LLM e Retrieval Augmented Generation (RAG) per le biblioteche. Sperimentazioni, prospettive e valorizzazione del patrimonio

Retrieval Augmented Generation (RAG) for Libraries. Research, Perspectives and Valorisation of Heritage

Angelo La Gorga
Università degli Studi di Torino
angelo.lagorga@unito.it

Roberto Testa
Università degli Studi di Torino
roberto.testa@unito.it

Lorenzo Verna
CTO e ricercatore indipendente in Intelligenza Artificiale
v.lorenzo@gmail.com

| abstract

Il contributo esplora le opportunità e i limiti connessi all'impiego dei Large Language Models (LLM) e, in particolare, dell'architettura Retrieval-Augmented Generation (RAG) nei contesti bibliotecari, con l'obiettivo di indagare il potenziale di queste tecnologie nel supportare la valorizzazione del patrimonio documentario. Dopo un inquadramento teorico dei principali modelli linguistici e delle architetture generative, si analizzano le applicazioni emergenti nei servizi bibliotecari, distinguendo tra ambiti di back-office e servizi rivolti all'utenza. A completamento della riflessione teorica, viene presentato un caso studio sperimentale sviluppato presso la Biblioteca "Arturo Graf" dell'Università di Torino, relativo al fondo "Emanuele Artom", introdotto al fine di verificare in che misura strumenti basati su LLM possano contribuire all'organizzazione, sintesi e trascrizione di materiali eterogenei, nonché alla costruzione di una base conoscitiva interrogabile secondo logiche semantiche e narrative.

The paper explores the opportunities and limitations associated with the use of Large Language Models (LLMs), and in particular the Retrieval-Augmented Generation (RAG) architecture, within library contexts, with the aim of investigating the potential of these technologies in supporting the enhancement of documentary heritage. Following a theoretical overview of the main language models and generative architectures, the study examines emerging applications in library services, distinguishing between back-office operations and user-facing services. Complementing the theoretical discussion, the paper presents an experimental case study developed at the "Arturo Graf" Library of the University of Turin, focusing on the "Emanuele Artom" collection. The case is introduced to assess the extent to which LLM-based tools can contribute to the organization, summarization, and transcription of heterogeneous materials, as well as to the construction of a knowledge base that can be queried through semantic and narrative approaches.

DOI 10.36158/97912566920714

1. Introduzione: il ruolo crescente dell'IA nelle biblioteche

Da alcuni anni le riflessioni relative ai possibili impatti dell'Intelligenza Artificiale (IA) sui processi interni e sui servizi offerti dalle biblioteche stanno caratterizzando in modo sempre più rilevante il dibattito di ambito biblioteconomico e sono accompagnate, e al contempo favorite, dalle numerose sperimentazioni sul campo di progetti innovativi (Morriello, 2024; Dinotola, 2024a). Tutto ciò ha portato sia

alla proliferazione della letteratura sul binomio "biblioteche e IA", declinato da vari punti di vista (AIB, 2024), sia all'elaborazione di documenti di ricerca e indirizzo, che, partendo dalla consapevolezza della complessità del tema, hanno fornito alcune raccomandazioni per aiutare i bibliotecari e le bibliotecarie ad affrontare in modo proattivo, critico e responsabile le sfide derivanti dall'impiego di sistemi basati sull'IA, che riguardano diverse questioni delicate, tra cui quelle legate agli aspetti tecnologici, all'etica e alla privacy (IFLA, 2020; Lana, 2023; Padilla, 2019). Dunque, chi si occupa di biblioteche è chiamato a sviluppare nuove conoscenze e competenze, da aggiornare costantemente; inoltre, risulta particolarmente utile l'adozione di un approccio aperto, meno autoreferenziale e interdisciplinare per aiutare a governare questa nuova complessità e a incidere positivamente tanto sul piano delle elaborazioni teoriche e sull'avanzamento della biblioteconomia, quanto sul piano pratico. In tale quadro, il presente contributo si propone di analizzare le potenzialità e i limiti dell'impiego dei Large Language Models (LLM) nelle biblioteche, con un focus specifico sull'architettura Retrieval-Augmented Generation (RAG), quale possibile risposta ad alcune delle principali criticità legate all'uso di sistemi generativi. L'obiettivo è duplice: da un lato, offrire un inquadramento teorico e tecnologico aggiornato, capace di rendere comprensibili i presupposti tecnici e metodologici che sottendono all'adozione di tali strumenti; dall'altro, valutare le loro implicazioni operative in ambito biblioteconomico. L'articolazione del contributo riflette questa duplice prospettiva. La prima sezione è dedicata alla definizione e contestualizzazione dell'intelligenza artificiale generativa, con un approfondimento sui LLM e sull'architettura RAG. La seconda sezione analizza le principali modalità di applicazione dei LLM nei contesti bibliotecari, distinguendo tra attività di back-office e servizi all'utenza, e offrendo una rassegna critica delle esperienze più significative emerse nella letteratura recente. La terza e ultima parte presenta un'ipotesi applicativa di strumenti LLM nella predisposizione di un'architettura RAG, nel contesto della valorizzazione dei fondi di persona.

2. I Large Language Models: definizione, architettura e limiti

I Large Language Models (LLM) costituiscono una delle evoluzioni più rilevanti nell'ambito dell'elaborazione del linguaggio naturale. Si tratta di modelli di machine learning addestrati su vasti insiemi di testi, attraverso tecniche di apprendimento supervisionato che consentono loro di prevedere la parola successiva in una sequenza e, più in generale, di generare e comprendere contenuti linguistici. Tra i principali esempi si possono citare GPT di OpenAI (Brown et al., 2020), BERT di Google (Devlin, Chang, Lee, & Toutanova, 2019) e PaLM (Chowdhery et al., 2022), noti per essere stati allenati su dataset composti da enormi quantità di dati testuali, provenienti da fonti eterogenee come il web, pubblicazioni scientifiche e opere librarie. Un elemento tecnico centrale che accomuna questi modelli è l'architettura Transformer, introdotta da Vaswani et al. (2017), che ha migliorato sensibilmente la gestione delle relazioni tra le parole, superando le limitazioni delle reti sequenziali tradizionali. Gli LLM si distinguono per alcune proprietà fondamentali, tra cui la dimensione e la scalabilità: GPT-3, ad esempio, conta 175 miliardi di parametri, una capacità che consente di rappresentare strutture linguistiche complesse e affrontare compiti che richiedono ragionamento e problem solving (Brown et al., 2020). Inoltre, presentano un'elevata capacità di generalizzazione, che li rende utilizzabili in una varietà di contesti, dalla traduzione automatica alla generazione di testi, codice e risposte in linguaggio naturale. Un'altra caratteristica distintiva

DOI 10.36158/97912566920714

è l'abilità di considerare il contesto complessivo di un testo, grazie al meccanismo di attenzione dell'architettura Transformer. Questo permette al modello di produrre output più coerenti e rilevanti, adattandosi in modo flessibile al contenuto e alla struttura delle sequenze testuali. A partire dai modelli più recenti, i LLM sono inoltre stati integrati in architetture multimodali, in grado di elaborare simultaneamente input testuali e visivi, ampliando così il ventaglio delle applicazioni potenziali. Negli ultimi anni, il campo ha visto un'accelerazione significativa, con l'introduzione di modelli sempre più avanzati, come GPT-4 di OpenAI e PaLM2 di Google, capaci di prestazioni migliorate in ambiti applicativi concreti. Tuttavia, permangono alcune criticità. In particolare, la conoscenza degli LLM è limitata al periodo di addestramento e non si aggiorna dinamicamente, il che può generare risposte imprecise o non allineate con informazioni più recenti. Inoltre, questi modelli tendono talvolta a produrre informazioni plausibili ma inesatte, un fenomeno noto come "allucinazione", che pone rilevanti questioni di affidabilità e verifica delle fonti.

3. L'architettura Retrieval-Augmented Generation (RAG)

Uno strumento possibile per utilizzare al meglio i LLM e superare alcune loro limitazioni è l'impiego della tecnica di RAG. Questa tecnica combina la potenza generativa degli LLM con l'accesso a fonti di informazione esterne e aggiornabili (Lewis et al., 2020). Quando un sistema basato su architettura RAG riceve una query utente, prima cerca informazioni rilevanti in un database esterno, poi fornisce queste informazioni al modello come contesto aggiuntivo per generare una risposta. Questo approccio si basa su tecniche avanzate di recupero delle informazioni e di rappresentazione semantica del testo, come quelle proposte da Reimers e Gurevych (2019). Un aspetto fondamentale dell'approccio RAG è la sua architettura a due fasi: *retrieval* e generation. Nella fase di *retrieval*, il sistema utilizza algoritmi di ricerca per identificare i documenti o i contenuti più pertinenti alla query dell'utente. Questi algoritmi possono basarsi su tecniche di *embedding* semantico, che permettono di catturare il significato delle parole e delle frasi, andando oltre la semplice corrispondenza lessicale. La fase di *generation*, invece, sfrutta la capacità del LLM di sintetizzare informazioni e generare risposte coerenti, utilizzando come input sia la query originale che le informazioni recuperate (Izacard & Grave, 2021). Un'ulteriore caratteristica della RAG è la sua possibilità di incorporare diverse fonti di conoscenza. Queste possono includere database strutturati, archivi di documenti, pagine web, o persino flussi di dati in tempo reale. Questa versatilità rende i sistemi basati su architettura RAG particolarmente adatti a scenari in cui è necessario accedere a informazioni aggiornate e specifiche per un dominio. Inoltre, la RAG offre la possibilità di implementare meccanismi di controllo e filtraggio delle fonti, contribuendo così a mitigare il problema delle "allucinazioni" tipiche degli LLM e a garantire una maggiore affidabilità delle risposte generate (Liu, Xiong, Sun, & Liu, 2020).

3.1. *Vantaggi e aspetti critici della RAG*

La tecnica RAG offre alcuni vantaggi rispetto al semplice *prompting* dei Large Language Models, in particolare il potenziale di fornire risposte più accurate, integrando informazioni aggiornate e verificabili nel processo di generazione. Questo approccio riduce il rischio di allucinazioni tipiche degli LLM, ancorando le risposte a fonti di informazione

DOI 10.36158/97912566920714

verificabili. La base di conoscenza esterna funge da "memoria" aggiuntiva per il modello, consentendogli di accedere a informazioni aggiornate e pertinenti che non erano necessariamente presenti nei dati di addestramento originali. La trasparenza è un aspetto critico nell'utilizzo di sistemi di intelligenza artificiale, soprattutto in contesti che richiedono responsabilità e verificabilità. Le tecniche RAG offrono un vantaggio significativo in questo ambito, rendendo possibile tracciare la fonte delle informazioni utilizzate per generare una risposta. Quando un sistema RAG produce un output, è in grado di fornire non solo la risposta generata, ma anche i riferimenti alle fonti specifiche da cui ha estratto le informazioni rilevanti. Questo livello di tracciabilità permette agli utenti e agli sviluppatori di verificare l'origine delle informazioni, valutarne l'affidabilità e comprendere il ragionamento alla base della risposta generata. La trasparenza offerta dalla RAG ha implicazioni importanti in termini di responsabilità e fiducia. Gli utenti possono avere una maggiore confidenza nelle risposte fornite, sapendo che sono basate su fonti identificabili e verificabili. Inoltre, questo aspetto facilita l'identificazione e la correzione di eventuali errori o bias presenti nel sistema. Non meno rilevante risulta la capacità di adattarsi a domini specifici. Mentre gli LLM sono addestrati su vasti corpus di dati generici, un sistema RAG può essere facilmente personalizzato per domini specialistici integrando fonti di conoscenza specifiche nel proprio database di conoscenza. Questa flessibilità permette di creare sistemi altamente specializzati senza la necessità di riaddestrare completamente il modello di base. La personalizzazione non si limita solo al contenuto, ma può estendersi anche allo stile di comunicazione e al livello di dettaglio delle risposte, permettendo di adattare il sistema alle esigenze specifiche di diversi gruppi di utenti o contesti applicativi. Infine una delle caratteristiche più vantaggiose dei sistemi basati su RAG è la possibilità di aggiornare la base di conoscenza in modo relativamente semplice e rapido. Questo permette di mantenere il sistema aggiornato con le informazioni più recenti senza dover ricorrere al costoso processo di riaddestramento del LLM sottostante. Uno degli aspetti più critici di un sistema RAG è, poi, la strategia di indicizzazione. Le tecniche tradizionali di indicizzazione basate su keyword sono state progressivamente sostituite da approcci basati su *embedding* semantici, che permettono di trasformare i documenti in vettori densi in uno spazio di alta dimensionalità (Reimers & Gurevych, 2019). Questo approccio, utilizzato da modelli come Sentence-BERT, consente di catturare la similarità semantica tra query e documenti, migliorando il ranking dei risultati. Tuttavia, l'accuratezza dell'indicizzazione vettoriale dipende dalla qualità degli *embeddings* e dalla capacità del modello di rappresentare il contesto. La selezione e l'assegnazione di priorità dei contenuti più rilevanti tra i risultati del *retrieval* rappresentano un ulteriore punto critico. Una volta recuperati i frammenti di conoscenza, è cruciale classificare i risultati in base alla loro pertinenza rispetto alla query dell'utente. I sistemi RAG spesso utilizzano un approccio in due fasi: un *retrieval* iniziale basato su similarità semantica, seguito da un *reranking* dei risultati migliori. Il *reranking* è solitamente eseguito utilizzando un modello più sofisticato, che può prendere in considerazione fattori aggiuntivi, come la coerenza tra il frammento recuperato e il contesto generato (Nogueira, Yang, Cho, & Lin, 2019). Stabilita l'architettura, le criticità non finiscono, in quanto emerge la necessità di trovare l'equilibrio ottimale tra l'utilizzo delle informazioni recuperate e le capacità generative dell'LLM. Un eccessivo affidamento al *retrieval* può portare a risposte frammentarie, mentre una dipendenza eccessiva dalla generazione può compromettere l'accuratezza. Bisogna inoltre tenere sempre in considerazione eventuali limiti computazionali o di performance. L'implementazione di sistemi basati su RAG, infatti, richiede notevoli risorse computazionali, superiori a quelle necessarie per un LLM isolato. Questo può introdurre latenze

DOI 10.36158/97912566920714

significative e limitare l'applicabilità in scenari che richiedono risposte in tempo reale. Infine, nonostante i miglioramenti in termini di trasparenza, l'interpretazione del processo decisionale in un sistema basato su RAG può rimanere complessa, soprattutto per risposte elaborate che combinano molteplici fonti di informazione.

4. LLM e RAG in Biblioteca: applicazioni e prospettive

Lo sviluppo e la conseguente adozione da parte delle biblioteche di strumenti *AI-based* può assumere diverse forme, che variano a seconda della natura, dell'identità e della mission di ogni specifico ente. La letteratura riguardante le diverse applicazioni è molto ricca se si presta attenzione all'area delle biblioteche accademiche. Numerosi sono gli strumenti che possono andare incontro a chi fa ricerca, laddove si vuole approfondire un determinato argomento. Un po' meno consistenti, dal punto di vista numerico, sono oggi gli studi relativi all'utilizzo di questi strumenti applicati nell'ambito delle biblioteche di pubblica lettura. Tra i più diffusi strumenti che talvolta possono essere implementati grazie all'utilizzo di IA si annoverano i chatbot, programmi che hanno lo scopo di conversare con utenti umani in linguaggio naturale (Kaushal & Yadav, 2022). I primi chatbot sono stati sviluppati negli anni '60 ed erano sistemi molto semplici, in grado di rispondere solo a un determinato numero di domande preimpostate. Nei primi anni 2000 riescono a eseguire richieste più complicate in quanto programmati con una serie di regole atte a determinare come rispondere a ogni specifico input: si tratta dei cosiddetti chatbot linguistici, o *rule-based* (Adetayo, 2023). Nel secondo decennio del nuovo millennio, l'avvento del *Natural Language Processing* (NLP) e delle tecniche di *Machine Learning* ha rappresentato un punto di svolta nello sviluppo dei chatbot, che iniziano così a essere allenati con enormi moli di dati. Successivamente, grazie all'implementazione di nuove tecniche di *Deep Learning*, i chatbot possono rispondere a tutta una serie di input e comandi espressi in linguaggio naturale con maggiore accuratezza (Verma, 2023): gli esempi più noti sono gli LLM quali ChatGPT e Claude (Frické, 2024). A questo punto può essere opportuno ipotizzare una distinzione tra:

1. l'utilizzo di tool basati su AI per funzioni di *back-office*, quali la catalogazione, la classificazione, la soggettazione, lo sviluppo e la gestione delle collezioni;
2. l'utilizzo per l'implementazione di servizi destinati direttamente agli utenti, come ad esempio il supporto nelle ricerche bibliografiche, la *Reader's Advisory* e, più in generale, tutti i servizi di *reference*.

Tra i più rilevanti LLM adoperati in quest'ambito – e di conseguenza maggiormente citato all'interno della letteratura sul tema – si trova ChatGPT, chatbot sviluppato da OpenAI (*GPT-3.5*, rilasciata nel novembre 2022, e *GPT-4*, rilasciata nel 2023), o sfruttando le API per creare chatbot personalizzati (OpenAI, 2024). Prima di illustrare le diverse possibilità di utilizzo dei LLM e della RAG in ambito bibliotecario, è opportuno sottolineare che le implementazioni presentate si configurano come soluzioni finalizzate all'ottimizzazione dei servizi informativi e al potenziamento dell'esperienza utente. Occorrerà, soprattutto per quanto concerne le attività di *back-office*, la presenza di una supervisione esterna che avrà lo scopo di intervenire per verificare che il flusso di lavoro proceda nella maniera corretta e che all'interno degli output prodotti dal sistema stesso non vi siano informazioni errate e ambiguità di varia natura.

DOI 10.36158/97912566920714

4.1. *Attività di back-office*

Brzustowicz (2023) analizza le implicazioni e le potenzialità di ChatGPT nell'attività di catalogazione, illustrandone le capacità attraverso un esperimento comparativo in cui al LLM viene richiesto di generare record MARC in formato RDA per diverse fonti, successivamente confrontati con quelli realizzati manualmente da catalogatori. Sono stati considerati materiali sia bibliografici sia discografici, in diverse lingue, presenti su WorldCat. Secondo lo studioso, con un training adeguato, ChatGPT è in grado di produrre buone catalogazioni, utilizzabili come punto di partenza e successivamente integrate o corrette dall'intervento umano. Il modello è anche capace di creare nuovi record catalografici a partire dalla fonte, generare descrizioni affidabili e produrre contenuti originali secondo diversi standard di metadatazione (Afjal, 2023; Corrado, 2021). ChatGPT può inoltre classificare informazioni non necessariamente legate a una collezione specifica, come i *call numbers*, ed estrarre metadati fondamentali quali titolo, autore, editore, data di pubblicazione, soggetto principale e altri elementi descrittivi. Queste funzioni sono rese possibili dalla presenza, nel training di ChatGPT, di dati provenienti da importanti cataloghi come OCLC, WorldCat, Library of Congress, British Library ed Europeana. Uno dei vantaggi principali di questo approccio è il risparmio di tempo nelle operazioni di copiatura durante la catalogazione, pur restando presenti alcune limitazioni e bias – dovuti, ad esempio, all'incompletezza o all'obsolescenza del materiale di addestramento, nonché a problematiche relative al copyright o alla qualità dei contenuti generati dall'IA. Studi successivi condotti da Lund e Wang (2023) e da Houston e Corrado (2023) hanno ulteriormente evidenziato la capacità di ChatGPT di generare metadati e descrizioni bibliografiche, come titoli, abstract e sintesi dei contenuti. Anche la catalogazione per copia può essere automatizzata, permettendo di modificare record esistenti piuttosto che crearli da zero. Analogamente, l'IA può essere impiegata per la classificazione documentaria, partendo da titoli e contenuti per generare soggetti e numeri di classificazione (Zakaria, 2023). L'intelligenza artificiale può anche supportare l'analisi di statistiche relative a lettura, prestiti e utilizzo delle risorse, fornendo interpretazioni utili per la generazione di liste di lettura o per l'aggiornamento automatico degli inventari, nell'ambito dello sviluppo delle collezioni (Zakaria, 2023). Inoltre, può contribuire all'analisi del comportamento degli utenti – i cui dati sono raccolti in forma anonima – al fine di identificare pattern, prevedere bisogni informativi e individuare le risorse più selezionate o richieste. In questo senso, il *Machine Learning*, sfruttando tali dati e integrandoli con analisi delle collezioni, può offrire indicazioni utili per le decisioni relative all'acquisto, al deposito o allo scarto di materiali (Frické, 2023).

4.2. *I servizi rivolti all'utente*

Un LLM sul quale sia stata svolta un'attività di training a partire da una base di dati ampia e relativa alle collezioni di una biblioteca può svolgere attività di *Reader's Advisory* fornendo consigli di lettura personalizzati e informazioni basate sulle abitudini di lettura, temi, generi e autori preferiti (Prathibha & Shilpa Rani, 2021). L'attività di *recommendation* può essere svolta fornendo al sistema informazioni estratte dal profilo dell'utente (che deve aver svolto delle attività registrate quali prestito, ricerche pregresse, consultazioni, creazioni di liste di preferiti). Un esempio è il prototipo Reading(&)Machine (Lamberti, Mellia, & Vivarelli, 2024), che sfrutta algoritmi di raccomandazione per suggerire letture; in alternativa, può essere l'utente stesso a specificare, durante il processo di interazione

DOI 10.36158/97912566920714

con l'IA, le sue preferenze per poi ottenere degli output attendibili e personalizzati in base alle proprie necessità, ad esempio delle liste di lettura su richiesta (Yang & Mason, 2023). Le stesse raccomandazioni possono essere utilizzate per lo sviluppo delle collezioni, coinvolgendo direttamente l'utente nella proposta di nuovi contenuti da selezionare, relativi a un argomento specifico da approfondire o che non è stato per qualche motivo incluso nella collezione (Houston & Corrado, 2023). Allo stesso modo, un utilizzo dell'IA può facilitare e stimolare l'approccio serendipico, offrendo all'utente suggerimenti inaspettati e spingendolo a percorrere strade e modalità di esplorazione alternative, come ad esempio quella della transmedialità (Dinotola, Testa, & Vivarelli, 2024). Una funzione dei LLM abbastanza comunemente utilizzata all'interno delle biblioteche è quella di *reference*. Il chatbot in questo senso può prendersi carico di una serie di semplici incarichi che tradizionalmente svolge il bibliotecario – ad esempio mostrare i servizi disponibili in biblioteca, gli orari di apertura, le modalità di accesso a documenti fisici e online, quindi il supporto nella ricerca base e tutte quelle altre attività tipicamente di *reference* – per permettere a quest'ultimo di svolgere compiti per cui è necessaria la presenza e l'impegno di una persona umana. Uno dei primi chatbot utilizzati nell'ambito delle biblioteche è Kornelia, sviluppato a Berna nel 2010, capace di fornire risposte immediate sui servizi della biblioteca e supportare nella ricerca della fonte corretta e nella risoluzione di problemi generali. Un altro esempio è Emma the Catbot, sviluppato dalla Mentor Public Library negli Stati Uniti nel 2009, utilizzato per rispondere a domande sulla biblioteca e sui servizi di *reference*. Nel 2012 Emma diventerà Infotabby, servizio open access che utilizza i metadati AIML (*Artificial Intelligence Markup Language*), in grado di connettersi a tutti i database della biblioteca e categorizzare e rispondere a domande (Adetayo, 2023). Tra le più recenti implementazioni si ricordano anche Kingbot (San Jose University) e Bizzy (University of Oklahoma), entrambi lanciati nel 2020. Il primo utilizza Dialogflow di Google attraverso il software proprietario Kommunicate per rispondere a richieste di *reference*, mentre il secondo utilizza Ivy, un software di *machine learning*, per rispondere a domande ordinarie (Lappalainen & Narayanan, 2023). Uno dei primi esempi di chatbot *AI-based* è Xiaotu, sviluppato dalla Tsinghua University Library nel 2011, che supporta gli utenti nei servizi di *reference* ed è in grado anche di acquisire nuova conoscenza a partire dalle domande e dalle risposte degli utenti (Lai, 2023). Un LLM generico non può sostituire il lavoro umano, e anche lo studio di Yang e Mason (2023) ha dimostrato che, nonostante possieda l'abilità tecnica di rispondere in modo formalmente corretto alle domande di *reference*, spesso manca di emozioni e della capacità di "leggere tra le righe" come solo una persona può fare. Lavorando in profondità su questioni più complesse, talvolta risulta impreciso, soprattutto nei casi in cui non dispone di informazioni specifiche o che richiedono aggiornamenti frequenti, come le politiche della biblioteca o la disponibilità di alcune risorse (Yang & Mason, 2023). In quest'ottica, uno degli esperimenti più avanzati riguarda AISHA, un chatbot (modello GPT) sviluppato dalla Zayed University Library. L'assistente virtuale è stato creato e addestrato su dataset che comprendono più di cento bibliografie e pagine web con informazioni sulla biblioteca, ed è stata sviluppata un'apposita interfaccia attraverso Streamlit, una libreria *open-source* di Python per le web app. Lo scopo di AISHA è offrire informazioni e supporto in una maniera innovativa e personalizzata agli studenti dell'università, fornendo risposte simili a quelle umane e capacità di traduzione automatica (Lappalainen & Narayanan, 2023). Alla base di AISHA si trova la tecnologia RAG, particolarmente utile per l'accesso all'informazione e il recupero di riferimenti bibliografici e documentari all'interno di sistemi quali archivi e biblioteche, oltre a rendere la risposta del chatbot più accurata e strut-

DOI 10.36158/97912566920714

turata (Di Marcantonio, 2024). Un LLM allenato appositamente sul corretto database e aggiornato regolarmente fornisce prestazioni migliori e più affidabili in queste tipologie di operazioni. Un ulteriore utilizzo interessante riguarda l'integrazione dei LLM all'interno dei *discovery tools* della biblioteca, con l'obiettivo di potenziare le funzionalità di ricerca, rendendola più aderente ai bisogni informativi dell'utente e offrendo nuove opportunità di esplorazione (Lappalainen & Narayanan, 2023). Integrato con cataloghi bibliotecari e banche dati accessibili online, un LLM può rendere l'esperienza di ricerca più fluida e completa (Panda & Kaur, 2023). Uno dei punti di forza dei LLM è la possibilità di offrire risposte immediate all'utente, basate su calcoli probabilistici e strutturate in forma di conversazione. L'interazione assume così le caratteristiche del dialogo naturale, differenziandosi dalle interfacce tradizionali *Google-like*, che restituiscono elenchi statici da scorrere (Panda & Kaur, 2023; Cox & Tzoc, 2023). Inoltre, l'approccio conversazionale può contribuire all'accessibilità, grazie a funzionalità come *speech-to-text* e *text-to-speech*, e offrire traduzioni automatiche che abbattono le barriere linguistiche (Lappalainen & Narayanan, 2023). Infine, un LLM può essere coinvolto in attività di fruizione e promozione dei servizi bibliotecari, come quiz letterari, musicali o cinematografici, curiosità su autori, argomenti e generi, mini-giochi di avventura narrativa, letture ad alta voce o interattive, e altre forme di intrattenimento volte a migliorare la *user experience* (Wani & Astunkar, 2024).

5. Verso la RAG: la valorizzazione dei fondi di persona come possibile campo di applicazione

In questa ultima sezione del contributo si è ritenuto utile fare riferimento a un caso applicativo che si configura come terreno di sperimentazione per l'impiego strutturale dei LLM nella progettazione di un'architettura RAG nell'ambito della valorizzazione dei fondi personali conservati nelle biblioteche accademiche, al fine di offrire una panoramica di un possibile approccio a tale pratica. In particolare questa sezione vuole riportare le attività in corso riferite alla valorizzazione del fondo Emanuele Artom, conservato presso la biblioteca "A. Graf" dell'Università di Torino, inserite nel contesto del progetto PNRR di trasformazione digitale della biblioteca, con la finalità di mostrare l'applicazione pratica di strategie facenti largo utilizzo dei LLM come strumenti di supporto per la predisposizione di un'architettura RAG finalizzata all'interrogazione interattiva del corpus documentario del fondo stesso. L'obiettivo finale del progetto è duplice: da un lato, testare le potenzialità della RAG nel migliorare l'accessibilità e la consultazione dei fondi di autore; dall'altro, proporre un modello metodologico replicabile per l'integrazione di dati eterogenei e la costruzione di narrazioni contestuali basate su relazioni semantiche e documentarie. La seguente sezione non si pone l'ancor prematuro compito di valutare il compimento degli obiettivi ultimi del progetto di ricerca, in quanto ancora in corso di elaborazione. Tuttavia si è ritenuto rilevante, nella struttura del contributo, fare riferimento alle fasi preliminari di predisposizione dell'architettura, in quanto esempi di applicazione sistematica di strumenti LLM a supporto delle attività di recupero e normalizzazione dei dati necessarie alla strutturazione della stessa. In particolare lo sviluppo e l'applicazione sistematica di strumenti LLM è maturata dalla presa in considerazione degli aspetti critici della valorizzazione di oggetti complessi come i fondi di persona, sia in generale, sia specifici dell'oggetto di studio individuato. Le *Linee guida sul trattamento dei fondi personali*, dell'Associazione Italiana Biblioteche (AIB), sottolineano l'importanza

DOI 10.36158/97912566920714

di affrontare i fondi personali in un'ottica integrata, mettendo in relazione materiale bibliografico, scritture archivistiche, documenti museali e altri oggetti (Associazione Italiana Biblioteche, 2019). L'individuo rappresenta l'elemento unificante di questi complessi documentari, che devono testimoniare interessi, attività e relazioni della persona nel proprio contesto storico e culturale. Questo approccio non deve limitarsi alla conservazione, ma deve garantire anche contestualizzazione e narrazione, rendendo accessibili contenuti e significati per le comunità di riferimento. La varietà di materiali (libri, documenti, manoscritti, fotografie, corrispondenze, appunti, oggetti museali) e la disomogeneità dei metadati generati da prassi descrittive eterogenee possono costituire ostacoli significativi (Sabba & Sardo, 2020) per un processo di contestualizzazione che implica la raccolta di informazioni biografiche, l'analisi delle reti relazionali con persone, istituzioni e luoghi significativi, nonché la ricostruzione del contesto di produzione dei documenti (Toccafondi, 2010; Barrera, 2006). L'indagine deve spesso superare i confini fisici della biblioteca, coinvolgendo anche archivi, musei e centri di documentazione suggerendo come solo un approccio interdisciplinare e transdisciplinare possa affrontare la complessità informativa e documentaria, al fine di restituire un quadro coerente della storia dell'autore e del suo patrimonio (Di Domenico & Sabba, 2020). La presa in considerazione di tali questioni cruciali che qui si è ritenuto richiamare velocemente, in riferimento allo specifico oggetto di studio, è avvenuta tramite lo svolgimento di un censimento delle risorse documentarie ascrivibili all'autore, valicando appunto i confini della biblioteca alla ricerca di tracce archivistiche ed elementi ulteriori, come le opere dell'autore stesso o quelle a lui riferite, che potessero arricchire la contestualizzazione storico culturale del fondo. Il censimento documentario relativo a Emanuele Artom ha permesso di identificare e raccogliere un insieme significativo di risorse distribuite presso diverse istituzioni, valutandone sia la consistenza che le modalità di descrizione e accesso, con una particolare attenzione allo stato delle risorse digitali disponibili (La Gorga, 2024; Marzal et al., 2025). Ai due bacini documentari principali, rappresentati dal fondo bibliografico conservato presso la biblioteca "A. Graf" e dall'archivio della famiglia Artom custodito dal Centro di Documentazione Ebraica Contemporanea (CDEC), si sono integrate ulteriori risorse documentarie, recuperate presso istituzioni di conservazione presenti sul territorio o da repository accessibili in rete, quali Internet Archive, Europeana, etc. La valutazione dei dati e dei documenti raccolti ha permesso di evidenziare possibili ostacoli alla realizzazione di una base dati integrata e coerente, punto di partenza fondamentale per la progettazione di un'architettura RAG. Da un lato, infatti, ci si è dovuti misurare con la presenza di inventari eterogenei e spesso descritti in formati non strutturati, come ad esempio file testuali. Dall'altro, invece, si è dovuto affrontare gli ostacoli rappresentati dal recupero efficace dei contenuti testuali dei vari libri e documenti, operazione non sempre immediata dovendo lavorare con digitalizzazioni provenienti da progetti pregressi e spesso, non solo non accompagnate da testo OCR, ma anche prodotte con risoluzioni e qualità delle immagini inadeguate all'estrazione dei testi attraverso l'uso di script tradizionali.

5.1. *Normalizzazione dati, estrazione e sintesi dei testi: prospettive d'uso dei LLM*

In questo paragrafo si desidera proporre una disamina degli strumenti LLM sviluppati per far fronte agli ostacoli specifici emersi dalla valutazione delle fonti documentarie raccolte. Considerato l'obiettivo del contributo non si farà riferimento all'intera metodologia applicata nell'elaborazione di un piano di valorizzazione del fondo, ma si presenteranno solamente gli strumenti LLM sviluppati e applicati per far fronte a operazioni altrimenti

DOI 10.36158/97912566920714

non automatizzabili. In primo luogo l'organizzazione e la razionalizzazione dell'inventario del fondo ha beneficiato dell'implementazione di tali strumenti, in particolare per definirne chiaramente i confini dovendo tenere conto di diverse fonti, non sempre concordanti. Si è pertanto ritenuto vantaggioso sviluppare uno strumento concepito per automatizzare l'estrazione di metadati bibliografici da documenti in formato PDF e organizzarli in un file strutturato csv/Excel, al fine di facilitare le operazioni di confronto automatizzate fra le diverse fonti a disposizione. L'applicazione si distingue per l'impiego di un modello di linguaggio multimodale, GPT-4o di OpenAI, sfruttato per le sue capacità di interpretazione sia testuale che visiva. La peculiarità dello strumento risiede infatti nella sua capacità di analizzare i documenti non solo come testo lineare, ma anche nella loro dimensione grafica e di impaginazione, elemento particolarmente rilevante per l'elaborazione di fonti complesse come inventari, tabelle, apparati critici o annotazioni manoscritte. Una delle caratteristiche più significative del sistema è l'analisi multimodale delle fonti: ogni pagina del PDF viene convertita in immagine, e tale rappresentazione viene inviata al modello GPT-4o. In questo modo, il processo di estrazione può tenere conto di elementi strutturali altrimenti non accessibili con approcci OCR tradizionali. Il risultato è una lettura più precisa e contestualizzata dell'informazione, in grado di adattarsi a casi d'uso eterogenei. L'estrazione dei dati è guidata da prompt personalizzabili, offrendo all'utente la possibilità di definire con precisione quali informazioni debbano essere rilevate. Questo approccio sfrutta la flessibilità semantica dei modelli linguistici avanzati, consentendo di personalizzare la procedura estrattiva in funzione della tipologia documentale e delle esigenze di ricerca. I dati ottenuti vengono infine organizzati automaticamente in un file dati strutturato, pronti per essere ulteriormente analizzati o archiviati[1].

Emerge poi la necessità di trattare i contenuti dei testi delle opere inventariate che dovranno essere integrati nella base dati di *retrieval* dell'architettura RAG, secondo logiche che, a nostro avviso, hanno meritato delle attente riflessioni. Dal punto di vista metodologico, il seguente approccio nasce da una riflessione critica sulla difficoltà di utilizzare i testi completi delle opere della biblioteca dell'autore in fase di *retrieval,* a causa dell'eccessivo rumore informativo che possono generare. Si è ipotizzata pertanto una strategia di sintesi incrementale dei testi al fine di ridurre la complessità informativa preservando, al contempo, la ricchezza semantica e la struttura tematica dei contenuti originari. Il tool di sintesi incrementale rappresenta, infatti, un'applicazione progettata per affrontare in modo strutturato ed efficiente l'elaborazione di testi lunghi e complessi, con l'obiettivo di produrre riassunti coerenti e tematicamente rilevanti, funzionali alla successiva indicizzazione e interrogazione in architetture basate su RAG. Il tool è pensato per superare i limiti di contesto degli attuali LLM, mediante una pipeline di sintesi distribuita e modulare, capace di trattare testi di grandi dimensioni senza compromettere la qualità semantica dell'output. In primo luogo, viene effettuata l'estrazione del contenuto testuale da documenti nei formati .txt, .pdf (utilizzando PyPDF2) ed .epub (tramite EbookLib e BeautifulSoup4). Successivamente, il testo viene sottoposto a una suddivisione intelligente in *chunk* da circa 10.000 parole, con attenzione alla conservazione della coerenza sintattica e tematica di paragrafi e frasi. Ogni porzione viene salvata come file .txt separato, per facilitarne la tracciabilità. La fase successiva prevede la sintesi incrementale di ciascun *chunk* tramite richieste al modello GPT-4.1-mini, impiegando prompt specificamente calibrati in funzione del dominio testuale. I riassunti ottenuti vengono salvati anch'essi in formato .txt, per poi essere aggregati in un unico documento coerente. Tale documento costitu-

1. La repository del progetto è pubblica al seguente indirizzo: https://github.com/AngeloLG/Estrazione_inventario.

DOI 10.36158/97912566920714

isce il punto di partenza per la sintesi finale, in cui l'intero testo aggregato viene rielaborato dal modello al fine di produrre un riassunto complessivo di circa 1.500 parole, salvato in formato Markdown e ottimizzato per l'impiego nei sistemi di *retrieval* semantico. In questo senso, il tool contribuisce in modo diretto all'ottimizzazione dei dati in ingresso per architetture RAG, migliorando l'efficacia della fase di recupero delle informazioni[2].

Anche la gestione del materiale archivistico ha potuto beneficiare di simili strategie, in particolare nelle operazioni di trascrizione dei documenti digitalizzati, che hanno comportato diverse sfide, alcune delle quali ancora aperte. In questo contesto, è stato sviluppato uno strumento finalizzato alla trasformazione del contenuto testuale presente in immagini in testo editabile, con un'attenzione specifica alla varietà dei documenti trattati (es. manoscritti, dattiloscritti, tipologie miste). Il sistema è progettato per orchestrare un flusso decisionale condizionato, capace di adattare dinamicamente le strategie di trascrizione a seconda della natura del documento in esame, ricorrendo ove necessario all'impiego di LLM via API. Uno degli elementi distintivi dell'applicazione è l'integrazione selettiva dei LLM nel processo di trascrizione, subordinata a una fase preliminare di classificazione automatica delle immagini. La prima fase prevede, infatti, la classificazione locale dell'immagine tramite un modello specializzato, in grado di identificare la tipologia documentaria (es. lettere, report, fatture). L'output di questa classificazione viene poi semplificato in categorie operative come manoscritto, dattiloscritto, o altro. Questo passaggio ha una funzione cruciale: determina quale motore di trascrizione debba essere attivato, ottimizzando così il bilanciamento tra accuratezza, costi e tempi di elaborazione. Nel caso di documenti dattiloscritti o di altro tipo non manoscritto, lo strumento procede inviando l'immagine codificata al modello GPT-4.1-mini di OpenAI, insieme al prompt selezionato. Il modello restituisce la trascrizione del testo contenuto nell'immagine, che viene successivamente salvata in un file .txt, mantenendo la corrispondenza tra nome del file immagine e file di output. L'implementazione include anche una logica dedicata ai documenti manoscritti, che originariamente avrebbe previsto l'impiego di modelli locali. Tuttavia, tale funzionalità risulta attualmente sospesa, in quanto i differenti test effettuati sia con modelli locali specializzati, sia con l'impiego di LLM attraverso API non hanno dato risultati accurati[3], in particolare a causa della peculiarità e varietà degli oggetti documentari da trascrivere, che complica le fasi di ingegnerizzazione dei prompt da usare per interagire con i modelli a disposizione. Di certo pertanto si mostra come un punto debole di questo approccio che riteniamo richieda ulteriori riflessioni.

6. Conclusioni

Il presente contributo ha inteso esplorare, con approccio teorico e applicativo, i vari ambiti di impiego dei LLM nella pratica biblioteconomica. L'articolazione del testo ha consentito di delineare progressivamente le caratteristiche dei modelli linguistici di ultima generazione, di evidenziarne le criticità, e di analizzare le possibilità offerte dall'integrazione con sistemi di *retrieval* semantico aggiornabile. In tale prospettiva, l'architettura RAG è emersa come possibile risposta a esigenze informative complesse, in grado di coniugare capacità generativa e affidabilità delle fonti, attraverso l'accesso controllato a basi di conoscenza esterne.

2. La repository del progetto è pubblica al seguente indirizzo: https://github.com/AngeloLG/Sintesi_incrementale.
3. La repository del progetto è pubblica al seguente indirizzo: https://github.com/AngeloLG/trascrizione_img.

DOI 10.36158/97912566920714

Le riflessioni sulle applicazioni bibliotecarie hanno voluto evidenziare una differenziazione tra impieghi a supporto delle attività di back-office e implementazioni orientate all'utenza, suggerendo scenari di adozione diversificati e suscettibili di evoluzione.

In questa cornice si inserisce l'ultima sezione del contributo, proposta come possibile approccio di integrazione operativa tra riflessione teorica e applicazione sul campo. L'analisi delle criticità legate alla valorizzazione dei fondi personali ha infatti suggerito come l'impiego mirato di strumenti basati su LLM possa supportare processi complessi di normalizzazione, estrazione, sintesi e trascrizione, contribuendo alla costruzione di basi conoscitive interrogabili secondo logiche semantiche e narrative.

I risultati preliminari di questa sperimentazione, insieme all'impianto complessivo del contributo, suggeriscono l'utilità di proseguire l'indagine in più direzioni. In primo luogo, si ritiene necessaria la predisposizione di ambienti in cui sia possibile coinvolgere utenti reali nella fase di verifica e messa alla prova dei sistemi, per raccogliere elementi utili alla valutazione dell'efficacia comunicativa e informativa delle soluzioni adottate. In tal senso, il confronto con le pratiche e i bisogni concreti rappresenta una condizione necessaria per orientare in modo sostenibile l'integrazione dell'intelligenza artificiale nei servizi bibliotecari. In questa direzione, uno sviluppo ipotizzato riguarda l'integrazione dell'architettura RAG all'interno di ambienti transmediali orientati all'interazione con l'utenza. Il progetto Transmedia Library Shelf Experience (Dinotola & Testa, 2025), in corso di definizione, propone una riconfigurazione dello scaffale bibliotecario come spazio narrativo ed esperienziale, capace di combinare elementi fisici e digitali in percorsi personalizzati e immersivi. In questo contesto si aprirebbe la possibilità di esplorare l'efficacia di interfacce conversazionali e intelligenti – basate su LLM e RAG – con lo scopo di facilitare la scoperta dei contenuti, la costruzione di connessioni semantiche e la generazione di esperienze di lettura orientate alla serendipità. Inoltre, sarà opportuno estendere la verifica dell'approccio ad altri contesti documentari, caratterizzati da differenti strutture e consistenze al fine di ottenere ulteriori elementi di valutazione.

Bibliografia

Adetayo, A.J. (2023). ChatGPT and librarians for reference consultations. *Internet Reference Services Quarterly*, 27(3), pp. 131-147. https://doi.org/10.1080/10875301.2023.2203681.

Afjal, M. (2023). ChatGPT and the AI revolution: A comprehensive investigation of its multidimensional impact and potential. *Library Hi Tech*. https://doi.org/10.1108/LHT-07-2023-0322.

Artom, E. (2008). *Diari di un partigiano ebreo. Gennaio 1940 – febbraio 1944* (G. Schwarz, a cura di). Bollati Boringhieri.

Associazione Italiana Biblioteche. (2019). *Linee guida sul trattamento dei fondi personali*. AIB. https://www.aib.it/documenti/linee-guida-sul-trattamento-dei-fondi-personali/.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., ... & Kaplan, J. (2022). Constitutional AI: Harmlessness from AI Feedback. *arXiv*. https://arxiv.org/abs/2212.08073.

Ballestra L. & Sonzini V. (2024). *Biblioteche e tecnologie al tempo dell'Intelligenza Artificiale. Atti del 62. Congresso nazionale AIB. Firenze, Biblioteca nazionale centrale, 16 e 17 novembre 2023*. Associazione italiana biblioteche. (in corso di stampa).

Barrera, G. (2006). Gli archivi di persone. In Pavone, C. (a cura di). *Storia d'Italia nel secolo ventesimo: Strumenti e fonti* (Vol. 3: *Le fonti documentarie*, pp. 617-657). Ministero

DOI 10.36158/97912566920714

per i Beni e le Attività Culturali. https://www.researchgate.net/publication/319188516_Gli_archivi_di_persone.

Bender, E.M., Gebru, T., McMillan-Major, A. & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922.

Brady, D.L. & Wang, T. (2023). Chatting about ChatGPT: How may AI and GPT impact academia and libraries? *Library Hi Tech News*, 40(3), pp. 26-29. https://doi.org/10.1108/LHTN-01-2023-0009.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, pp. 1877-1901.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *arXiv*. https://doi.org/10.48550/arXiv.2005.14165.

Brzustowicz, R. (2023). From ChatGPT to CatGPT: The implications of artificial intelligence on library cataloging. *Information Technology and Libraries*, 42(3). https://doi.org/10.5860/ital.v42i3.16295.

Intelligence and Virtual Reality (AIVR) (pp. 13-20). IEEE. https://ieeexplore.ieee.org/document/10773731.

Chen, X. (2023). ChatGPT and its possible impact on library reference services. *Internet Reference Services Quarterly*, 27(2), pp. 121-129. https://doi.org/10.1080/10875301.2023.2181262.

Corrado, Edward M. (2021). Artificial Intelligence: The Possibilities for Metadata Creation. *Technical Services Quarterly,* 38(4), pp. 395-405.

Di Marcantonio, G. (2024). Artificial Intelligence, Large Language Models (LLMs), and Retrieval-Augmented Generation (RAG). New tools for accessing archival and bibliographic resources. *Bibliothecae.It*, 13(1), pp. 146-173. https://doi.org/10.6092/issn.2283-9364/19982.

Dinotola, S. (2024a). L'intelligenza artificiale in biblioteca: Iniziative e sperimentazioni nel contesto italiano. In Ponzani, V. & Battaggia, M. (a cura di). *Rapporto sulle biblioteche italiane 2021-2023* (pp. 57-65). Associazione italiana biblioteche.

Dinotola, S. (2024b). Costruire, valutare, comunicare le collezioni secondo un approccio rinnovato: Dal modello concettuale alla ricerca applicata. In Vivarelli, M. & Dinotola, S., (a cura di). *Sul confine: Le collezioni delle biblioteche tra gestione, produzione editoriale, esperienze di lettura*. Ledizioni.

Dinotola, S., Testa, R. & Vivarelli, M. (2024). Verso lo scaffale narrativo e transmediale. *Biblioteche Oggi*, 42(5), pp. 26-36. http://dx.doi.org/10.3302/0392-8586-202405-026-1.

Dinotola, S., Testa, R. (2025). Transmedia library shelf experience: Innovative research approaches. *Journal of Librarianship and Information Science, 0*(2025), 0. https://doi.org/10.1177/09610006241310908.

Devlin, J., Chang, M., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*. https://arxiv.org/abs/1810.04805.

Fondazione CDEC. (s.d.). *Archivio Emanuele Artom*. https://digital-library.cdec.it.

Frické, M. (2024). *Artificial intelligence and librarianship: Notes for teaching*. SoftOption Ltd.

DOI 10.36158/97912566920714

Gao, J., Galley, M. & Li, L. (2019). Neural approaches to conversational AI. *Foundations and Trends® in Information Retrieval*, 13(2-3), pp. 127-298. https://doi.org/10.1561/1500000074.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.

Houston, A.B. & Corrado, E.M. (2023). Embracing ChatGPT: Implications of emergent language models for academia and libraries. *Technical Services Quarterly*, 40(2), pp. 76-91. https://doi.org/10.1080/07317131.2023.2187110.

International Federation of Library Associations and Institutions (IFLA). (2023). *IFLA statement on libraries and artificial intelligence*. https://repository.ifla.org/items/8c05d706-498b-42c2-a93a-3d47f69f7646.

Izacard, G. & Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 874-880). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.74.

Kaushal, V. & Yadav, R. (2022). The role of chatbots in academic libraries: An experience-based perspective. *Journal of the Australian Library and Information Association,* 71(3), pp. 215-232. https://doi.org/10.1080/24750158.2022.2106403.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint*. https://arxiv.org/abs/2001.08361.

Kingma, D.P. & Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint*. https://arxiv.org/abs/1312.6114.

Lai, K. (2023). How well does ChatGPT handle reference inquiries? An analysis based on question types and question complexities. *College and Research Libraries*, 84(6), pp. 974-995. https://doi.org/10.5860/crl.84.6.974.

Lamberti, F., Mellia, M. & Vivarelli, M. (a cura di). (2024). *Biblioteche, lettura, intelligenza artificiale: Struttura e contesto del progetto Reading(&)Machine*. Editrice Bibliografica.

Lana, M. (2023). Leggere l'IFLA statement on libraries and Artificial Intelligence al tempo di ChatGPT. *Biblioteche oggi Trends*, 9(1), pp. 4-12. https://doi.org/10.3302/2421-3810-202301-006-1.

Lappalainen, Y. & Narayanan, N. (2023). Aisha: A custom AI library chatbot using the ChatGPT API. *Journal of Web Librarianship*, 17(3), pp. 37-58. https://doi.org/10.1080/19322909.2023.2221477.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S. & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv*. https://doi.org/10.48550/arXiv.2005.11401.

Liu, Z., Xiong, C., Sun, M. & Liu, Z. (2020). Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7342-7351). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.655.

Marcus, G. & Davis, E. (2020). GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. *MIT Technology Review*. https://www.technologyreview.com/2020/08/22/1007539/gpt-3-openai-language-ai-opinion.

Marzal, M.Á., La Gorga, A. & Vivarelli, M. (2025). El enriquecimiento del valor de los fondos personales en las bibliotecas universitarias. *Revista Española de Documentación Científica*, 48(1), p. 1812. https://doi.org/10.3989/redc.2025.1.1653.

DOI 10.36158/97912566920714

Morriello, R. (2024). Intelligenza artificiale nelle biblioteche: Stato dell'arte ed esperienze di applicazione. In A. Capaccioni & P. Castellucci (a cura di), *II Seminario italo-spagnolo di biblioteconomia e documentazione, Roma, 4-5 novembre 2022* (pp. 33-46). Ledizioni.

OpenAI. (2023). GPT-4 technical report. *arXiv preprint*. https://arxiv.org/abs/2303.08774.

Padilla, T. (2019). Responsible operations: Data science, machine learning, and AI in libraries. *OCLC Research*. https://doi.org/10.25333/xk7z-9g97.

Panda, S. & Kaur, N. (2023). Exploring the viability of ChatGPT as an alternative to traditional chatbot systems in library and information centers. *Library Hi Tech News*, 40(3), pp. 22-25. https://doi.org/10.1108/LHTN-02-2023-0032.

Prathibha, S.N. & Shilpa Rani, N.R. (2021). ChatGPT: A boon to library services. *LIS Links Newsletter*, 7(1), pp. 8-13. http://newsletter.lislinks.com.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 8748-8763). PMLR. https://proceedings.mlr.press/v139/radford21a.html.

Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982-3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410.

Sabba, F. & Sardo, L. (2020). I fondi personali e la terza missione: proposta di buone pratiche. In Di Domenico, G. & Sabba, F. (a cura di). *Il privilegio della parola scritta* (pp. 427-446). AIB.

Strubell, E., Ganesh, A. & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *arXiv preprint*. https://arxiv.org/abs/1906.02243.

Toccafondi, D. (2010). Gli archivi letterari del Novecento: un laboratorio per la collaborazione tra professionisti. In Desideri, L. & Zagra, G. (a cura di). *Conservare il Novecento. Gli archivi culturali* (pp. 39-46). AIB.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems,* 30.

Verma, M. (2023). Novel study on AI-based chatbot (ChatGPT) impacts on the traditional library management. *International Journal of Trend in Scientific Research and Development (IJTSRD)*, 7(1), pp. 961-964. https://www.ijtsrd.com/papers/ijtsrd52767.pdf.

Vivarelli, M. (2022). Reading practices in the public library space. *DigitCult – Scientific Journal on Digital Cultures*, 7(2), pp. 7-22. https://doi.org/10.36158/97888929562231.

Wani, A.G. & Astunkar, G.S. (2024). Open artificial intelligence (AI) of ChatGPT for library services and library science professionals. *Library Scholar*, 4(1), pp. 1-10.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. *arXiv preprint*. https://arxiv.org/abs/2112.04359.

Yang, S.Q. & Mason, S. (2023). Beyond the algorithm: Understanding how ChatGPT handles complex library queries. *Internet Reference Services Quarterly*. https://doi.org/10.1080/10875301.2023.2291441.

Zakaria, N. & Sani, M.K.J.A. (2024). Implications of ChatGPT in library services: A systematic review. *Environment-Behaviour Proceedings Journal*, 9(SI18), pp. 263-270. https://doi.org/10.21834/e-bpj.v9iSI18.5487.

DOI 10.36158/97912566920714