«History of Education & Children's Literature», XX, 1 (2025), pp. 537-550
ISSN 1971-1093 (print) / ISSN 1971-1131 (online) / eum (Edizioni Università di Macerata, Italy) 2025

Intelligenza artificiale generativa e fonti storico-educative: prospettive metodologiche

Florindo Palladino
Department of Humanities, Social Sciences and Education
University of Molise
Campobasso (Italy)
florindo.palladino@unimol.it

Generative Artificial Intelligence and Historical-Educational Sources: Methodological Perspectives

Abstract: The paper examines the methodological potential of generative artificial intelligence in historical-educational research, addressing a subject that has so far remained unexplored. After outlining the functioning of Large Language Models (LLM) and Retrieval Augmented Generation (RAG) systems, it analyses their most promising applications, including the automatic transcription of manuscripts, the translation of texts in ancient languages, and the processing of extensive documentary corpora. Through case studies, the research demonstrates how RAG architectures can effectively overcome the limitations of LLMs in analysing large collections of historical sources. Finally, a structured methodological framework is proposed to integrate these technologies into historical-educational research, establishing an operational protocol for documentary analysis.

Eet/Tee keywords: Generative Artificial Intelligence; Large Language Models; Retrieval Augmented Generation; History of Education; Historical Research Methodology.

Introduzione

L'intelligenza artificiale generativa, fondata sui *Large Language Models* (modelli di linguaggio di ampie dimensioni, d'ora in poi LLM), costituisce un ambito di innovazione tecnologica che pone sfide rilevanti agli storici. Nel panorama internazionale, i primi tentativi di analisi del rapporto tra intelligenza artificiale e ricerca storica sono emersi in occasione dei panel tematici presentati durante il 137° convegno annuale dell'*American Historical Association*

(gennaio 2024)[1]. I panel sono stati preceduti da un forum ospitato dalla rivista *American Historical Review*, intitolato *Artificial Intelligence and the Practice of History*. L'iniziativa, concepita con l'intento di offrire uno spazio di riflessione critica sulle implicazioni delle tecnologie emergenti, prendeva spunto dalla constatazione che, con la progressiva digitalizzazione del patrimonio documentale, «gli algoritmi di apprendimento automatico diventeranno strumenti essenziali per la ricerca storica, influenzando non solo l'attività sociale, ma anche il processo di costruzione della conoscenza storica»[2].

Le sfide poste dalle tecnologie emergenti non sono quindi di natura semplicemente tecnica, ma investono: il piano epistemologico (in che modo si ridefiniscono i processi di costruzione e di validazione della conoscenza storica nell'era dell'IA); il piano metodologico (come integrare i nuovi strumenti nei protocolli di ricerca storica) e quello etico (qual è il ruolo dello storico e la sua responsabilità nell'uso di tali strumenti).

Nell'ambito specifico della *History of Education* si osserva, tuttavia, una marcata lacuna, come attestano le ultime edizioni della *International Standing Conference for the History of Education* (ISCHE), dove mancano contributi dedicati a questo tema. Eppure la digitalizzazione, che sta generando una trasformazione rilevante nelle modalità di accesso ai documenti, riguarda anche l'ambito storico-educativo[3]. Dal punto di vista della pratica storica, i modelli di linguaggio potrebbero costituire strumenti utili, permettendo allo storico dell'educazione di analizzare vaste collezioni di fonti ricorrendo a interrogazioni nel linguaggio naturale.

Il presente contributo si concentrerà, quindi, sull'aspetto esclusivamente metodologico, illustrando possibili modalità per integrare le nuove tecnologie di intelligenza artificiale generativa nei protocolli di ricerca storico-educativa, ponendo al contempo in rilievo la questione cruciale della loro affidabilità. L'intento è quello di illustrare soltanto le procedure metodologiche replicabili dagli storici dell'educazione, evitando intenzionalmente di presentare quelle

[1] Cfr. *137th Annual Meeting Program*, consultabile sul sito web del convegno al seguente indirizzo: <https://www.historians.org/past-meeting/137th-annual-meeting/> (ultimo accesso: 28.02.2025).

[2] R.D. Meadows, J. Sternfeld, *Artificial Intelligence and the Practice of History: A Forum*, «The American Historical Review», vol. 128, n. 3, 2023, pp. 1345-1349.

[3] Basti qui ricordare il progetto *Google Books*. Il colosso di Mountain View ha stipulato una serie di accordi con alcune delle biblioteche più prestigiose al mondo per digitalizzare le loro collezioni librarie e renderle disponibili online, trasformando *Google Books* nella più imponente biblioteca digitale globale, con milioni di testi accessibili sulla sua piattaforma. Per il contesto italiano, merita particolare attenzione l'accordo siglato nel marzo 2010 tra il Ministero per i Beni e le Attività Culturali (MiBAC) e Google, che prevedeva la digitalizzazione di un milione di volumi di alto valore storico conservati presso le Biblioteche Nazionali Centrali di Roma e Firenze. Per una ricostruzione storica documentata e un'analisi critica accurata sul progetto *Google Book* si rimanda a: A. Jacquesson, *Google livres et le futur des bibliothèques numériques*, Paris, Éditions du cercle de la librairie, 2010.

tecniche che richiederebbero, ad oggi, risorse computazionali e competenze informatiche avanzate[4].

Dopo un'introduzione al funzionamento dei *Large Language Models* e dei sistemi di *Retrieval Augmented Generation* (Generazione aumenta dal recupero, d'ora in poi RAG) – focalizzata sugli aspetti funzionali e strutturali maggiormente rilevanti per l'analisi delle fonti storiche – passeremo in rassegna gli studi pionieristici che hanno incorporato tali tecnologie nei protocolli di indagine storica, ricavandone indicazioni rilevanti. In particolare, esamineremo applicazioni degli LLM in due ambiti chiave anche per lo storico dell'educazione: la trascrizione automatica di fonti manoscritte e la traduzione di documenti redatti in lingue antiche. Successivamente, discuteremo il grado di conoscenza storica implicita negli LLM, ossia 'quanto sanno' di storia sulla base del loro addestramento. Nella parte seguente, analizzeremo invece gli approcci basati sulla Generazione aumentata dal recupero, concepita per mitigare alcuni limiti intrinseci dei modelli linguistici, illustrandone l'utilizzo per l'analisi di vasti corpora documentali. Infine, nella sezione conclusiva, proporremo un quadro metodologico per l'impiego consapevole e proficuo dell'IA generativa nello studio delle fonti storico-educative.

1. *I Large Language Models*

Gli LLM rappresentano una delle innovazioni più rilevanti nel campo del *Natural Language Processing*, ramo dell'informatica che studia l'interazione tra linguaggio umano e sistemi computazionali. Tali modelli si distinguono per la capacità di elaborare e generare testo in linguaggio naturale con un livello di coerenza e pertinenza contestuale senza precedenti[5].

Seguendo Cristianini, per comprendere il loro funzionamento è utile distinguere tre livelli concettuali: l'*agente* con cui interagiamo (ad esempio, il noto ChatGPT di OpenAI); il *modello di linguaggio* sottostante, definito LLM, che per il modello di OpenAI è attualmente il GPT-4.5 (*Generative Pre-trained*

[4] Ci riferiamo, in particolare, alle tecniche di *Fine-tuning* che, pur dimostrando notevole efficacia nell'ambito degli studi storici, richiedono la costituzione di team interdisciplinari composti da storici e informatici. Nonostante l'interdisciplinarietà sia il paradigma metodologico dominante nel contesto dell'intelligenza artificiale applicata alle discipline umanistiche, riteniamo opportuno che lo storico sviluppi competenze specifiche per l'utilizzo di tali tecnologie, una posizione apparentemente controcorrente rispetto alla tendenza attuale, ma che si fonda su precise ragioni metodologiche che saranno articolate nella sezione conclusiva del presente contributo.

[5] Per una introduzione che esplora le fondamenta matematiche dell'intelligenza artificiale, in una prospettiva storica, rimandiamo al volume di A. Ananthaswamy, *Perché le macchine imparano. L'eleganza della matematica dietro all'AI*, Milano, Apogeo, 2024.

Transformer, versione 4.5); e l'*algoritmo* che crea il modello a partire da grandi volumi di dati testuali (il Transformer)[6].

Alla base degli LLM vi è un modello probabilistico del linguaggio, che apprende regolarità e correlazioni statistiche dai testi su cui addestrato. Durante la fase di addestramento, i modelli apprendono a predire l'elemento linguistico successivo (il *token*) in una sequenza testuale, sviluppando progressivamente una 'comprensione' implicita delle regole grammaticali e sintattiche e delle relazioni semantiche. Il processo, definito *pre-training*, avviene in modo autosupervisionato, ossia senza intervento umano diretto, consentendo al modello di acquisire 'conoscenze' in modo autonomo. In una successiva fase, definita di *fine-tuning* (ottimizzazione), il modello viene specializzato su dataset più mirati e con obiettivi specifici, perfezionando ulteriormente le sue capacità.

Ciò che ha colto di sorpresa la comunità scientifica degli informatici è costatare come questi modelli, pur essendo stati progettati primariamente per prevedere il *token*[7] successivo in una sequenza linguistica, manifestano inaspettate capacità in ambiti quali «il ragionamento logico, l'espressione creativa e la capacità deduttiva», unitamente a una conoscenza enciclopedica che spazia dalla «letteratura alla medicina fino alla programmazione informatica»[8].

Recentemente, si è anche assistito all'evoluzione dai modelli esclusivamente testuali a modelli in grado di elaborare e generare contenuti in diverse modalità: testo, immagini, audio e video, denominati funzionalmente *Large Multimodal Models*.

[6] Cfr. N. Cristianini, *Machina sapiens. L'algoritmo che ci ha rubato il segreto della conoscenza*, Bologna, il Mulino, 2024. L'architettura Transformer, presentata nel 2017 da un team di ricercatori di Google, ha rivoluzionando il campo dell'elaborazione del linguaggio naturale e creato le basi per tutti i moderni LLM. Il suo meccanismo di auto-attenzione consente ai modelli di analizzare simultaneamente l'intero testo, assegnando un peso differenziato a ciascuna parola in relazione al contesto. L'approccio supera i limiti delle precedenti architetture neurali, migliorando l'abilità di catturare relazioni semantiche anche tra parole distanti. L'articolo seminale che descrive il Transformer è: A. Vaswani, N. Shazeer, N. Parmar *et alii*, *Attention is all you need*, «Advances in Neural Information Processing Systems», vol. 30, 2017, pp. 5998-6008.

[7] Il termine *token* indica l'unità minima di testo elaborata da un modello di linguaggio. Un token può corrispondere a una parola intera, a una parte di parola o persino a un singolo carattere, a seconda della segmentazione adottata dal modello. È stato calcolato che 100 token corrispondono approssimativamente a 75 parole. Per essere processati, i token vengono convertiti in rappresentazioni numeriche attraverso un processo chiamato *embedding*, che associa a ciascun token un vettore in uno spazio matematico ad alta dimensione, una trasformazione che consente al modello di catturare relazioni semantiche e sintattiche tra le parole e di generare risposte coerenti con il contesto.

[8] Cfr. S. Bubeck, V. Chadrasekaran, R. Eldan *et alii*, *Sparks of artificial general intelligence: Early experiments with gpt-4*, 2023. Il noto documento redatto dai ricercatori di Microsoft e disponibile in rete, prende in rassegna le capacità del modello di Openai GPT-4, in grado di svolgere con successo compiti nuovi e complessi e manifestare ampie conoscenze in diversi domini. Tale comportamento, caratterizzato da un salto qualitativo improvviso che emerge solo superando una determinata soglia di scala (in termini di parametri o quantità di addestramento), rappresenta quelle che vengono definite in letteratura come «capacità emergenti».

Parallelamente, sono state potenziate le capacità di ragionamento degli LLM, attraverso tecniche specifiche in fase di addestramento, che hanno dato origine a modelli ottimizzati per affrontare compiti di ragionamento complesso, i cosiddetti *Reasoning Models*. Infine, sono comparsi i modelli agentici, sistemi basati sugli LLM e progettati per agire autonomamente in ambienti digitali, eseguire sequenze di azioni e perseguire obiettivi complessi con supervisione minima[9].

Nonostante queste impressionanti capacità ed evoluzioni, gli LLM presentano limiti strutturali rilevanti per la ricerca storica. Innanzitutto l'affidabilità: i modelli linguistici possono generare affermazioni plausibili ma non necessariamente verificate, il che è particolarmente critico dove è imprescindibile l'accuratezza, come negli studi storici. Inoltre, la cosiddetta finestra di contesto (C*ontext Window*), vincola il numero di token elaborabili simulatamente, ostacolando l'analisi di vaste collezioni documentarie[10]. Infine, gli LLM non sono nativamente progettati per il recupero di informazioni specifiche da grandi raccolte i dati, bensì produrre testo coerente basato sull'input ricevuto.

Un approccio promettente per mitigare queste limitazioni è offerto dall'architettura *Retrieval Augmented Generation*, attualmente considerata una delle soluzioni più efficaci anche per l'analisi di corpora documentali.

2. *I sistemi di Retrieval Augmented Generation*

I sistemi RAG rappresentano un'importante innovazione nell'ambito dell'intelligenza artificiale generativa, poiché consentono di ovviare a diversi limiti insiti negli LLM. Sono, infatti, progettati per integrare un meccanismo di recupero di informazioni da fonti esterne (*knowledge base*), consentendo agli LLM di accedere a documenti pertinenti per generare risposte più affidabili e contestualizzate[11].

[9] Un esempio autorevole di modello agentico, progettato per la ricerca scientifica da Google DeepMind, è rappresentato da *AI Co-Scientist*, in grado di assistere i ricercatori di ambito biomedico in tutte le fasi della ricerca (J. Gottweis, W.H. Weng, A. Daryin *et alii*, *Towards an AI co-scientist*, 2025, <https://arxiv.org/abs/2502.18864> (ultimo accesso: 12.02.2025)).

[10] La finestra di contesto nei modelli di linguaggio indica il numero massimo di token elaborabili in un singolo prompt o durante una sessione interattiva. Essa definisce essenzialmente la quantità di testo che il modello può 'ricordare' e analizzare contemporaneamente per produrre risposte coerenti e contestualmente rilevanti. Le dimensioni di queste finestre variano significativamente tra i diversi modelli. GPT-4, per esempio, supporta fino a 128.000 token (equivalenti a circa 96.000 parole), mentre Claude 3.5 Sonnet di Anthropic gestisce fino a 200.000 token. Il modello Gemini 2.0 di Google, riesce a processare fino a 2 milioni di token. Per offrire un paragone concreto, considerando che la rivista in cui appare questo articolo contiene mediamente 400 parole per pagina, la finestra di contesto di GPT-4 equivale a 240 pagine, quella di Claude 3.5 a 375 pagine, mentre Gemini 2.0 può elaborare l'equivalente di 3.750 pagine.

[11] La RAG è stata introdotta nel 2020, e presentata come un'architettura che migliora significativamente le prestazioni degli LLM in compiti che richiedono conoscenze fattuali precise (P.

Nel contesto della ricerca storico-educativa, l'architettura RAG risulta particolarmente vantaggiosa per l'analisi di vaste collezioni di fonti. Le raccolte di vaste dimensioni e continuità temporale – ad esempio, la collezione completa di un giornale scolastico, la serie di cataloghi di un editore scolastico, i corpora normativi, le collezioni di manualistica scolastica, e così via – rappresentano risorse di straordinario valore ma pongono significative sfide metodologiche per la loro analisi esaustiva. La mole supera non solo le possibilità di un'analisi manuale sistematica, ma anche le capacità di elaborazione tramite gli LLM, che sono vincolati da una finestra di contesto limitata che non consente di processare simultaneamente un intero corpus di fonti, ma, soprattutto, limitati dal fatto che non sono progettati per «recuperare informazioni», ossia non sono sistemi di *Information Retrieval*.

I sistemi RAG superano tali limitazioni, permettendo di recuperare selettivamente le informazioni più pertinenti a ogni specifica domanda posta dall'utente, che verranno poi elaborate da un LLM.

Nello specifico, i sistemi RAG integrano in un unico processo tre componenti fondamentali che lavorano in sequenza:

1) Il sistema inizia con il recupero (*Retrieval*) delle informazioni più pertinenti dai documenti forniti dall'utente. Quando lo storico pone una domanda, questa componente cerca nei documenti disponibili le informazioni più rilevanti per rispondere[12].
2) Segue il meccanismo di incremento (*Augmentation*) che arricchisce la domanda posta dall'utente con le informazioni recuperate. Questa fase combina i documenti selezionati con la domanda iniziale, creando un «prompt arricchito» che contiene sia la richiesta che le informazioni contestuali utili.
3) Infine, la *generazione* della risposta avviene tramite un LLM, che crea contenuti basandosi sul «prompt arricchito», producendo una risposta informata sulla base delle informazioni recuperate dai documenti esterni.

L'intera sequenza garantisce che le risposte generate siano non solo pertinenti alla domanda, ma anche verificabili attraverso i documenti di riferimento. I documenti citati vengono infatti richiamati a supporto delle affermazioni generate dal modello linguistico, un aspetto metodologicamente decisivo in ambito storico, dove l'accuratezza e la tracciabilità della fonte sono requisiti

Lewis, E. Perez, A. Piktus *et alii*, *Retrieval-augmented generation for knowledge-intensive NLP tasks*, «Advances in Neural Information Processing Systems», vol. 33, 2020, pp. 9459-9474).

12 Nel processo di recupero, un modello di embedding trasforma sia la query che i documenti in rappresentazioni vettoriali multidimensionali che catturano il significato semantico del testo. Il sistema calcola matematicamente la similarità coseno tra il vettore della query e i vettori dei documenti precedentemente indicizzati nel corpus. I documenti i cui vettori presentano la maggiore similarità con il vettore della query vengono identificati come i più rilevanti. Questo processo di matching vettoriale permette di superare le limitazioni delle ricerche basate su semplice corrispondenza di parole chiave

fondamentali. La figura 1 illustra parte del risultato finale del processo, utilizzando come fonte documentale 34 annate del periodico *I diritti della scuola*, relative al periodo compreso tra 1899 e il 1940, accessibili e scaricabili tramite l'Emeroteca virtuale della Biblioteca Nazionale di Roma. La domanda posta è relativa all'analisi degli articoli che trattano della disciplina scolastica; la risposta data dal modello linguistico è corredata dai riferimenti alle fonti originarie.

Pic. 1. Esempio di risposta generata dal modello linguistico con riferimenti alle fonti

Analogamente agli LLM, anche le architetture RAG negli ultimi tempi hanno registrato una rapida evoluzione: dalle prime implementazioni con architetture relativamente semplici, si è passati a sistemi più avanzati che ottimizzano sia la fase di recupero delle informazioni sia la loro integrazione nel processo generativo. Allo stato dell'arte, si sperimentano sistemi RAG modulari, che offrono componenti specializzati configurabili per adattarsi a esigenze specifiche[13]. Tale flessibilità risulta particolarmente preziosa nell'ambito della ricerca storico-educativa, dove le fonti variano notevolmente per tipologia e struttura.

3. *Applicazioni dei Large Language Models nella ricerca storica*

Le sperimentazioni sull'uso degli LLM in ambito storico si stanno moltiplicando, rivelando interessanti potenzialità. I casi d'uso che seguono mostrano come, anche senza architettura RAG, gli LLM possano offrire contributi rile-

[13] Y. Gao, Y. Xiong, X. Gao *et alii*, *Retrieval-augmented generation for large language models: A survey*, 2023, <https://arxiv.org/abs/2312.10997> (ultimo accesso: 12.02.2025). Lo studio di Gao e colleghi propone una tassonomia dei sistemi RAG, classificandoli in Naive, Advanced e Modulare, e analizza le tecniche per ciascuna categoria.

vanti anche per la ricerca storico-educativa, in particolare per la trascrizione di documenti manoscritti e per la traduzione di testi redatti in lingue antiche.

3.1. *Trascrizione di fonti manoscritte*

L'evoluzione degli LLM verso capacità multimodali ha aperto nuove prospettive per la trascrizione di fonti manoscritte, un aspetto di straordinaria importanza per la ricerca storica e storico-educativa. L'introduzione dei modelli linguistici multimodali come GPT-4, Claude Sonnet 3.5 e Gemini 1.5, avvenuta nel corso del 2023, ha segnato un punto di svolta, consentendo di superare molte delle limitazioni dei tradizionali sistemi di conversione del testo scritto a mano.

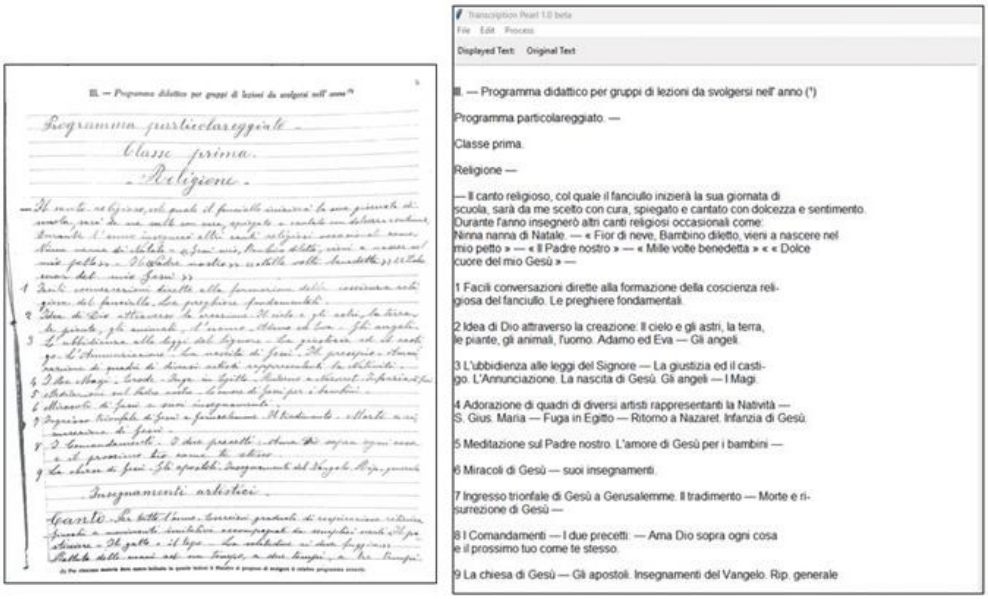
A differenza dei sistemi di *Handwritten Text Recognition* (HTR), che necessitano spesso di estensive fasi di pre-elaborazione delle immagini e addestramenti specifici per singoli stili di scrittura, i modelli linguistici multimodali possono trascrivere accuratamente documenti manoscritti senza necessità di addestramento aggiuntivo, grazie alla loro capacità di integrare l'analisi visiva dei caratteri con una comprensione del contesto linguistico, caratteristica particolarmente preziosa per lo storico dell'educazione, che lavora abitualmente su fonti manoscritte brevi, spesso molto diverse tra loro per tipo di grafia e condizioni materiali di conservazione.

Un contributo metodologico di rilievo è stato presentato dagli storici dell'università canadese "Wilfried Laurier", che hanno proposto un sistema bifasico di trascrizione[14], che impiega due distinti modelli linguistici: nella prima fase, un LLM multimodale genera una trascrizione preliminare del documento, mentre nella seconda fase un modello distinto analizza simultaneamente il manoscritto originale e la trascrizione iniziale, procedendo a un processo di correzione e affinamento. L'adozione di due modelli separati si è rivelata metodologicamente cruciale, poiché i tentativi di «auto-correzione» effettuati mediante il medesimo modello non hanno prodotto miglioramenti significativi.

La trascrizione effettuata su un corpus di documenti in lingua inglese datati

[14] M. Humphries, L.C. Leddy, Q. Downton *et alii*, *Unlocking the Archives: Large Language Models Achieve State-of-the-Art Performance on the Transcription of Handwritten Historical Documents*, «Digital Scholarship in the Humanities», vol. 39, n. 1, 2024, pp. 78-96. Humphries e colleghi offrono una descrizione approfondita del sistema bifasico, illustrandone chiaramente il fondamento teorico e l'efficacia dimostrata dai risultati empirici. Particolarmente utile è la descrizione puntuale della procedura di prompting adottata per massimizzare l'accuratezza degli LLM multimodali nella trascrizione e correzione di testi storici manoscritti. Il contributo è ulteriormente valorizzato dalla condivisione open-source del software sviluppato dai ricercatori, denominato *Transcription Pearl*, che include un'interfaccia intuitiva e impostazioni facilmente configurabili dagli utenti.

tra il 1761 e il 1827, caratterizzati da 33 differenti grafie, ha raggiunto livelli di accuratezza prossimi a quelli della trascrizione umana esperta, superando le prestazioni di noti software HTR, come Transkribus, dimostrando il potenziale di questa metodologia per l'analisi delle fonti manoscritte su larga scala e a costi contenuti. La figura 2 mostra un esempio di trascrizione automatica di una pagina di registro scolastico italiano degli anni Trenta del Novecento, ottenuta mediante il software sviluppato dagli storici canadesi.



Pic. 2. Esempio di trascrizione automatica effettuata con il software *Transcription Pearl*

3.2. *Traduzione di documenti redatti in lingue antiche*

Sul versante della traduzione di fonti storiche, un recente studio condotto presso l'Università di Zurigo documenta l'applicazione degli LLM utilizzati come strumenti di traduzione di un corpus documentario del XVI secolo, costituito da oltre 3.000 lettere in *Early New High German*, una variante storica della lingua tedesca diffusa tra il 1350 e il 1650[15].

La procedura proposta dagli autori non prevede fasi di addestramento specifico dei modelli (*fine tuning*), generalmente complesse per uno storico che

15 M. Volk, D.P. Fischer, P. Scheurer, R. Schwitter, P.B. Ströbel, *LLM-based Translation Across 500 Years. The Case for Early New High German*, in *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, Vienna, Association for Computational Linguistics, 2024, pp. 368-375.

non dispone di un supporto informatico specializzato. Piuttosto, gli autori dello studio hanno utilizzato i modelli di linguaggio (GPT-4 e Gemini) nella loro configurazione standard, con una procedura basata su un'accurata strutturazione dei *prompt* (l'input testuale fornito al modello linguistico), arricchiti con informazioni lessicali.

I risultati della ricerca confermano la maggiore efficacia degli LLM multimodali di gestire variazioni lessicali delle lingue storiche, rispetto ai sistemi di traduzione automatica come DeepL e Google Translate, determinata soprattutto dalla capacità degli LLM di sfruttare due livelli distinti di informazioni contestuali: quello esplicitamente fornito al modello tramite prompt arricchiti con informazioni puntuali, e quello implicito nel testo stesso, derivante dalla comprensione profonda del contesto linguistico e semantico proprio del tedesco protomoderno.

4. *La valutazione del grado di conoscenza storica dei Large Language Models*

Valutare il grado di conoscenza storica posseduto dai modelli linguistici non è una semplice curiosità, ma una reale esigenza metodologica. In questa direzione si inserisce il recente studio realizzato da un gruppo interdisciplinare composto da storici e informatici dell'Università di Poitiers (Francia), che ha analizzato in modo sistematico le capacità e, soprattutto, i limiti attuali degli LLM nell'ambito della conoscenza storica[16]. Focalizzata sul periodo moderno e sul territorio del Poitou – antica provincia francese –, la ricerca rappresenta uno dei primi tentativi di verifica empirica dell'affidabilità di tali modelli, in totale 14, nella restituzione accurata di fatti ed eventi.

La metodologia di valutazione si è basata su un rigido sistema di valutazione, articolato in 62 quesiti storici declinati in 268 query, equamente ripartite tra formulazioni in linguaggio naturale e in forma di parole chiave. Un aspetto particolarmente rilevante è l'impiego di 14 diversi LLM in configurazione standard, facendo leva esclusivamente sulla loro conoscenza preesistente.

I risultati sono piuttosto critici: l'accuratezza media si è fermata 37,6%, evidenziando limitazioni sostanziali nell'affidabilità dei modelli[17]. È altresì rilevate l'analisi differenziata per domini tematici, dalla quale emerge che, an-

[16] M. Chartier, N. Dakkoune, G. Bourgeois, S. Jean, *HiBenchLLM: Historical Inquiry Benchmarking for Large Language Models*, «Data & Knowledge Engineering», vol. 156, 2025, p. 102383.

[17] Emergono tuttavia significative differenze tra i diversi modelli: Gemini raggiunge il valore più elevato (70,34%), seguito da Copilot (56,53%), ChatGPT con GPT-4 (53,54%) e GPT-3.5-Turbo (50,93%), mentre altri modelli mostrano prestazioni nettamente inferiori, fino al minimo del 4,10% registrato da Falcon.

che in presenza di eventi storici ben documentati, l'accuratezza non supera il 40,9%, invece per argomenti più complessi si abbassa ulteriormente al 23,2%. La ricerca, adottando una metodologia rigorosa, dimostra che le prestazioni dei modelli calano sensibilmente all'aumentare della complessità.

Da un punto di vista metodologico, questi dati suggeriscono la necessità di un approccio estremamente prudente nell'impiego degli strumenti di intelligenza artificiale generativa nella ricerca storica, soprattutto in ambiti specialistici come la *History of Education*. In questo settore disciplinare, le limitazioni riscontrate potrebbero risultare ancora più marcate, a causa di un probabile squilibrio nei dati di addestramento, in cui la storia dell'educazione è di certo sottorappresentata rispetto ad altri ambiti della ricerca, come ad esempio la storia della letteratura.

Le prospettive future indicate dagli autori dello studio fanno riferimento alla necessità di sviluppare metodologie più raffinate per integrare i modelli di linguaggio nei processi di indagine stoica. Tale orientamento introduce la sezione successiva, dedicata all'integrazione degli LLM in architetture RAG, approccio che, anziché affidarsi unicamente alla conoscenza interna dei modelli, la arricchisce attraverso l'accesso diretto a fonti documentali, offrendo una possibile risposta agli evidenti limiti di accuratezza.

5. Applicazioni di sistemi di Retrieval Augmented Generation nella ricerca storica

La ricerca di Tran, González-Gallardo e Doucet si colloca nel filone di studi volto a valutare l'accuratezza di un sistema RAG, in particolare di una versione progettata per l'analisi di collezioni di stampa periodica[18].

Per testare l'efficacia dell'architettura RAG progettata, il team di ricerca ha utilizzato un corpus particolarmente impegnativo: 4.836 articoli tratti da periodici francesi, finlandesi e tedeschi del XIX e inizio XX secolo, su temi ricorrenti e tra loro correlati. La scelta ha permesso di porre domande storiche complesse, così da verificare la capacità del sistema di recuperare e integrare

[18] T.T. Tran, C.E. González-Gallardo, A. Doucet, *Retrieval Augmented Generation for Historical Newspapers*, in *ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, Hong Kong, 2024. Il lavoro presenta un'implementazione completa di un sistema RAG per l'analisi di periodici storici. Dal punto di vista tecnico, il sistema utilizza un'architettura sofisticata che combina un modulo di recupero semantico basato sul modello multilingue E5, un sistema di reranking che integra il punteggio Cohere con un'analisi delle entità nominate, e un modello generativo (LLaMA3) ottimizzato per la sintesi di informazioni storiche. Questa architettura consente di superare i limiti dei sistemi tradizionali di Named Entity Recognition, fornendo risultati più accurati anche in presenza di variazioni ortografiche e errori di trascrizione tipici dei documenti storici digitalizzati.

contenuti coerenti in presenza di diversità linguistica e stratificazione temporale.

I risultati sono stati valutati con un insieme integrato di metriche, sia quantitative e sia qualitative. Le prime hanno misurato in modo oggettive la pertinenza e l'accuratezza delle sintesi generate, evidenziando un buon recupero delle informazioni; le seconde hanno approfondito la capacità del sistema di gestire le peculiarità dei documenti storici, che si è mostrato capace di salvaguardare la coerenza semantica e la fedeltà alle fonti originali.

Restando nell'ambito della RAG, il contributo di Garcia e Weilbach, dal titolo suggestivo *Se le fonti potessero parlare*, approfondisce le potenzialità di un'architettura ottimizzata per due scopi principali: da un lato, valutare la capacità di un sistema RAG, debitamente progettato, di integrare il recupero di informazioni con l'elaborazione testuale per rispondere a interrogativi storici complessi di natura argomentativa, descrittiva e 'integrativa'; dall'altro, testare la sua efficacia nell'estrazione e nella strutturazione di dati fattuali[19].

Sul primo versante, gli autori hanno selezionato ottantasei monografie accademiche, di rilevanza consolidata in ambiti di ricerca inerenti le migrazione irlandesi, la schiavitù e gli studi cubani. Il corpus è stato impiegato per verificare le capacità del sistema di produrre sintesi storiografiche, elaborare interpretazioni e fornire risposte coerenti con i paradigmi della disciplina storica.

Il secondo ambito di ricerca ha esaminato la capacità del sistema di estrarre e organizzare informazioni strutturate da fonti primarie, focalizzandosi sui nove volumi della *Historia de Familias Cubanas* di Francisco Javier de Santa Cruz y Mallen, in particolare per testare la capacità di recupero di dati anagrafici, relazioni di parentela e cronologie, elementi fondamentali per la ricostruzione delle dinamiche migratorie e sociali.

I risultati presentati dagli autori dello studio confermano il potenziale dell'architettura RAG nell'ambito della ricerca storica. Sul versante dell'analisi di interrogativi complessi relativi a fonti secondarie, il sistema ha restituito risposte con coerenza interpretativa e profondità argomentativa; sull'altro versante, il sistema ha dimostrato una notevole efficacia nel recupero e nella categorizzazione dei dati genealogici, automatizzando in parte operazioni che tradizionalmente richiedono un'ampia mole di lavoro manuale.

[19] G. Gonzalez Garcia, C. Weilbach, *If the Sources Could Talk: Evaluating Large Language Models for Research Assistance in History*, 2023, <https://arxiv.org/abs/2310.10808> (ultimo accesso: 12.02.2025).

6. *Verso un quadro metodologico per l'analisi delle fonti storico-educative*

Per rendere operative nel campo della ricerca storico-educativa le indicazioni emerse dagli studi analizzati, è necessario adottare un quadro metodologico di riferimento che consenta di identificare chiaramente le tipologie di domande di ricerca che possono essere poste nel contesto degli strumenti di analisi basati sull'intelligenza artificiale generativa, nonché di anticipare quali risultati siano ragionevolmente ottenibili sulla base delle caratteristiche dei dati investigati.

Facendo riferimento allo studio di Chartier e colleghi[20], l'indagine storica supportata da strumenti di intelligenza artificiale generativa può essere formalmente articolata in tre dimensioni fondamentali, che ne delineano il campo operativo:

- *domande chiuse vs domande aperte*: le prime mirano a informazioni fattuali, le seconde richiedono spiegazioni più articolate;
- *domande quantitative vs domande qualitative*: le domande quantitative puntano alla raccolta di dati misurabili, mentre quelle qualitative mirano aspetti descrittivi e interpretativi di eventi o fenomeni storici;
- *la tipologia di risposta attesa*, che varia da dati quantitativi e elenchi di dati, sino a descrizioni sempre più dettagliate per interrogativi aperti e complessi.

Dall'intreccio di queste tre dimensioni, emergono cinque tipologie di domande, ordinabili per complessità crescente: 1. domanda quantitativa chiusa, in cui si cercano dati numerici; 2. domanda qualitativa chiusa, finalizzata all'identificazione di metadati specifici; 3. domanda qualitativa chiusa, basata sulla raccolta di elenchi di dati; 4. domanda qualitativa aperta, centrata su definizioni o descrizioni sintetiche; 5. domanda qualitativa aperta, che richiede una trattazione più dettagliata e articolata di un problema. Nella tabella seguente (tabella 1) proponiamo un esempio concreto, applicato al caso specifico del giornale scolastico *I diritti della Scuola*, evidenziando per ciascuna tipologia di domanda i dati attesi e i quesiti esemplificativi.

[20] Chartier, Dakkoune, Bourgeois, Jean, *HiBenchLLM: Historical Inquiry Benchmarking for Large Language Models*, cit.

Tipologia di domanda	Dati attesi	Esempi di domande
Quantitativa (chiusa)	Dati numerici	Quante volte la rivista ha trattato il tema della retribuzione degli insegnanti tra il 1910 e il 1930? Quanti articoli dedicati alla scuola rurale sono stati pubblicati tra il 1899 e il 1922?
Qualitativa (chiusa)	Metadati	Chi era il direttore della rivista nel 1922? In che anno la rivista ha cambiato la sua periodicità di pubblicazione?
Qualitativa (chiusa)	Elenco di dati	Chi sono i collaboratori della rivista tra il 1918 e il 1922? Quali categorie di insegnanti erano maggiormente rappresentate negli articoli della rivista negli anni '20?
Qualitativa (aperta)	Definizione/ Descrizione	Qual era la posizione della rivista sullo status giuridico degli insegnanti nel primo dopoguerra? Come la rivista affrontava il tema della formazione degli insegnanti negli anni '20 e '30?
Qualitativa (aperta)	Descrizione dettagliata di un problema	In che modo la rivista ha affrontato il dibattito sulla laicità dell'istruzione tra il 1906 e il 1908? Come gli articoli della rivista hanno discusso i punti di forza e le criticità della Riforma Gentile? Quali argomenti venivano utilizzati per sostenerla o criticarla?

Tabella 1. Esemplificazione delle Tipologie di domande e dati attesi

L'adozione di questa tipologia di framework consente di strutturare con chiarezza l'indagine supportata dalle architetture basate sugli LLM, esplicitando sin dall'inizio il tipo di informazione atteso e fornendo criteri per valutare l'adeguatezza e accuratezza delle risposte ottenute.

Infine, è importante evidenziare alcune limitazioni tecniche attuali nell'impiego dei sistemi basati sull'intelligenza artificiale generativa. Non esiste, ad oggi, una soluzione tecnologica universale in grado di gestire tutte le tipologie di fonti storiche, a causa della loro natura intrinsecamente eterogenea. Ad esempio, manuali scolastici, riviste specialistiche e cataloghi scolastici richiedono adattamenti specifici degli strumenti impiegati. Pertanto, diventa indispensabile che lo storico dell'educazione acquisisca competenze tecniche adeguate per utilizzare efficacemente queste tecnologie nella ricerca, riducendo la necessità di ricorrere a supporti informatici esterni.