

LLM e Wikipedia 2.0

Susanna Giaccaï

A gennaio 2023, a chiusura di un articolo su *Bibelot* scrivevo¹: «Nel frattempo ChatGPT si è evoluta in GPT-4, di cui vedremo le potenzialità nei prossimi mesi o forse solo settimane».

Allora ChatGPT era poco conosciuta ma, già poche settimane dopo, gli articoli iniziarono a moltiplicarsi. Sono passati quasi 2 anni e mezzo e i cambiamenti prodotti dall'intelligenza artificiale generativa (i modelli linguistici di grandi dimensioni (in inglese *Large Language Models*, d'ora in avanti LLM)² sono enormi; la sua capillarità inizia a invadere letteralmente le nostre vite. La troviamo nelle nostre chat in WhatsApp, gli studenti la usano per le ricerche scolastiche, la troviamo negli articoli dei giornali. Dal 18 marzo scorso il quotidiano *Il foglio* ha fatto un esperimento interessante, pubblicando da quel giorno due edizioni cartacee del quotidiano, una scritta da giornalisti e una scritta interamente da un LLM e proseguendo con una versione online fatta con l'AI che affianca quella scritta dai giornalisti.

Abbiamo inoltre scoperto che l'uso dei LLM ha, e sempre di più avrà, un fortissimo impatto ambientale sia per quanto riguarda l'energia necessaria al funzionamento dei server sia per il consumo dell'acqua per raffreddare i circuiti. Viene calcolato che nel 2030 si consumerà l'11% in più di energia rispetto al 2023 e, nello stesso anno, saranno prodotti cinque milioni di tonnellate di rifiuti elettronici in più a causa dell'espansione dei *data center*. L'acqua per il raffreddamento dei *data center* quadruplicherà da 175 miliardi di litri nel 2023 a 664 miliardi di litri entro il 2030.

1 Susanna Giaccaï, *ChatGPT e Wikipedia* in *Bibelot*, 2023, n. 1-2, p. 5. <<https://tinyurl.com/y8f83c5u>>

2 Si veda la voce in Wikipedia <https://it.wikipedia.org/wiki/Intelligenza_artificiale_generativa>

Il 31 maggio 2025 la rivista online *Wired* titolava *ChatGPT e gli altri large language models si stanno mangiando internet*³ Cosa sta accadendo? I grandi motori di ricerca non offrono più ai propri utenti una lista di link in cui cercare la risposta alla loro ricerca di informazioni, ma, utilizzando propri sistemi interni di LLM, offrono all'utente direttamente la risposta, non favorendo più la sua navigazione nel web. Per questo motivo Wikipedia in questi ultimi mesi ha perso il 6% della propria utenza, così come l'hanno persa diversi social media; così come la perderanno molti siti minori non più resi visibili dalle ricerche fatte con Google. I siti informativi sono stati usati come fonti di apprendimento di ChatGPT di OpenAI, Gemini di Google e Llama di Meta e *similia*. I siti web di Archivi, Musei e Biblioteche sono stati sottoposti ad un massiccio *web scraping*⁴ da parte di robot per catturare informazione con cui addestrare gli LLM. Al momento in cui ci si rende conto dello stress cui sono sottoposti da questa attività i nostri siti a scapito dell'accesso degli utenti, si tentano rimedi e contromisure, cercando comunque sempre una via di mezzo tra il blocco dei bot e la garanzia di libero accesso al sito da parte dell'utente.

Da questo *scraping* non è stata esente naturalmente Wikipedia, la più grande enciclopedia mondiale, redatta in oltre 342 lingue e contenente oltre 64 milioni di voci enciclopediche in tutti gli ambiti della conoscenza. Si tratta di materiale preziosissimo per alimentare le conoscenze di questi LLM. Il consumo di banda provocato dai *web scraping* ha messo a dura prova i server di Wikipedia. Nel gennaio 2024 il 50% di banda dei suoi server è usato da questi bot creando seri problemi all'accesso degli utenti:

«I nostri contenuti sono gratuiti, la nostra infrastruttura no. Dobbiamo agire ora per ristabilire un sano equilibrio, così da poter dedicare le nostre risorse ingegneristiche al supporto e alla priorità dei progetti Wikimedia, dei nostri collaboratori e dell'accesso umano alla conoscenza», si esprimevano i responsabili della Wikimedia Foundation⁵.

Dopo aver pensato di ridurre l'accesso ai bot, nell'aprile 25 è stato invece deciso di collaborare con i progetti di AI creando una copia di Wikipedia strutturata apposta per

3 Si veda l'articolo di Andrea Daniele Signorelli <<https://www.wired.it/article/chatgpt-llm-fine-internet/>>

4 Si veda la spiegazione in <https://it.wikipedia.org/wiki/Web_scraping>

5 Cfr. [Tanya Weaver](#), *AI scraper bots putting costly strain on Wikimedia infrastructure*, <<https://tinyurl.com/2bwaz3tp>>

facilitare il lavoro di *scraping* e riutilizzo dei dati, nello spirito di promozione della cultura libera che caratterizza la comunità wikipediana, facilitando così anche il lavoro alle piccole aziende che intendono sfruttare l'AI per creare nuovi servizi⁶.

Del resto da tempo Wikipedia usa tecniche di AI per la continua attività di manutenzione delle voci: correzione di errori di scrittura, blocco dei vandalismi ecc. Recentemente era anche stata attivata una funzionalità di creazione con LLM di riassunti delle voci, ma le proteste della comunità l'ha poi bloccata⁷.

In questo contesto naturalmente i volontari wikipediani si confrontano tra di loro discutendo delle problematiche connesse alla scrittura di voci realizzate con LLM.

Un articolo dell'ottobre 2024 segnalava che «*we estimate that 4.36% of 2,909 English Wikipedia articles created in August 2024 contain significant AI-generated content*»⁸. Il 19 aprile 2025 il volontario Frinate scriveva in una pagina di discussione del Bar dei wikipediani:

ormai il problema dell'uso dell'IA sulle voci su it.wiki diventa sempre più rilevante. [...] già lo scorso agosto tra il 3 e il 5% dei nuovi *edit* su it.wiki era fatto da IA. In base all'esperienza empirica sia personale che leggendo i *chan* degli admin, questa percentuale è sicuramente salita.⁹

I problemi sollevati dall'uso della AI per scrivere le voci in Wikipedia sono principalmente i seguenti:

1 Le fonti: è documentata la tendenza attuale dell'AI a non citare le fonti oppure ad inventarsele od avere delle allucinazioni. Per i Wikipediani citare le fonti è un obbligo fondamentale nella scrittura delle voci per consentire all'utente di verificare quanto scritto e per consentirgli di approfondire la ricerca. Le fonti sono anche essenziali per attestare che la voce non sia una opera originale.

6 Fabio Ricceri, *Meno scraping, più sharing: la nuova mossa di Wikipedia per l'AI*, in Riviata AI, 19 aprile 2025. Wikipedia mette in pausa i riassunti AI dopo le proteste degli editor <<https://tinyurl.com/ye25cdwh>>

7 Diego Barbera, Wikipedia mette in pausa i riassunti AI dopo le proteste degli editor, <<https://tinyurl.ee/OdDEL>>

8 Creston Brooks, Samuel Eggert, Denis Peskoff, *The Rise of AI-Generated Content in Wikipedia* <<https://tinyurl.com/25uz2i8u>>

9 Frinate, *Una nuova policy sull'Intelligenza artificiale?* <<https://tinyurl.com/3yaafx7>>

Wikipedia infatti è un'enciclopedia e in quanto tale i suoi contenuti devono essere derivati da testi già editi.

Essa nei 5 Pilastri, le sue regole basilari, viene definita così:

Wikipedia comprende caratteristiche delle enciclopedie 'generaliste', delle enciclopedie 'specialistiche' e degli almanacchi. Wikipedia non è una raccolta indiscriminata di informazioni. Non è una fonte primaria, ma piuttosto uno strumento di divulgazione secondaria e terziaria; non è un dizionario, né un palco per comizi, né un giornale; non è neanche un luogo nel quale fare promozione, né un banco di prova per l'anarchia o la democrazia; non è neppure uno spazio web utilizzabile indiscriminatamente, né un posto nel quale inserire le proprie opinioni, esperienze o argomentazioni soggettive; tutti i contributori devono sforzarsi di seguire le politiche comunitariamente basate sulla verificabilità e sul divieto di ricerche originali.¹⁰

Va inoltre tenuto conto che Wikipedia è una fonte secondaria e in quanto tale non è ritenuto corretto citarla come fonte nelle voci di Wikipedia.

2 Diritto d'autore: Wikipedia ha una grande attenzione al rispetto del diritto di autore. Non accetta testi copiati, ma richiede che il testo sia riformulato: è necessario cioè utilizzare parole e frasi proprie per esprimere i concetti espressi dall'autore. Non ci sono garanzie però che ChatGPT e le altre AI generative rispettino questo criterio, di conseguenza non è chiaro se sia garantito il diritto d'autore.

Inoltre non è garantito che le voci prodotte dalle LLM non siano ricerche originali, non siano, come dovrebbero invece essere, riformulazione di testi già editi.

3 Bias o pregiudizi: i sistemi di intelligenza artificiale generativa sono addestrati con risorse informative e da queste derivano impostazioni e opinioni. Se le risorse provengono da una cultura occidentale la AI risponderà con quella impostazione culturale, magari riportando anche eventuali opinioni razziste, maschiliste, omofobe ecc.

¹⁰Si veda la voce *Cinque pilastri* <https://it.wikipedia.org/wiki/Wikipedia:Cinque_pilastri>

«I pregiudizi algoritmici sono una delle principali ombre che pesano sul futuro (già in molti casi 'presente') dei sistemi di intelligenza artificiale»¹¹.

Inoltre se su Wikipedia una voce può essere corretta da un altro volontario eliminando affermazioni errate, razziste ecc. Ciò non è possibile con LLM, i cui parametri interpretativi della realtà sono dati dal suo gestore, troppo spesso un monopolista privato.

Da diverse settimane la comunità wikipediana italiana sta lavorando alla stesura di linee guida che indirizzino i redattori nell'uso di LLM per la scrittura di voci. La discussione iniziata al Bar¹² il 19 aprile dopo aver raggiunto il consenso sull'impostazione, si è spostata a fine maggio nella sotto sezione *Wikipedia*, quella usata per i manuali, per stendere la voce *Wikipedia: Intelligenza artificiale: guida all'uso delle LLM* per scrivere le voci in Wikipedia¹³. Il manuale naturalmente è affiancato dalla relativa pagina di discussione in cui continua lo scambio di idee su come scrivere il manuale¹⁴. In testa al manuale si legge:

Il succo di questa pagina: Evita di usare strumenti che facciano uso di intelligenza artificiale (IA) per generare nuovo contenuto originale, per generare note, o per scrivere da zero commenti. Fonti che facciano uso di testi prodotti da IA non sono attendibili. I risultati ottenuti da IA non sono utilizzabili come prova a sostegno di un'argomentazione in una discussione. Nei casi in cui l'uso di IA è ammesso, controlla sempre il risultato prima di pubblicare.

Vedremo quindi nei prossimi mesi quale sarà l'impatto delle LLM su Wikipedia, sia nella redazione delle sue voci che nel possibile calo delle sue consultazioni, segnando in questo caso il declino di questa straordinaria creazione culturale collettiva.

Susanna Giaccai

Collaboratrice in Wikipedia, Wikisource, Wikidata, Commons e biblioteca

giaccai@pm.me

¹¹Bias in Treccani vocabolario <[https://www.treccani.it/vocabolario/neo-bias_\(Neologismi\)/?search=bias%2F](https://www.treccani.it/vocabolario/neo-bias_(Neologismi)/?search=bias%2F)>

¹²Si veda la pagina <<https://tinyurl.com/53xc94hs>>

¹³Si veda la pagina <<https://tinyurl.com/3bjmparc>>

¹⁴Si veda la pagina <https://it.wikipedia.org/wiki/Discussioni_Wikipedia:Intelligenza_artificiale>