Umanistica Digitale – ISSN: 2532–8816 – n. 24, 2026
Fabio Ciotti – *Una Agenda sull'AI generativa per le Digital Humanities*
DOI: https://doi.org/10.60923/issn.2532-8816/25557

Una Agenda sull'AI generativa per le Digital Humanities

Fabio Ciotti

Università di Roma - Tor Vergata, Italia
fabio.ciotti@uniroma2.it

Abstract

L'impatto trasformativo dell'Intelligenza Artificiale generativa ci impone, come comunità italiana delle Digital Humanities, di predisporre un'agenda strategica di riflessione collettiva. In questo editoriale propongo un contributo in tale direzione, individuando tre direttrici di analisi interconnesse: (1) la teoria dell'IA generativa, che richiede approcci esplicativi non riducibili alle sole spiegazioni matematico-informatiche, incorporando descrizioni funzionali e nozioni intenzionali per comprendere le proprietà linguistiche e cognitive emergenti dei Large Language Models; (2) l'IA come tecnologia socio-culturale, considerando i LLM come modelli della memoria culturale che a loro volta retroagiscono sui sistemi culturali stessi; (3) l'IA come metodo nella ricerca umanistica, che riorganizza profondamente i workflow epistemici permettendo l'operazionalizzazione di quasi ogni aspetto della ricerca, dalla costruzione dei corpora alla produzione di interpretazioni ermeneutiche. I sistemi di IA generativa rappresentano una rivoluzione epistemologica che impone di riconcettualizzare assunzioni fondamentali su significato, cultura, conoscenza e interpretazione. La tradizione metodologica delle Digital Humanities italiane può svolgere un ruolo critico essenziale nello sviluppo di protocolli d'uso controllato e pratiche di cooperazione tra IA e agency umana, orientandoci anche come studiosi che hanno una responsabilità sociale.

Parole chiave: Intelligenza Artificiale generativa, Epistemologia delle Digital Humanities, Large Language Models, Impatto sociale delle DH

The transformative impact of generative Artificial Intelligence compels us, as the Italian Digital Humanities community, to establish a strategic agenda for collective reflection. In this op-ed, I propose a contribution in this direction by identifying three interconnected lines of analysis: (1) the theory of generative AI, which requires explanatory approaches not reducible to mathematical-computational explanations alone, incorporating functional descriptions and intentional notions to understand the emergent linguistic and cognitive properties of Large Language Models; (2) AI as a socio-cultural technology, considering LLMs as models of cultural memory that in turn act back upon cultural systems themselves; (3) AI as a method in humanistic research, which profoundly reorganizes epistemic workflows by enabling the operationalization of nearly every aspect of research, from corpus construction to the production of hermeneutic interpretations. Generative AI systems represent an epistemological revolution that demands a reconceptualization of fundamental assumptions about meaning, culture, knowledge, and interpretation. The methodological tradition of Italian Digital Humanities can play an essential critical role

in developing protocols for controlled use and practices of cooperation between AI and human agency, guiding us also as scholars who bear social responsibility.

Keywords: Generative Artificial Intelligence, Epistemology of Digital Humanities, Large Language Models, Social impact of DH

1. Introduzione[1]

La comunità di studi che nel corso degli anni ha dato vita alla "galassia" delle Digital Humanities (Ciotti, 2023a) si è sempre confrontata con l'evoluzione, in gran parte autonoma, dei modelli, dei sistemi e delle infrastrutture tecno-informatiche (l'introduzione del personal computing, poi delle interfacce grafiche, e ancora l'avvento del Web, la digitalizzazione di massa, il Web semantico e i linked data…). Assunzioni teoriche, framework epistemologici e metodologici, domini di ricerca che hanno popolato la storia concettuale del nostro campo di studi e di pratiche sono sempre state negoziate con quella sfera a un tempo aliena e familiare, ponendoci peraltro nel difficile ruolo di mediatori epistemici e culturali tra il complesso dei saperi umanistici e quello dei saperi tecno-informatici (Buzzetti, 2019). Percorsi ed esiti di questa mediazione sono stati molteplici, contraddittori e fortemente differenziati anche in base all'ambiente linguistico e socioculturale in cui si sono situati, sebbene la fase più recente delle Digital Humanities sia stata di fatto caratterizzata da una sostanziale globalizzazione e omogeneizzazione internazionale, tanto da rendere assai meno nette quelle stesse differenze di tradizioni culturali (Horvath et al., 2022).

La storia del presente ci pone di fronte ancora una volta a una innovazione tecno-scientifica il cui potere trasformativo promette di rivelarsi ancora più profondo e pervasivo di tutti quelli cui abbiamo assistito, e che siamo stati in grado di metabolizzare concettualmente e metodologicamente, nello scorso cinquantennio: l'esplosione dell'Intelligenza Artificiale generativa. La questione terminologica, alla quale come umanisti digitali siamo assai avvezzi, ci potrebbe portare a criticare, distinguere, puntualizzare, sulla correttezza, adeguatezza, legittimità di questo termine; ma mi sembra di poter dire che di fronte all'enormità della cosa, mettersi a disquisire sulla parola sia futile, e comunque meno urgente del capire e riflettere su come collocarci di fronte a questa travolgente novità scientifica e tecnologica. Si tratta insomma di discutere collettivamente su quale debba essere l'Agenda sull'IA per il nostro ambito di studi. E ci sono buone ragioni per farlo, pur nella diversità delle posizioni su questioni anche fondazionali, perché tra tutte le grandi innovazioni tecnologiche questa dell'IA avrà di sicuro un grande impatto su quella sfera di pratiche e artefatti di cui ci occupiamo come umanisti, la sfera dei linguaggi e della cultura; perché essa avrà certamente un impatto enorme nei nostri stessi metodi, approcci e saperi come umanisti digitali, mettendo di fatto in questione la nostra stessa specificità; ma soprattutto perché potrebbe essere quella con le più profonde conseguenze in generale, dico a livello di storia della civilizzazione.

[1] Questo testo è un Editoriale e non è stato sottoposto a peer review: i suoi contenuti sono responsabilità dell'autore e non rispecchiano necessariamente la posizione del Comitato Editoriale, che pure li ha discussi collettivamente. Si tratta della rielaborazione dell'omonimo paper presentato al convegno AIUCD 2026, qui arricchita da alcuni spunti emersi nella discussione in quella sede. Mi auguro che la pubblicazione offra uno stimolo di riflessione su natura, impatto e conseguenze della IA generativa alla comunità scientifica a cui *Umanistica Digitale* si rivolge, e che apra un confronto più ampio, del quale la rivista si impegna a dare conto.

Agli albori di questa esplosione, chi scrive aveva già posto all'attenzione della comunità DH italiana le possibilità che i nuovi modelli linguistici generativi avrebbero potuto offrire per i nostri studi (Ciotti, 2023b). Oggi dopo tre anni di sviluppo e diffusione pervasiva, l'impatto trasformativo è ormai evidente, e non a caso si registrano, oltre a decine di studi, esperimenti, riflessioni teoriche, anche alcuni tentativi di elaborazione strategica collettiva: il rapporto *Doing AI Differently: Rethinking the Foundations of AI via the Humanities* (Hemment e Kommers 2025), il white paper *Computational Hermeneutics: Evaluating Generative AI as a Cultural Technology* (Kommers et al. 2025), il documento *Provocations from the Humanities for Generative AI Research* (Klein et al. 2025), l'*AI Manifesto* del Luxembourg Centre for Contemporary and Digital History (C2DH 2025), e da ultimo il framework *AI and the Humanities: A Framework for Language and Literary Scholarship* elaborato dall'AI and Research Working Group della Modern Language Association (MLA AI and Research Working Group 2026). A queste iniziative internazionali si affianca, nel nostro contesto, l'Osservatorio italiano sull'Intelligenza Artificiale e le Digital Humanities costituito come SIG della AIUCD (https://aiucd.github.io/DH-AI).

Obiettivo di questo articolo è fornire un contributo a questa riflessione collettiva, uno stimolo a farsene protagonista rivolto alla nostra comunità scientifica. Uno stimolo certamente orientato, fondato su precise posizioni teoriche e metodologiche, ma anche frutto di una lunga esperienza nel campo delle Digital Humanities, perché ritengo che la nostra Associazione scientifica e la comunità che in essa si riconosce possa e debba essere protagonista di questo dibattito, e che anzi possa e debba porsi ancora una volta come riferimento per tutto il mondo dei saperi umanistici.

IA e DH: tre direttrici per la riflessione

Inizio con due considerazioni preliminari. La prima, condivisa peraltro da molti studiosi e gruppi di ricerca, è che pensare strategicamente sull'impatto dell'Intelligenza Artificiale generativa sulle Digital Humanities significa anche pensare simmetricamente sul contributo che le DH possono offrire alla ricerca sull'IA, sia che si considerino questi sistemi come tecnologie culturali, sia che le si intendano come agenti cognitivi di un qualche tipo nell'ambito di una ontologia inclusiva del mentale. La seconda è che la riflessione sulla IA deve necessariamente articolarsi su molteplici livelli di discorso, poiché il suo impatto si estende dalle questioni cognitive a quelle epistemologiche, sociali, pedagogiche, etiche, politiche e ovviamente tecnologiche e informatiche. Non è possibile per limiti di competenze individuali e di spazio concesso a un articolo come questo, coprire tale molteplicità. Pertanto, mi concentrerò in questa sede su tre direttrici, ambiti di discorso distinti ma ovviamente interconnessi: (1) la riflessione teorica e sperimentale sull'IA generativa e sulla natura e proprietà dei Large Language Models; (2) la riflessione sull'IA come tecnologia socioculturale e sui suoi impatti; (3) la riflessione epistemologica sull'IA come metodo nella ricerca umanistica.

Teoria dell'IA generativa

Il punto di partenza per questo ambito di discorso è una constatazione ormai difficilmente contestabile: le spiegazioni di tipo esclusivamente informatico e matematico dei meta-algoritmi e dei processi computazionali degli LLM, pur necessarie, non sono sufficienti a spiegare i loro comportamenti reali, né possiedono una capacità esplicativa o predittiva rispetto alle proprietà linguistiche e cognitive emergenti che osserviamo empiricamente. Un modello linguistico basato su reti neurali esplora lo spazio dei significati mediante rappresentazioni distribuite e continue (codificate come vettori di numeri reali), che si trasformano attraversando i vari strati del sistema e possono codificare non solo la prosecuzione immediata della catena linguistica di input, ma

anche vincoli a lungo termine, candidati alternativi e forme locali di pianificazione. Questo aspetto del funzionamento degli LLM emerge anche empiricamente, come mostrano i recenti lavori di interpretabilità meccanicistica mediante *sparse autoencoder* condotti nei laboratori di Anthropic che hanno identificato sia l'esistenza di feature interpretabili semanticamente all'interno del *residual stream* di modelli come Claude Sonnet e Opus (Templeton et al. 2024), sia la presenza di strutture emergenti dedicate alla pianificazione anticipata: il modello, durante l'inferenza, costruisce rappresentazioni di token e obiettivi futuri prima di emetterli, e mostra forme di pianificazione anticipatrice nella generazione (Lindsey et al. 2025).

Un parallelismo con le neuroscienze può essere istruttivo sotto questo rispetto. Così come una teoria del comportamento umano non può essere ridotta alla biochimica o alla neurofisiologia, pur includendole, allo stesso modo una teoria degli LLM non può arrestarsi al livello architetturale o ingegneristico implementativo. Essa deve incorporare descrizioni funzionali e, in senso controllato e non ingenuo, anche nozioni intenzionali: parlare di competenze, di capacità, di generalizzazione, di sensibilità contestuale. Ho già più volte suggerito come un inquadramento teorico efficace da questo punto di vista sia la nozione di sistemi intenzionali di Daniel Dennett (Dennett 1989), ovvero sistemi che possono essere interpretati adottando l'atteggiamento intenzionale, la strategia di spiegare il comportamento di un'entità trattandola come se fosse un agente razionale che governa la sua scelta di azione mediante una considerazione delle sue "credenze" e dei suoi "desideri". Non intendo ovviamente sostenere che i modelli linguistici attuali siano effettivamente e intrinsecamente dotati di tutte le proprietà che vorremmo attribuire alla mente umana (facoltà di linguaggio, raziocinio, modello del mondo, teoria della mente, coscienza, empatia...), ma che non c'è nulla di misterioso e irriducibile o nascosto in una insondabile interiorità che permetta di spiegare tali proprietà: tutto ciò che abbiamo sono i comportamenti del sistema, e se l'esame di tali comportamenti ci induce ragionevolmente ad attribuire una di tali facoltà a un agente artificiale, ciò è quanto basta per tale ascrizione. È appena il caso di osservare che questo principio, già al cuore dell'argomento di Turing sull'*imitation game* (Turing, 1950), vale simmetricamente per la negazione: i medesimi criteri comportamentali in base ai quali attribuiamo una facoltà sono quelli in base ai quali la neghiamo, e proprio per questo la questione non è soltanto epistemologica, ma investe anche, come si vedrà nella parte conclusiva, il piano etico delle nostre attribuzioni reciproche.

Si potrebbe obiettare che proprio i risultati di interpretabilità citati sopra, identificando strutture interne ai modelli, quali feature nel residual stream e circuiti dedicati, mostrino come ciò che sembrava esigere un vocabolario intenzionale sia in realtà riconducibile a una descrizione puramente meccanicistica, e dunque rendano superfluo il livello esplicativo intenzionale. L'obiezione, però, fraintende la natura dell'atteggiamento intenzionale, che non è una congettura su entità nascoste da confermare o smentire mediante l'ispezione dei meccanismi, ma una strategia predittiva che individua regolarità reali nel comportamento del sistema. La scoperta di meccanismi e processi di livello computazionale (peraltro anche essi emergenti perché non sono progettati nel senso algoritmico classico) non elimina il livello intenzionale: mostra piuttosto che le regolarità che descriviamo parlando di obiettivi, di pianificazione e di sensibilità contestuale possiedono una controparte identificabile, esattamente come la psicologia non è confutata ma sostanziata dalla neurofisiologia che ne descrive i correlati.

Questi argomenti potrebbero sembrare liminari rispetto al dibattito proprio dei vari bracci della galassia DH, ma non lo sono. Lo scopo fondamentale del nostro campo, nella sua diversità, è stato quello di rendere conto, integrando metodi computazionali, della natura e del significato dell'agire comunicativo e culturale umano. In questo quadro si collocano naturalmente questioni e domande come le seguenti: quali teorie del significato e della comprensione sono compatibili con le evidenze empiriche sulle performance linguistiche degli LLM? Che cosa sono esattamente

i modelli del mondo, o la teoria della mente, e fino a che punto possiamo ascrivere queste proprietà ai sistemi AI? Che tipo di processi di ragionamento, se lo sono, possiamo attribuire ai modelli linguistici con reasoning avanzato che sono stati in grado di conseguire risultati stupefacenti nel dominio della ricerca matematica, in modo sorprendente rispetto alle aspettative iniziali (OpenAI 2026; Tsoukalas et al. 2026; Klowden e Tao 2026)?

Da questo punto di vista, aggiungo, gli LLM non sono solo oggetti da spiegare, ma anche strumenti epistemici per ripensare la natura della cognizione, del linguaggio e della comunicazione umana: il loro studio può fornire nuove evidenze per valutare, validare o mettere in discussione le teorie classiche della cognizione, del linguaggio e dei sistemi culturali che su di esso si basano. Vorrei indicare due articoli di recente pubblicazione che articolano questa idea. Il primo, «Whither symbols in the era of advanced neural networks?» (Griffiths et al. 2026), affronta uno dei fondamenti di gran parte delle teorie cognitive del Novecento: la tesi che il pensiero corrisponda all'elaborazione di informazione espressa in simboli discreti, le rappresentazioni, in base a un insieme finito di regole esplicite di tipo fondamentalmente logico o quasi-logico. Lo sviluppo dei modelli linguistici contemporanei costituisce una evidenza empirica che mette in crisi tale tesi, e ne sposta l'onere della prova. Va detto che la questione resta controversa, poiché il grado e la robustezza della generalizzazione composizionale esibita dalle reti neurali sono tuttora oggetto di dibattito, e non mancano studi che ne segnalano la fragilità su classi specifiche di costrutti. Ma il punto teoricamente rilevante è che i modelli linguistici basati su reti neurali artificiali esibiscono esibiscono proprietà per cui si riteneva necessario assumere il rappresentazionalismo logicista, la composizionalità, la produttività e altre ancora, senza che le si sia dotate strutturalmente di un'architettura simbolica esplicita, e questo basta a privare tale architettura del suo statuto di precondizione necessaria, retrocedendola da assunto a priori dei processi neuro-cognitivi a ipotesi fra le altre, da valutare empiricamente.

Il secondo lavoro, «How Linguistics Learned to Stop Worrying and Love the Language Models» (Futrell e Mahowald 2025), sostiene apertamente che i moderni LLM abbiano manifestato una competenza grammaticale capace di catturare strutture gerarchiche e dipendenze a lunga distanza che un tempo si ritenevano inaccessibili ai metodi puramente statistici. I due autori osservano che se sistemi privi di vincoli grammaticali predefiniti generalizzano correttamente su fenomeni sintattici complessi, l'assunto che l'acquisizione di una simile competenza presupponga una dotazione innata di vincoli grammaticali formali alla Chomsky perde la sua necessità logica. Anche l'argomento della povertà dello stimolo diventa un fatto empirico contingente: per ora gli LLM apprendono da quantità di dati linguistici di ordini di grandezza superiori a quelle cui è esposto un bambino, ma questo non implica che l'apprendimento empirico e statistico del linguaggio sia in generale impossibile. Gli LLM sono sistemi fortemente iper-parametrizzati che apprendono generalizzazioni linguistiche corrette in virtù di soft-bias, cioè di una propensione implicita a preferire le funzioni più semplici compatibili con i dati linguistici, indotta dalla regolarizzazione operante nel processo di addestramento, e non grazie a restrizioni formali imposte a priori. La struttura linguistica, in questo contesto assume uno statuto teorico descritto dalla nozione dennettiana di *real pattern* (Dennett 1991): una astrazione di livello superiore scientificamente legittima perché riduce il carico di lavoro descrittivo necessario a rendere conto di processi di elaborazione altrimenti opachi, e ciò vale sia per i sistemi biologici sia per quelli artificiali.

Ho voluto presentare e discutere questi due articoli, dedicati alla natura delle rappresentazioni mentali e del linguaggio e al loro rapporto con l'emergenza dei modelli linguistici probabilistici, non per assumerne le tesi come acquisizioni dimostrate, pur condividendole, ma perché valgono come spunto di discussione e, soprattutto, come esemplificazione di un più generale processo di riflessione teorica e di ricerca sperimentale di cui sostengo la necessità. I sistemi di IA generativa, infatti, ci stimolano a rivedere molti dei nostri saperi su questioni fondazionali: non solo la

conoscenza del linguaggio e del significato, ma anche quei tratti centrali che riteniamo caratterizzino la natura umana, quali il ragionamento, la cultura, la creatività, la conoscenza, la natura e la struttura dei rapporti sociali. Non solo: la trasformazione promossa dall'IA investe anche la riflessione normativa e l'agire sociale, con i loro presupposti etici. E qui mi pare di poter dire che una visione dell'etica dell'IA ridotta a ingegneria organizzativa e buone pratiche sia estremamente debole, e che una fondazione concettuale solida le sia indispensabile. Credo dunque che sia possibile costruire una seria teoria etica "su e per" l'intelligenza artificiale, e a maggior ragione sviluppare un approccio di *ethics by design*, solo a condizione di chiarire lo statuto e la natura di questi sistemi, e di costruire una teoria filosoficamente solida relativa a quali concetti di responsabilità, autonomia, agentività, affidabilità e autorità epistemica siano applicabili a queste entità (De Caro & Giovanola, 2025). Anche l'etica dell'IA, in questo senso, è una delle questioni fondazionali che questi sistemi ci costringono a ripensare, non un'appendice applicativa e operativa.

IA generativa come tecnologia socioculturale

Indipendentemente dalle teorie che si vogliano assumere circa la loro natura cognitiva, è chiaro che i grandi modelli linguistici e multimodali (su cui occorrerebbe una riflessione specifica che qui non posso affrontare, poiché il caso multimodale non è una semplice estensione del testuale, ma pone problemi semiotici qualitativamente diversi), poiché sono addestrati su vastissimi insiemi di oggetti semiotici e culturali, a loro volta veicoli di credenze, valori, norme, stereotipi e narrazioni, sono dei portentosi modelli dell'ambiente culturale in cui la nostra specie opera, che altri chiama infosfera o noosfera: io preferisco la nozione lotmaniana di semiosfera (Lotman, 1990, 1992). Essi funzionano come dispositivi di compressione e riorganizzazione della memoria culturale, ma al contempo seguono criteri di modellizzazione che riflettono asimmetrie linguistiche, assiologiche e culturali. In modo simmetrico, questi sistemi retroagiscono sui sistemi culturali. Inseriti nei circuiti della comunicazione sociale, modificano la velocità, la scala e la forma della produzione discorsiva. La generazione automatica di testi e immagini influisce sui processi di diffusione delle credenze, sulla stabilizzazione delle narrazioni condivise e, in ultima istanza, sui comportamenti collettivi. Infine, l'IA generativa riorganizza i workflow dell'industria culturale e della formazione: editoria, giornalismo, audiovisivi, istruzione. Cambiano le filiere produttive, le forme di gatekeeping, le nozioni di autorialità e responsabilità, il senso stesso di pratiche come la scrittura, la valutazione e l'apprendimento.

Questa prospettiva tecno-culturale sull'IA generativa è stata elaborata, da e verso direzioni teoriche diverse, da numerosi studiosi, che lo scienziato politico Henry Farrell (Farrell, 2025; Farrell et al., 2025) ha cercato di organizzare in modo sistematico in quattro correnti: Gopnikismo, Interazionismo, teoria del Role-play e Strutturalismo. Il Gopnikismo, che fa riferimento alle tesi della psicologa evolutiva Alison Gopnik (Yiu et al., 2023), considera i LLM come motori di trasmissione culturale fondamentalmente disconnessi dalla realtà empirica: addestrati esclusivamente su dati testuali prodotti dall'attività culturale umana, essi operano interamente all'interno del dominio simbolico, e questa disconnessione dall'esperienza percettiva e dalla causalità fisica genera un'asimmetria caratteristica, per cui i LLM dimostrano notevole facilità nel riorganizzare, sintetizzare e trasmettere la conoscenza culturale esistente, eppure mancano della capacità di autentica innovazione o scoperta causale. La prospettiva interazionista, che Farrell pone in continuità con le teorie della cultura di Sperber e con gli studi di *cultural evolution* (Sperber, 1996), sposta il fuoco dalle proprietà intrinseche dei sistemi alla loro relazione con l'architettura cognitiva degli interpreti umani, e permette di capire come alcuni contenuti culturali, funzionando da attrattori, conseguano adozione e persistenza diffuse perché fanno presa su moduli mentali evolutisi per altri scopi: in questo quadro i LLM possono generare output particolarmente risonanti con l'architettura psico-cognitiva umana, e dunque nuovi

attrattori culturali. La teoria del *role-play*, che si fonda sul lavoro di Murray Shanahan (Shanahan et al., 2023) e che è stata di recente confermata sperimentalmente da alcuni lavori del team di Anthropic (Chen et al., 2025; Lu et al., 2026), descrive un LLM come una sovrapposizione di simulacri all'interno di un multiverso di personaggi possibili, una *superposizione* di tutte le *personae*, voci, situazioni e stereotipi culturali presenti nel corpus di addestramento, la cui operazione fondamentale non è il ragionamento da credenze stabili, bensì la performance di ruoli finzionali tratti da questo vasto repertorio. Il sistema non seleziona un'identità fissa all'inizializzazione, ma collassa dinamicamente la sua distribuzione sui personaggi possibili man mano che si accumulano indizi contestuali, e data la dominanza statistica di tropi convenzionali e scenari stereotipati nei dati di addestramento, questo processo converge frequentemente su risposte stereotipe o culturalmente sovradeterminate. L'interpretazione strutturalista, infine, connette le architetture degli LLM alla tradizione che da Saussure e Hjelmslev arriva alla semiotica strutturalista novecentesca, e trova la sua espressione più rilevante nel libro di Leif Weatherby (Weatherby, 2025), il quale sostiene che i LLM operino una separazione del linguaggio dalla cognizione individuale, realizzando in forma computazionale la visione strutturalista della *langue* come distinta dalla *parole*, del sistema dall'uso, del fatto sociale dallo stato psicologico; su un terreno concettuale simile si muove anche l'ottimo libro di Claudio Paolucci *Nati cyborg* (Paolucci, 2025), che propone una teoria degli LLM come agenti semiotici non umani sulla base della linea di pensiero che da Peirce arriva a Eco, incrociata con la teoria della mente estesa di Clark.

Si muovono in questa direzione alcuni contributi teorici che io stesso ho pubblicato di recente (Ciotti, 2024, 2025), dove propongo che i modelli linguistici possano essere intesi come istanziazioni computazionali della nozione di semiosfera di Jurij Lotman. Significativamente, Lotman stesso in un articolo del 1979 intitolato «La cultura come intelletto collettivo e i problemi dell'intelligenza artificiale», scrisse: "Dobbiamo sottolineare che l'intelletto collettivo, come modello per l'intelletto artificiale, ha diversi vantaggi rispetto all'intelletto individuale. Poiché l'intelletto collettivo è un meccanismo creato dalla storia dell'umanità, è molto più esplicito, le sue procedure sono manifeste nel linguaggio della cultura e sono registrate in numerosi testi, a differenza dei linguaggi nascosti del cervello umano." (Lotman, 1979). Anche questa prospettiva ha rilevanti implicazioni metodologiche: se i LLM modellizzano la semiosfera, essi forniscono agli studiosi strumenti senza precedenti per analizzare struttura e dinamiche culturali, e l'indagine umanistica tradizionale, vincolata dalle limitazioni soggettive degli interpreti individuali e dalla difficoltà di operazionalizzare concetti culturali, può ora avvalersi di modelli computazionali che esternalizzano e rendono manipolabili i processi stessi della significazione culturale.

IA generativa come metodo nella ricerca umanistica

Veniamo dunque a esaminare le implicazioni metodologiche dell'IA generativa. È ormai evidente anche ai più scettici che questi sistemi siano in grado di riorganizzare in modo profondo i workflow e le stesse pratiche epistemiche fondamentali della ricerca umanistica, digitale e no. La loro presenza, a differenza di metodi e approcci precedenti, è ubiqua e soprattutto permette di operazionalizzare (concetto qui da intendere in senso lato) e delegare a sistemi non umani ogni processo e aspetto della ricerca: la costruzione dei corpora e dei dataset, l'annotazione, l'esplorazione preliminare dei dati, la generazione di ipotesi interpretative e la loro verifica, la sintesi comparativa, persino la confezione finale del prodotto scientifico. Vorrei inoltre osservare che gli LLM, come infrastrutture di ricerca e come apparati metodologici, occupano una posizione non riducibile alla tradizionale dicotomia tra metodi quantitativi e qualitativi, tra close reading e distant reading, tra formalizzazione e libertà ermeneutica. Anzi, si potrebbe dire che l'uso degli LLM intercetti il punto più delicato della ricerca umanistica: la produzione di interpretazioni locali (annotazione di fenomeni linguistici, stilistici, tematici etc.) e di generalizzazioni o esplicazioni ermeneutiche, che non si limitano all'estrazione o classificazione dell'informazione, ma implicano giudizi di rilevanza, coerenza, plausibilità e valore. Si tratta

ovviamente di saggiare fino a che punto questa intersezione sia efficace, di fronte alla complessità degli oggetti culturali e della loro natura intrinsecamente storica. Infine, un ulteriore e centrale ambito di riflessione e di sperimentazione è la relazione che si instaura tra i modelli di IA e i numerosi approcci e modelli di ricerca, qualitativi e quantitativi, formali o non formali, che la nostra comunità scientifica ha sviluppato e applicato negli scorsi decenni: come armonizzare la tradizione teorica e operativa della codifica testuale mediante linguaggi di markup, o l'adozione estensiva di strumenti come le ontologie formali e i linked data, con le *affordances* di questi nuovi agenti e strumenti?

Ormai di lavori che includono LLM in esperimenti di analisi in numerosi ambiti umanistici ve ne sono molteplici. Ne presento uno molto significativo dal punto di vista metodologico e che mi sembra adeguato a sostenere queste mie osservazioni, anche se non recentissimo: «Using Large Language Models for Understanding Narrative Discourse» di Piper e Bagga (Piper e Bagga 2024). L'obiettivo di questo lavoro è verificare se modelli linguistici possano annotare in modo affidabile alcune categorie elementari del discorso narrativo, ispirate alla narratologia di Genette. Gli autori isolano quindici tratti distribuiti lungo i tre assi canonici di tempo, modo e voce: presenza di anacronie, specificità della collocazione temporale, dominanza dei tempi verbali, indicatori di focalizzazione interna, dialogo diretto, successione degli eventi contro introspezione, conflittualità, grado di astrazione, emozionalità, uso del simbolismo, ai quali si aggiunge una feature-trabocchetto fittizia introdotta per misurare la propensione del modello ai falsi positivi. La sperimentazione prevede l'uso di un prompt standardizzato in cui il modello, investito del ruolo di «esperto interprete di storie», riceve un estratto di centocinquanta-duecento parole e una definizione della categoria da identificare, e deve restituire un valore ternario (assente, debole, forte). Sono stati testati i modelli disponibili al momento della stesura, tutti ormai ritirati: GPT-4 e LLaMA-2 8B, Mistral 7B e Mixtral 56B; GPT-4 è stato inoltre impiegato come *silver annotator* per produrre circa 4800 brani etichettati, sui quali LLaMA e Mistral sono stati raffinati tramite LoRA con quantizzazione a otto bit, ottenendo un modello open-source dalle prestazioni assimilabili a quelle di GPT-4. La valutazione è stata condotta su un sottoinsieme di 150 brani annotati manualmente da tre laureandi in letteratura, assumendo due criteri di accordo: maggioritario stretto e min-match, ossia la concordanza con almeno uno degli annotatori umani, criterio che riconosce la pluralità interpretativa intrinseca al dato narrativo.

I risultati sono di notevole interesse: GPT-4 raggiunge un F1-weighted (una misura aggregata delle prestazioni di classificazione che combina precisione e richiamo, pesandone la media in base alla distribuzione delle classi) di circa 0.79 in regime maggioritario e di 0.95 in min-match, mentre il modello Llama3-8B raffinato si attesta poco sotto (F1 ≈ 0.76 nel criterio maggioritario); le feature più nettamente codificabili, come dialogo, anacronia e tempo passato, ottengono valori prossimi all'unità con altissimo accordo intersoggettivo, mentre quelle più sfumate, come conflittualità e astrazione, oscillano attorno a 0.5-0.6, in linea con la concordanza umana di base. È significativo che nessun modello marchi sistematicamente la feature-trabocchetto, segno di una stabilità interpretativa non banale. L'esperimento mostra dunque che gli LLM possono classificare e rendere disponibili per analisi quantitative su larga scala proprietà e caratteristiche testuali a lungo considerate refrattarie all'annotazione automatica, come l'eventfulness o il grado di conflittualità.

Se le potenzialità sono sorprendenti – e lo sono vieppiù con la ininterrotta crescita delle capacità dei modelli: nel momento in cui scrivo è stata appena rilasciata l'ultima generazione dei modelli Anthropic, Claude Fable 5. Per apprezzarne il livello rimando alla puntuale analisi che ne fa Ethan Mollick nel suo blog (2026b) – il dibattito scientifico, e l'esperienza pratica di ogni studioso che abbia avuto modo di utilizzare questi sistemi in reali applicazioni di ricerca, mostrano come essi abbiano anche grandi limiti, come possano fallire clamorosamente in molti task e come

richiedano una continua supervisione e controllo da parte di esperti. Tra i molti lavori che hanno esplorato questi limiti nei nostri ambiti, riporto i risultati di questo recente esperimento condotto dal team del progetto ERC Deep Culture diretto da Tobias Blanke (Deep Culture Team 2026), in ambito di ricerca storica. Prendendo le mosse dalla celebre tesi di Richard Sutton sulla *Bitter Lesson* (https://www.cs.utexas.edu/~eunsol/courses/data/bitter_lesson.pdf), secondo cui i metodi fondati sulla crescita di scala in potenza computazione e dimensione quantitativa dei dataset tendono a soppiantare gli approcci ancorati alla conoscenza esperta di dominio, sono stati replicati tre studi basati su metodi di machine learning precedenti all'introduzione degli LLM sostituendoli con LLM open-source (GPT-oss-120b, Qwen-2.5-VL, Mistral-small-3.2). I risultati mostrano come i sistemi generativi abbiano performance non uniformi a seconda dei task: nella classificazione binaria gli LLM raggiungono il 100% di accuratezza con soli quattro esempi dimostrativi, eguagliando le pipeline dedicate; nell'estrazione di parole chiave la replica è parziale, con un divario residuo di cinque-nove punti percentuali rispetto a KeyBERT, ma i tre modelli convergono nel 91% dei documenti su almeno una parola chiave condivisa, mostrando il temuto «effetto alveare» che rischia di ridurre la pluralità interpretativa costitutiva delle scienze umane; l'attribuzione temporale, infine, fallisce nettamente, con un errore medio assoluto di circa venti anni contro lo 0,74 dell'esperimento originale con BERT, un sistematico *forward-dating bias* e almeno un caso di collasso del modello sull'anno mediano del corpus. La conclusione degli autori è che l'efficacia degli LLM è alta quando la conoscenza rilevante è culturalmente generale e ampiamente rappresentata nei dati di pre-addestramento, e decresce sensibilmente quando il compito analitico richiede competenze locali, distribuzionalmente rare o storicamente specifiche. Va osservato ovviamente che la differenza di prestazione dei modelli *top tier* rispetto ai modelli testati è tale che i risultati potrebbero essere nettamente migliori, come in molti ambiti in cui sono stati testati. Ma qui interviene una seconda considerazione, quella dei costi. Adottare workflow con modelli di frontiera mediante accesso alle loro API ha dei costi che crescono molto rapidamente, rendendo inaccessibili questi metodi per chi non dispone di ingenti finanziamenti di ricerca. Questo è un problema che è ormai non superabile, visti i costi di sviluppo e gestione dei modelli top, ma una rifunzionalizzazione dei numerosi sforzi e investimenti indirizzati nell'ultimo decennio verso la creazione di infrastrutture di ricerca, potrebbe avere certamente un ruolo nella mitigazione dello stesso.

L'efficacia, ma anche i limiti, dei modelli come agenti nei workflow di ricerca di cui ho parlato, ci conducono a una questione metodologica di fondo, troppo spesso elusa, quella della verifica. In entrambi i casi la valutazione del modello poggia su un riferimento umano, l'annotazione manuale di laureandi nel primo, la pipeline pre-LLM con il suo *gold standard* nel secondo, e questo consente di misurare l'accordo e di parlare con qualche rigore di prestazione. Ma la pratica industriale della valutazione dei modelli, fondata su benchmark che misurano accuratezza, recupero o concordanza rispetto a risposte attese, è strutturalmente inadeguata a catturare ciò che le scienze umane chiedono ai loro oggetti. I benchmark generalisti premiano compiti a risposta determinata, dove esiste una soluzione corretta contro cui misurarsi, mentre gli aspetti più qualificanti dell'indagine umanistica risiedono precisamente là dove un *gold standard* non è disponibile né, in linea di principio, costruibile: nell'interpretazione di un testo che ammette letture plurime e legittimamente confliggenti, nella valutazione del valore estetico o funzionale di un manufatto, nella ricostruzione del senso storicamente situato di una serie di eventi. Quando un LLM produce una interpretazione di questo genere, la domanda che ci dobbiamo porre è con quali criteri ne possiamo saggiare la plausibilità rispetto ai dati di partenza, la coerenza interna, la produttività esplicativa. La costruzione di criteri di validazione adeguati a oggetti che resistono per natura a una risposta determinata è forse il contributo metodologico più specifico che la tradizione delle Digital Humanities può recare alla ricerca sull'IA, e non solo riceverne.

Sul piano più strettamente legato a metodi e workflow, i recentissimi sviluppi dei modelli agentivi e di piattaforme come Claude Code e OpenAI Codex spingono ulteriormente i limiti di quanto

i modelli possono fare (Mollick, 2026a). Se fino a pochi mesi fa si poteva al massimo parlare di *vibe coding*, la scrittura di codice eseguita da modelli linguistici (che comunque nel contesto della ricerca umanistica digitale/computazionale rappresenta un enorme supporto), oggi con tali agenti è possibile delegare un intero workflow di ricerca sperimentale a un modello agente, con livelli di autonomia sperimentale prima non immaginabili. Ovviamente si tratta di capire i limiti che l'introduzione estesa di LLM nei protocolli di ricerca in ambito umanistico possano avere, le loro reali capacità di esplorare lo spazio delle possibilità teorico-esplicative, e il ruolo da affidare al controllo del/i ricercatore/i umano/i nel circolo ermeneutico: questioni non solo aperte ma anche soggette a una continua ridefinizione per via della continua innovazione tecnologica. La medesima cautela si deve osservare nel confidare eccessivamente della capacità dei modelli dotati di tool per la *deep research* o di strumenti verticali come Notebook LM di effettuare quel tipo di attività di ricerca che delimita lo stato dell'arte e identifica le questioni e le ipotesi già esplorate e verificate nella letteratura scientifica di un dato ambito di studi.

Questi che ho trattato sono solo esempi che mostrano come le potenzialità dei modelli linguistici siano considerevoli, ma anche come la tradizione di studi delle Digital Humanities, e in particolare quella italiana, per il suo forte orientamento metodologico, possa svolgere una funzione critica essenziale al fine di progettare protocolli di uso controllato, definire criteri di validazione, costruire pratiche di cooperazione tra analisi automatica e interpretazione umana.

Una rivoluzione epistemologica da capire e guidare: verso un'agenda

Riassumendo gli argomenti che ho esposto finora, l'IA generativa ci costringe a riconcettualizzare le nostre assunzioni fondamentali su cosa siano il significato, la cultura, la conoscenza, l'interpretazione, la metodologia. Le Digital Humanities per decenni si sono dovute misurare con la resistenza degli oggetti di studio dei saperi umanistici all'operazionalizzazione computazionale. Concetti come genere, stile, ideologia, struttura narrativa, variante, intertestualità, valore estetico, fatto storico, e infiniti altri che popolano gli studi umanistici resistono sia alla formalizzazione attraverso la rappresentazione simbolica classica o sistemi di regole esplicite, sia all'operazionalizzazione quantitativa, alla riduzione a misure e numeri da trattare con modelli statistico-matematici. I sistemi di IA generativa impiegano paradigmi di rappresentazione ed elaborazione fondamentalmente diversi. Piuttosto che simboli espliciti manipolati secondo regole formali, o quantità misurabili, questi sistemi codificano linguaggio, conoscenza e cultura in rappresentazioni continue sub-simboliche, vettori ad alta dimensionalità di numeri reali privi di sottostruttura discreta. Le singole dimensioni non possiedono interpretabilità semantica inerente; il significato emerge olisticamente dalle relazioni geometriche attraverso l'intero spazio. Questo formato rappresentazionale si dimostra notevolmente capace di catturare fenomeni culturali refrattari alla formalizzazione. Inoltre, i modelli linguistici permettono, come visto, di esternalizzare i processi interpretativi tradizionalmente confinati alla cognizione umana soggettiva. La comprensione ermeneutica, l'interpretazione del significato all'interno dei testi culturali, ha storicamente risieduto nelle menti dei singoli studiosi, accessibile solo attraverso l'introspezione e comunicabile solo attraverso linguaggi descrittivi. I LLM esternalizzano questi processi, ma allo stesso tempo li rendono tanto poco perspicui quanto quelli umani.

L'emergere dell'IA generativa segna quindi non meramente uno sviluppo tecnologico ma una trasformazione epistemologica che richiede sofisticazione teorica e metodologica. Gli studiosi devono sviluppare cornici rigorose per comprendere cosa i sistemi intelligenti sono e non sono,

come si relazionano alla cognizione culturale umana, quali tipi di conoscenza possono fornire e quali limitazioni ne vincolano l'applicazione. Si tratta di un lavoro che deve procedere sia sul piano concettuale sia su quello empirico, attraverso studi sistematici su come le rappresentazioni dei LLM corrispondano alle concettualizzazioni culturali umane, dove divergano e quali bias o distorsioni introducano. Infine, dal punto di vista metodologico è necessario sviluppare best practices per impiegare i LLM nella ricerca mantenendo consapevolezza critica delle loro limitazioni e potenziali applicazioni erronee.

A questa esigenza si può dare una prima risposta concreta e collettiva. Anche facendo tesoro della discussione che si è sviluppata durante il recente convegno AIUCD 2026, propongo dunque che la nostra comunità scientifica si faccia promotrice della redazione di un libro bianco, o di un corpo di linee guida condivise, dedicato ai metodi, ai casi d'uso e ai criteri di validazione per l'impiego dei modelli generativi nella ricerca umanistica. Non si tratta di predisporre una pubblicazione scientifica intesa in senso tradizionale, ma di un documento vivo e periodicamente rivisto, che censisca gli usi documentati e ne discuta apertamente potenzialità e limiti, che proponga protocolli di uso controllato e criteri di trasparenza e riproducibilità, che affronti i nodi della provenienza dei dati e della dipendenza dalle infrastrutture proprietarie, e che offra così alla comunità non una serie di prescrizioni ma una mappa orientativa, costruita dal basso e aperta alla revisione. Uno strumento simile risponderebbe a un bisogno reale e già avvertito, quello di non lasciare il singolo ricercatore solo di fronte a scelte metodologiche ed etiche di grande portata, e collocherebbe al tempo stesso la nostra comunità nel ruolo che le compete, quello di interlocutore attivo e non meramente recettivo nel dibattito sull'IA.

La tradizione teorica e culturale dell'Informatica Umanistica (Ciotti, 2018) di cui la nostra comunità scientifica è erede, ci pone nella condizione di assumere un ruolo guida in questa impresa, che è anche una ennesima (forse l'ultima) possibilità di riaffermare la necessità e la centralità sociale dei saperi umanistici. Già la storia passata ha mostrato come la ricerca umanistica, anche nei settori più esoterici (e a prima vista privi di impatto sociale e di sostenibilità economica, sulla base della razionalità economicista che oggi domina anche l'istruzione superiore) se si mette in gioco seriamente e punta all'innovazione metodologica e disciplinare può non solo invertire la narrativa del declino ma contribuire allo sviluppo dei nuovi ecosistemi digitali; perché i problemi veramente complessi con cui la ricerca sull'Intelligenza Artificiale si deve misurare sono spesso quelli di cui si occupano le discipline umanistiche: storicità e contestualità dei significati; diversità delle lingue e delle culture; stratificazione sociale e culturale del linguaggio; rapporto tra verità, significato e discorso, solo per ricordarne alcuni.

La posta in gioco si estende oltre la ricerca accademica per abbracciare implicazioni culturali e sociali più ampie. In primo luogo, nella sfera che non ho qui trattato, solo perché la sua rilevanza e difficoltà è al di là delle mie capacità di singolo studioso: quella della formazione. Ma più in generale, se è vero che i sistemi di IA generativa di oggi e quelli a venire operano sui sistemi culturali-semiotici, essi possiedono una capacità senza precedenti di plasmare il discorso sociale, influenzare la formazione delle credenze e riconfigurare i paesaggi culturali. A ciò si aggiunge la questione etica annunciata in apertura sul piano teorico, che l'indistinguibilità comportamentale sollevata dall'argomento di Turing rende ineludibile: se i criteri in base ai quali attribuiamo o neghiamo competenza, comprensione e agentività sono comportamentali, allora le nostre attribuzioni reciproche, verso le macchine e, per riflesso, verso gli altri esseri umani, non sono atti neutri, ma scelte gravide di conseguenze morali, e una teoria della valutazione adeguata deve farsi carico anche di questa dimensione. Comprendere questi sistemi diventa essenziale non solo per la ricerca delle DH e quella umanistica in generale, ma per assumere un atteggiamento responsabile verso tecnologie che stanno trasformando la cultura e la società in cui viviamo. Per questo penso che, come comunità, sia nostra responsabilità predisporre un'agenda che sia anche

la mappa per il nostro futuro prossimo come studiosi e come intellettuali socialmente responsabili.

Reference

Buzzetti, Dino. 2019. «The Origins of Humanities Computing and the Digital Humanities Turn». Tradotto da Massimo Lollini. *Humanist Studies & the Digital Age* 6 (1): 32–58. https://doi.org/10.5399/uo/hsda.6.1.3.

C2DH. 2025. «AI Manifesto». https://www.uni.lu/c2dh-en/articles/ai-manifesto/.

Chen, Runjin, Andy Arditi, Henry Sleight, Owain Evans, e Jack Lindsey. 2025. «Persona Vectors: Monitoring and Controlling Character Traits in Language Models». Preprint, arXiv. https://doi.org/10.48550/arXiv.2507.21509.

Ciotti, Fabio. 2018. «From Informatica Umanistica to Digital Humanities and Return: A Conceptual History of Italian DH». *Testo e Senso* 19. https://testoesenso.it/index.php/testoesenso/article/view/398.

Ciotti, Fabio. 2023a. «Introduzione. La galassia delle Digital Humanities». In *Digital Humanities. Metodi, strumenti, saperi*. Roma: Carocci.

Ciotti, Fabio. 2023b. «Minerva e il pappagallo. IA generativa e modelli linguistici nel laboratorio dell'umanista digitale». *Testo e Senso* 26: 289–315. https://doi.org/10.58015/2036-2293/671.

Ciotti, Fabio. 2024. «Gli LLM come lettori modello artificiali». In *Me.Te. Digitali. Mediterraneo in rete tra testi e contesti. Proceedings del XIII Convegno Annuale AIUCD2024*, 342–344. Bologna: AIUCD. https://doi.org/10.6092/unibo/amsacta/7927.

Ciotti, Fabio. 2025. «Lettori artificiali e macchine pigre: Cosa i Large Language Model possono farci capire sulla cooperazione interpretativa». In *Letteratura e intelligenza artificiale. Un dialogo interdisciplinare*, 109–146. Roma: Lithos.

De Caro, Mario, e Benedetta Giovanola. 2025. *Intelligenze. Etica e politica dell'IA*. Bologna: Il Mulino.

Deep Culture Team. 2026. «The Occasional Bitter Lesson: AI Engineering and Expert Knowledge in the Digital Humanities». Deep Culture Research Notes. https://deep-culture.org/the-occasional-bitter-lesson-ai-engineering-and-expert-knowledge-in-the-digital-humanities/.

Dennett, Daniel C. 1989. *The Intentional Stance*. Cambridge, MA: MIT Press.

Dennett, Daniel C. 1991. «Real Patterns». *The Journal of Philosophy* 88 (1): 27–51.

Farrell, Henry. 2025. «Large Language Models Are Cultural Technologies. What Might That Mean?» *Programmable Mutter*, 7 gennaio. https://www.programmablemutter.com/p/large-language-models-are-cultural.

Farrell, Henry, Alison Gopnik, Cosma Shalizi, e James Evans. 2025. «Large AI Models Are Cultural and Social Technologies». *Science* 387 (6739): 1153–1156. https://doi.org/10.1126/science.adt9819.

Futrell, Richard, e Kyle Mahowald. 2025. «How Linguistics Learned to Stop Worrying and Love the Language Models». *Behavioral and Brain Sciences*, 1–98. https://doi.org/10.1017/S0140525X2510112X.

Griffiths, Thomas L., Brenden M. Lake, R. Thomas McCoy, Ellie Pavlick, e Taylor W. Webb. 2026. «Whither Symbols in the Era of Advanced Neural Networks?» *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2026.02.003.

Hemment, Drew, e Cody Kommers. 2025. *Doing AI Differently: Rethinking the Foundations of AI via the Humanities*. https://doi.org/10.5281/zenodo.16421295.

Horvath, Agnes, Anastasia Bonch-Osmolovskaya, Alison Goodrich, Adriana Lombana-Bermudez, Anita Gurumurthy, Boris Orekhov, Caitlin Clark, et al. 2022. *Global Debates in the Digital Humanities*. A cura di Domenico Fiormonte, Sukanta Chaudhuri e Paola Ricaurte. Vol. 8. Minneapolis: University of Minnesota Press. https://doi.org/10.5749/9781452968919.

Klein, Lauren, Meredith Martin, André Brock, Maria Antoniak, Melanie Walsh, Jessica Marie Johnson, Lauren Tilton, e David Mimno. 2025. «Provocations from the Humanities for Generative AI Research». Versione 2. Preprint, arXiv. https://doi.org/10.48550/arXiv.2502.19190.

Klowden, Tanya, e Terence Tao. 2026. «Mathematical Methods and Human Thought in the Age of AI». Versione 1. Preprint, arXiv. https://doi.org/10.48550/arXiv.2603.26524.

Kommers, Cody, Ruth Ahnert, Maria Antoniak, Emmanouil Benetos, Steve Benford, Mercedes Bunz, Baptiste Caramiaux, et al. 2025. «Computational Hermeneutics: Evaluating Generative AI as a Cultural Technology». Preprint, SSRN. https://doi.org/10.2139/ssrn.5409144.

Lindsey, Jack, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, et al. 2025. «On the Biology of a Large Language Model». Transformer Circuits Thread. https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Lotman, Jurij M. 1979. «Culture as Collective Intellect and the Problems of Artificial Intelligence». In *Russian Poetics in Translation*, vol. 6, 84–96.

Lotman, Jurij M. 1990. *Universe of the Mind: A Semiotic Theory of Culture*. Bloomington: Indiana University Press.

Lotman, Jurij M. 1992. *La semiosfera: L'asimmetria e il dialogo nelle strutture pensanti*. A cura di Simonetta Salvestroni. 2a ed. Venezia: Marsilio.

Lu, Christina, Jack Gallagher, Jonathan Michala, Kyle Fish, e Jack Lindsey. 2026. «The Assistant Axis: Situating and Stabilizing the Default Persona of Language Models». Versione 1. Preprint, arXiv. https://doi.org/10.48550/arXiv.2601.10387.

MLA AI and Research Working Group. 2026. *AI and the Humanities: A Framework for Language and Literary Scholarship*. New York: Modern Language Association of America. https://mlaai.hcommons.org/.

Mollick, Ethan. 2026a. «Claude Code and What Comes Next». *One Useful Thing*, 18 febbraio. https://www.oneusefulthing.org/p/claude-code-and-what-comes-next.

Mollick, Ethan. 2026b. «What it feels like to work with Mythos». *One Useful Thing*, 10 giugno. https://www.oneusefulthing.org/p/what-it-feels-like-to-work-with-mythos.

OpenAI. 2026. «An OpenAI Model Has Disproved a Central Conjecture in Discrete Geometry». 20 maggio. https://openai.com/index/model-disproves-discrete-geometry-conjecture/.

Paolucci, Claudio. 2025. Nati cyborg. Cosa l'intelligenza artificiale generativa ci dice dell'essere umano. Roma: Luca Sossella Editore.

Piper, Andrew, e Sunyam Bagga. 2024. «Using Large Language Models for Understanding Narrative Discourse». In *Proceedings of the 6th Workshop on Narrative Understanding*, a cura di Yash Kumar Lal, Elizabeth Clark, Mohit Iyyer, et al. Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wnu-1.4.

Shanahan, Murray, Kyle McDonell, e Laria Reynolds. 2023. «Role Play with Large Language Models». *Nature* 623 (7987): 493–498. https://doi.org/10.1038/s41586-023-06647-8.

Sperber, Dan. 1996. *Explaining Culture: A Naturalistic Approach*. Oxford: Basil Blackwell.

Templeton, Adly, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, et al. 2024. «Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet». Transformer Circuits Thread, 21 maggio. https://transformer-circuits.pub/2024/scaling-monosemanticity/.

Turing, Alan M. 1950. «Computing Machinery and Intelligence». *Mind* 59 (236): 433–460. https://doi.org/10.1093/mind/LIX.236.433.

Tsoukalas, George, Anton Kovsharov, Sergey Shirobokov, Anja Surina, Moritz Firsching, Gergely Bérczi, Francisco J. R. Ruiz, et al. 2026. «Advancing Mathematics Research with AI-Driven Formal Proof Search». Preprint, arXiv. https://arxiv.org/abs/2605.22763.

Underwood, Ted. 2021. «Mapping the Latent Spaces of Culture». *The Stone and the Shell*, 21 ottobre. https://tedunderwood.com/2021/10/21/latent-spaces-of-culture/.

Underwood, Ted. 2025. «The Impact of Language Models on the Humanities and Vice Versa». *Nature Computational Science* 5 (9): 695–697. https://doi.org/10.1038/s43588-025-00819-4.

Underwood, Ted, Laura K. Nelson, e Matthew Wilkens. 2025. «Can Language Models Represent the Past without Anachronism?» Versione 1. Preprint, arXiv. https://doi.org/10.48550/arXiv.2505.00030.

Weatherby, Leif. 2025. *Language Machines: Cultural AI and the End of Remainder Humanism*. Minneapolis: University of Minnesota Press.

Yiu, Eunice, Eliza Kosoy, e Alison Gopnik. 2023. «Transmission Versus Truth, Imitation Versus Innovation: What Children Can Do That Large Language and Language-and-Vision Models Cannot (Yet)». *Perspectives on Psychological Science* 19 (5): 874–883. https://doi.org/10.1177/17456916231201401.