UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Filosofia, Sociologia, Pedagogia e Psicologia Applicata – FISPPA

Corso di Laurea in

Scienze Psicologiche Sociali e del Lavoro (L3)

Tesi di laurea triennale

Artificial General Intelligence: sfide e prospettive future

Artificial General Intelligence: challenges and future perspectives

Relatore:

Prof. Marco Zorzi

Laureando: ***Edoardo Gracis***

Matricola: ***2012811***

Anno Accademico 2022/2023

INDICE

INTRODUZIONE

Cos'è l'intelligenza? Ad oggi non c'è ancora una risposta definitiva. Il senso scientifico ne dà per certa l'esistenza, ma, non essendo percepibile, può solo osservarla indirettamente e costruire teorie sulla base di ipotesi. Storicamente è stata oggetto di studio di più discipline, ma è la Psicologia, a partire dalla fine del XIX secolo, ad aver creato diverse teorie sulla sua architettura ed i relativi strumenti di valutazione per misurarla. È importante notare come tutte le teorie contemporanee più importanti vedono l'intelligenza come costrutto scomponibile in più abilità (Teoria dell'Intelligenza Generale di Spearman, 1904; Teoria delle Abilità Primarie di Thurstone, 1938; Teoria dell'intelligenza di Anderson, 1992), tipologie (Teoria delle Intelligenze Multiple di Gardner, 1993) o categorie (Teoria Triarchica di Sternberg, 2011). Un fattore critico dell'intelligenza è che storicamente viene dato più peso solo ad alcune delle sue dimensioni. In particolare, gli strumenti di valutazione spesso utilizzano compiti che prevedono l'uso del linguaggio naturale (ad es. compiti di memoria o produzioni scritte) oppure di linguaggi formali (come la matematica e la chimica), e ciò presuppone una mediazione linguistica e culturale che penalizza gli individui meno educati (Terman, 1916). Alcuni autori, riconoscendo questo limite, hanno creato strumenti che valutino altre abilità che prescindano dall'educazione, come quelle logiche e visuo-spaziali (Wechsler, 1939), ma ne rimangono alcune per cui, nonostante siano teorizzate, non esistono strumenti rigorosi di valutazione (ad es. le abilità musicali).

Dunque quando possiamo dire che un umano è intelligente? Come appena argomentato, i classici test di intelligenza valutano solo alcune delle abilità umane, asserendo non tanto se un umano è intelligente o meno, ma *quanto* è intelligente secondo la scala di misura specifica per ciascun test. Tuttavia ci sono innumerevoli altre abilità che possono far guadagnare ad un individuo, seppur in maniera non scientifica ma a discrezione dei membri della comunità, l'etichetta di "intelligente". Una persona che sa identificare l'altezza assoluta di un suono (capacità nota come "orecchio assoluto") può essere considerata intelligente? Leggendo la Divina Commedia possiamo affermare che Dante Alighieri sia stato una persona intelligente? Se sì, qual è il criterio secondo il quale affermiamo che l'autore di un testo scritto è intelligente o meno? Ma soprattutto: se una macchina producesse lo stesso testo di un autore intelligente, verrebbe considerata ugualmente intelligente?

Nonostante queste domande non abbiano una risposta certa, il problema della varietà e l'ambiguità dei criteri in base a cui attribuire l'etichetta "intelligente" può essere visto come prova dell'unica chiara caratteristica dell'intelligenza: la sua eterogeneità.

CAPITOLO 1 – DUE SISTEMI INTELLIGENTI A CONFRONTO

Data questa eterogeneità non c'è da stupirsi se alcune delle caratteristiche dell'intelligenza si possono ritrovare anche negli animali o addirittura nelle macchine. E pare ancora più ovvio se si considera che tutti e tre i tipi di sistemi intelligenti condividono lo stesso principio di funzionamento: la rete neurale. È giusto precisare che l'Intelligenza Artificiale include anche altri tipi di architetture, ma quelle più all'avanguardia si ispirano proprio alle reti neurali biologiche nel cervello. Quest'organo è un sistema complessissimo, ma che in ultima analisi si basa su un meccanismo molto semplice: il neurone. Ognuna di queste cellule è una singola unità di calcolo che elabora il segnale in entrata e, a seconda della sua funzione di attivazione, lo propaga ad altri neuroni oppure lo blocca. Nel cervello umano ci sono circa 85 miliardi di neuroni e, a seconda della tipologia, ciascuno di essi può avere da poche a decine di migliaia di connessioni, che in totale si stimano essere tra le 10^{13} e 10^{15}. Numero che, per quanto strabiliante, è stato quasi raggiunto dalle attuali reti neurali (GPT-4 ha 1.7×10^{12} connessioni), benché, considerato da solo, non costituisca un indicatore affidabile della complessità della rete. Infatti, il cervello stesso non è interamente connesso, ma composto da più parti che si coordinano in maniera efficace ed efficiente nell'elaborazione di stimoli e risposte.

Nonostante questi due sistemi intelligenti condividano lo stesso principio neurale di base, ci sono alcune caratteristiche che li rendono nettamente diversi per architettura e funzionamento. I neuroni animali sono collocati fisicamente in uno spazio biologico che da una parte provvede al loro funzionamento fornendo nutrienti ed eliminandone gli scarti, ma dall'altra implica anche dei limiti fisici come quello della connettività (generalmente i neuroni stabiliscono connessioni con gli altri neuroni vicini, ad esempio organizzandosi in colonne corticali). I neuroni artificiali, invece, sono collocati in uno spazio digitale che necessità di elettricità e potenza di calcolo, e che non pone limiti alla connettività (ogni singolo neurone può potenzialmente connettersi a qualsiasi altro neurone della rete). Inoltre, mentre nei neuroni biologici il segnale si propaga in una sola direzione lungo l'assone, in quelli artificiali la connessione tra due neuroni può essere utilizzata per propagare il segnale anche all'indietro (meccanismo della back-propagation). Per queste ragioni, in letteratura è prassi distinguere le caratteristiche e funzionamenti dei neuroni artificiali tra *biologicamente plausibili* e *non biologicamente plausibili*. All'aumentare del numero di neuroni e connessioni, ovvero della *scala* delle reti neurali, queste differenze fondamentali si possono amplificare fino a generare modalità qualitativamente differenti di elaborazione delle informazioni. Quali siano le reali implicazioni di queste differenze rimane ancora in gran parte un mistero.

1.1 Natural Language Processing (NLP)

I dati testuali, così come i video o i messaggi vocali, sono un tipo di dati sequenziali in quanto, per poter cogliere il significato della parola che si sta elaborando, bisogna tenere traccia delle sue relazioni con le altre parole della sequenza (l'ordine che hanno nella frase, il genere, se sono al singolare o plurale, ma anche molti altre). Ciò è fondamentale ad esempio per disambiguare parole polisemiche (per stabilire il significato di "penne" nella frase "Butta le penne che sono..." bisogna sapere se la frase finisce con "scariche" o "affamato") oppure per tradurre da una lingua all'altra (nel tradurre dall'inglese all'italiano una frase che contenga il pronome "it", bisogna sapere se quel pronome è riferito ad una parola al femminile o al maschile).

Fino al 2017 esistevano due tipi di reti in grado di elaborare dati sequenziali: i modelli feed-forward basati su finestre temporali e le reti ricorrenti. I primi analizzano simultaneamente un certo numero di dati in sequenza (che indica la larghezza della finestra temporale) per produrre un solo output (come ad esempio la NetTalk di Sejnowski & Rosenberg, 1987). Tuttavia la larghezza della finestra temporale va stabilita trovando un compromesso tra performance (una finestra troppo piccola non coglie le dipendenze distanti nella sequenza) e efficienza computazionale (per ogni elemento in più considerato c'è bisogno di un'altra serie di neuroni input che aumenta la complessità della rete). Le reti ricorrenti, invece, analizzano un elemento della sequenza alla volta, ma prima di passare a quello successivo producono, insieme all'output, una "traccia" dell'elaborazione passata. Questa traccia viene elaborata insieme al nuovo input e tiene conto non solo dell'elaborazione immediatamente prima, ma anche di quelle precedenti. Tuttavia il problema è che, con il procedere delle elaborazioni, la traccia di quelle lontane si fa sempre più debole fino a svanire, limitando così l'insieme di dati che le reti ricorrenti riescono a considerare ad ogni elaborazione (anche se si è cercato di affrontare il problema migliorando l'architettura come nel caso delle Long-Short Memory Networks introdotte da Hochreiter & Schmidhuber, 1997).

Un grande passo avanti è stato fatto dal team Google di Vaswani et al. che nel 2017 hanno introdotto una nuova architettura neurale chiamata Transformer. Nel processare dati sequenziali, il Transformer utilizza un meccanismo attentivo grazie al quale è grado di prendere in considerazione un insieme di dati nettamente maggiore rispetto alle architetture precedenti. Ciò gli consente di cogliere relazioni tra parole anche molto distanti tra loro e di costruire rappresentazioni più ricche e generali dei dati. Tuttavia, queste architetture non lavorano direttamente con i dati testuali, ma con vettori ottenuti a partire dalle parole tramite due processi: Tokenization e Word Embedding.

1.1.1 Tokenization e Word Embedding

Il processo di Tokenization trasforma le parole in token, ossia di sequenze di lettere (o simboli grammaticali) che possono formare una parola o parte di essa. Il motivo per cui conviene processare sequenze di lettere corte è che in questo modo si riduce la grandezza del vocabolario che la rete neurale deve apprendere (che solitamente si aggira attorno ai 10.000 tokens), risparmiando memoria e migliorando l'efficienza computazionale. Infatti, per ogni token in più nel vocabolario la rete deve apprendere tutte le sue relazioni con gli altri tokens. Per questo motivo è più conveniente partire da un vocabolario ristretto e combinare i token per poter ottenere parole più complesse (ciò è particolarmente vantaggioso per lingue agglutinanti come il tedesco e il turco).

Una volta ottenuti i tokens, la funzione di Word Embedding li trasforma in vettori che hanno dalle 300 alle 500 dimensioni a seconda di quante caratteristiche della parola si vogliono catturare: più sono numerose, più è ricca la rappresentazione della parola (ossia la word embedding), ma servono anche più risorse computazionali. Uno degli approcci più utilizzati per ottenere i valori di queste dimensioni è il Word2vec (Mikolov et al., 2013) che si basa su due reti neurali: la continuos bag-of-words (CBOW) e la Skip-Gram. La CBOW ha lo scopo di predire un token a partire da una serie di token contesto, mentre la Skip-Gram ha lo scopo di predire i token contesto a partire da un token specifico. In questo modo, ogni testo usato nell'addestramento fornisce sia i tokens da cui partire sia quelli che la rete deve imparare a predire (target). Ciò consente di stabilire automaticamente se la predizione è corretta o sbagliata, per poi aggiustare di conseguenza i parametri della rete (un tipo di apprendimento chiamato auto-supervisionato).

Grazie al Word2vec e ad altri approcci, è possibile costruire rapidamente algoritmi di word embedding che, grazie al loro addestramento su enormi quantità di testo, riescono a generare rappresentazioni estremamente descrittive e accurate. Queste rappresentazioni dettagliate offrono una serie di vantaggi:

- Si preservano le somiglianze contestuali: parole simili per uso e significato hanno dimensioni simili e sono dunque vicine nello spazio vettoriale. Ciò consente alla rete di navigare per sinonimi e classi, e quindi di creare nuove combinazioni sensate di parole ("il muro è bianco" si può trasformare in "il tetto è rosso").
- Si possono fare analogie sottoforma di operazioni vettoriali (se al vettore "regina" sottraggo il vettore "donna" e aggiungo quello "uomo" il risultato è il vettore "re", così come "Sushi" – "Giappone" + "Italia" = "Pizza").
- Allineando gli spazi vettoriali contenenti le parole di due lingue diverse si possono scoprire nuove traduzioni (quando due parole che non erano mai state messe in relazione si

sovrappongono) e si evidenziano le parole che invece sono isolate (e dunque hanno un significato unico in quella lingua) (Zou et al., 2013).

- Quando si creano embeddings condivisi (diversi tipi di dati integrati nello stesso spazio vettoriale) la precisione dei vettori permette di fare deduzioni sorprendenti: in due diversi studi si è addestrato una rete convoluzionale a classificare le immagini producendo un vettore-immagine che si situi in prossimità del vettore-parola corrispondente (immagini di cani vengono mappate vicino alla parola "cane"). Dal momento che anche i vettori-immagini riescono a catturare le somiglianze tra i dati (il vettore-immagine di "cane" è più vicino a quello di "cavallo" che a quello di "camion"), quando si presenta alla rete un'immagine per cui non è stata addestrata (ad esempio un gatto), il vettore-immagine risultante non viene mappato casualmente, ma in prossimità dei vettori-immagine simili. Il fatto sorprendente è che la rete, in alcuni casi, mappa il nuovo vettore-immagine vicino al vettore-parola corrispondente, di fatto dimostrando di saper classificare anche immagini per cui non è stata addestrata (Socher et al., 2013; Frome et al., 2013).

1.1.2 Transformers e Large Language Models

I Large Language Models sono massicce reti neurali basate su Transformers in grado di processare linguaggio naturale. Ogni LLM viene addestrato per svolgere uno o più compiti (ad es. generare testo a partire da un prompt o tradurre da una lingua all'altra) in base ai quali viene scelta una libreria token che il modello deve apprendere a trasformare nei relativi word embeddings. Una volta terminata la fase di apprendimento, la libreria di word embeddings e i parametri dei Transformers vengono congelati, impedendo che il modello cambi continuamente (infatti si dice che il LLM è aggiornato fino ad una certa data).

L'architettura del Transformer è formata da una pila di encoders e da una di egual numero di decoders: i primi sono formati da uno strato Multi-Head Attention e da uno Feed-Forward, i secondi sono formati sempre da questi due strati ma in mezzo ce n'è un terzo di Encoder-Decoder Attention.

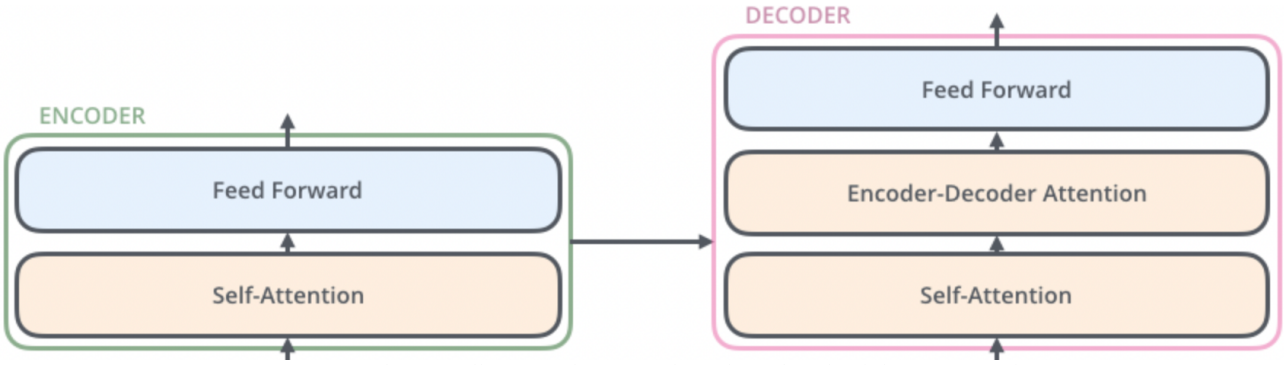

Figura 1. Gli strati all'interno di un encoder e di un decoder (Alammar, 2018)

Dato un prompt, l'informazione fluisce nel LLM per fasi:

1. Il testo in input viene suddiviso in tokens e ne viene selezionato un certo numero in base alla finestra del modello. Il testo oltre questo limite verrà processato nella fase successiva, col rischio però di non cogliere le relazioni con la porzione di testo precedente. Per questa ragione è importante che la finestra del Transformer sia il più ampia possibile (lo stato dell'arte è rappresentato da GPT-4 con una finestra di 32.000 tokens, ossia circa 50 pagine).
2. Si "traducono" i tokens (che per semplicità d'ora in avanti chiamerò "parole") in vettori e ad ognuno di questi si somma un altro vettore che codifichi la sua posizione nella sequenza (positional encoding). In questo modo la rete tiene traccia dell'ordine delle parole nella frase.
3. I vettori passano al primo encoder della pila dove, per ogni parola, il primo strato di Multi-Head Attention presta attenzione alle altre posizioni della sequenza per trovare indizi che aiutino a costruire una migliore codifica della parola. In questo modo, man mano che il vettore parola passa da un encoder all'altro, viene arricchito con informazioni riguardanti le sue relazioni con le altre parole del testo (nel precedente esempio "Butta le penne che sono affamato", per disambiguare il vettore "penne" il modello deve codificarlo prestando attenzione al vettore "affamato"). Il termine Multi-Head indica la possibilità di avere più Attention Heads che prestino attenzione in diversi modi alle relazioni tra parole, codificando i vettori "da più punti di vista".
4. Quando la pila di encoders ha codificato tutta la sequenza input, produce una serie di matrici che verranno passate ai decoders per la fase di generazione del testo.
5. La pila di decoders prende come contesto sia le matrici degli encoders sia tutte le parole generate precedentemente (tranne nel primo passaggio in cui ci sono solo le matrici). Lo strato Encoder-Decoder Attention permette al decoder di focalizzarsi su diverse parti della sequenza, svincolandolo dall'ordine della frase in input (ciò è importante, ad esempio, traducendo dall'italiano all'inglese in cui l'ordine del sostantivo e dell'aggettivo sono invertiti). Al termine dell'elaborazione, la pila di decoders produce un vettore.
6. Questo vettore finale viene elaborato dallo strato Linear che produce un vettore molto più grande che abbia tante dimensioni quante le parole del vocabolario del modello. Successivamente lo strato Softmax trasforma i valori delle dimensioni in probabilità e seleziona stocasticamente quale parola produrre in output. L'algoritmo di selezione stocastica, specifico per ogni modello, è fondamentale per far sì che non vengano prodotte sempre e solo le combinazioni più probabili di parole, ma che si esplorino anche altre combinazioni.

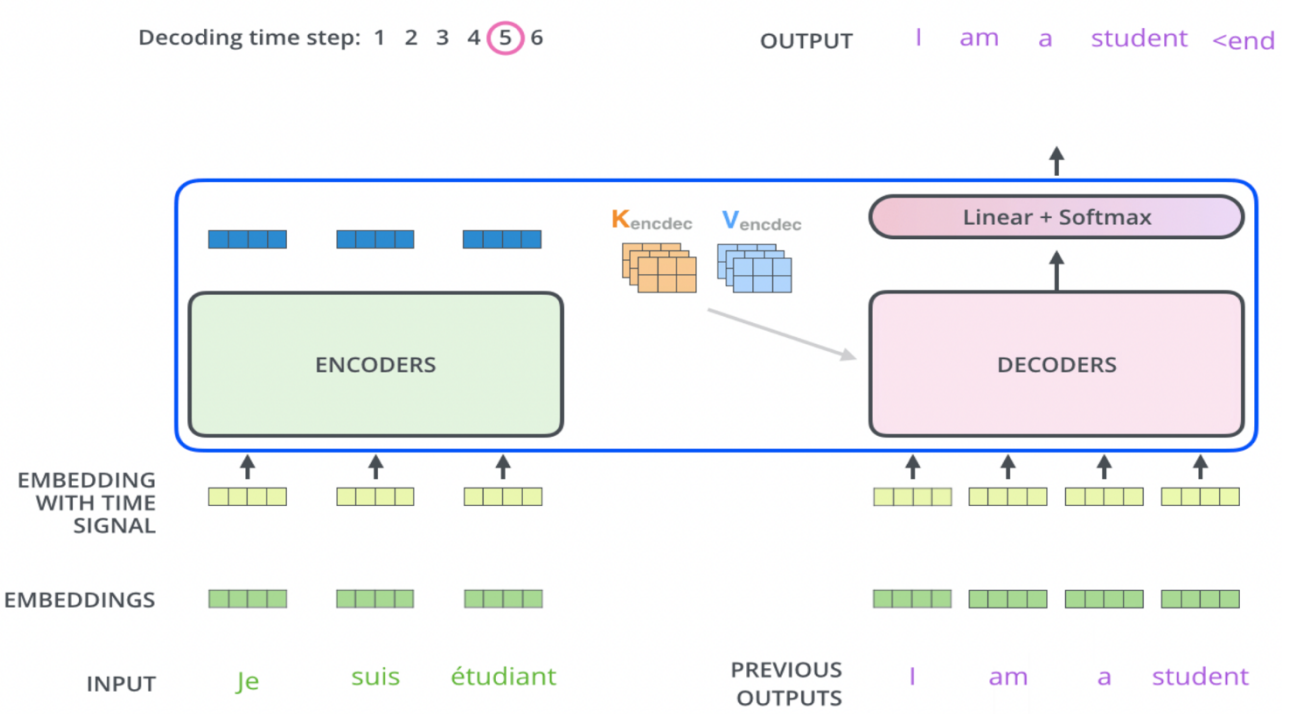

Figura 2.Ultimo time step nel processamento dell'input "Je suis étudiant" dal francese all'inglese. I vettori output dei decoders (in blu) vengono elaborati per formare due set di matrici (arancioni e azzurre) che rimangono uguali per tutte le successive elaborazioni (Alammar, 2018)

Ogni Large Language Model (LLM) è formato da più Transformers che vengono addestrati non solo a produrre testo, ma anche a farlo in una maniera precisa ed eticamente corretta grazie ad una metodologia chiamata Reinforcement Learning from Human Feedback (RLHF) in cui operatori umani indicano alla rete quale delle risposte generate sia più desiderabile e aderente alla richiesta in input (prompt). La rete risultante è un complesso modello statistico di come si correlano le frasi e le parole viste nei dati di addestramento e, grazie a queste correlazioni, è in grado di generare linguaggio naturale sorprendentemente simile a quello umano. Tuttavia la complessità delle interazioni tra le diverse parti della rete e del funzionamento delle Attention Heads rende molto complesso il comprendere a fondo *come* il modello riesca a produrre testo.

1.2 Due diverse modalità di elaborazione

All'inizio del capitolo si è discusso come, anche se entrambe le reti neurali biologiche e artificiali condividono lo stesso principio neurale di base, esistono delle differenze fondamentali che si amplificano all'aumentare della scala (a parte casi speciali di studi che richiedono reti neurali artificiali biologicamente plausibili). Le risultati modalità di elaborazione differiscono per molti aspetti, tra cui quelli relativi all'elaborazione del linguaggio.

Da una parte il cervello umano: 1) costruisce, a partire dalle esperienze fatte durante il corso della vita, un vocabolario di concetti (intesi come modelli mentali dell'ambiente esterno e del Sé) che sono localizzati in specifiche aree corticali (Huth et al., 2016); 2) gestisce l'attenzione mediante "reti attenzionali" (Petersen e Posner, 2012); 3) elaborano il linguaggio navigando tra i concetti mediante meccanismi causali, di inferenza e di deduzione.

Dall'altra i LLMs: 1) costruiscono un vocabolario di Word Embeddings osservando, per ogni parola, tutte le relazioni con le altre parole incontrate nell'immensa quantità di testo analizzata; 2) gestiscono l'attenzione mediante Attention Heads che, durante la fase di apprendimento, si auto-specializzano a prestare attenzione a diverse parti del testo; 3) elaborano il linguaggio in maniera sequenziale mediante operazioni vettoriali e attivazioni delle varie reti neurali di cui sono composti.

1.2.1 I modelli autoregressivi

I LLMs sono modelli definiti autoregressivi, in quanto la parola in output dipende dalle parole prodotte precedentemente e da una variabile stocastica (che, come si è discusso nel paragrafo sui Transformers, permette ai LLMs di non ripetersi e di esplorare nuove combinazioni di parole). Questo funzionamento puramente sequenziale di predizione della prossima parola permette di generare testo fluido e coerente, ma allo stesso tempo impedisce al modello di revisionare ciò che ha prodotto, di pianificare in anticipo come organizzare la risposta e di compiere simulazioni "mentali".

Nel tentativo di spiegare meglio questo limite, Bubeck et al. propongono un'analogia tra il modello e due costrutti psicologici introdotti da Kanheman: il Pensiero Veloce e il Pensiero Lento. Il primo è una modalità di pensiero "automatica, intuitiva e senza sforzo, ma anche più incline ad errori e bias", ed il secondo è una modalità di pensiero "controllata, razionale e dispendiosa, ma anche più accurata ed affidabile" (Kanheman, 2011). Mentre gli umani utilizzano una combinazione delle due modalità a seconda della situazione, i LLMs possono essere ritenuti capaci di eseguire solamente operazioni "veloci" ad una portata impressionante, ma gli mancano la componente "lenta" che supervisioni l'elaborazione (Bubeck et al., 2023).

In alcuni casi la mancanza di questa "voce interna" che permetta di ragionare prima di generare il testo può essere ovviata specificando al modello di risolvere il problema "passo a passo". Tuttavia, all'aumentare della complessità dei problemi, questa strategia di prompting non è più sufficiente, e l'unico modo per arrivare alla soluzione è attraverso una comprensione profonda e globale del problema (come nel caso di alcune dimostrazioni matematiche).

L'assenza di pianificazione è evidente anche nei problemi di "generazione di testo vincolata, in cui si chiede al modello di generare contenuto testuale secondo specifiche istruzioni che contengono vincoli strutturali" (Bubeck et al., 2023). Quando questi vincoli sono globali (coinvolgono parti di testo distanti fra loro) il modello non riesce a rispettare le istruzioni (come nel caso del prompt "Scrivi una breve storia in cui l'ultima frase contenga metà delle parole della prima frase").

1.2.2 Il dibattito sulla comprensione dei LLM

La comprensione umana può essere vista come un fenomeno fondato sull'organizzazione dei concetti in un insieme coerente e gerarchico di relazioni causali, che permette alle persone di apprendere astraendo conoscenza dalle esperienze ("Le piante hanno bisogno di acqua per sopravvivere"), di semplificare e dare un senso al mondo ("Vado ad un matrimonio quindi mi vesto elegante"), e di fare anticipazioni, generalizzazioni e analogie ("Se disinfetto la ferita, guarirò prima") (Mitchell e Krakauer, 2023). Gli umani sembrano avere una spinta innata per questo tipo di comprensione che consente di risparmiare risorse cognitive grazie a modelli minimali e parsimoniosi che richiedono poche informazioni per funzionare.

I LLMs, diversamente, non fanno esperienza del mondo, non creano dei modelli interni abbastanza ricchi ed efficienti da semplificare in maniera significativa l'ambiente, e non fanno anticipazioni e generalizzazioni (ma riescono a fare analogie in una certa misura). Questi modelli imparano a conoscere come ogni parola si colleghi con le altre, ma si può dire che questa conoscenza generi comprensione?

Questa incertezza ha fatto sì che gli esperti di Natural-Language-Processing si dividessero di fronte all'ipotetica capacità di questi modelli di comprendere il linguaggio naturale. Quando, nel 2022, alla comunità NLP è stato sottoposto un sondaggio riguardante la capacità dei LLM di "comprendere il linguaggio naturale in una maniera non-triviale", i risultati hanno mostrato la netta divisione tra quelli a favore (51%) e quelli contro (49%) (Michael et al., 2022). Le argomentazioni degli esperti a favore si concentrano soprattutto sull'evidenza che aumentare la dimensione delle reti migliori la performance, la robustezza e diminuisca gli errori "unhumanlike". Secondo questa logica, aumentando sufficientemente le dimensioni si potrebbe arrivare a modelli dotati di intelligenza e comprensione di livello umano o superiore (alcuni esperti sostengono addirittura che potrebbero acquisire consapevolezza).

Al contrario, l'altra fazione sostiene che i LLM non possano comprendere il linguaggio perché l'addestramento per dati testuali gli insegna la *forma* del linguaggio, non il *significato*. Comprendere il significato di "solletico" significa collegare questa parola ad una sensazione, non a un'altra parola, e queste reti non hanno né esperienza né modelli mentali del mondo (Mitchell e Krakauer, 2023). Sebbene siano in grado di generare linguaggio simile a quello degli umani per fluidità e correttezza grammaticale (competenza linguistica formale), sono prive dei concetti necessari per poter comprendere ed utilizzare il linguaggio nel mondo reale (abilità funzionali del linguaggio) (Mahowald et al., 2023). È indiscutibile che con l'evolversi di questi sistemi emergano alcuni comportamenti intelligenti, ma, come si chiede Sejnowski, "se non è umana, qual è la natura della loro intelligenza?" (Sejnowski, 2022).

I ricercatori su questo lato del dibattito vedono gli LLM come repertori compressi di conoscenza umana più simili a librerie e enciclopedie che a sistemi intelligenti. Dal momento che la loro capacità di elaborare questa conoscenza è solo frutto di relazioni statistiche tra parole piuttosto che relazioni causali tra concetti, risulta fuorviante utilizzare i costrutti di "intelligenza", "comprensione" e "libero arbitrio" nello stesso modo in cui vengono utilizzati in psicologia (Mitchell e Krakauer, 2023).

1.2.3 Abilità emergenti

In un articolo del 2022, un team congiunto di ricercatori provenienti da Google, DeepMind e da università americane ha cercato di fare luce sul fenomeno delle abilità emergenti nei LLM (Wei et al., 2022). In questa ricerca l'emersione viene descritta come il fenomeno secondo cui "i cambiamenti quantitativi in un sistema risultano in cambiamenti qualitativi nel comportamento". Per cambiamenti quantitativi vengono intesi non solo il numero dei parametri (connessioni tra i neuroni), ma anche la quantità di risorse computazionali e la dimensione del dataset di addestramento. In tutti i test sottoposti ai vari LLM si nota come la performance delle reti, superata una certa dimensione soglia, smette di essere casuale ed incrementa vorticosamente. Anche se rimane da scoprire come e perché le dimensioni sblocchino le abilità emergenti, la gran quantità di risultati ottenuti da reti neurali artificiali nettamente diverse tra loro ha permesso di scoprire altri fattori, oltre ai cambiamenti quantitativi sopra menzionati, correlati al fenomeno: il tipo di architettura, la qualità dei dati, le procedure di addestramento e le tecniche di prompting. Tutto ciò è utile per capire da un lato se l'aumento delle dimensioni e dei dataset di addestramento è tutto ciò che serve per permettere ai LLM di risolvere i compiti ancora fuori dalla loro portata (frontier tasks), e dall'altro come fare a progettare modelli più piccoli e computazionalmente economici che performino allo stesso livello, o superiore, dei modelli più grandi.

CAPITOLO 2 – STRUMENTI DI VALUTAZIONE DEI LLMs

A partire dalle reti neurali più semplici, l'aumento della scala genera cambiamenti qualitativi (come l'acquisizione di abilità emergenti o la specializzazione di neuroni o di porzioni di rete) che rendono progressivamente più difficile comprendere il reale funzionamento del sistema. Nel campo del Machine Learning i ricercatori sono spesso portati a fare un compromesso tra accuratezza e intelligibilità: i sistemi con un funzionamento accessibile e decifrabile hanno un'accuratezza limitata, mentre quelli più accurati sono visti come scatole nere (Caruana et al., 2015). L'intelligibilità è una caratteristica della rete importante per l'avanzamento della ricerca di base e per le applicazioni nel mondo reale:

- Distinguere il contributo dei singoli blocchi della rete permetterebbe di identificare e correggere gli errori, amplificare i punti di forza, e più in generale orientare la ricerca verso la scoperta di architetture più potenti (Clune, 2020).
- La conoscenza dei meccanismi neurali alla base delle abilità consentirebbe di progettare in anticipo un modello che possegga quelle desiderate.
- La possibilità di comprendere i processi dietro ad una risposta o decisione è una condizione fondamentale per le applicazioni nel mondo reale (ad esempio in campo medico, giuridico ed economico).

Per queste ragioni una buona parte della ricerca è dedicata alla costruzione di strumenti di valutazione che aiutino a gettare luce sui meccanismi interni di questi modelli. Con l'avvento dei LLMs si è aperta la strada ad una nuova, vasta gamma di compiti veicolabili attraverso il linguaggio naturale, grazie alla quale i modelli possono essere testati in numerosi domini (dai processi di ragionamento a quelli di presa di decisione, dalla 'comprensione' dei codici informatici alla scrittura di programmi, dalla conoscenza delle discipline alla generazione di nuovo materiale, e molto altro). La ricchezza e diversità di questi domini fornisce numerosi elementi per poter argomentare quali possano essere i processi alla base delle risposte dei modelli.

Nel presente capitolo esporrò tre approcci allo studio dei LLM (ma anche di altri tipi di rete neurali) ed i loro relativi punti di forza. Ho deciso di discutere i loro limiti e le sfide insite nel loro utilizzo alla fine del capitolo per la fluidità dell'argomentazione.

2.1 Benchmarks

Nel campo dell'intelligenza artificiale, per valutare e comparare la performance dei vari modelli su problemi più o meno specifici, vengono generalmente utilizzati dei test standardizzati chiamati benchmarks. Thiyagalingam et al. discutono i fattori che bisogna considerare nel costruire strumenti di valutazione scientifici adeguati per l'intelligenza artificiale, tra cui: il focus della valutazione (ad es. il raggiungimento di target scientifici o la comparazione tra diversi sistemi hardware), la scelta delle metriche (che influenza la comparabilità dei risultati con quelli di altri benchmarks) oppure le caratteristiche dei dati da includere (quanto numerosi, quanto distribuiti, riguardanti quali argomenti, etc.) (Thiyagalingam et al., 2022).

Fino a poco tempo le reti neurali erano in grado di processare un unico tipo di input e venivano dunque definite *unimodali*. Per questa ragione svolgevano dei compiti specifici sulla base dei quali venivano costruiti i benchmarks (ad es. due dei benchmarks più famosi nel campo del riconoscimento di immagini sono MNIST e ImageNet). Il recente sviluppo di reti neurali sempre più avanzate ha reso possibile l'elaborazione di più tipi di input, facendole guadagnare il termine di *multimodali*. Dal momento che queste reti possono svolgere una quantità maggiore di compiti, possono essere testate su più benchmarks (o benchamarks comprensivi di più tipi di problemi) e, in alcuni casi, i loro risultati hanno addirittura superato quelli delle reti costruite appositamente per un certo tipo di compito.

2.2 Approcci psicologici

Un altro approccio per migliorare la comprensione di questi modelli prevede l'utilizzo del ricco materiale teorico e metodologico sviluppato dalla Psicologia Cognitiva per studiare la mente umana. Questo materiale include teorie ed esperimenti riguardanti sia fenomeni psicologici di alto livello (osservabili principalmente attraverso il linguaggio, come la presa di decisioni o il ragionamento causale), sia di basso livello (veicolati non solo tramite il linguaggio ma anche i sensi, come le euristiche e i bias).

Come primo esempio riporto un recente studio un cui il LLM GPT-3 viene sottoposto ad una serie di problemi tratti da test di psicologia cognitiva volti ad indagare, *negli umani*, abilità cognitive di alto livello quali: presa di decisione, ricerca di informazioni, ragionamento casuale e la riflessione (Binz e Schulz, 2023). I problemi appartengono a "quattro esperimenti classici ritenuti rappresentativi della letteratura di psicologia cognitiva" e sono costruiti in maniera tale da poter distinguere, per ciascuna abilità cognitiva, le varie strategie (ma anche euristiche e bias) con cui un individuo può giungere alla soluzione (esplorazione direzionata e/o randomica, apprendimento libero o basato su modelli, e altre). L'articolo si conclude discutendo i "sorprendenti risultati [...] che potrebbero indicare –almeno in

alcuni casi– che GPT-3 non è solo un pappagallo stocastico ma potrebbe passare come un soggetto valido per alcuni degli esperimenti somministrati". Tuttavia, come viene ribadito per ben due volte nella citazione precedente, i casi di successo sono solo parziali, accompagnati dai "meno sorprendenti casi di fallimento".

Il secondo esempio riguarda un fenomeno psicologico di basso livello studiato in Psicologia dello Sviluppo: lo Shape Bias, ossia la tendenza, nell'imparare nuove parole, ad assegnare lo stesso nome ad oggetti simili per forma piuttosto che per colore, consistenza o dimensione (Landau et al., 1988). Basandosi su questo studio originale, Ritter et al. (2017) creano un dataset e condizioni sperimentali per testare l'ipotesi se anche le reti neurali profonde, addestrate per l'apprendimento one-shot di parole, mostrano lo stesso bias. Il fenomeno viene studiato sia comparando due diverse architetture neurali (Matching Networks e Inception Networks) sia confrontando la stessa architettura ma con diverse inizializzazioni di pesi. I risultati confermano che i modelli, nonostante non vengano esplicitamente ottimizzati a categorizzare la forma, sviluppano spontaneamente uno Shape Bias. Portano inoltre ad altre due scoperte: 1) l'alta variabilità dei risultati, *anche tra modelli architettonicamente identici*, dimostra che i pesi iniziali hanno un effetto sul grado con cui le reti neurali sviluppano lo Shape Bias; 2) nella condizione in cui la Matching Network eredita i dati elaborati precedentemente dall'Inception Network, i risultati rivelano che viene ereditato anche lo Shape Bias, dimostrando come questo possa propagarsi attraverso i componenti di un modello.

Mentre i benchmarks forniscono solo dati riguardanti la capacità del sistema intelligente di svolgere uno o più compiti (anche se, come visto nel paragrafo precedente, si possono considerare più metriche che diano più informazioni), l'approccio psicologico fornisce dati anche sulle modalità (o strategie) con cui viene svolto. Ciò permette di argomentare i possibili meccanismi dietro ad una certa abilità manifesta.

2.3 Nuovi approcci

Alcuni gruppi di ricerca ritengono che i benchmarks e i test psicologici non siano sufficienti a valutare le vaste capacità dei LLMs, come il caso del team di ricerca Microsoft che nell'aprile 2023 ha pubblicato un'indagine sulla versione unimodale di GPT-4. L'approccio sviluppato in questa indagine si basa su "compiti innovativi e complessi" che fanno leva sulla creatività e curiosità per dimostrare che "il modello non si limita a memorizzare, ma che ha una comprensione profonda e flessibile di concetti, competenze e domini" (Bubeck et. al., 2023). Gli autori riconoscono che questo approccio sia in qualche misura soggettivo e informale, ma lo ritengono "un primo passo necessario per

apprezzare le notevoli capacità dei LLMs più avanzati". Per fare ciò i compiti presentati si concentrano su 5 aree:

1. **Padronanza del linguaggio naturale** = vengono testate le capacità di:
    a. sintetizzare informazioni da differenti domini o modalità attraverso compiti come quello di "produrre codice che generi immagini casuali nello stile del pittore Kandinsky" oppure di "scrivere la dimostrazione che esistono infiniti numeri primi nello stile letterario di Shakespeare".
    b. applicare conoscenze e abilità in diversi contesti e discipline, dove per "diversi" s'intende che non erano presenti nell'addestramento. Infatti, nonostante abbia avuto un addestramento puramente testuale, si può chiedere a GPT-4 di generare immagini scrivendo codice in Scalable Vector Graphics (SVG). Compiti come quello di disegnare figure utilizzando lettere dell'alfabeto, oppure di generare immagini seguendo istruzioni dettagliate, dimostrano che il modello è in grado di gestire concetti visuali. Riesce inoltre a manipolare, seppur in maniera rudimentale, i concetti musicali.
    c. fare distinzioni tra diversi stimoli, concetti e situazioni chiedendo al modello di individuare in un testo le Informazioni di Identificazione Personale (PII) (che non hanno una chiara definizione ma sono spesso contesto-specifiche e dunque difficili da identificare), oppure di generare risposte a domande specifiche per misurarne la veridicità e la somiglianza alla "gold answer".
2. **Programmazione** = GPT-4 ha delle eccellenti abilità di programmazione sia in termini di generazione di codice (facendogli risolvere problemi di programmazione in scenari realistici, come estrarre e visualizzare dati, scrivere giochi 3D, interfacciarsi con linguaggio LATEX), sia di comprensione di codice esistente (è in grado di testare la sicurezza di un software, di predire l'output di un programma, eseguire codice Python e pseudo-codice, ovvero codice scritto con espressioni vaghe o informali).
3. **Matematica** = oltre ai classici benchmarks, in questo approccio vengono proposte a GPT-4 conversazioni matematiche teoriche (manipolazione concetti) e pratiche (situazioni realistiche). Nel mondo del NLP, la matematica è una delle aree più problematiche per i LLMs, ma se si concepisce la comprensione matematica come un insieme di più aspetti, i problemi si limitano solo ad alcuni di essi:
    a. il Ragionamento Creativo permette di identificare quali passaggi sono necessari per arrivare alla soluzione (GPT-4 lo dimostra ad un alto livello);

b. la Competenza Tecnica identifica l'abilità di eseguire procedure, come isolare un termine in un'equazione (GPT-4 dimostra di conoscere le procedure, ma compie spesso errori basilari)

c. il Ragionamento Critico permette di individuare, per ogni passaggio, i sottocomponenti e le relazioni con il resto dell'argomentazione (l'area di maggiore difficoltà di GPT-4, possibilmente a causa sia dei dati di addestramento che spesso includono il problema e la soluzione ma non il processo, sia del suo funzionamento lineare di predizione della prossima parola senza revisione del suo output precedente).

4. **Interazione col mondo** = gli autori definiscono l'Interattività come "l'abilità di comunicare e reagire agli stimoli provenienti da altri agenti, strumenti e contesti", strettamente collegata all'apprendimento dall'esperienza e all'applicazione della conoscenza (aspetti strettamente collegati con l'intelligenza). Nell'articolo vengono indagate due dimensioni dell'Interattività:

   a. l'Utilizzo di strumenti informatici (come API, motori di ricerca o addirittura altre reti neurali) è una strategia molto efficace nel superare i limiti e le debolezze tipiche dei LLMs. Nei compiti presentati (penetrare in un computer, generare un ambiente virtuale (Figura 3), gestire calendario e mail, navigare internet, ed altri) GPT-4 dimostra di saper comprendere il compito, identificare gli strumenti disponibili, usarli adeguatamente e rispondere al loro feedback.

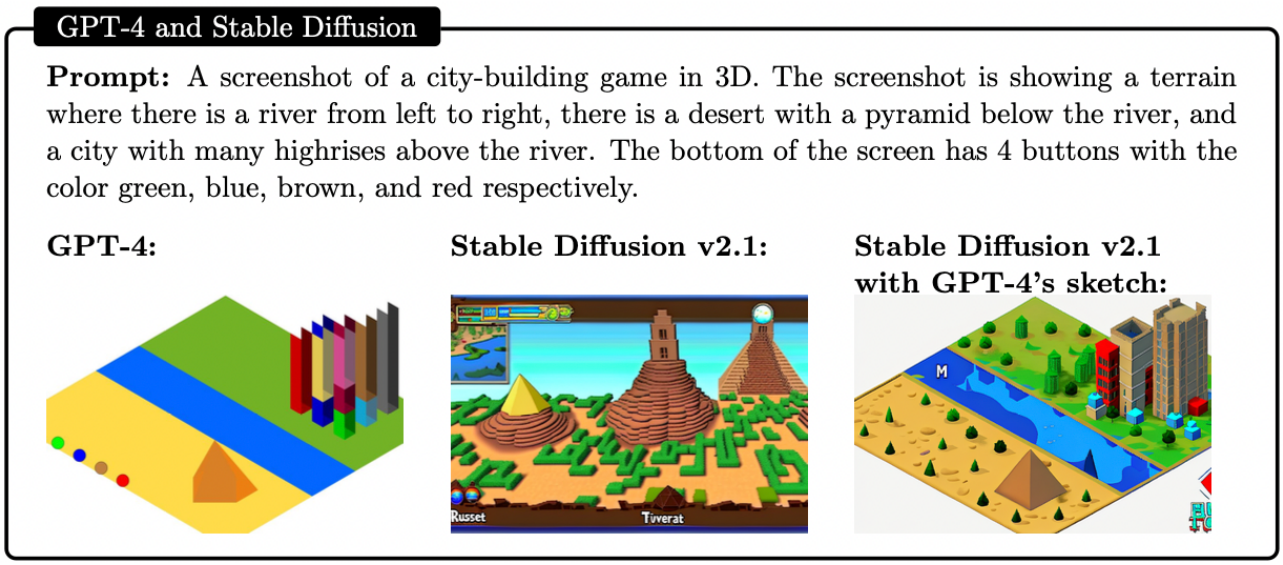

Figura 3. GPT-4 utilizza Stable Diffusion per risolvere la richiesta con più precisione

   b. con Interazione [Incorporata/impersonata] (Embodied Interaction) s'intende l'interazione non mediata da strumenti, ma diretta tra un agente e l'ambiente. Mentre gli umani usano molteplici forme di linguaggio con cui comunicare (verbale, fisico, musicale, etc.), i LLMs hanno unicamente quello verbale. Possono dunque interagire con ambienti realistici o simulati attraverso il linguaggio naturale in compiti come:

navigare una mappa, giocare a text-based games, risolvere problemi del mondo reale (come il trovare una perdita in una casa).

5. **Comprensione dell'umano** = in questa sezione si indaga la Teoria della Mente, ovvero "l'abilità di attribuire stati mentali [...] e di comprendere come questi influenzino il comportamento e la comunicazione (Wellman, 1992). Questa abilità è essenziale per:
    a. comprendere i contesti sociali: quando gli vengono presentate situazioni sociali di diversa complessità, GPT-4 riesce ad individuare la radice del problema e fornisce suggerimenti che favoriscano la cooperazione verso uno scopo comune.
    b. spiegare il proprio comportamento (sia a sé stessi come forma di ragionamento, sia agli altri per farsi capire): al modello vengono fatte delle domande (ad es. "Che anno è?" o "Traduci questa frase in portoghese") e gli viene chiesto di spiegare perché abbia dato la sua risposta. Due possibili modi per valutare la qualità di una spiegazione sono considerare se questa è coerente con la risposta, anche quando questa è sbagliata (*output consistency*: una spiegazione coerente con la risposta "Siamo nel 1400" potrebbe essere "Perché sono un'IA medievale"), oppure se è coerente con il processo che ha generato la risposta (*process consistency*: se il modello afferma che la traduzione standard di "insegnante" al portoghese è femminile, quando gli si chiede di tradurre una frase con tale parola dovrebbe tradurla al femminile). GPT-4 mostra alta *output consistency* ma bassa *process consistency*.

Nelle conclusioni dell'articolo il Team Microsoft sostiene che "GPT-4 possegga una forma di intelligenza *generale* [grazie a] le sue capacità mentali (come ragionamento, creatività e deduzione), la gamma di argomenti in cui è esperta (come letteratura, medicina e informatica), e la varietà di compiti che riesce a risolvere (ad es. giocare, usare strumenti, spiegarsi, ...)".

2.4 Limiti e sfide

2.4.1 degli strumenti testuali

Il limite più citato in letteratura riguardo agli strumenti di valutazione contenenti problemi testuali è che il LLM in questione possa aver già incontrato quegli stessi problemi insieme alle loro soluzione durante l'apprendimento, e che quindi produca una risposta "a memoria". Un altro limite, per alcuni versi anche più insidioso, è lo Shortcut Learning: fenomeno secondo il quale un sistema si affida a correlazioni spurie nei dati per poter ottenere una buona performance in un determinato benchmark (McCoy et al., 2019; Geirhos et al., 2020). Il termine 'spurie' si riferisce al fatto che il modello scopra solo le correlazioni tra le parole dei problemi presentati e non quelle tra i concetti del mondo reale

(che porterebbero ad un'effettiva 'comprensione della realtà'). Normalmente queste correlazioni non sono riconoscibili dagli umani, ma i LLM, grazie ai loro meccanismi statistici, hanno una predisposizione ad individuarle ed utilizzarle per risolvere i problemi riuscendo ad ottenere una performance quasi perfetta senza i processi di ragionamento e comprensione che gli autori dei benchmark si prefissano di misurare.

Fino ad oggi, le tecniche impiegate per contrastare questi problemi si concentrano soprattutto sulla modifica del prompt: cambiare l'ordine delle parole oppure le parole stesse (sostituendole, togliendone alcune specifiche o aggiungendone di nuove). Prendiamo come esempio il benchmark Argument Reasoning Comprehension Task (Habernal et al., 2018). In ogni problema viene fornita un'argomentazione (composta da un'affermazione e una motivazione) e due enunciati: l'obiettivo è determinare quale dei due enunciati è consistente con l'argomentazione. Si riporta un esempio dal dataset:

> **Argomentazione:** Rifiutare le cure mediche causa malattia e morte. Le credenze religiose dei genitori non gli dovrebbe permettere di rifiutare le cure mediche per i propri figli.
>
> **Enunciato A:** Dio ha creato la vita.
>
> **Enunciato B:** Dio ha creato la malattia e la morte.

Inizialmente il modello BERT aveva ottenuto performance quasi a livello umano, ma una volta scoperto che la presenza di alcune parole nelle frasi (ad esempio "non") poteva aiutare a predire la risposta corretta, si è alterato il dataset per prevenire queste correlazioni e la performance del modello è calata quasi a quella ottenibile rispondendo casualmente (Niven e Kao, 2019). Nonostante i modelli possano navigare per sinonimi e catturare relazioni tra parole lontane tra loro, questi esempi dimostrano che alterazioni anche piccole del contesto in input possono influenzare in maniera negativa tutto il processo di generazione dell'output.

Entrambi fenomeni sono ben conosciuti e studiati, tuttavia la comprensione parziale di quali siano i loro meccanismi e fattori influenzanti rende complicato individuarli, contrastarli e capire se e come siano state effettivamente risolti, lasciando senza risposta domande come "Perché alcuni cambiamenti portano a differenze di performance ed altri no?". Anche se in alcuni benchmark ne sono state individuate e corrette alcune, ce ne sono possibilmente molte altre ancora da scoprire che compromettono i risultati degli studi e dunque invalidano le conclusioni tratte sul funzionamento dei modelli.

2.4.2 degli strumenti psicologici

Nonostante l'approccio psicologico fornisca informazioni che "completano i benchmarks esistenti in maniere significative" (Binz e Schulz, 2023), bisogna ragionare sulla natura di queste informazioni, in virtù della quale può essere più o meno adeguato trarre certe conclusioni sui meccanismi neurali, sia biologici che artificiali.

Di fatto, anche per i processi di basso livello ('più semplici' di quelli di alto livello nei termini in cui si basano su reti neurali più piccole e quindi coinvolgono meno meccanismi) è necessario tenere a mente che gli strumenti psicologici da soli "non forniscono approfondimenti sui meccanismi neurali" ma che, "così come la psicologia cognitiva spesso fa per le neuroscienze, [in informatica] gli approfondimenti a livello computazionale possono fornire un punto di partenza per la ricerca a livello meccanicistico" (Ritter et al., 2017). È il caso degli studi sullo Shape Bias (paragrafo 2.2) che fino ad oggi hanno 'solo' riscontrato questa preferenza per la forma sia negli umani che nelle macchine, ma non sono ancora state tratte conclusioni su quali meccanismi neurali (sia nella corteccia celebrale che nelle reti neurali artificiali) riescano a spiegare la preferenza per la forma rispetto al colore, consistenza o dimensione.

Un altro modo per approcciarsi ai processi di basso livello consiste nel costruire una rete neurale che, se sottoposta ad un test psicologico, manifesti lo stesso comportamento riscontrato per soggetti umani sottoposti allo stesso test. Seguendo questa linea di ricerca, alcuni studi nel campo della percezione di lettere e parole hanno ottenuto sorprendenti risultati, riuscendo a svelare alcuni meccanismi neurali come l'effetto contesto (McClelland e Rumelhart, 1981), o scoprendo fattori influenzanti come la percezione di immagini naturali (Changizi et al., 2006; Testolin et al., 2017).

I processi di alto livello sono una questione ben più insidiosa principalmente per due ragioni: la prima è che negli umani, i sub-strati neurali che li supportano sono decisamente più complessi ed estesi (e quindi più difficili da riprodurre artificialmente); la seconda è che la quasi totalità degli strumenti volti ad indagarli comportano l'utilizzo del linguaggio naturale. La seconda ragione, oltre ad esporci ai rischi spiegati nel paragrafo precedente, ci pone di fronte anche ad un'altra questione: se da una parte un LLM è un singolo modello, dall'altra è anche vero che è addestrato per imitare testo scritto da tanti diversi individui. Va dunque considerato come singolo partecipante o più? (Binz e Schulz, 2023).

I modelli in grado di produrre comportamenti complessi (dietro ai quali si suppone ci siano processi di alto livello, anche se non si sa di che tipo) sono costituiti da più reti neurali interconnesse tra loro per formare massicce architetture con un numero di parametri che arriva all'ordine dei triliardi. A queste grandezze, e con architetture nettamente diverse tra loro, è chiaro che le modalità di

elaborazione dei dati siano qualitativamente diverse tra macchine e umani. Dunque, “a causa delle differenze sostanziali tra LLMs e umani come oggetti di studio in psicologia, potrebbe essere inappropriato assumere che le risposte di un LLM possano essere analizzate così come gli psicologi cognitivi analizzerebbero il comportamento umano con lo stesso compito” (Shiffrin e Mitchell, 2023; Mitchell e Krakauer, 2023). Infatti, per gli psicologi cognitivi, il comportamento umano orientato verso un obiettivo può essere visto come l’applicazione di una o più strategie. Tralasciando le implicazioni epistemologiche della Psicologia secondo le quali ogni teoria può esporre la propria serie di strategie, non esiste nessun criterio rigoroso in base al quale si possa asserire che un modello utilizzi le stesse strategie di un individuo. Risultati come quelli riportati da Binz e Schulz dai quali si ricava, ad esempio, che “il modello utilizza l’esplorazione randomica, ma non in maniera strategica, mentre non impiega affatto l’esplorazione direzionata”, non contemplano la possibilità che il modello utilizzi strategie diverse rispetto agli umani.

CAPITOLO 3 – LE TRE STRADE VERSO L'AGI

Verso dove sta procedendo lo sviluppo di nuove intelligenze artificiali? Con il computer la strada era relativamente delineata: più potenza di calcolo in meno spazio possibile. Raramente venivano inventate componenti nuove, e la maggior parte della ricerca si concentrava su come rendere più performanti e più piccole le componenti già esistenti. Al contrario, i modelli di intelligenza artificiale hanno un solo componente essenziale: il neurone. Da qui in poi ci sono pressoché infinite possibilità di organizzarlo e addestrarlo. Infatti, ad oggi, sviluppare reti neurali più avanzate comporta o la scoperta di nuove architetture (più grandi ma anche più efficienti) o lo sviluppo di nuove metodologie di addestramento (algoritmi d'addestramento, ordine di presentazione dei training, qualità vs quantità dei dati, etc.). È sufficiente la scoperta di una nuova architettura più efficiente, come il Transformer, e tutto il mondo dell'Intelligenza Artificiale accelera improvvisamente. Per queste ragioni, nel campo dell'Intelligenza Artificiale è difficile immaginarsi un traguardo tecnologico ben definito che riesca ad orientare la ricerca e lo sviluppo.

Ciononostante, già a partire dagli anni '80, filosofi e ricercatori hanno iniziato a distinguere le intelligenze artificiali che eccellono in compiti specifici (la "weak AI", ossia la (quasi) totalità dei modelli creati ad oggi) dalle altre forme superiori che, nel corso degli anni, hanno ricevuto più definizioni: "strong AI", "human-level AI", "Universal Artificial Intelligence", ma attualmente vi si riferisce principalmente con il nome di "Artificial General Intelligence" (AGI) (Searle, 1980; Hutter, 2005; Legg, 2008; Goertzel, 2014; Bubeck et al., 2023). Ciascuna definizione rimanda a determinate caratteristiche e abilità che il modello deve possedere, ma tra le più importanti c'è sicuramente la capacità di generalizzazione, dalla quale deriva la definizione di AGI. Il modello che la possiede sarebbe in grado di astrarre conoscenza dall'esperienza, creare modelli del mondo e connettere più concetti tra loro, e dunque potrebbe potenzialmente risolvere una vastissima gamma di compiti. Non si è ancora riusciti a creare reti neurali che posseggano quest'abilità, ma alcuni autori hanno già avanzato possibili "architetture e paradigmi d'addestramento grazie ai quali poter costruire agenti intelligenti autonomi" (LeCun, 2022).

Tralasciando il problema del *cosa* si stia cercando di costruire, Jeff Clune si domanda invece il *come* riuscirci (Clune, 2019). La sua idea è che esistano tre strade percorribili per poter costruire un'AGI, ciascuna caratterizzata da grandi potenzialità ma anche da limiti economici e di tempo.

3.1 The Mimic path

Concettualmente la più chiara delle tre, questa strada fa riferimento al "tentativo di attuare un processo di reverse engineering sull'intelligenza animale per capire come funzioni così da poterla ricostruire

computazionalmente". Il vantaggio più grande sarebbe il disporre di una copia dettagliata del cervello da poter liberamente osservare e manipolare per studiare l'intelligenza umana.

Tuttavia questa strada è probabilmente la più lunga, rallentata dalla difficoltà nel creare tecnologie in grado di osservare ciò che avviene dentro ai cervelli in funzione. Inoltre, perché la copia sia veramente fedele, non si potrebbe nemmeno beneficiare delle 'scorciatoie' che sono possibili computazionalmente ma non biologicamente (ad es. per identificare la presenza di barre verticali in un'immagine, una rete neurale artificiale necessita di un solo neurone specializzato che analizzi tutta l'immagine, mentre nel cervello ce n'è uno per ogni porzione di retina da analizzare).

3.2 The Manual path

Attualmente la più perseguita, questa strada include tutti i tentativi di costruire manualmente reti neurali sempre più avanzate. Il termine "manuale" si riferisce al fatto che i ricercatori devono fare una serie di scelte riguardanti gli iper-parametri della rete (architettura, connettività, funzioni di attivazione, learning rate, etc.), i dati di addestramento (quantità, qualità, variabilità, etc.) e la procedura di addestramento (durata, ordine di presentazione, suddivisione tra training set e test set, etc.). Quando si costruisce una weak AI queste possibili scelte sono numerose ma comunque limitate, e, una volta selezionato un determinato compito, la ricerca riesce a sviluppare modelli progressivamente più efficienti. Il problema si pone quando ci si sposta verso la strong AI che, per poter risolvere più tipi di compiti, deve necessariamente essere composta da più moduli, aggiungendo alle scelte già citate anche quelle su come combinare tra loro questi moduli. Oggi i LLMs rappresentano un primo passo verso la strong AI sia perché la capacità di elaborare linguaggio naturale permette l'accesso a molti più compiti, sia perché gli ultimi modelli ormai non elaborano più solo il linguaggio ma anche le immagini e i suoni (sono quindi multimodali). Ciononostante siamo ancora molto lontani dalla strong AI, infatti i ricercatori parlano al massimo di "scintille di Intelligenza Artificiale Generale" (Bubeck et al., 2023).

Il primo ostacolo di questo approccio è che, già a partire dalla fase di progettazione, bisogna scegliere quali abilità e caratteristiche possiederà il modello, il che vuol dire dare una definizione di intelligenza. Se ciò era già difficile per gli umani, la cui intelligenza esiste già e va 'solo' osservata e descritta, per le macchine lo è ancora di più in quanto bisogna costruirla da zero. Una volta concordata la definizione si avvia un processo a due fasi:

1. **Selezione dei Building Blocks**: così Clune chiama tutti gli elementi necessari a costruire manualmente l'AGI, che, come detto qui sopra, vanno dagli iper-parametri della rete ai dati e procedure di addestramento. Ma non basta fare una selezione, bisogna anche scoprire quale

sia la giusta variante di ciascuno di essi. Ad oggi la maggior parte degli articoli scientifici è volta a introdurre o perfezionare un singolo Building Block e, per dare un'idea della loro numerosità, Clune ne riporta una lista "parziale" di 58. A fronte di questa fase possono sorgere delle domande: Siamo già in possesso dei moduli necessari per costruire un'AGI? E se non lo siamo, quanti altri ne servono e quanto tempo impiegheremo per scoprirli tutti? In base a quali criteri si può dire di aver raggiunto la giusta variante dei moduli?

2. **Combinazione dei Building Blocks**: le difficoltà di questa fase stanno nella numerosità di possibili combinazioni tra cui scegliere, nella compatibilità dei moduli vincolata al loro funzionamento, e "nelle sfide ingegneristiche e scientifiche inerenti al lavorare con complessi sistemi di decine o centinaia di parti interagenti, ciascuna delle quali è a sua volta un complesso modulo di machine learning simile ad una scatola nera". Altri studi riportano anche dei casi in cui bias originati in un modulo vengono trasmessi agli altri moduli (Ritter et al., 2017). Per tutte queste ragioni, trovare errori e scoprire quali regolazioni dover compiere per far funzionare questo sistema richiede tantissimo tempo e sforzo.

In aggiunta a tutte queste difficoltà, bisogna considerare anche quella prettamente organizzativa. Infatti questa strada è percorribile solo con la collaborazione di un team sufficientemente grande che lavori assieme per un periodo di tempo molto probabilmente nell'ordine degli anni. Ciò, oltre a richiedere ingenti risorse economiche, non si sposa bene con la cultura scientifica, la quale procede, anche nelle organizzazioni come OpenAI o DeepMind, per team ristretti che lavorano separatamente per riuscire a pubblicare articoli in un periodo di tempo necessariamente limitato. Nonostante queste considerazioni, Clune non esclude la possibilità che la strada manuale sia la prima ad arrivare a produrre un'AGI, ma aggiunge che, anche se non lo fosse, sarebbe comunque utile da perseguire.

3.3 The Artificial Intelligence–Generating Algorithms (AI-GAs) path

Questa strada segue un pattern ricorrente nel Machine Learning e nella ricerca sull'Intelligenza Artificiale: quando la comunità scientifica vuole costruire un sistema intelligente prova prima a progettarlo interamente a mano, poi ne costruisce solo alcune parti lasciando che sia il sistema a capire come combinarle al meglio, ed ultimamente, con sufficienti dati e potere computazionale, riesce ad 'apprendere' l'intero sistema. Ciò è possibile grazie allo sviluppo di algoritmi che cerchino da soli non solo gli iper-parametri ottimali, ma anche nuove architetture neurali e quindi, potenzialmente, potrebbero imparare tutto il necessario per creare un'AGI.

Per quanto ambiziosa possa sembrare questa strada, è l'unica che abbiamo la certezza funzioni perché di fatto ha già generato un'intelligenza generale: noi umani. Infatti l'evoluzione darwiniana non è

altro che un semplicissimo algoritmo che per 3,5 miliardi di anni ha accompagnato lo sviluppo delle forme di vita da semplici molecole organiche fino a creare la mente umana. In questo senso, l'evoluzione darwiniana è il primo algoritmo generatore di intelligenza generale.

Nei processi evolutivi ci sono tre componenti essenziali: l'agente, l'ambiente e l'algoritmo evolutivo. L'ambiente fornisce le risorse e le sfide che l'agente deve imparare a sfruttare e risolvere per passare la selezione dell'algoritmo evolutivo. Nel caso dell'evoluzione biologica queste tre componenti sono strettamente interconnesse e interdipendenti, ma se riprodotte artificialmente possono assumere innumerevoli forme. Per scoprire quali di queste forme porterebbero alla creazione di un'intelligenza generale bisogna astrarre le "condizioni necessarie, sufficienti, e catalizzanti" che si sono create sulla Terra e riprodurle artificialmente attraverso i tre componenti:

1. **Agente:** a differenza delle prime forme di vita che venivano selezionate passivamente in base a quanto si adattassero all'ambiente, gli agenti più evoluti sono in grado di adattarsi attivamente per favorire la propria evoluzione. Ciò implica che un agente debba poter apprendere diversi tipi di informazione ed elaborare diversi tipi di comportamenti adattivi, e ciò è possibile negli animali grazie alle loro complessissime reti neurali. È per questa ragione che le tecnologie allo stato dell'arte sono proprio le reti neurali artificiali, le cui architetture sono caratteristiche determinanti per il livello con cui riescono ad apprendere ed elaborare le informazioni e i comportamenti. Per riuscire a trovare architetture sempre più efficienti e performanti si sono creati gli Algoritmi Meta-Learning, ossia algoritmi che apprendono come produrre migliori apprendisti (piuttosto che progettarli a mano).
2. **Ambiente:** perché un agente evolva, l'ambiente gli deve presentare le sfide giuste nell'ordine giusto. Ciò vale sia attraverso le generazioni (se gli organismi non fossero usciti dall'acqua non sarebbe stato possibile sviluppare le dita) sia nel corso della vita (se non avessimo imparato l'alfabeto non sapremmo scrivere). Il problema è che, se l'ambiente è fisso e limitato, l'agente può evolvere solo fino ad un certo punto. Per produrre "ambienti efficaci per l'apprendimento" Clune propone di adottare l'Approccio Open-Ended che prevede l'uso di algoritmi in grado di generare una serie in continua espansione di sfide. In questo modo gli agenti devono costantemente trovare nuove soluzioni evolvendosi in maniere inaspettate (infatti la natura non ha generato le sfide con l'obiettivo di ottenere un'intelligenza generale, ma questa è stata 'scoperta' dagli agenti come soluzione alle sfide dell'ambiente).
3. **Algoritmo evolutivo:** è la funzione di ricompensa che stabilisce se la risposta dell'agente ad una sfida viene considerata un successo o un fallimento. In natura è strettamente legato ai limiti biologici degli esseri viventi (sopravvive chi corre più veloce, chi resiste più tempo senza cibo, etc.), ma se riprodotto in un ambiente virtuale potrebbe potenzialmente

promuovere qualsiasi tipo di comportamento e caratteristica (sopravvive chi fa meno errori matematici, chi legge più velocemente, etc.). Due delle condizioni fondamentali perché l'algoritmo evolutivo porti ad agenti più avanzati sono che esso promuova la diversità (sia all'interno di una specie che tra specie diverse) e la performance (trovare maniere più efficienti ed efficaci per superare una sfida). La soluzione più di successo prevede la combinazione di queste due condizioni, e ciò corrisponderebbe ad algoritmi in grado di creare un set di diverse possibili soluzioni e cercare la miglior maniera possibile per arrivare a ciascuna di queste soluzioni.

Dal momento che la versione artificiale di queste componenti non deve rispondere a nessun limite biologico, i potenziali processi evolutivi che ne scaturirebbero potrebbero arrivare a generare forme di intelligenza completamente diverse dalla nostra. Tuttavia il limite principale di questo approccio è che richiederebbe quantità esorbitanti di potenza di calcolo per riuscire a simulare ciò che l'evoluzione ha fatto in miliardi di anni sulla Terra.

CONCLUSIONI

Per coordinare il lavoro dei ricercatori e per evitare disinformazione e mal interpretazioni da parte del senso comune, è importante fare chiarezza sui termini psicologici che vengono utilizzati per descrivere comportamenti e caratteristiche delle macchine e degli umani (ma anche degli animali). Le profonde differenze tra i due sistemi rendono necessaria la contestualizzazione di costrutti come "intelligenza", "comprensione", "coscienza" e molti altri, formulati dalla Psicologia per descrivere fenomeni strettamente legati al medium biologico che li supporta. Infatti, anche non comprendendone a fondo il funzionamento, è chiaro che il cervello abbia sviluppato queste abilità in seguito a una lunga evoluzione che ha selezionato i sistemi intelligenti che meglio si adattavano all'ambiente. Qualsiasi fenomeno mentale umano è il risultato di meccanismi e processi che si sono evoluti in virtù del rapporto tra agente e ambiente.

L'avvento di nuovi agenti in grado di generare fenomeni altrettanto complessi, ma legati ad un medium artificiale (e quindi indipendenti dall'evoluzione), pone di fronte ad una decisione: utilizzare gli stessi termini per descrivere i comportamenti sia umani che artificiali ("ha risolto il problema in una maniera intelligente...") ma contestualizzandone l'uso (ad es. aggiungendo per gli umani "... grazie alla capacità di manipolare concetti", e per le macchine "... grazie alla capacità di cogliere correlazioni statistiche"); oppure costruire nuovi termini che rendano conto di modalità di elaborazione diverse tra loro.

A prescindere da quanto a fondo si comprendano questi due sistemi intelligenti, i modelli artificiali sviluppati fino ad oggi dimostrano effettivamente una sorprendente gamma di abilità, ma la prospettiva dell'Intelligenza Generale Artificiale rimane ancora relativamente lontana. Per fare progressi nel campo è necessario sviluppare nuove tipologie di benchmark e strumenti d'indagine che consentano di ricavare una conoscenza più precisa di questi sistemi. Ciò aprirebbe la strada alla ricerca su come migliorarne i punti di forza, attenuarne quelli di debolezza, ma anche su come poter integrare assieme due forme di intelligenza così distinte tra loro.

BIBLIOGRAFIA

Alammar, J. (2018). The Illustrated Transformer. https://jalammar.github.io/illustrated-transformer/

Anderson, M. (1992). *Intelligence and development: A cognitive theory*. Blackwell Publishing

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6).

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 1721–1730.

Changizi, M. A., Zhang, Q., Ye, H., & Shimojo, S. (2006). The structures of letters and symbols throughout human history are selected to match those found in objects in natural scenes. *The American Naturalist*, 167(5), E117–E139.

Clune, J. (2019). AI-GAs: AI-generating algorithms, an alternate paradigm for producing general artificial intelligence. *arXiv preprint arXiv:1905.10985*.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., & Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26.

Gardner, H. (1993). *Multiple intelligences: The theory in practice*. Basic books.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11), 665-673.

Goertzel, B. (2014). Artificial general intelligence: Concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1), 1.

Habernal, I., Wachsmuth, H., Gurevych, I., & Stein, B. (2017). The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. *arXiv preprint arXiv:1708.01425*.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600), Articolo 7600.

Hutter, M. (2005). *Universal Artificial Intellegence*. Springer.

Kahneman, D., Lovallo, D., & Sibony, O. (2011). Before You Make That Big Decision.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive development*, 3(3), 299–321.

LeCun, Y. (2022). A path towards autonomous machine intelligence. *Open Review*, 62.

Legg, S. (2008). *Machine super intelligence*.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: A cognitive perspective. *arXiv preprint arXiv:2301.06627*.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological review*, 88(5), 375.

McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Michael, J., Holtzman, A., Parrish, A., Mueller, A., Wang, A., Chen, A., Madaan, D., Nangia, N., Pang, R. Y., Phang, J., & Bowman, S. R. (2022). What Do NLP Researchers Believe? Results of the NLP Community Metasurvey. *arXiv preprint arXiv:2208.12852*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv*:*1301.3781*.

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13), e2215907120.

Niven, T., & Kao, H.-Y. (2019). Probing Neural Network Comprehension of Natural Language Arguments. *arXiv preprint arXiv*:*1907.07355*.

OpenAI. (2023). GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Petersen, S. E., & Posner, M. I. (2012). The Attention System of the Human Brain: 20 Years After. *Annual Review of Neuroscience*, 35(1), 73–89.

Ritter, S., Barrett, D. G., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. *International conference on machine learning*, 2940–2949.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, 3(3), 417–424.

Sejnowski, T. J. (2023). Large Language Models and the Reverse Turing Test. *Neural Computation*, 35(3), 309–342.

Sejnowski, T. J., & Rosenberg, C. R. (1987). Parallel networks that learn to pronounce English text. *Complex systems*, 1(1), 145–168.

Shiffrin, R., & Mitchell, M. (2023). Probing the psychology of AI models. *Proceedings of the National Academy of Sciences*, 120(10), e2300963120.

Socher, R., Ganjoo, M., Manning, C. D., & Ng, A. (2013). Zero-Shot Learning Through Cross-Modal Transfer. *Advances in Neural Information Processing Systems*, 26.

Sternberg, R. J., & Kaufman, S. B. (2011). *The Cambridge handbook of intelligence*. Cambridge University Press.

Terman, L. M. (1916). *The uses of intelligence tests*.

Testolin, A., Stoianov, I., & Zorzi, M. (2017). Letter perception emerges from unsupervised deep learning and recycling of natural image features. *Nature human behaviour*, 1(9), 657–664.

Thiyagalingam, J., Shankar, M., Fox, G., & Hey, T. (2022). Scientific machine learning benchmarks. *Nature Reviews Physics*, 4(6), Articolo 6.

Thurstone, L. L. (1938). *Primary mental abilities*. Psychometric monographs.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wechsler, D. (1939). *The nature of intelligence*.

Wellman, H. M. (1992). *The child's theory of mind*. The MIT Press.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent Abilities of Large Language Models. *arXiv preprint arXiv:2206.07682*.

Zou, W. Y., Socher, R., Cer, D., & Manning, C. D. (2013). Bilingual Word Embeddings for Phrase-Based Machine Translation. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1393–1398). Association for Computational Linguistics.