo esplicitando nel prompt l'obiettivo che si vuole raggiungere. È un approccio utile quando non è possibile fornire esempi o quando si vuole risolvere problemi in contesti nuovi o inesplorati. Tuttavia, lo *zero-shot learning* richiede che il prompt sia formulato con estrema chiarezza e precisione, poiché il modello può solo fare affidamento sulla sua comprensione generale del linguaggio, acquisita in fase di addestramento.
- *One-shot learning* e *few-shot learning:* la qualità nella risposta di un LLM migliora lì dove si inserisca all'interno del prompt uno o più esempi da seguire per la generazione dell'output[19]. Quindi, se ad esempio si vogliono generare descrizioni di prodotti, il prompt potrebbe includere due o tre esempi di descrizioni ben fatte prima di chiedere una nuova descrizione per un prodotto specifico. Questo aiuta il LLM a capire cosa ci si aspetta da esso, migliorando la coerenza e la qualità della risposta. Si parla allora di *In-Context Learning* perché il modello risolve il problema grazie alle informazioni contenute nel contesto della conversazione.
- *Chain of Thought*: è un approccio di diverso tipo, in cui la sfida da superare è riuscire a far leva sulle emergenti capacità di ragionamento del LLM, scomponendo un problema complesso in parti semplici e accompagnando il modello nella risoluzione dei singoli step.

La scrittura dei prompt richiede competenza, perché non si basa solo su una comprensione tecnica del funzionamento della macchina, ma anche su una sensibilità linguistica nel formulare richieste chiare, concise e orientate agli obiettivi desiderati (per quanto ci siano ad oggi diverse ricerche che dimostrano come i LLM siano capaci di scrivere prompt per se stessi e ottimizzarli[20]). Inoltre, un prompt strutturato in un modo ben definito può limitare la variabilità nella risposta in termini qualitativi.

19 Tom B. Brown, *Language Models are Few-Shot Learners*. 2020, <https://arxiv.org/abs/2005.14165> (Ultima consultazione: 23 settembre 2024).

20 Chrisantha Fernando [et al.], *Promptbreeder: Self-Referential Self-Improvement Via Prompt Evolution*. 2023, <https://arxiv.org/abs/2309.16797> (Ultima consultazione: 17 settembre 2024); Chengrun Yang [et al.], *Large Language Models as Optimizers*. 2023, <https://arxiv.org/abs/2309.03409> (Ultima consultazione: 17 settembre 2024).

La continua ricerca in questo campo ha portato alla creazione di tecniche e strumenti sempre più raffinati, che evidenziano l'importanza di un approccio strutturato e metodico alla formulazione dei prompt. A fronte di modelli che aumentano in modo evidente il numero di token disponibili per prompt[21], la sfida è quella di avere LLM capaci di risolvere task specifici senza il bisogno di intervenire sui parametri di addestramento (*fine-tuning*) ma soltanto facendo leva sul prompting. Un esempio di ciò è descritto nell'articolo di Nori [et al.] (2023)[22], *Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine*, in cui unendo insieme più tecniche di *prompt engineering* si è arrivati alla definizione di un prompt (Medprompt) in grado di potenziare le capacità insite in GPT-4 nella risoluzione di test di medicina, con un tasso di errore inferiore a modelli customizzati.

Campi di ricerca

La ricerca nell'ambito dell'intelligenza artificiale generativa sta accelerando a un ritmo vertiginoso, aprendo nuove strade che promettono di ampliarne le capacità e le applicazioni. Tra le innovazioni più promettenti emergono il *Retrieval Augmented Generation* (RAG) e l'*Agentic AI*.

RAG

Il *Retrieval Augmented Generation* (RAG)[23] è un approccio che combina le capacità di generazione del linguaggio naturale con la potenza del recupero di informazioni. Una soluzione ibrida che mira a migliorare la precisione, la rilevanza e la contestualizzazione dei contenuti

21 Si consideri il caso di Gemini Advanced, il modello di punta di Google, che ha ampliato la finestra contestuale a 1 milione di token e la possibilità di gestire caricamenti fino a 1500 pagine: <https://gemini.google/advanced/?hl=it> (Ultima consultazione: 17 settembre 2024).

22 Harsha Nori [et al.], *Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine*. 2023, <https://arxiv.org/abs/2311.16452> (Ultima consultazione: 17 settembre 2024).

23 Yunfan Gao [et al.], *Retrieval-Augmented Generation for Large Language Models: A Survey*. 2023, <https://arxiv.org/abs/2312.10997> (Ultima consultazione: 17 settembre 2024).

generati dall'IA, superando alcune delle limitazioni intrinseche dei modelli puramente generativi.

In un sistema RAG, il processo di generazione del testo non si basa esclusivamente sulla conoscenza interna del modello, che è stata acquisita durante l'addestramento su vasti dataset. Invece, il sistema ha la capacità di accedere dinamicamente a fonti esterne di informazioni durante la generazione del contenuto, grazie a un meccanismo di recupero di informazioni (*retrieval*) che gli permette di cercare e integrare dati pertinenti da database o documenti specifici. Ad esempio, quando viene posta una domanda complessa, la macchina può utilizzare il modulo di *retrieval* per cercare in un archivio di articoli accademici, notizie o altre risorse rilevanti, recuperando le informazioni necessarie e poi utilizzandole per generare una risposta accurata e aggiornata.

L'architettura RAG si distingue per la sua capacità di combinare in modo sinergico due componenti chiave: un modello di *retrieval* e un modello generativo. Il modello di retrieval è responsabile per l'individuazione di documenti rilevanti o estratti di testo che possono rispondere alla query dell'utente. Una volta recuperate le informazioni, il modello generativo le utilizza come contesto per creare una risposta che non solo risponde alla domanda, ma lo fa in modo specifico e informato. Inoltre, il LLM è in grado di indicare l'esatta fonte utilizzata per la generazione della risposta, aprendo alla possibilità di verifiche.

Uno dei vantaggi principali del RAG è la sua flessibilità. Il sistema può essere aggiornato senza necessità di riaddestrarlo. Invece di dover incorporare tutte le nuove informazioni direttamente nel modello durante l'addestramento, è sufficiente aggiornare il database di informazioni da cui esegue il *retrieval*, mantenendo il tutto sempre allineato con le conoscenze più recenti, riducendo i costi e i tempi di aggiornamento.

Tuttavia, l'implementazione del RAG non è semplice. Una delle principali difficoltà è la selezione delle fonti di informazioni utilizzate per il *retrieval*: per generare contenuti affidabili e accurati, è essenziale che le fonti siano di alta qualità e prive di bias. La selezione errata delle fonti può portare alla generazione di contenuti fuorvianti o addirittura dannosi. Un'altra sfida è l'integrazione coerente delle informazioni recuperate: l'IA deve essere in grado di comprendere e combinare i dati provenienti da fonti diverse in modo che la risposta finale sia coesa e logica.

Inoltre, l'interazione tra il modello di *retrieval* e quello generativo deve essere attentamente calibrata per evitare che le informazioni

recuperate siano utilizzate in modo inappropriato o decontestualizzato. Questo richiede un'attenta progettazione dell'architettura del sistema e una formazione specifica per garantire che le risposte siano valide.

Agentic AI

Per chiarire il tema del paragrafo, partiamo da un esempio concreto di un'attività e immaginiamo che un autore voglia usare un LLM per la stesura di un testo da destinare a una pubblicazione. Di per sé, scrivere è un'attività complessa che può essere scomposta in un certo numero di step più semplici.

1. L'autore, definito l'argomento del testo, delinea una traccia dei temi che vuole affrontare.
2. Si documenta sull'argomento, cercando fonti attendibili a cui fare riferimento.
3. Sulla base delle informazioni reperite, se non ritiene necessario modificare la traccia, l'autore procede alla stesura della prima bozza.
4. L'autore legge la bozza per assicurarsi che da un punto di vista contenutistico e stilistico il testo scritto rispetti gli standard di qualità necessari.
5. Si apportano le modifiche individuate nella fase precedente.
6. Il processo di rilettura e correzione continua finché non si è soddisfatti dell'esito.

Immaginando adesso di voler utilizzare un LLM per svolgere il compito, ci si confronta con diversi livelli di complessità.

- Una prima soluzione, la più semplice, potrebbe essere quella di eseguire un'unica chiamata a un LLM per ottenere come output il testo che ci interessa, aggiungendo le informazioni rilevanti nel prompt.
- Un secondo livello di complessità consiste nell'eseguire più chiamate a un LLM, per ogni step che abbiamo qui individuato.
- A un terzo livello, potremmo lasciare che sia lo stesso LLM a decidere che azione intraprendere e che strumento (tool) usare per il suo svolgimento. Ad esempio, il LLM potrebbe richiamare la funzione di web search per individuare fonti da aggiungere al testo.

- A un quarto livello, il LLM esegue gli step e riproduce il processo in loop. Tuttavia, spetta ancora all'utente umano definire la sequenza degli step per la risoluzione del problema.
- L'ultimo livello prevede la possibilità che l'intero sistema operi autonomamente, assorbendo gli step precedenti (*Ai agent*).

Ognuno di questi livelli costituisce un esempio di architettura a complessità crescente[24], capace di affrontare task sempre più difficili. L'opzione dell'*AI agent*, quindi di un sistema che, data una knowledge base e un set di istruzioni, opera autonomamente e iterativamente per la risoluzione di un task, sfruttando le potenzialità del LLM, si sta affermando. La forza dell'approccio risiede nel fatto di spezzettare il problema in parti più semplici, utilizzando il LLM per la risoluzione dei singoli step, fino ad arrivare alla completa automazione del task.

Posto che costruire simili sistemi è tutt'altro che semplice e che quando parliamo di *Agentic AI* facciamo riferimento al momento più a uno spettro di architetture che operano a diversi livelli di complessità e autonomia, ci sono alcuni pattern di utilizzo la cui applicazione può facilitare la riuscita dell'operazione[25].

- *Reflection*: chiedere al LLM di valutare i suoi output, per evidenziare punti di forza e debolezza e individuare azioni utili per ottimizzare gli esiti delle chiamate.
- *Tool use*: fornire al LLM strumenti per la risoluzione di task (per navigare il web, per scrivere ed eseguire codice, ecc.).
- *Planning*: chiedere al LLM di pianificare gli step necessari alla risoluzione di un problema.
- *Collaborazione multiagente*: costruire un sistema con più agenti, che preveda la possibilità di orchestrare gli sforzi di ognuno di loro, proprio come fosse un team.

24 Altrove si fa riferimento al concetto di «architettura cognitiva», come descritto in *OpenAI's Bet on a Cognitive Architecture*. «Langchain blog», 28 novembre 2023, <https://blog.langchain.dev/openais-bet-on-a-cognitive-architecture/> (Ultima consultazione: 17 settembre 2024).

25 Andrew Ng, *Agentic Design Patterns Part 1. Four AI agent strategies that improve GPT-4 and GPT-3.5 performance*. «The Batch», 20 marzo 2024, <https://www.deeplearning.ai/the-batch/how-agents-can-improve-llm-performance/> (Ultima consultazione: 17 settembre 2024).

LLM e editoria

Abbiamo visto finora i *Large Language Models* dal punto di vista tecnico, spiegando in modo generale da dove vengono fuori, come funzionano e cosa possono fare. La domanda che però adesso dovremmo porci è che utilizzo si potrebbe fare di questi strumenti nella vita di tutti i giorni di una casa editrice, perché è ovvio che una macchina in grado di 'hackerare' il linguaggio naturale umano non può non essere di interesse per un'industria che di parole vive.

Per rispondere è utile dare uno sguardo a quanto è stato fatto finora. Il report *A third Transformation? Generative AI and Scholarly Publishing* pubblicato il 30 ottobre 2024 da Ithaka S+R[26], anche se focalizzato sull'editoria accademica, è utile per avere una panoramica aggiornata del contesto e individuare alcune direttive di sviluppo che sono comuni ad altri segmenti dell'industria editoriale. Il report si basa sulle osservazioni raccolte da interviste semi-strutturate nei confronti di dodici rappresentati scelti tra importanti editori e organizzazioni editoriali in senso lato (Ithaka S+R include in questa definizione anche fondazioni, società accademiche, biblioteche), selezionati a partire dalla loro preparazione sui principali trend che caratterizzano il mercato dell'editoria accademica, con un occhio attento nei riguardi dell'IA generativa. Il documento, come sottolineato dal titolo, si interroga sull'entità della trasformazione tecnologica che l'IA sta introducendo già adesso nell'industria editoriale accademica, con alcune riflessioni sugli aspetti strategici più sensibili e le opportunità che potrebbero aprirsi. Partiamo però da un chiarimento:

> Innovation in the generative AI space has been rapid over the past 24 months, and the pace of iteration has outstripped publishing organizations' abilities to adapt underlying business models. Individuals we spoke to across the sector acknowledged that organizations need to invest more time in understanding generative AI technology more deeply because it

26 Tracy Bergstrom – Dylan Ruediger, *A Third Transformation?: Generative AI and Scholarly Publishing*. 30 ottobre 2024, <https://sr.ithaka.org/publications/a-third-transformation/> (Ultima consultazione: 12 novembre 2024). Questo documento è da considerare come un addendum al più ampio report: Tracy Bergstrom – Oya Y. Rieger – Roger C. Schonfeld, *The Second Digital Transformation of Scholarly Publishing*. 29 gennaio 2024, <https://sr.ithaka.org/publications/the-second-digital-transformation-of-scholarly-publishing/> (Ultima consultazione: 12 novembre 2024).

> has the potential to upend a number of underlying systems in the future. At the time of writing, we did not learn of concrete examples of instances in which publishing organizations were utilizing generative AI tools in ways that substantially increased revenue in regard to their backend processes (Tracy Bergstrom – Dylan Ruediger, *A Third Transformation?,* cit., p. 24).

L'AI generativa si evolve molto rapidamente, è difficile tenere il passo e continua a cogliere alla sprovvista. Di fronte a uno scenario incerto, si fatica a individuare modelli di business affidabili. Al momento le strade intraprese sono due e di diverso tipo. Da un lato ci sono alcuni grandi gruppi editoriali che hanno operato scelte di chiusura e protezione. Ad esempio, a dicembre 2023 il New York Times ha citato in giudizio OpenAI e Microsoft accusandoli di avere utilizzato suoi contenuti protetti da copyright in modo fraudolento per l'addestramento di modelli[27]. Taylor & Francis e Elsevier, invece, hanno rilasciato nuove policy nelle quali si proibisce l'uso da parte degli autori di utilizzare l'AI generativa per scopi che potrebbero inficiare l'autorialità dell'opera[28]. Ma dall'altro lato, c'è chi ha scelto di stringere accordi commerciali per cedere i propri contenuti in licenza. Nell'ambito dell'editoria accademica è il caso di Wiley e (curiosamente) Taylor & Francis[29], ma anche grandi

27 Michael M. Grynbaum – Ryan Mac, *The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work*. «New York Times», 27 dicembre 2023, <https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html> (Ultima consultazione: 12 novembre 2024).

28 Taylor & Francis, *AI Policy*. 2024, <https://taylorandfrancis.com/our-policies/ai-policy/> (Ultima consultazione: 12 novembre 2024); Elsevier, *The use of generative AI and AI-assisted technologies in writing for Elsevier*. 2024, <https://www.elsevier.com/about/policies-and-standards/the-use-of-generative-ai-and-ai-assisted-technologies-in-writing-for-elsevier> (Ultima consultazione: 12 novembre 2024).

29 Christa Dutton, *Two Major Academic Publishers Signed Deals with AI Companies. Some Professors Are Outraged*. «The Chronicle of Higher Education», 2024, <https://www.chronicle.com/article/two-major-academic-publishers-signed-deals-with-aicompanies-some-professors-are-outraged> (Ultima consultazione: 12 novembre 2024); Kathryn Palmer, *Taylor & Francis AI Deal Sets 'Worrying Precedent' for Academic Publishing*. «Inside Higher Ed», 2024, <https://www.insidehighered.com/news/faculty-issues/research/2024/07/29/taylorfrancis-ai-deal-sets-worrying-precedent> (Ultima consultazione: 12 novembre 2024).

gruppi di testate giornalistiche come GEDI in Italia[30] e The Atlantic[31] negli Stati Uniti si sono mossi in questa direzione. L'idea di concordare una licenza di utilizzo ha il vantaggio di dare visibilità ai contenuti dell'editore sollevandolo dall'onere dello sviluppo di un'applicazione specifica e monetizzando l'uso dei propri dati. Detto ciò, diventa cruciale scegliere oculatamente quali contenuti cedere in licenza, per non dilapidare il proprio vantaggio strategico. Inoltre, è improbabile che questa scelta si riveli conveniente per editori medio-piccoli, che non possono vantare lo stesso peso contrattuale di un grande gruppo.

Di fronte alle promesse dell'IA generativa ci si muove con cautela, per i dubbi legati al copyright dei contenuti generati e per le criticità di uno strumento che per sua stessa natura soffre di allucinazioni. Nell'immediato le aziende (non solo in campo editoriale) sperimentano, introducendo internamente tool in modalità più sicure come ChatGPT Team o Enterprise[32], che mettono a disposizione servizi di IA generativa con maggiori garanzie di rispetto della privacy e gestione dei dati immessi in fase di inferenza. Avere tra le mani uno strumento che ti permette di interrogare documenti e fare ricerca su argomenti in modo discorsivo (superando il modello dei motori di ricerca), generare testi da questi documenti (abstract, riassunti, parole chiave), generare immagini in relazione ai contenuti della conversazione significa avere accanto un'assistente che potenzialmente accelera i tempi di lavoro e che mostra tutta la sua utilità in redazione, ma anche per l'ufficio stampa/marketing che dovrà promuovere la pubblicazione.

Tuttavia, se è vero che portare in azienda strumenti di questo tipo è un investimento interessante anche in ottica di formazione dei propri dipendenti, l'impatto dell'IA generativa su una casa editrice supera l'efficientamento dei flussi di lavoro interni per andarsi a inserire in una dinamica di sviluppo nuova. Sempre facendo riferimento al report di Ithaka S+R sull'editoria accademica:

30 GEDI, *OpeanAI | GEDI*. 26 settembre 2024, <https://www.gedi.it/it/media/sala-stampa/openai-e-gedi-annunciano-una-partnership-strategica-rendere-accessibili-contenuti> (Ultima consultazione: 12 novembre 2024).

31 The Atlantic, *The Atlantic announces product and content partnership with OpenAI*. 29 maggio 2024, <https://www.theatlantic.com/press-releases/archive/2024/05/atlantic-product-content-partnership-openai/678529/> (Ultima consultazione: 12 novembre 2024).

32 OpenAI, *Redefine work in the age of AI. 2024*, <https://openai.com/chatgpt/enterprise/> (Ultima consultazione: 12 novembre 2024).

> As we noted in our earlier report, scholarly publishing as a whole is in the midst of a long-term shift away from a model centered on editorial work towards one based on services and platforms. We expect generative AI to accelerate this trend: publishing organizations are already engaged in strategic planning about how to map generative AI services to support the workflows of readers, authors, and editorial staff (Tracy Bergstrom – Dylan Ruediger, *A Third Transformation?,* cit., p. 3).

Questa transizione da editore tradizionale a piattaforma erogatrice di servizi non riguarda solo l'editoria accademica, ma può interessare anche altri ambiti del mercato (scolastica, formazione professionale, manualistica, ecc.). Già da tempo gli editori hanno superato l'idea di offrire come prodotto il singolo il libro cartaceo o l'ebook, aggiungendo un corredo di contenuti aggiuntivi disponibili online. Con l'IA generativa i contenuti diventano non solo interattivi, ma anche 'intelligenti' (con tutte le limitazioni del caso, già precedentemente spiegate nel corso dell'articolo). È chiaro che in questo caso l'editore può anche non limitarsi a cedere i propri contenuti in licenza a terzi, ma essere diretto protagonista dello sviluppo delle applicazioni. Con un occhio a quanto viene fatto fuori, per non rischiare di investire su strumenti generici destinati a farsi schiacciare da tool messi a punto dai grandi colossi dell'intelligenza artificiale (OpenAI, Google, Amazon). Ecco allora che approcci come il RAG si dimostrano utili nell'ottica di costruire sistemi a partire dalla propria base di conoscenza per fare qualcosa di 'unico' e riconoscibile.

Conclusioni

I LLM sorprendono per la loro capacità di generare testo. Ma sono affidabili? Possono essere utilizzati per affrontare problemi complessi? Che grado di ragionamento mostrano nel loro operato? Sono un appariscente giocattolo o possono trovare un uso professionale?

Che i LLM dimostrino capacità di ragionamento è ancora una discussione aperta[33]. Posto che come sempre la difficoltà sta innanzitutto

33 Se da un lato c'è chi vede i LLM come potenti machine per il recupero di informazioni (Subbarao Kambhampati, *Can large language models reason and plan?*, «Annals of the New York Academy of Sciences», 1534 (2024), 1, pp. 15-18), dall'altro c'è chi vede in questi strumenti l'insorgere di nuovo sistema operativo, che in pochi

nel dare una definizione univoca e condivisa di un concetto così ampio, per comodità di discorso con la parola 'ragionamento' ci riferiamo in questo contesto all'insieme dei processi logici che ci permettono di affrontare un problema complesso, scomporlo nelle sue componenti più semplici ed eseguire gli step necessari alla sua risoluzione. Quindi, i LLM, posti di fronte a problemi complessi, non sembrano avere ancora le capacità di analisi necessarie per pianificare la loro risoluzione in modo concreto; e anche lì dove propongono una credibile scomposizione del problema in singoli step, non è affatto certo che riescano a eseguirli in modo adeguato fino alla risoluzione del problema stesso. I LLM, più che ragionare, sembrano estrarre informazioni dalle proprie rappresentazioni interne e adattare l'output sulla base delle informazioni di contesto fornite.

Per quanto possa sembrare limitante, non è affatto detto che al LLM servano capacità di ragionamento per essere utilizzabile. Ciò che soluzioni come il RAG o gli agenti dimostrano è la possibilità di affrontare i limiti del modello costruendo architetture ad ampio respiro in cui il LLM è solo una componente di una struttura modulare e ibrida. Così si sposta il focus d'attenzione dal LLM e dalla sua potenza di calcolo al problema che si desidera affrontare e al tipo di architettura necessaria per riuscire nell'intento: non è il LLM a dover affrontare il problema come una sorta di factotum digitale, ma è il sistema che gli si costruisce intorno ad assorbire il peso della complessità. In una logica di questo tipo, i dati sui quali il sistema lavora mantengono un ruolo da protagonista, dal momento che l'architettura è costruita su misura per loro. Stiamo parlando di un tipo di flessibilità molto interessante, perché facilita un'integrazione efficace dell'IA in contesti diversi, dando valore ai dati sui quali si lavora, più che sul modello in sé. Si tratta di una soluzione conveniente su più fronti.

---

anni saprà ragionare a lungo sui problemi, migliorare autonomamente e comunicare con altri LLM (Andrej Karpathy, *Intro to Large Language Models*. «YouTube», 23 novembre 2023, <https://www.youtube.com/watch?v=zjkBMFhNj_g&ab_channel=AndrejKarpathy> [Ultima consultazione: 23 settembre 2024]). Chiudiamo con una veloce nota sul fatto che OpenAI ha lanciato una nuova serie di modelli il 12 settembre 2024 che è stata addestrata a pensare più a lungo prima di rispondere, prendendo spunto dalla teoria dei due sistemi di pensiero umano di Daniel Kahneman che distingue il Sistema 1 (che si attiva automaticamente per dare una risposta velocemente, ma in modo superficiale) dal Sistema 2 (che procede più lentamente ma è in grado di andare più in profondità). I primi risultati dichiarati da OpenAI sono promettenti (<https://openai.com/index/introducing-openai-o1-preview/> [Ultima consultazione: 23 settembre 2024]).

Innanzitutto, perché oggi il mondo dell'IA è in continua ebollizione e la velocità con la quale si rinnovano e si arricchiscono i modelli e le soluzioni architetturali disponibili non permette di stare al passo con tutto e non è detto che tutto abbia un futuro. In secondo luogo, i dati non sono soltanto una sequenza di bit, ma sono l'essenza dell'azienda o dell'ente che li ha prodotti, in quanto recano traccia della loro storia, del loro operato e del loro know-how. Trattare i dati in modo adeguato, porli al centro dell'attenzione, significa averne capito la ricchezza e la sfida che l'applicazione dell'IA (non solo generativa) pone veramente: trovare la soluzione migliore per partire dai dati e costruire conoscenza ancorata all'esperienza che ha prodotto quei dati stessi.