Sono solo parole

ChatGPT: anatomia e raccomandazioni per l'uso

Tommaso Caselli, Antonio Lieto^, Malvina Nissim*, Viviana Patti^*

**CLCG, University of Groningen*
^Dipartimento di Informatica, Università di Torino

Abstract

ChatGPT has revolusionised the way people view and interact with language-based artificial agents. But is it a real revolution? And are people using ChatGPT with appropriate knowledge of its inner workings, its abilities, and potential risks? We think ChatGPT is very much in need of some proper contextualisation. In this short contribution we show how ChatGPT has come to life, both historically and technically, describing in detail the anatomy of large language models, and on the basis of this we clarify what ChatGPT can (be expected to) do, and what it cannot. We also discuss its limitations, specifically related to its intrinsic inability to be factual in what it generates, its reflection of societal biases, and the broader ethical implications of its use.

Keywords: ChatGPT, Natural Language Processing, Language Generation, Large Language Models, Generalized Pretrained Transformers, Fair Artificial Intelligence

1. Introduzione

Nel febbraio 2019 Open AI, nata come organizzazione senza fini di lucro con un focus sulla ricerca sull'intelligenza artificiale ed evolutasi nel tempo come azienda commerciale, annunciava allo stesso tempo la creazione di un modello generativo della lingua in grado di produrre testi estremamente simili a quelli naturali, prevalentemente in lingua inglese, e l'intenzione di rilasciarne solo una versione ridotta per paura delle conseguenze che un tale modello messo a disposizione di tutti potesse avere.[1] Le preoccupazioni, tra le altre, erano espresse in termini di generazione su larga scala di *fake news* e commenti offensivi e tossici su social media, oppure farsi passare per altri mimando il loro stile di scrittura. Il modello in questione si chiama GPT-2, dove GPT sta per *Generalized Pretrained Transformer*, ed è stato addestrato su 40 gigabytes di testi di varia natura setacciati online sfruttando, in termini di architettura, la parte *decoder* del modello *Transformer* (Vaswani *et al.*, 2017). GPT-2 appartiene alla famiglia di nuovi *Generative Language Models* basati su reti neurali profonde e nella comunità di *Natural Language Processing* (NLP) è stato percepito come il primo modello avanzato di generazione automatica della lingua non addestrato per un compito specifico ma semplicemente per imparare a scrivere. È stato infatti il primo modello generativo generico in grado di eseguire una varietà di compiti quali, tra gli altri, traduzione automatica, riassunti, riscrittura con trasferimento di stile, composizione di brevi racconti, ricevendo in input semplicemente l'inizio di un testo o la descrizione di un compito da portare a termine, invece che esempi strutturati e annotati apposta per addestrare il modello a eseguire il compito specifico. A seguire, nella famiglia GPT sono stati siluppati modelli ancora più grandi in termini di numero di parametri e di quantità di dati di addestramento: GPT-3, GPT-3.5, GPT-4, dove gli ultimi due sono alla base del modello

[1] "Due to concerns about large language models being used to generate deceptive, biased, or abusive language at scale, we are only releasing a much smaller version of GPT-2 along with sampling code. We are not releasing the dataset, training code, or GPT-2 model weights." https://openai.com/research/better-language-models, February 14, 2019.

conversazionale ChatGPT che OpenAI ha rilasciato per uso pubblico a fine novembre 2022, avendo apparentemente superato le preoccupazioni espresse anni prima in merito al rilascio di GPT-2. ChatGPT è un modello di generazione automatica di testo in grado di gestire una conversazione (scritta) con l'utente sullo stile dei chatbot. Questa capacità di interazione si somma a tutte le altre che accomunano i vari modelli GPT*.

Chiunque abbia provato ad usare ChatGPT sarà rimasto sorpreso dalla naturalezza dell'interazione, la fluidità e la coerenza dei testi generati. Ovviamente non c'è magia ma tanta lingua, tanta statistica e soprattutto un percorso di lunga data nella ricerca in NLP. Ci preme infatti chiarire fin da subito che ChatGPT non è rivoluzionario da un punto di vista tecnologico, ma è stato sviluppato usando metodi noti già utilizzati per sviluppare modelli predittivi e generativi. L'averne aperto l'accesso senza limitazioni a un pubblico non esperto e senza una progressiva esposizione a partire da modelli con prestazioni generative più scadenti, conosciuti solo dalla comunità scientifica, ha contribuito non soltanto al suo successo, ma anche a suscitare curiosità e stupore misti a inquietudine e preoccupazione, certamente anche dovuti a una mancanza di conoscenza del suo funzionamento. Prima di addentrarci nei meandri di ChatGPT e simili modelli conversazionali, sia da un punto di vista di sviluppo che di utilizzo, capacità, limiti e impatto, capiamo come funziona un modello di generazione della lingua (*generative language model*), il cuore degli agenti conversazionali odierni.

2. Il cuore di ChatGPT: come funziona, cosa fa e cosa non fa un modello di generazione della lingua

La famiglia di modelli di generazione della lingua via via più sofisticati che anima ChatGPT prevede l'uso di una serie di reti neurali artificiali profonde (i *Transformers*) che costruiscono un modello statistico del linguaggio (Vaswani *et al.*, 2017). Alla base di questi modelli troviamo rappresentazioni vettoriali dense basate su numeri reali, $\mathbb{R}$. Questa modalità di rappresentazione, già usata in maniera più o meno efficace a partire dalla fine degli anni '90 con l'avvento dei modelli statistici basati su *features*, permette la codifica, il processamento, e la trasformazione di qualsiasi tipo di informazioni, anche ciò che pertiene alle lingue naturali.

La profondità di una rete neurale si riferisce al numero di nodi, o parametri, che la compongono e alla loro stratificazione su diversi livelli. La profondità e il numero di parametri usati dal modello per poter compiere delle predizioni sono i due ingredienti chiave alla base del loro successo. Per dare un'idea dell'ordine di grandezza, GPT3.5, su cui è basata la prima *release* di ChatGPT, ha ben 96 livelli di profondità e 175 miliardi di parametri. Per fare un confronto, una rete neurale come un *Long Short-Term Memory* con 100 unità (nodi) e vettori in input di 100 dimensioni, ha 80.400 parametri; il primo modello di linguaggio basato su *Transformer*, BERT (*Bidirectional Encoder Representations from Transformers;* Devlin et al., 2019), sviluppato da Google nel 2018, ha 12 livelli e 110 milioni di parametri. Ogni nodo è una 'funzione' che riceve dati in ingresso, compie delle trasformazioni (attraverso una serie di operazioni di algebra lineare e non lineare) dei dati per poi passarli a dei nodi recettori. Se questo fosse l'unico elemento distintivo, i *Transformer* sembrerebbero solo una versione "con ormoni" dei primi modelli statistici usati in ambito generativo basati sulle Catene Markoviane, o di reti neurali più semplici, come i *Long Short-Term Memory* (uni- o bi-direzionali). In realtà, il successo dei *Transformer* sta, in grande parte, nel meccanismo dell'attenzione (*attention*). Grazie all'attenzione "una rappresentazione in una posizione è calcolata come una combinazione ponderata di rappresentazioni provenienti da altre posizioni"[2] (Manning, 2022: 130).

Intuitivamente, il meccanismo dell'attenzione permette di ottenere rappresentazioni vettoriali che codificano non solo il "significato" di un oggetto informativo in una determinata posizione in una sequenza, ma anche quella degli altri oggetti informativi che fanno parte della sequenza stessa.

---

[2] Il testo originale recita "*a representation at a position is computed as a weighted combination of representations from other positions*."

Entrando nel dettaglio tecnico, assumiamo di avere in input una frase come "*Il gatto rincorse il topo e se lo mangiò*", che passiamo come input a un *Transformer*. Il meccanismo dell'attenzione ottimizza la relazione tra ogni parola della nostra frase di esempio (elementi di una sequenza di input) e tutte le altre parole che si trovano nella frase (ovvero, gli altri elementi della sequenza di input). Se prendiamo in analisi il pronome oggetto "*lo"*, la rappresentazione vettoriale che otteniamo applicando il meccanismo dell'attenzione è tale per cui il vettore risultante codifica le informazioni relative alla parola stessa e, allo stesso tempo, presenta pesi differenziati in base all'importanza che le altre parole nella sequenza hanno per rappresentare il significato del pronome. Questo vuol dire che la parola "*topo*", a cui il pronome si riferisce, ha un peso maggiore rispetto alla parola "*gatto*".

L'assegnazione dei pesi per le diverse rappresentazioni vettoriali avviene attraverso la moltiplicazione di tre matrici: *Query* (Q), *Keys* (K), *Values* (V). L'input della sequenza viene inizialmente moltiplicato per Q e per K ottenendo così un'autocorrelazione tra le sequenze di input e quali di questi elementi, all'interno della stessa sequenza, sono più attinenti tra di loro (i.e., la nostra frase di esempio). Successivamente, il risultato di questa moltiplicazione tra K e V viene passato a un layer di *softmax* che permette, da un lato, di ottenere una distribuzione di probabilità finita (i.e., la somma di tutti i valori è sempre 1), e, dall'altro, di selezionare l'accoppiamento più pertinente rispetto alla specifica parola. Questi accoppiamenti vengono successivamente moltiplicati per la matrice V (valori), che dà un punteggio a quanto è stato precedentemente selezionato come elemento pertinente per decidere quanta attenzione "prestare" a ogni altro elemento della sequenza. Questo meccanismo in un *Transformer* si ripete per N volte e il risultato di ognuna di queste moltiplicazioni di matrici può essere concatenato permettendo di costruire pile (*stack*) di meccanismi di attenzione. Per esempio, il modello BERT citato poco sopra, si presenta in due versioni: una base, con 12 meccanismi di attenzione, e una grande (*large*), con 24. Il vantaggio del meccanismo di attenzione è la possibilità di variare le rappresentazioni della stessa parola a seconda del contesto di occorrenza; in altre parole, in base alla specifica sequenza di input (frase o paragrafo, con una lunghezza massima di parole che varia a seconda dello specifico modello) in cui una parola occorre, la sua rappresentazione vettoriale sarà sempre diversa.

Uno degli aspetti più affascinanti della creazione dei modelli di linguaggio generativi è il loro addestramento. Prima ancora di imparare a svolgere un compito ben preciso, per esempio discriminare messaggi d'odio, il modello è addestrato senza nessun intervento umano esplicito (in maniera non supervisionata) a svolgere un task di predizione molto semplice. Nel caso di modelli di linguaggio generativo, questo task di base consiste nel predire la parola successiva in una sequenza, tenendo conto di tutte le parole che sono presenti in precedenza. Altri modelli usano altri obiettivi di apprendimento auto-supervisionato come predire una parola mascherata (e.g., "*Il gatto rincorre il [MASK]*" - soluzione: *topo*). Nella fase di addestramento, il modello riceve un feedback rispetto alle predizioni fatte. Gli errori sono occasioni per ricalcolare i vari pesi dei vettori e essere indirizzato verso la soluzione corretta (meccanismo della "*back-propagation*"). Sulla base di questo paradigma di apprendimento, e grazie all'uso di enormi e molto varie quantità di testi, i *large language models* (LLMs) imparano *pattern* e generalizzazioni, acquisendo un bagaglio di conoscenza implicita linguistica (e non): dalle regolarità morfosintattiche, a informazioni di tipo semantico o enciclopedico, fino a tratti socioculturali espressi linguisticamente nei testi a cui il modello è esposto (Nissim e Pannitto, 2022).

Come già detto, questi modelli hanno bisogno di grandi quantità di dati per ottenere quella conoscenza implicita di una lingua che permette loro di ottenere risultati allo stato dell'arte una volta specializzati per un compito preciso. Per dare un'idea dei numeri di cui stiamo parlando, GPT-3 è stato ottenuto usando 300 miliardi di parole per un totale di 570 Gb, combinando diverse fonti di testo – incluse pagine Web tra il 2016 e il 2019 ottenute dal database *Common Crawl*.

Modelli di linguaggio più recenti, come ChatGPT, hanno iniziato a nutrirsi, in fase di addestramento, anche di documenti scritti in codice, ovvero in diversi linguaggi formali di programmazione, oltre che della consueta mole di dati in linguaggio naturale. La presenza di

linguaggio formale sembra essere alla base di un ulteriore salto nelle capacità di codificare conoscenza implicita da parte di questi modelli, inclusa la possibilità di compiere semplici passi di ragionamento.[3] Attenzione: non si sta dicendo che questi modelli abbiano acquisito capacità di ragionamento, ma solo che siano in grado di riprodurre - in determinati compiti di natura inferenziale - un output che se fosse stato prodotto da un essere umano, avrebbe richiesto l'utilizzo dei meccanismi cognitivi che sono alla base di processi di ragionamento.

La qualità e la rappresentatività dei dati usati sono fondamentali per determinare la qualità del modello stesso. Documentare i dati, ovvero rendere esplicite le fonti che sono state usate, è un passaggio fondamentale per comprendere il comportamento dei modelli stessi quando vengono usati per compiti specifici una volta addestrati. A oggi, una documentazione completa ed esaustiva di tutti i dati usati per addestrare ChatGPT non è disponibile. Un esempio che va in direzione opposta è quello del modello BLOOM (Scao *et al.*, 2022), i cui sviluppatori hanno fatto della trasparenza dei dati usati un elemento fondante dell'intero progetto.

La rappresentatività dei dati non riguarda solo la ricchezza delle fonti ma anche delle lingue stesse. In questo senso, lingue che faticano a occupare una significativa porzione dello spazio online - sia perché non scritte sia perché il materiale in formato digitale pronto all'uso (tendenzialmente materiale privo di copyright) è scarso - tendono a essere escluse. Questo ha un impatto nello sviluppo di modelli multilingue: per quanto sia stato dimostrato che l'applicazione diretta di questi modelli su lingue sconosciute al modello, in quanto mai viste in fase di training, possa dare risultati più o meno buoni, la possibilità di integrare del materiale in fase di addestramento resta fondamentale per migliorare le prestazioni ed evitare distorsioni nella rappresentazione di una lingua e della sua comunità di parlanti. A tale proposito, la lingua italiana si trova in una posizione relativamente buona rispetto alla disponibilità di LLMs, sia per quanto riguarda modelli multilingue che nei dati di addestramento sono stati esposti anche all'italiano (mBERT, Pires et al., 2019; XLM-R, Conneau et al., 2020; GPT-3, Brown et al., 2020; mDeBERTa, He et al., 2021), sia per quanto riguarda modelli monolingue, anche per diversi domini o generi testuali (AlBERTo, Polignano *et al*., 2019; $\text{BERT}_{\text{BASE}}$ Italian XXL e GilBERTo[4]; UmBERTo, Parisi *et al*., 2020; BERToldo, Palmero Aprosio *et al*., 2022). La copertura è piuttosto adeguata anche per la presenza di modelli generativi monolingue, che hanno visto un'evoluzione nelle architetture e prestazioni (GePpeTto, De Mattei *et al*., 2020; IT5, Sarti e Nissim, 2022; Camoscio, Santilli, 2023).

3. Da *large language model* ad agente conversazionale

Gli LLMs basati su *Transformer*, come quelli della famiglia GPT, ma non soltanto, vengono spesso definiti "base", "generici" (*general purpose*), o "fondamentali" (*foundational*) perché, come abbiamo visto, pur nella loro complessità, non sono addestrati per un compito specifico quanto piuttosto per processare e generare lingua. Questa loro caratteristica li rende adatti ad essere ulteriormente specializzati (*fine-tuned*) grazie ad un progressivo addestramento che viene fatto su dati (possibilmente annotati) relativi allo specifico compito che si desidera che il modello impari a svolgere. Per esempio, si può partire da un modello base e continuare ad addestrarlo con dati annotati manualmente per classificare testi come contenenti o meno messaggi di odio. Il grande vantaggio di questa strategia, opposta ad una strategia che addestra un modello di apprendimento da zero o in maniera puramente supervisionata a partire soltanto dai dati manualmente annotati, è che la base è appunto un modello che già "conosce" bene la lingua, e ha bisogno di essere solo raffinato per risolvere un problema specifico, rendendo possibile l'utilizzo di un numero limitato di esempi che non sarebbero assolutamente sufficienti per addestrare un modello puramente supervisionato.

---

[3] Questo aspetto è stato evidenziato per la prima volta in questo *living article*: https://yaofu.notion.site/How-does-GPT-Obtain-its-Ability-Tracing-Emergent-Abilities-of-Language-Models-to-their-Sources-b9a57ac0fcf74f30a1ab9e3e36fa1dc1

[4] Per questo modelli non esistono pubblicazioni di riferimento, la documentazione è disponibile qui: https://huggingface.co/dbmdz/bert-base-italian-xxl-cased; https://github.com/idb-ita/GilBERTo

Nel caso di ChatGPT, il compito specifico da svolgere è semplicemente "conversare". Quindi, a partire dal modello base GPT-3.5,[5] ChatGPT è stato successivamente addestrato in modalità *fine-tuning* con conversazioni prodotte da persone che, con lo scopo specifico di produrre dati da fornire al modello, simulano conversazioni tra il modello stesso e un agente umano.

Per spingere ulteriormente ChatGPT a conversare "come una persona" e allo stesso tempo per integrare nel modello una strategia di controllo non soltanto di qualità generica ma anche relativa ad aspetti più delicati legati a linguaggio potenzialmente offensivo, *bias*, etc. (si veda Sezione 5), OpenAI ha incluso un ulteriore livello di addestramento basato sul paradigma di "apprendimento per rinforzo con feedback umano" (*Reinforcement Learning with Human Feedback – RLHF*). In questa ultima fase, il cui scopo come già detto è quello di avvicinare il più possibile la produzione di testo alle aspettative e intenzioni di un ideale parlante umano, gli sviluppatori hanno collezionato delle interazioni realmente avvenute tra persone e ChatGPT, e rigenerato automaticamente alcune delle risposte del modello. Le diverse versioni sono state poi sottoposte a giudizio umano e ordinate per preferenza. Attraverso molteplici iterazioni di questo processo, il modello è stato così ulteriormente addestrato ad allinearsi nelle sue produzioni alle preferenze umane.

Quindi, per riassumere in un'unica frase, ChatGPT è un chatbot avanzato basato su un modello di linguaggio generativo di tipo *Transformer* con dati in molteplici lingue ma a dominanza inglese, raffinato con esempi di conversazioni e con apprendimento per rinforzo con feedback umano. Chiarito *cos'è*, risulta anche più semplice chiarire *cosa non è*:

- ChatGPT *non è* rivoluzionario da un punto di vista tecnologico. ChatGPT è stato sviluppato usando metodi che sono noti e sono stati già usati. Sicuramente, l'averne aperto l'accesso al pubblico senza limitazioni e senza una particolare promozione delle sue capacità sono due elementi che hanno contribuito al suo successo e allo stupore generalizzato su quanto sia 'bravo'.
- ChatGPT *non è* un database: per quanto parte dei dati usati per addestrarlo possano essere memorizzati e quindi poi riprodotti dal modello, ChatGPT è in grado di generare istanze di testo completamente nuove.
- ChatGPT *non è* un motore di ricerca come Google Search (anche se sistemi basati su LLMs, come ad esempio Bard di Google o la nuova versione di Bing, sono e saranno utilizzati sempre di più come supporto alla tradizionale funzione di ricerca). In quanto modello di linguaggio, ChatGPT *non* ha accesso *live* al Web per recuperare informazioni. La sua 'conoscenza del mondo' è soltanto un *byproduct* del suo addestramento su base linguistica (ha acquisito informazione nel processo di imparare a scrivere, sostanzialmente), ed è limitata ai dati usati per addestrarlo. Tutto ciò che produce è 'inventato' ma, per questioni probabilistiche, può essere vero o realistico.
- ChatGPT *non è* 'obiettivo' né un oracolo. Come tutti i modelli addestrati su dati, è soggetto a - e quindi incorpora - gli stessi *bias* che sono presenti nei dati di training. Considerato che i dati di training sono testi che sono espressione delle culture e delle società che si esprimono online, i *bias* socioculturali presenti in ChatGPT non sono particolarmente diversi da quelli veicolati nei testi; anzi i *bias* presenti nelle società stesse sono potenzialmente accentuati dal modello probabilistico.

I compiti in cui i modelli come ChatGPT eccellono sono quelli generazione di testo e riscrittura con controllo e manipolazione di stile. Si possono ottenere testi coerenti scritti 'nello stile di': Stile comunicazione elettronica amministrativa? Stile biblico? A là Proust? Come se fosse una canzone dei Queen? E se lo scrivesse un bambino? Parimenti, questi modelli producono solitamente buoni risultati quando si chiede loro di generare un riassunto su un argomento enciclopedico. Certamente, questa capacità apre a dibattiti sulle capacità creative e sull'originalità del testo prodotto. Si ottengono,

[5] Al momento della pubblicazione ChatGPT è stato anche addestrato sul successore di GPT-3.5: GPT-4. Il modello conversazionale basato su quest'ultimo può essere selezionato però solo da utenti premium. La versione aperta di ChatGPT è basata su GPT-3.5.

inoltre, risultati solitamente molto soddisfacenti nei compiti di traduzione, anche se, quando si tratta di lingue a cui i modelli sono stati meno esposti, le generazioni senza senso (o "allucinate", si veda Sezione 4) aumentano. Lo stesso si dica quando si tratta di varietà linguistiche come i dialetti: non è difficile stuzzicare ChatGPT a scrivere poesie o racconti in falso (ma plausibile) piemontese!

4. Allucinazioni e Limiti

Oltre alle elevate prestazioni esibite in attività di generazione testuale (sia per compiti che fanno riferimento al linguaggio naturale, sia per quelli inerenti alla generazione di porzioni di programmi), gli LLMs generativi, basati su architettura *Transformer*, sono passati alla ribalta delle cronache, tanto nella letteratura scientifica quanto in quella di carattere divulgativo, per il cosiddetto fenomeno delle "allucinazioni" (*hallucinations* in inglese). In sostanza, con questa espressione - molto dibattuta, si veda (Klein, 2023) - si fa riferimento al fatto che uno dei limiti di sistemi come ChatGPT, GPT-4, Bard etc. riguardi la possibilità di produrre sequenze testuali grammaticalmente fluenti e plausibili a corredo, tuttavia, di risposte completamente false (ma spesso verosimili) o inventate di sana pianta (ad es. su fenomeni mai avvenuti). A completamento di questo fenomeno "allucinatorio", anche il supporto delle affermazioni (false) generate viene corredato da riferimenti fasulli (per una rassegna del fenomeno si veda Ji, *et al.* 2023). In buona sostanza, il non avere una nozione formale di "verità", o "falsità" delle espressioni linguistiche maneggiate, bensì un suo surrogato di tipo probabilistico, fa sì che questi sistemi non siano in grado di fare - a monte - un controllo sulla natura fallace o meno di quanto viene generato. Questo tipo di limite formale naturalmente esplode quando si testano tali sistemi su questioni di natura logica (anche elementare). I limiti relativi alla capacità di elaborazione di input testuali che richiedono diverse forme inferenziali sono documentati in una lunga serie di rassegne pubblicate negli ultimi mesi e relativi a GPT-3, ChatGPT e GPT-4 (si veda, ad esempio, Jin et al 2023; Borij 2023; Valmeekam et al. 2022).

È da notare che, nell'ultima versione di ChatGPT (selezionando GPT-4 come "motore" del dialogo, opzione al momento accessibile solo a pagamento), molti degli errori precedentemente riportati non vengono più commessi. Tuttavia, questo miglioramento non si può ascrivere in toto ai miglioramenti tecnici ottenuti con il RLHF bensì al fatto che l'azienda, come indicato dalla stessa OpenAI[6], ha iniziato ad utilizzare - da GPT-4 in avanti - in modo massiccio un forte controllo e intervento manuale da parte di annotatori umani per la revisione delle risposte fornite dal sistema. Questo forte intervento manuale, finalizzato per motivi di business ad ottenere un prodotto commerciale sempre più performante, di fatto inquina l'analisi delle prestazioni attribuite all'architettura neurale sottostante al sistema. Tanto è vero che, di recente, si è proposto di evitare di usare questi sistemi proprietari come "baseline" per valutare le prestazioni di sistemi di NLP.[7] L'utilizzo sistematico di intervento manuale supervisionato va a discapito della narrativa che vede il miglioramento delle nuove versioni di questo sistema come conseguenza dell'aumento della "dimensione" del modello e degli iperparametri utilizzati (i quali sarebbero i fattori determinanti alla base del fenomeno di "emergenza" di nuove capacità; si veda ad es. Schaeffer et al. 2023). E va anche contro alla narrativa che vede l'attribuzione a questi modelli generativi di abilità cognitive tipiche di entità biologiche (per esempio, intenzionalità, o il fatto di avere o meno una "teoria della mente").

Nello specifico, negli ultimi mesi una grande fonte di confusione legata all'interpretazione dei comportamenti esibiti dai modelli generativi è stata dovuta all'assunzione, infondata, che tali sistemi siano sistemi di Intelligenza Artificiale Forte (*Strong AI*). Tale espressione, *Strong AI*, è stata originariamente introdotta dal filosofo John Searle per identificare la posizione che assume che modelli computazionali (incarnati o meno) possano avere una "mente", una "coscienza", etc. esattamente come gli esseri umani. A fare da contraltare a questa visione, vi è l'espressione *Weak AI*

---

[6] Si vedano ad esempio le FAQ di ChatGPT: https://help.openai.com/en/articles/6783457-what-is-chatgpt

[7] https://hackingsemantics.xyz/2023/closed-baselines/

(Intelligenza Artificiale debole), che sintetizza la posizione secondo cui modelli artificiali possono simulare il comportamento e le abilità esibite da un essere umano senza pretendere di possedere effettivamente le capacità biologiche che sottostanno all'esibizione di quel comportamento nel sistema naturale. In Lieto (2021), è stato illustrato come modelli come GPT (e di conseguenza ChatGPT) siano perfettamente allineati con l'ipotesi della *Weak AI* e che essi possono essere descritti (al più) come rappresentazioni superficiali, imprecise e con un alto grado di implausibilità biologica rispetto ai cervelli biologici. Nello specifico, tali sistemi sono progettati secondo una prospettiva "funzionalista": sono in grado, nella maggior parte dei casi, di riprodurre superficialmente lo stesso tipo di output di un essere umano ma i meccanismi che determinano tale output sono completamente differenti da quelli che producono un risultato analogo in modelli biologici. Come conseguenza di questo stato di cose, non è possibile, né scientificamente legittimo, ascrivere teorie e relativi meccanismi usati per spiegare l'output di un sistema biologico per interpretare l'output, ancorché del medesimo tipo, generato da questi modelli artificiali del linguaggio. Le differenze e le asimmetrie tra queste classi di sistemi è, infatti, enorme. A questo proposito è anche importante sottolineare che il processo di apprendimento da parte di ChatGPT non simula in alcun modo l'apprendimento del linguaggio da parte di un bambino. Per quanto si possano aprire spunti di riflessione interessanti (legati all'argomento della povertà dello stimolo), ChatGPT non è esposto a interazioni sociali o comunicative multimodali come un essere umano nel suo naturale sviluppo linguistico.

5. Impatto e aspetti etici, prospettive future e auspici

I modelli di linguaggio generativi sono – a oggi – rappresentazioni eccellenti della dissociazione tra linguaggio e pensiero. Non a caso sono stati soprannominati 'pappagalli stocastici' (Bender et al., 2021). Se da un lato il fatto che non si possano ascrivere a modelli della famiglia GPT e ChatGPT capacità *human-like* dovrebbe servire a comprendere la natura talvolta forzata (e capziosa) di certi richiami apocalittici[8], dall'altro lato ciò non implica, naturalmente, che non ci siano rischi e pericoli etici che possono scaturire dall'uso di questa tecnologia. Pensiamo al possibile impatto di questi sistemi sul lavoro (per esempio:come evolverà la professione dell'informatico e in particolare come formeremo i nuovi programmatori?), sul tipo di "mondo" che rappresentano, sui *bias* che contengono[9], sul loro possibile utilizzo per generare automaticamente disinformazione pilotata. Il testo di ingaggio (*prompt*) della conversazione con questi modelli può essere manipolato per poi ottenere il risultato voluto. Un modello può generare un contenuto che sembra scritto da una persona e che è fattualmente errato, o fuorviante, e tale contenuto può venire diffuso in maniera massiccia sui social media. Chi deve essere ritenuto responsabile dell'eventuale disinformazione che si genera su larga scala in questo caso? È evidente che si tratta di strumenti potenti che possono essere utilizzati per facilitare la diffusione di notizie false e alimentare campagne di disinformazione.

Parimenti un modello potrebbe generare testi offensivi, con abusi verbali e contenuti razzisti e, di nuovo, con possibilità di diffusione e di eventuale supporto alla creazione di campagne coordinate online mirate a colpire persone o gruppi di persone specifiche, anche appartenenti a categorie vulnerabili. Non si può negare che gli sviluppatori di ChatGPT abbiano posto attenzione a evitare che contenuti tossici e violenti avessero (troppo) spazio nell'insieme dei dati di pre-addestramento. E certamente, nell'ambito del paradigma di *RLFH,* non conosce sosta il lavoro di annotatori umani mirato ad affinare via via il modello evitando che durante le interazioni con gli utenti ChatGPT risponda a richieste capziose di generare contenuti falsi o dannosi. Uno sforzo i cui effetti sono tuttavia in gran parte solo superficiali, vista la facilità con cui questi filtri possono essere aggirati (in gergo *jailbreak*), e che solleva spinose questioni relative a censura e enorme soggettività nelle scelte di quali contenuti moderare o eliminare. Inoltre, il costo sociale di questa operazione,

[8] Si veda: https://joanna-bryson.blogspot.com/2023/05/sam-altman-is-speaking-in-munich-today.html

[9] Si veda (Nissim e Pannitto, 2022), Capitolo 6.

sembra avere ricevuto poca attenzione, e diverse inchieste hanno portato alla luce scenari inquietanti di lavoratori pagati meno di $2 l'ora perché eliminassero manualmente contenuti tossici[10]. Oltre a essere pura estrazione di surplus dalla forza lavoro di altri esseri umani, i danni psicologici di un'esposizione prolungata a contenuti tossici possono essere una conseguenza non banale e duratura di questo tipo di lavoro, sollevando ulteriori questioni etiche relative allo sviluppo e all'uso di questi modelli (Roberts, 2017; Steiger et al,. 2021).

Il problema della sostenibilità ambientale, legato ai noti elevatissimi costi di addestramento dei LLMs, è un altro aspetto molto rilevante da considerare, e su questa linea sta crescendo nella comunità di NLP l'interesse per la ricerca di architetture sostenibili (Molinaro et *al.*, 2023).

È certamente di questi aspetti che bisognerà occuparsi per evitare un uso nefasto delle tecnologie del linguaggio nell'ambito della nostra società, insieme a una campagna continuativa di educazione al funzionamento e all'uso di queste tecnologie. In quest'ottica, è bene ricordare l'importanza e la necessità di un controllo e di un intervento delle istituzioni statali e sovranazionali per mantenere viva la prospettiva futura di una ricerca pubblica per lo sviluppo di tecnologie del linguaggio in maniera *open*. Nella consapevolezza che modelli come ChatGPT rappresentano un importante valore aggiunto per le tecnologie di Intelligenza Artificiale, è auspicabile che la comunità di ricerca nazionale e internazionale promuova sforzi congiunti nella direzione, da un lato, della trasparenza scientifica, in termini di replicabilità dei modelli e degli esperimenti e della libertà di accesso a risorse e dati, e, dall'altro, della creazione di infrastrutture computazionali, come auspicato in recenti *call for action* sia a livello europeo che nazionale[11].

Riferimenti bibliografici

Bender, E. M., Gebru, T. McMillan-Major, A., Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623.

Borji, A. (2023). A categorical archive of ChatGPT failures. *arXiv preprint arXiv:2302.03494*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, pp.1877-1901.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

De Mattei, L., Cafagna, M., Dell'Orletta, F., Nissim, M., & Guerini, M. (2021). GePpeTto Carves Italian into a Language Model. *Proceedings of the Seventh Italian Conference on Computational Linguistics*, CLiC-it 2020, Bologna, Italy, CEUR Workshop Proceedings, vol. 2769. CEUR-WS.org.

[10] https://www.theverge.com/features/23764584/ai-artificial-intelligence-data-notation-labor-scale-surge-remotasks-openai-chatbots

[11] Si vedano a questo proposito la *call for action* a livello europeo documentata qui: https://ircai.org/wp-content/uploads/2023/06/v3-eurogpt-press-release-A4-Document.pdf , oltre al recente comunicato della Associazione Italiana di Linguistica Computazionale (https://www.ai-lc.it/en/ailc-per-unintelligenza-artificiale-responsabile-e-aperta-3/).

Devlin, J., Chang, M-W., Lee, K., and Toutanova, K.. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

He, P., Gao, J., & Chen, W. (2021). DeBERTav3: Improving DeBERTa Using ELECTRA-style Pre-training with Gradient-disentangled Embedding Sharing. *arXiv preprint arXiv:2111.09543*.

Klein, N. (2023), AI Machines aren't hallucinating, *The Guardian*, May 8th, 2023: https://www.theguardian.com/commentisfree/2023/may/08/ai-machines-hallucinating-naomi-klein

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1-38.

Jin, Z., Liu, J., Lyu, Z., Poff, S., Sachan, M., Mihalcea, R., & Schölkopf, B. (2023). Can Large Language Models Infer Causation from Correlation?. *arXiv preprint arXiv:2306.05836*.

Lieto, A. (2021). *Cognitive design for artificial minds*. Routledge.

Manning, C. D. (2022) Human Language Understanding & Reasoning. *Daedalus*, 151 (2): 127–138.

Molinaro, L., Tatano, R., Busto, E., Fiandrotti, A., Basile, V., Patti, V. (2023) DelBERTo: A Deep Lightweight Transformer for Sentiment Analysis. *Advances in Artificial Intelligence - Proceedings of XXIst International Conference of the Italian Association for Artificial Intelligence, AIxIA 2022, Lecture Notes in Computer Science*, vol: 13796: 443-456. Springer.

Nissim, M., Pannitto, L. (2022). *Che cos'è la linguistica computazionale*. Carocci, Roma.

Palmero Aprosio, A., Menini, S., & Tonelli, S. (2022). BERToldo, the Historical BERT for Italian. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*.

Parisi, L., Francia, S., Magnani, P. (2020). UmBERTo: an Italian Language Model trained with Whole Word Masking. https://github.com/musixmatchresearch/umberto. GitHub.

Pires, T., Schlinger, E., & Garrette, D. (2019). How Multilingual is Multilingual BERT?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Polignano, M., Basile, P., De Gemmis, M., Semeraro, G., & Basile, V. (2019). AlBERTo: Italian BERT language understanding model for NLP challenging tasks based on tweets. In *CEUR Workshop Proceedings* (Vol. 2481, pp. 1-6). CEUR.

Sarah T Roberts (2017). Social media's silent filter. *The Atlantic*.

Santilli, A. (2023). Camoscio: An Italian Instruction-Tuned LLaMA. https://github.com/teelinsan/camoscio. GitHub.

Sarti, G., Nissim, M.. (2022). IT5: Large-Scale Text-to-Text Pretraining for Italian Language Understanding and Generation. *arXiv preprint arXiv:*2203.03759.

Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are emergent abilities of Large Language Models a mirage?. *arXiv preprint arXiv:2304.15004*.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., ... & Manica, M. (2022). BLOOM: A 176b-parameter Open-Access Multilingual Language Model. *arXiv preprint arXiv:* 2211.05100.

Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease (2021). The psychological Well-Being of content moderators: The emotional labor of commercial moderation and avenues for improving support. In *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, number Article 34. ACM.

Valmeekam, K., Olmo, A., Sreedharan, S., & Kambhampati, S. (2022). Large Language Models Still Can't Plan (A Benchmark for LLMs on Planning and Reasoning about Change). *arXiv preprint arXiv:2206.10498*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.