Modelli linguistici di grandi dimensioni e ricerca scientifica:

opportunità, rischi e controlli

Rocco De Nicola

Istituto di Informatica e Telematica (CNR), Pisa · Gran Sasso Science Institute, L'Aquila

Gli LLM sono utili quando velocizzano lavoro ripetitivo e preparatorio. Diventano rischiosi quando il loro output entra direttamente nel processo di validazione o nella decisione scientifica. La questione da porre non è se ma come integrarli nel processo scientifico senza comprometterne l'integrità.

Nel discorso pubblico, intelligenza artificiale e Large Language Model vengono spesso usati come sinonimi. Non lo sono. L'IA è il campo più ampio che comprende tecniche diverse — dai sistemi simbolici all'apprendimento automatico, dalla visione artificiale alla robotica, fino ai modelli generativi. Gli LLM sono una classe specifica di questi sistemi: modelli addestrati su grandi quantità di testo e, sempre più spesso, di dati multimodali, capaci di produrre risposte linguisticamente plausibili, codice, sintesi, classificazioni e proposte operative.

Questa collocazione è importante per la ricerca scientifica. Molti strumenti di IA erano già presenti nei laboratori come metodi di classificazione, ottimizzazione, simulazione o analisi di pattern. Gli LLM cambiano il punto di accesso: non richiedono soltanto dati strutturati o interfacce specialistiche, ma consentono di trasformare domande, protocolli, articoli e codice in oggetti manipolabili attraverso il linguaggio naturale. Questa potenza di interfaccia spiega sia il loro interesse sia la loro ambiguità: possono accelerare passaggi cognitivi diversi, ma possono anche rendere meno visibile la distanza tra una formulazione plausibile e una conoscenza verificata.

Un *Large Language Model* (LLM) è un sistema probabilistico di generazione del testo: riceve un contesto in ingresso e produce, token dopo token, la continuazione statisticamente più plausibile. Non è un database, non è un motore di ricerca, non è un sistema esperto. Questa distinzione, spesso trascurata nella retorica divulgativa, ha conseguenze profonde quando si tratta di integrare questi strumenti nei processi della ricerca scientifica.

Il problema centrale è il disaccoppiamento tra plausibilità linguistica e affidabilità operativa. Un testo fluido e ben strutturato non è necessariamente un testo corretto. Il modello può

generare affermazioni convincenti prive di fondamento nelle fonti disponibili — il fenomeno comunemente noto come «allucinazione» — con un grado di confidenza formale indistinguibile da quello con cui cita dati reali. In ambito scientifico, dove la verità fattuale è il criterio ultimo, questo disaccoppiamento non è un dettaglio tecnico: è il nodo centrale di qualsiasi politica d'uso responsabile.

Questo articolo analizza dove e come gli LLM possono effettivamente supportare la ricerca, quali rischi introducono nelle diverse fasi della pipeline scientifica e quali controlli minimi sono necessari per preservare integrità e tracciabilità. L'obiettivo non è esprimere un giudizio aprioristico — né favorevole né contrario — ma offrire una mappa operativa per un'adozione consapevole.

1. Quattro ruoli dell'IA nella scienza

Una tassonomia proposta in letteratura distingue diverse modalità ricorrenti con cui l'intelligenza artificiale viene impiegata nei processi scientifici [2]. Non si tratta di categorie rigide né mutuamente esclusive. Ciò che li distingue è la natura del contributo richiesto al modelloe, di conseguenza, il tipo di rischio che ciascun ruolo introduce nel processo di produzione della conoscenza.

L'IA come oracolo

Nel ruolo di **oracolo**, l'intelligenza artificiale viene impiegata per sintetizzare grandi quantità di letteratura scientifica, identificare pattern latenti in corpus testuali o bibliometrici, e generare ipotesi di ricerca che un singolo ricercatore difficilmente avrebbe formulato a partire da una lettura manuale. È il ruolo che ha suscitato il maggiore entusiasmo divulgativo: la promessa di un sistema capace di «leggere tutto» e suggerire connessioni inedite tra campi distanti è intellettualmente seduttiva.

Il rischio specifico, tuttavia, è strutturale. Come argomentano Messeri e Crockett [2], i modelli linguistici tendono a generare quella che gli autori chiamano un'*illusione di comprensione*: la fluidità e la coerenza dell'output inducono il ricercatore a sopravvalutare la profondità dell'analisi sottostante. Le ipotesi prodotte non emergono dall'analisi della realtà, ma dall'analisi di testi il cui messaggio può essere determinato da specifiche priorità o pregiudizi. Un modello addestrato prevalentemente su letteratura anglofona, su riviste ad alto impact factor, su domini ben finanziati, tenderà a considerare «plausibili» ipotesi che si collocano all'interno di quei paradigmi dominanti. L'IA-oracolo rischia dunque di amplificare il mainstream piuttosto che sfidarlo, producendo non nuovi insight, ma variazioni di quanto già pubblicato.

Luo et al. [3] approfondiscono questa dinamica distinguendo due modalità operative dell'IA-oracolo: approcci *literature-based*, in cui il modello sintetizza corpus testuali per identificare connessioni latenti tra domini, e approcci *data-driven*, in cui il modello cerca pattern direttamente nei dati sperimentali. In entrambi i casi, gli autori sottolineano la necessità di criteri espliciti per valutare le ipotesi generate (novità, validità logica, chiarezza e significatività empirica) criteri che il modello non applica autonomamente, ma che devono essere imposti dall'esterno dal ricercatore. L'entusiasmo per la capacità generativa dell'IA-oracolo non deve oscurare questa responsabilità valutativa, che resta irriducibilmente umana.

L'IA come surrogato

Come **surrogato**, l'IA sostituisce esperimenti reali attraverso la generazione di dati sintetici o l'esecuzione di simulazioni computazionali. Questo ruolo ha una lunga storia in fisica teorica e in chimica computazionale, ma si è enormemente esteso con la diffusione dei modelli generativi: oggi si parla di generazione di immagini mediche sintetiche, per addestrare classificatori; di molecole plausibili secondo certe regole chimiche, per esplorare spazi chimici; di pazienti virtuali, per testare protocolli clinici.

Il vantaggio è reale: la generazione sintetica accelera la ricerca in contesti dove i dati reali sono scarsi, costosi o eticamente problematici da raccogliere. Il rischio, tuttavia, è altrettanto reale: il NIST AI 600-1 [1] identifica tra le criticità dell'IA generativa la *confabulazione*, cioè la produzione di contenuti falsi o erronei presentati con sicurezza, e osserva che tali output derivano dall'approssimazione della distribuzione statistica dei dati di training. In domini ad alta competenza contestuale, come quello biomedico, questo può indurre gli utenti a trattare regolarità statistiche del dataset come se fossero evidenze empiriche. A ciò si aggiungono i rischi di bias, omogeneizzazione e over-reliance, per cui il surrogato computazionale può consolidare assunzioni implicite anziché sottoporle a verifica.

Luo et al. [3] documentano concretamente questa estensione del ruolo di surrogato: sistemi basati su LLM vengono oggi impiegati nella scoperta di farmaci, nella progettazione di molecole, nella sintesi di protocolli sperimentali per chimica e biologia, e persino nell'automazione di interi flussi di lavoro di laboratorio. Gli autori avvertono però che la pianificazione autonoma resta fragile perché i modelli possono proporre procedure non eseguibili o fallire in presenza di vincoli specialistici non rappresentati nel corpus di addestramento.

L'IA come strumento quantitativo

Nel ruolo di **strumento quantitativo**, l'IA esegue analisi su larga scala (classificazione, regressione, clusterizzazione, estrazione di pattern) su dataset che sarebbero intrattabili per il

solo ricercatore umano. È il ruolo più consolidato e, in molti contesti, il meno controverso: l'utilizzo di reti neurali per l'analisi di sequenze genomiche, di immagini astronomiche o di segnali neurofisiologici è ormai prassi consolidata in numerose discipline.

Eppure anche qui sono presenti rischi. La scalabilità dell'IA non distingue tra segnale robusto ed errore sistematico. I modelli linguistici impiegati in pipeline analitiche complesse possono propagare e amplificare errori sistematici attraverso stadi successivi di elaborazione, rendendo difficile la loro individuazione a valle. Più il dataset è grande, più l'analisi è automatizzata, più diventa difficile individuare il punto in cui il sistema ha cominciato a misurare qualcosa di diverso da ciò che il ricercatore intendeva misurare.

Luo et al. [3] sottolineano però la mancanza di benchmark affidabili per valutare l'accuratezza di sistemi quantitativi nei diversi domini: la diffusione dell'IA ha preceduto lo sviluppo delle infrastrutture di valutazione necessarie per controllarne l'affidabilità. Per il ricercatore, questo rende necessario un obbligo metodologico: le assunzioni statistiche incorporate nello strumento devono essere esplicitate, i risultati devono essere confrontati con metodi alternativi e gli errori sistematici devono essere cercati attivamente.

L'IA come arbitro

Il caso più critico, e quello su cui la riflessione è ancora più urgente, è il ruolo di **arbitro**: l'IA viene chiamata a valutare il merito scientifico di contributi, ipotesi o metodologie — assistendo la revisione tra pari, classificando la qualità di un manoscritto, segnalando potenziali problemi metodologici, sostituendo parzialmente il revisore umano.

In questo ruolo si esternalizza non un compito computazionale, ma il giudizio stesso: la decisione su cosa conta come evidenza sufficiente, su quale grado di novità giustifica la pubblicazione, su quale metodologia è considerata rigorosa in un dato campo. La preoccupazione è condivisa da istituzioni di diversa natura: l'UNESCO [6] critica esplicitamente la delega del giudizio valutativo a sistemi automatizzati in contesti educativi e di ricerca; l'AI Act europeo [7] classifica i sistemi di valutazione automatizzata tra quelli ad alto rischio, richiedendo supervisione umana significativa; le linee guida dell'ERA Forum della Commissione Europea [9] raccomandano esplicitamente che i sistemi di IA generativa non assumano un ruolo sostanziale nella valutazione del merito scientifico dei contributi di ricerca. Il modello tenderà a valutare positivamente ciò che assomiglia a ciò che già ha appreso a riconoscere come «buona scienza». Il consenso scientifico rischia così di cristallizzarsi non attorno a ciò che l'evidenza supporta, ma attorno a ciò che il modello considera plausibile, un cortocircuito in cui il passato valuta il futuro.

Questo non significa che l'IA non possa avere un ruolo utile nell'ecosistema della revisione scientifica. Ma la distinzione tra assistenza procedurale e delega è cruciale, e mantenerla richiede una consapevolezza esplicita da parte della comunità scientifica. Luo et al. [3] tracciano una distinzione operativa utile tra due configurazioni del ruolo di arbitro. Nella prima, il sistema genera valutazioni in modo quasi autonomo; nella seconda, assiste il revisore umano con strumenti di supporto: riassunti strutturati del manoscritto, verifica di coerenza interna, controllo delle citazioni, individuazione di errori formali, supporto alla stesura del giudizio.

Nel complesso, la rassegna di Luo et al. [3] conferma e articola il quadro dei quattro ruoli: gli LLM sono strumenti avanzati di produttività scientifica, non sostituti completi del ricercatore. Il loro contributo è reale in tutte e quattro le modalità, ma rimane limitato da allucinazioni, debolezza nella verifica empirica, difficoltà con la competenza di dominio specialistica, problemi etici irrisolti e mancanza di benchmark affidabili. La direzione futura più promettente è l'integrazione degli LLM con sistemi di verifica indipendente, strumenti specialistici, laboratori automatizzati e processi *human-in-the-loop*: un'automazione che migliori la ricerca senza comprometterne l'integrità, e che mantenga il controllo umano nei punti critici della pipeline.

Queste quattro modalità non esauriscono il panorama, e la loro separazione analitica non deve oscurare il fatto che i rischi si moltiplicano quando i ruoli si combinano. Messeri e Crockett [2] descrivono questo scenario come una potenziale «ecologia degli strumenti AI» in cui la distribuzione del lavoro cognitivo tra sistemi diversi riduce progressivamente i punti di controllo umano sull'intero ciclo epistemico. Un sistema che genera ipotesi, le testa su dati sintetici, analizza i risultati in modo automatizzato e valuta la rilevanza delle proprie scoperte ha concentrato in sé l'intero processo scientifico — ed è precisamente questo scenario che rende urgente una riflessione strutturata sull'uso dell'IA nella ricerca.

Un caso esemplare è quello di Giorgio Parisi e Francesco Zamponi [10], che dichiarano esplicitamente come la dimostrazione del teorema centrale del loro lavoro, osservata numericamente da oltre dodici anni ma mai dimostrata analiticamente, sia stata ottenuta attraverso un'interazione strutturata con Claude e successivamente verificata dai due autori. Il processo non è stato una semplice interrogazione: dalla trascrizione resa pubblica da Parisi emerge un dialogo di circa quaranta turni, in cui il modello propone una strategia dimostrativa, il ricercatore ne individua un errore, il modello corregge, il ricercatore reindirizza. In un passaggio cruciale, Claude aveva utilizzato il principio del massimo per provare la non-negatività di una certa funzione: Parisi ha riconosciuto che l'argomento non era applicabile in quel contesto e ha proposto un approccio alternativo. La prova finale è emersa da questa iterazione.

Il caso è istruttivo non perché mostri che un LLM può fare matematica avanzata, ma perché rende visibile la struttura epistemica che rende il risultato accettabile: il modello ha contribuito alla formulazione di passaggi dimostrativi e all'esplorazione di strategie, ma la validità di ciascun passaggio è stata sottoposta a verifica indipendente da parte di esperti in grado di riconoscerne gli errori. La correttezza matematica non derivava dalla plausibilità dell'output, ma dalla possibilità di sottoporlo a controllo rigoroso. Questo è esattamente il punto di confine tra uso assistivo e delega epistemica: non dove si trova il contributo del modello nel processo, ma se quel contributo rimane tracciabile, verificabile e ancorato a responsabilità umane definite.

Il caso è istruttivo perché mostra che la distinzione tra uso assistivo e delega epistemica non coincide più con la separazione tra "scrivere" e "scoprire". Quando il contributo del modello è sostanziale, la questione non riguarda solo se l'LLM debba essere citato: riguarda quale parte del processo scientifico sia stata affidata al sistema, quali controlli indipendenti siano stati applicati e fino a che punto la comunità possa ricostruire la genealogia della scoperta.

La discussione pubblica tende a polarizzarsi tra celebrazione della creatività artificiale e scetticismo radicale. Per una governance responsabile della ricerca il punto è più preciso: un risultato ottenuto con il contributo di un LLM è scientificamente accettabile solo se il percorso che lo ha prodotto rimane tracciabile, verificabile e attribuibile a responsabilità umane definite.

2. La pipeline della ricerca

La ricerca scientifica è una sequenza di passaggi distinti. In forma schematica, essa procede dalla formulazione di una domanda o di un problema conoscitivo, alla revisione della letteratura esistente, alla costruzione di ipotesi e disegni metodologici, alla raccolta e gestione dei dati, all'analisi, all'interpretazione dei risultati, alla scrittura e comunicazione degli esiti, fino alla revisione tra pari e alla pubblicazione. Il processo non è sempre lineare: nuovi risultati possono modificare le ipotesi iniziali, generare nuove domande o richiedere ulteriori esperimenti.

Alla luce di questa articolazione, i rischi degli LLM non sono uniformi lungo la pipeline scientifica, ma variano significativamente a seconda della fase in cui il modello viene impiegato e del costo dell'errore associato. Un errore nella riformulazione di una query bibliografica ha conseguenze diverse da un errore nell'analisi statistica, nella gestione di dati sensibili o nella valutazione confidenziale di un manoscritto. Per questo, l'uso degli LLM nella ricerca deve essere valutato non in astratto, ma in relazione alla funzione svolta, al grado di

automazione introdotto e alla possibilità di verifica indipendente. Le sezioni seguenti analizzano i possibili impatti di un LLM in ciascuna delle fasi principali.

Ricerca bibliografica

La fase di revisione della letteratura beneficia dell'uso degli LLM per la riformulazione di query, la sintesi di abstract, il clustering tematico e lo screening preliminare dei risultati. In questa fase il modello può aiutare a esplorare rapidamente un campo, individuare parole chiave alternative e organizzare grandi quantità di testi. Tuttavia, i rischi principali sono le citazioni inventate, il bias di selezione, l'appiattimento delle controversie e la perdita del nesso tra affermazione e fonte. Il controllo minimo richiede retrieval su fonti curate, verifica manuale sulle fonti primarie e tracciamento sistematico della provenienza di ogni affermazione rilevante.

Metodo e ipotesi

Nella definizione del metodo e delle ipotesi, un LLM può supportare il brainstorming, produrre checklist di minacce alla validità e suggerire varianti di disegno sperimentale. Può inoltre aiutare a rendere esplicite alternative teoriche, variabili confondenti e possibili criteri di falsificazione. Il rischio, tuttavia, è l'illusione di ampiezza esplorativa: il modello può generare molte opzioni senza distinguere adeguatamente tra ipotesi falsificabili, ipotesi vaghe e mere riformulazioni linguistiche, confondendo correlazione e causalità. I controlli includono l'esplicitazione delle assunzioni, la formulazione di predizioni differenziali, il confronto con la letteratura primaria e la verifica della coerenza tra domanda, metodo, dati attesi e analisi previste.

Gestione dei dati

Nella gestione dei dati, gli LLM possono accelerare la documentazione, la costruzione di data dictionary, la generazione di codice per trasformazioni ripetitive e la produzione di esempi sintetici per testare pipeline. Questa fase è però particolarmente delicata perché gli errori introdotti a monte possono propagarsi invisibilmente in tutte le analisi successive. I rischi principali riguardano il leakage di dati sensibili, la contaminazione tra set di training, validazione e test, e le distorsioni introdotte durante il cleaning: codifiche scorrette, imputazioni implausibili, normalizzazioni inappropriate o filtri che escludono casi in modo sistematico. È fondamentale classificare i dati prima di qualsiasi interazione con il modello, distinguere dati pubblici, confidenziali e sensibili, e usare solo strumenti approvati per le categorie soggette a vincoli etici, legali o istituzionali.

Coding e analisi

Nel coding e nell'analisi, il supporto degli LLM copre completamento, debugging, test unitari, documentazione e traduzione tra linguaggi o librerie. Il vantaggio operativo è evidente, ma il rischio tecnico è elevato: il codice generato può contenere vulnerabilità non rilevate, dipendenze obsolete, pacchetti inesistenti [5], errori nelle trasformazioni dei dati o analisi statistiche non validate. La ricerca empirica [4] conferma che l'assistenza AI può aumentare la produttività, ma non elimina la possibilità di introdurre vulnerabilità. I rischi di sicurezza sono ulteriormente documentati dall'OWASP GenAI Security Project [8], che segnala tra le vulnerabilità più frequenti l'esecuzione di codice generato da LLM. I controlli minimi sono test automatizzati, revisione profonda del codice, analisi statica, controllo delle dipendenze.

Scrittura e comunicazione

Nella scrittura scientifica, gli LLM accelerano la produzione di bozze, la ristrutturazione argomentativa, l'editing, la traduzione e la sintesi. Possono migliorare chiarezza, coerenza e leggibilità, soprattutto nelle fasi di revisione formale. I rischi principali sono però affermazioni non supportate, citazioni riempitive, omogeneizzazione dello stile, perdita di precisione terminologica. Il rischio non è soltanto che il testo contenga errori, ma che renda più difficile distinguere tra interpretazione, evidenza e speculazione. I controlli richiedono una separazione netta tra editing linguistico e produzione di contenuto scientifico, fact-checking puntuale, conservazione del nesso tra claim e fonte primaria, e revisione umana delle formulazioni che incidono su risultati, limiti e conclusioni.

Peer review e pubblicazione

La peer review è uno dei meccanismi centrali attraverso cui la comunità scientifica valuta la qualità, la robustezza e la credibilità dei risultati prima della pubblicazione. Pertanto la fase di peer review e pubblicazione presenta i rischi più seri perché riguarda materiali non pubblici, giudizi valutativi e decisioni che incidono direttamente sulla reputazione scientifica e sulla circolazione della conoscenza. Gli LLM possono aiutare a controllare chiarezza, struttura e completezza di una revisione, ma il loro uso come arbitro sostanziale della qualità scientifica introduce rischi di violazione della confidenzialità, bias implicito, appiattimento del giudizio critico e delega impropria della responsabilità valutativa. Il divieto di caricare manoscritti, dati, revisioni o materiali riservati su servizi non approvati non è una semplice raccomandazione prudenziale, ma un vincolo di integrità e riservatezza. La supervisione umana, il rispetto delle policy editoriali e la distinzione tra supporto linguistico e valutazione scientifica completano il quadro.

3. Per una governance responsabile degli LLM

L'uso degli LLM nella ricerca scientifica comporta un rischio sistemico sottile: l'illusione di comprensione. Ricevere risposte fluide e persuasive può far credere di aver capito davvero un problema, senza che questo emerga nel risultato finale. È quindi fondamentale distinguere tra un uso consapevole — per apprendere, verificare e chiarire — e un uso delegante, privo di reale comprensione. Quattro principi definiscono le condizioni di un'adozione responsabile; a ciascuno corrisponde un insieme di verifiche operative.

Provenance e versionamento. Ogni interazione rilevante con un LLM dovrebbe essere registrata: prompt, input forniti, versione del modello, data, set documentale di riferimento, test e verifiche eseguiti. Prima di consegnare un elaborato o pubblicare un risultato, occorre poter rispondere alla domanda «quale modello, con quale input, in quale data ha prodotto questo output?».

Separazione tra bozza e risultato. L'LLM può proporre alternative, accelerare il lavoro preparatorio e chiarire passaggi intermedi, ma il passaggio da output generato a risultato scientifico richiede prove esterne e validazione indipendente. Ogni affermazione rilevante deve essere riconducibile a una fonte primaria consultabile; un'affermazione non verificabile non può essere presentata come dato di fatto. Il codice generato deve essere testato, analizzato staticamente e revisionato prima dell'uso.

Disclosure e responsabilità. Il modello non è autore e la responsabilità intellettuale resta umana; l'uso dell'AI va dichiarato dove richiesto da riviste, atenei, corsi o processi editoriali. Prima di sottomettere un contributo occorre verificare le policy specifiche della sede di pubblicazione o del corso.

Protezione dei materiali riservati. Manoscritti non pubblicati, dati sensibili e revisioni confidenziali non devono essere caricati su strumenti non approvati dall'istituzione. Occorre classificare i dati prima di qualsiasi interazione con il modello e verificare che lo strumento utilizzato sia approvato per quel livello di sensibilità.

Il primo è la provenance e il versionamento: ogni interazione rilevante con un LLM dovrebbe essere registrata, includendo prompt, input forniti, versione del modello, data, set documentale di riferimento, test e verifiche eseguiti. Il secondo è la separazione tra bozza e risultato: l'LLM può proporre alternative, accelerare il lavoro preparatorio e chiarire passaggi intermedi, ma il passaggio da output generato a risultato scientifico richiede prove esterne e validazione indipendente. Il terzo controllo riguarda disclosure e responsabilità: il modello non è autore e la responsabilità intellettuale resta umana; l'uso dell'AI va dichiarato dove richiesto da riviste, atenei, corsi o processi editoriali. Il quarto è la protezione dei materiali

riservati: manoscritti non pubblicati, dati sensibili e revisioni confidenziali non devono essere caricati su strumenti non approvati dall'istituzione.

Per atenei e gruppi di ricerca, le scelte estreme, ovvero divieto assoluto o adozione senza criteri, comportano entrambe rischi significativi. Il divieto assoluto tende a spostare l'uso degli LLM verso pratiche informali e non supervisionate, riducendo la capacità dell'istituzione di definire standard condivisi. L'adozione indiscriminata espone dati, processi e risultati a vulnerabilità che spesso emergono solo a posteriori. Una strategia più realistica si articola su tre livelli: un primo livello di uso assistivo a basso rischio, come revisione di testi, sintesi di fonti verificabili e documentazione non sensibile; un secondo livello di integrazioni controllate, ad esempio sistemi RAG su archivi curati con privilegi minimi, da coordinare a livello di gruppo o dipartimento; un terzo livello di adozione strutturata, con logging standardizzato, verifica continua delle prestazioni e governance istituzionale.

In questo quadro, il riferimento normativo europeo è l'AI Act, Regolamento UE 2024/1689 [7], che fornisce il contesto per la classificazione del rischio, il procurement e gli obblighi di trasparenza. Le linee guida UNESCO per l'istruzione superiore [6] offrono inoltre indicazioni operative su uso, disclosure e protezione dei dati. A livello europeo, le linee guida "living" della Commissione Europea sull'uso responsabile dell'IA generativa nella ricerca [9] — elaborate nell'ambito dell'ERA Forum e aggiornate alla terza versione nel maggio 2026 — forniscono raccomandazioni operative specificamente rivolte a ricercatori, enti di ricerca e organismi di finanziamento, articolate attorno ai principi di affidabilità, onestà, rispetto e responsabilità.

4. Conclusioni

L'integrazione degli LLM nella ricerca scientifica non è un semplice aggiornamento strumentale: è una trasformazione socio-tecnica che ridefinisce i costi del lavoro cognitivo, i criteri di attribuzione della responsabilità intellettuale e le superfici di rischio. Questi strumenti producono valore concreto quando accelerano compiti ripetitivi, onerosi o preparatori; diventano critici quando sostituiscono attività che richiedono validazione umana, giudizio esperto o verifica delle fonti.

Il caso Parisi-Zamponi offre in questo senso una bussola. Non dimostra che gli LLM possono sostituire il ricercatore, né che debbano essere tenuti ai margini del processo scientifico. Mostra invece che un risultato ottenuto con il contributo sostanziale di un modello è scientificamente accettabile solo se il percorso che lo ha prodotto rimane tracciabile,

verificabile e attribuibile a responsabilità umane definite. È questa condizione — non la potenza del modello, non la fluidità dell'output — il criterio discriminante tra un'adozione matura e una rischiosa.

La differenza, in ultima analisi, non dipende principalmente dallo strumento scelto, ma dalla solidità del workflow che lo circonda: fonti primarie affidabili, procedure di validazione, revisione tra pari, tracciabilità e responsabilità esplicite.

Gli LLM evolveranno rapidamente. La responsabilità ultima, però, resterà umana — e sarà tanto più efficace quanto prima le istituzioni scientifiche avranno elaborato politiche chiare, condivise e verificabili.

Riferimenti

1. NIST AI 600-1 (2024). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. National Institute of Standards and Technology. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf?utm_source=chatgpt.com
2. Messeri L., Crockett M.J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627, 49–58. https://www.nature.com/articles/s41586-024-07146-0
3. Luo Z. et al. (2025). LLM4SR: A Survey on Large Language Models for Scientific Research. https://arxiv.org/pdf/2501.04306?utm_source=chatgpt.com
4. Perry N. et al. (2023). Do Users Write More Insecure Code with AI Assistants? *ACM CCS '23*. arXiv:2211.03622
5. Spracklen J. et al. (2025). We Have a Package for You! A Comprehensive Analysis of Package Hallucinations by Code Generating LLMs. *USENIX Security '25*. https://www.usenix.org/conference/usenixsecurity25/presentation/spracklen?utm_source=chatgpt.com
6. UNESCO (2023). *Guidance for Generative AI in Education and Research*. https://unesco.org.uk/site/assets/files/10375/guidance_for_generative_ai_in_education_and_research.pdf?utm_source=chatgpt.com
7. Unione Europea (2024). Regolamento (UE) 2024/1689 — AI Act. eur-lex.europa.eu

8. OWASP GenAI Security Project (2025). *OWASP Top 10 for Large Language Model Applications (v2025)*. https://owasp.org/www-project-top-10-for-large-language-model-applications/?utm_source=chatgpt.com

9. Commissione Europea (2026). *Living guidelines on the responsible use of generative AI in research*. ERA Forum Stakeholders' document, Third version, May 2026. https://research-and-innovation.ec.europa.eu/

10. Giorgio Parisi e Francesco Zamponi (2026). *A proof of an identity for the critical exponents of jamming*. https://arxiv.org/abs/2606.03300.