GERARCHIE SEMANTICHE E DEGRADAZIONE DELL'INFORMAZIONE NEI MODELLI DI IA GENERATIVA: UN APPROCCIO BASATO SULLA TEORIA DELLA COMPLESSITÀ

Antonello Rizzi, Enrico De Santis, Edoardo Bruno, Francesca Ronci
Dipartimento di Ingegneria dell'Informazione, Elettronica e Telecomunicazioni

Università di Roma "La Sapienza"
Via Eudossiana 18, 00184 Roma

L'Intelligenza Artificiale (IA) generativa ha conosciuto negli ultimi anni uno sviluppo accelerato, imponendosi come paradigma dominante nella produzione automatica di testi, immagini e contenuti multimodali. Tuttavia, parallelamente agli straordinari risultati ottenuti, si è manifestata una crescente necessità di comprendere, analizzare e inquadrare i modelli generativi all'interno di una cornice teorica che includa l'approccio sistemico alla base dei sistemi complessi. In questa linea si inserisce la nostra indagine, che ha affrontato la questione da due prospettive convergenti: la decostruzione gerarchica dei modelli Transformer e l'analisi del collasso informativo in architetture ricorsive.

Nel primo caso, l'obiettivo è stato comprendere come la gerarchia interna di un modello Transformer, in particolare GPT-2, contribuisca alla qualità del testo generato. Tramite un processo sistematico di ablazione, gli strati del modello vengono disattivati progressivamente, partendo dagli ultimi (più vicini all'output) fino ai primi (più vicini all'input). Ogni configurazione viene valutata secondo una metrica composita che include indici linguistici (BLEU, Dale-Chall), metriche di coerenza testuale (Text Flow) e valutazioni semantiche ottenute tramite LLM (GPT-4 Legacy). I risultati rivelano un comportamento stratificato dell'informazione, dove la coerenza locale del testo viene compromessa fin dalle prime ablazioni degli strati terminali, mentre la comprensibilità globale e la coesione semantica resistono più a lungo, decrescendo significativamente solo dopo la rimozione degli strati centrali. Ciò suggerisce che nei Transformer l'organizzazione semantica segue una gerarchia non lineare, dove la profondità corrisponde a livelli di rappresentazione sempre più astratti e distribuiti [1]. Tale struttura gerarchica, rivelata attraverso la degradazione controllata del modello, consente una nuova lettura della «funzione cognitiva» dei Transformer che non sono visti più come semplici encoder di similarità tra token, ma come sistemi capaci di comportamenti emergenti ed in grado di organizzare concetti secondo gradienti semantici. La dissoluzione della qualità del testo non è infatti lineare, ma mostra discontinuità e soglie critiche, segnali tipici dei sistemi complessi che attraversano transizioni di fase. L'adozione di indici linguistici ispirati dalla psicologia cognitiva e dalle teorie dell'informazione e dei sistemi caotici rafforza ulteriormente questo quadro interpretativo, suggerendo un parallelismo tra la struttura funzionale degli LLM e l'organizzazione multilivello del linguaggio umano [3]. La seconda direttrice ha esplorato gli effetti della retroazione nei sistemi generativi, simulando il caso in cui un modello viene iterativamente addestrato su testi generati da sé stesso. Tale scenario, noto come «model collapse» [3], è stato affrontato tramite un esperimento ricorsivo basato su LSTM *stateful*, partendo da un corpus scritti da autori umani («Alice in Wonderland») e osservando l'evoluzione della qualità testuale su 73 generazioni. L'analisi è stata condotta impiegando indici semantici, sintattici e inerenti alla teoria della complessità, quali la similarità coseno, la lunghezza media di frase e parola, il

type-token ratio, le frequenze di elementi grammaticali, il ROUGE score, l'entropia approssimata, la complessità di Kolmogorov, lo spettro di singolarità e il coefficiente di Hurst. I risultati mostrano un quadro composito. Specificatamente, da un lato, si osserva una perdita progressiva della diversità lessicale e della complessità sintattica (riduzione di TTR, nomi e verbi, lunghezza delle frasi); dall'altro, aumentano indici associati a comportamento stocastico e imprevedibilità (entropia, complessità di Kolmogorov). Particolarmente significativa è la riduzione del coefficiente di Hurst, indicativa della perdita di correlazioni a lungo raggio, fondamentali per la costruzione del significato nel linguaggio naturale. In parallelo, si osserva un incremento della larghezza dello spettro di singolarità, interpretabile come aumento di complessità locale e disomogeneità, a fronte di una perdita di struttura globale [4]. La ricorsività agisce dunque come un amplificatore delle dinamiche entropiche, dove i pattern frequenti vengono rinforzati, mentre quelli rari vengono progressivamente eliminati, fino a una degenerazione statistica della distribuzione lessicale verso configurazioni *delta-like*. Un modello differenziale proposto in parallelo formalizza questo comportamento, descrivendo la degradazione della qualità testuale in funzione della proporzione tra dati umani e generati e introducendo una costante α di sensibilità al contenuto sintetico. Il tempo di collasso del modello risulta dipendere criticamente dal rapporto tra produzione testuale umana e artificiale, suggerendo scenari in cui, in assenza di interventi correttivi, anche modelli performanti possono degradare se esposti esclusivamente a contenuti generati da IA [4]. La sinergia tra i due approcci fornisce una visione unificata e olistica dove, da un lato, i Transformer manifestano una gerarchia funzionale interna, le cui dinamiche possono essere interpretate alla luce della teoria dei sistemi complessi; dall'altro, i processi generativi iterativi rivelano la fragilità strutturale dei modelli se privati di input esterni informativamente ricchi. L'integrazione di misure derivate dalla teoria multifrattale, dall'entropia e dalla complessità algoritmica permette una quantificazione rigorosa del degrado semantico e sintattico. Questi strumenti rafforzano la necessità di mantenere la diversità nei dati di addestramento e suggeriscono anche nuove metriche per il monitoraggio e il controllo della qualità linguistica nei sistemi generativi.

Alla luce di tali evidenze, il nostro contributo si propone di rafforzare il ponte tra l'analisi strutturale dei modelli generativi e la teoria dei sistemi complessi, promuovendo una visione epistemologicamente fondata dell'IA generativa. In un contesto in cui l'informazione prodotta dai modelli alimenta il proprio processo evolutivo, la questione non è solo tecnica ma ontologica: quale è la natura della conoscenza generata da sistemi che apprendono da sé stessi? La risposta passa attraverso una comprensione profonda della loro struttura interna, della loro dinamica evolutiva e della loro relazione con la complessità del linguaggio.

[1] E. De Santis, Martino A., A. Rizzi, et al., «2025: A GPT Odyssey. Deconstructing Intelligence by Gradual Dissolution of a Transformer», in *Proc. IEEE Int. Joint Conf. on Neural Networks. (IJCNN)*, Rome, Italy, Jun. 30–Jul. 5, 2025.
[2] E. De Santis, A. Martino, and A. Rizzi, «Human versus machine intelligence: Assessing natural language generation models through complex systems theory», *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 46, no. 7, pp. 4812–4829, 2024.
[3] I. Shumailov et al., «The Curse of Recursion: Training on Generated Data Makes Models Forget», arXiv preprint arXiv:2305.17493, 2024.
[4] E. De Santis, A. Martino, A. Rizzi, et al., «The End of Knowledge? A Recursive Complexity-based Study on LLM Collapse and Model Drift», in *Proc. IEEE Int. Joint Conf. on Neural Networks. (IJCNN)*, Rome, Italy, Jun. 30–Jul. 5, 2025.