

Evaluating Retrieval-Augmented Generation for personal collections: architecture, models and criteria*

Angelo La Gorga^(a), Lorenzo Verna^(b)

a) University Carlos III of Madrid, <https://orcid.org/0009-0002-0402-9129>

a) Sapienza University of Rome, <https://orcid.org/0009-0006-1593-5812>

Contact: Angelo La Gorga, angelo.lagorga@unito.it; Lorenzo Verna, v.lorenzo@gmail.com

Received: 17 January 2026; **Accepted:** 30 May 2026; **First Published:** 15 September 2026

ABSTRACT

This article presents an ongoing experiment on the use of Retrieval-Augmented Generation (RAG) architectures in library contexts, considering them not as straightforward extensions of information retrieval techniques but as document mediation devices that require explicit methodological reflection. The contribution focuses in particular on the design and analysis of an integrated evaluation framework, conceived as a structural component of system development rather than as an ex post performance check. Using the personal archive of Emanuele Artom as a case study, characterized by bibliographic and archival heterogeneity, the article examines how the quality of generated responses emerges from the interaction between knowledge base modeling, retrieval strategies and evaluation criteria. The proposed framework combines automatic metrics, LLM-as-a-judge approaches, and human-in-the-loop processes, treating divergences between automated evaluation and expert judgment as diagnostic tools for analyzing system behaviour. Preliminary results indicate that core categories of document mediation, such as relevance, completeness, citability and transparency, cannot be fully reduced to computational parameters but, instead, require continuous negotiation between automated models and disciplinary expertise. From this perspective, RAG is framed as an epistemically unstable research object, whose reliability and governability depend on the robustness and reflexivity of the evaluation processes embedded in its development.

KEYWORDS

Retrieval-Augmented Generation (RAG); Academic Libraries; Digital collection; Personal collection; AI evaluation.

* While the authors worked collaboratively on the text, they specifically focused on the following areas. Angelo La Gorga: Conceptualization, Methodology, Software, Investigation, Data curation, Validation, Formal analysis, Visualization, Writing – original draft. Lorenzo Verna: Conceptualization, Methodology, Software, Validation, Writing – review & editing.

© 2026, The Author(s). This is an open access article, free of all copyright, that anyone can freely read, download, copy, distribute, print, search, or link to the full texts or use them for any other lawful purpose. This article is made available under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. JLIS.it is a journal of the SAGAS Department, University of Florence, Italy, published by EUM, Edizioni Università di Macerata, Italy, and FUP, Firenze University Press, Italy.

Introduction

In recent years, artificial intelligence (AI) has assumed a central role in library and information science discourse, not only as a set of technologies aimed at increasing process efficiency, but above all as a catalyst for structural transformations affecting the entire spectrum of documentary mediation. Recent research in the library field highlights how the introduction of AI-based systems can lead to a redefinition of the ways in which information related to the broader domain of cultural heritage is accessed, queried, and interpreted, impacting the very identity of services and opening up the exploration of new models of participatory knowledge production (Roncaglia 2023a). All this takes place in a context in which libraries are increasingly being reconceptualized as sites of knowledge production and validation, rather than solely as spaces for access to and preservation of resources, in line with orientations identified in studies on professional practices and service development strategies (Cox et al. 2018; Cox 2024). In particular, academic libraries emerge as privileged contexts for experimentation, as they are situated at the intersection between the demands of scientific rigor and the need for interpretative mediation, that is, within a space where technology cannot be adopted without simultaneously interrogating its epistemological premises and methodological implications (Tammaro 2024; Harisanty et al. 2025). This framework further fuels the need for a definition of AI that is meaningful for the library domain, one that enables the design of tools that are not only technically effective but also functionally consistent with the documentary and cognitive values and functions of libraries (Cox and Mazumdar 2024). In parallel, programmatic documents such as the *Statement on Libraries and Artificial Intelligence* issued by IFLA emphasize the urgency of adopting a critical approach to the development and use of AI, placing particular emphasis on issues of transparency and traceability of algorithmic decision-making (IFLA 2020; Lana 2023). These concerns intersect with broader cultural and social dynamics, as demonstrated by recent analyses of public perceptions of AI, which highlight the urgency of developing intelligible and verifiable systems that are consistent with the forms of trust governing access to knowledge in institutional contexts (IPSOS 2024). Within this perspective lies the growing interest in the adoption of Retrieval-Augmented Generation (RAG) architectures, in which the generative capabilities of large language models (LLMs) are anchored to bounded and verifiable data sources. Such systems have a particularly significant impact on two critical issues in generalist language models: citation uncertainty and the production of “plausible” but inaccurate content, which are countered by the requirement to use appropriate and reliable sources and metadata (Roncaglia 2024; Gao et al. 2024). For this reason, experimentation with RAG architectures is of particular interest to documentary institutions, where verifiability is not an ancillary technical feature but rather the very condition of the epistemic legitimacy of informational mediation and heritage narration. The architecture discussed in this article emerges within this research horizon and is developed as part of a PhD project funded by the PNRR (Next Generation EU), embedded within the *Technologies for Cultural Heritage* (T4C) program at the University of Turin. It is also situated within the PRIN project *NEREIDE – New Reading Experiences in Digital Ecosystems* (2022–2025),¹ and aims to experiment with innovative forms of

¹ The NEREIDE project (*NEw Reading Experiences In the Digital Ecosystem*) aims to investigate new reading practices and habits related to the evolution of the digital ecosystem. The project does not focus exclusively or primarily on “digital

digital enhancement of the Historical Library “Arturo Graf” of the University of Turin, while investigating how RAG can contribute to the construction of interpretative pathways grounded in complex and non-homogeneous corpora. This alternative model for representing the informational content of different types of bibliographic and documentary resources neither replaces nor modifies the formal standards of archival or bibliographic representation, which remain fully operative and distinct within their respective disciplinary domains. Rather, it is positioned at a meta-bibliographic and meta-documentary level, in the sense that it makes ex post use of vertical representation models that conform to principles, models and standards that are normalized and validated through practice, without interfering with the internal structure of the original descriptions. Within this framework, the “meta” level should be understood in an operational rather than hierarchical sense: not as a superstructure that subsumes or supersedes existing bibliographic and documentary systems but as an integrative mediation space that enables different modes of interaction. Within this “meta” level, the model thus makes it possible to establish horizontal and transversal relationships of narrative connection among heterogeneous bibliographic and documentary data, preserving the autonomy of individual descriptive systems while fostering dialogue along a functional and communicative axis. In metaphorical terms, the overall architecture can be described as a space in which *the architect and the oracle* are called upon to engage in operational dialogue, echoing the image proposed by Roncaglia, who identifies the tension between formal design and interpretative questioning as one of the defining elements of cultural systems mediated by artificial intelligence (Roncaglia 2023b). By analogy, this approach resonates with reflections on metacataloguing as a space for descriptive integration between different disciplinary traditions, as outlined by Valacchi, while remaining situated on a distinct plane and pursuing different objectives (Valacchi 2022). While metacataloguing aims at a possible synthesis of domain-specific languages, the model proposed here is instead configured as a device for narrative and analytical mediation, oriented toward supporting exploratory and interpretative research processes, without calling into question the solidity or centrality of the underlying descriptive systems. At this stage, the architecture is analyzed not in relation to user interface development, which will be the subject of subsequent reflections, but rather as a cognitive and technological infrastructure whose functioning must be critically assessed before mediation toward users can take place. The choice of the personal collection of Emanuele Artom, preserved in heterogeneous forms and distributed across institutions such as the Historical Library “Arturo Graf,” the CDEC Foundation, the Historical Archive of the University of Turin, the National Museum of the Risorgimento, and the ISTORE-TO,² directly responds to this need: to test the ability of RAG systems to operate on materials that, by their very nature, require interpretation, recomposition and historical contextualization, in line with the *Linee guida sul trattamento dei fondi personali* (AIB 2019). In this sense, the research situates itself within the evolving role of academic libraries in the Third Mission, where the digital enhancement of heritage no longer coincides with accessibility, but rather with the construction of interpretative competencies and publicly oriented narrative processes (Sabba 2019; Cassella 2020).

reading” (i.e., reading on digital devices), but rather on the interaction between different reading practices (both on paper and on screen) and their social, cultural, and informational contexts. <https://filosofiacomunicazionepettacolo.uniroma3.it/ricerca/progetti-di-ricerca/prin/>

² For a detailed overview of the distribution of documentary resources and their census, see Marzal et al. (2025).

To evaluate this architecture, the article proposes an analytical model that combines automatic metrics and expert judgment, acknowledging the interpretative nature of documentary relevance as extensively discussed in information retrieval theory (Voorhees 2000).

The article therefore develops along a trajectory that reflects the progressive construction of the entire system. First, it outlines the process of building the knowledge base related to the Artom collection, highlighting the implications that the heterogeneous nature of the sources entails for a semantic retrieval model. It then presents the adopted RAG architecture, analyzing its technical components and the ways in which they enable the transformation of bibliographic and documentary data into verifiable narrative material. A third section is devoted to defining the evaluation model, conceived not as a final or external phase of development, but as an integrated, design-oriented and methodological component necessary to govern system behaviour. Finally, the analysis of preliminary results from automated evaluation and expert assessment aims to demonstrate how the dialogue between metrics and human judgment constitutes a crucial step for platform tuning and, more broadly, for validating its narrative potential when applied to personal archives and collections.

Knowledge base modelling

The construction of the knowledge base (KB) intended for the RAG platform is oriented toward the creation of an experimental representation of documentary information, conceived to support transversal narrative and thematic paths. While keeping the original bibliographic and archival descriptions fully intact and autonomous within their respective domains, the KB is therefore configured as a parallel view: not descriptive but relational and narrative. In a RAG environment, the primary unit of interrogation does not coincide with the representation, even if relational, of the individual document, but rather with “meanings”, understood as configurations or paths of sense that an LLM can infer and articulate when it is fed with content originating from authoritative and controlled documentary domains, without altering them. From the initial phase, it was necessary to define an internal classification system in order to clearly distinguish the different categories of bibliographic and documentary resources, according to their relationship with the subject of the collection:

1. the books of the Artom bibliographic collection, derived from the inventory of the Historical Library “Arturo Graf”;³
2. the documents originating from the CDEC archival collection, selected for biographical and contextual relevance;⁴
3. the works written by Emanuele Artom, including diaries and other texts published or recovered in digital format;
4. the works on Artom, including historical, commemorative and critical contributions.

³ The descriptions of the individual copies, prepared by Elena Prandi for her master’s thesis, are currently accessible through the University catalogue (Prandi 2014) .

⁴ The documents of the collection and their descriptions are fully available on the CDEC Digital Library website: <https://digital-library.cdec.it/cdec-opac/>.

This simple subdivision does not have the function of an actual bibliographic or archival taxonomy but makes it possible to integrate heterogeneous materials within a single interrogable architecture, while at the same time maintaining the conceptual distinction between models of representation of different nature. The bibliographic segment is by far the most extensive: the Artom collection consists of 823 volumes, each of which has been described on the basis of metadata extracted from the University catalogue and normalized within the KB. With regard to the textual dimension, the digitized volumes, except for those by and about Artom, were processed through a procedure of thematic incremental summarization,⁵ allowing the treatment of even very extensive volumes. The need to reduce informational noise, while at the same time preserving the relevant conceptual structures of each work, is confirmed by contemporary studies on the treatment of long documents in retrieval-augmented systems. Recent research such as *BooookScore* (Chang et al. 2024) shows how the combination of chunking⁶ and the production of incremental summaries is an effective strategy for making complex monographic texts interrogable, preserving thematic coherence and reducing semantic ambiguity. The form of the result is analogous to that observable in Appendix A, where the summary of the novel *La capanna dello zio Tom* shows the type of output obtained: a thematically organized reformulation, capable of reducing informational noise and of making explicit the salient conceptual cores of the work. The works written by Artom and those about him were instead integrated into the KB in their entire textual form. Their integrity is considered indispensable for the reconstruction of the intellectual and political biography of the author and for the analysis of the processes of reception of his figure. The digitized texts were retrieved through authoritative digital sources, in particular from the *Turin Public Domain* project on Internet Archive, and described according to current bibliographic standards, so as to ensure formal coherence with the bibliographic segment of the Artom collection. In this case, the modelling intervention focuses on the integration structure, not on the contents, which are preserved in their entirety.

With regard to the selection of materials originating from the CDEC archival collection, several documentary units directly relevant to the biographical and narrative purposes of the experimentation were identified. In particular, the following were included: *Ultime volontà* (*Last will*), *Note Storiche* (*Historical notes*), *Certificato di morte* (*Death certificate*), the documentation of the *Dal Dosso–Malanga–Peccolo Besso* trial, and a selection of the correspondence of Amalia Segre Artom, mainly relating to the years following her son's death. For these units, beyond the textual content, only the titles and the narrative descriptions provided by the archival domain were retained. This alternative view, devoid of descriptive and normative ambitions, is intended to make these materials integrable within the narrative paths of the KB, where what is brought into relation are not archival or bibliographic structures, but the informational contents of the documents and

⁵ In the context of this contribution, *incremental summarization* refers to the procedure adopted for processing extended texts intended for the retrieval base of the RAG architecture. This procedure is based on the decomposition of content into coherent units, the production of partial summaries, and their subsequent aggregated reworking, aimed at reducing informational noise while preserving relevant thematic and semantic structures. For a detailed discussion of the procedure and its methodological implications, see La Gorga et al. (2025).

⁶ A content segmentation technique consisting in dividing an extended text into smaller, semantically coherent units (chunks) that can be managed individually. In the context of information systems and AI, chunking facilitates indexing, targeted information retrieval, and automated text processing.

their function in reconstructing the socio-cultural context in which the author operated. Overall, the constructed KB takes the form of a representational model aimed at the enhancement of the personal archive and collection through the preparation of thematic, conceptual and narrative paths, and is therefore not attributable to the classical functions of information retrieval, which nevertheless remain fully available through traditional documentary mediation tools. Its purpose is not to identify or describe documents according to traditional library or archival criteria but to enable interrogations that traverse the materials in order to recompose the cultural and biographical fabric of Emanuele Artom.

From a technical point of view, the KB is hosted in PostgreSQL⁷ and articulated into two complementary components. On the one hand, the *pgvector* extension allows for the management of semantic embeddings and thus vector-based querying, based on the textual chunks that constitute the primary unit of semantic retrieval. On the other hand, the SQL database preserves bibliographic metadata associated with each document in normalized form. The adopted architecture, with a relational database for metadata and a vector database for semantic embeddings, reflects contemporary practices for integrated knowledge bases. Studies such as CHASE (Ma et al. 2025) demonstrate that it is technically possible to perform hybrid queries on structured and unstructured data, optimizing both semantic accuracy and query efficiency. Other recent models for multimodal data (Chandra et al. 2025) confirm that the coexistence of vector similarity and relational queries represents a robust approach suited to complex and stratified collections. Moreover, open-source tools such as PostgreSQL with vector extensions practically demonstrate how to integrate keyword search, metadata filtering, and semantic search within a single infrastructure. This paradigm therefore offers the dual advantage of preserving traditional descriptive structures, necessary for queries compliant with bibliographic and archival models and enabling exploratory research across texts and contents.

The architecture: Retrieval and Generation

The introduction of a RAG architecture within the context of this project does not imply a revision of consolidated practices of bibliographic and documentary mediation, nor does it call into question the role of tools traditionally devoted to the description, organization and access to resources. Catalogues, discovery systems and informational apparatuses, based on formalized descriptive models, continue to constitute the primary infrastructure through which bibliographic and documentary heritage is made accessible and interrogable according to the principles of the library and archival tradition. Within this perspective, the RAG environment is configured as an additional experimental layer, designed to complement these tools and to expand modes of mediation without replacing them. The developed platform provides a prototype that allows users to access the contents of the KB through interaction in natural language. This mode of access is not conceived as a new form of bibliographic search, nor as an alternative to structured querying mechanisms, but rather as a space for guided exploration of the constituted collection. The use of

⁷ PostgreSQL is an open-source relational database management system, widely used for the storage and structured management of complex data in research applications and information services.

natural language makes it possible to formulate discursive requests and to obtain responses constructed from controlled selections of the corpus, fostering a form of engagement oriented toward understanding and contextualization of information. It is at this level that a distinction becomes necessary with respect to information retrieval processes as defined in current conceptual models of bibliographic and documentary description. The querying procedures activated in the RAG environment do not aim at the retrieval of bibliographic entities nor at the return of records, but at the elaboration of responses that treat heritage as a structured set of texts and metadata, brought into relation according to specific informational requests. In coherence with this approach, the role of metadata is also interpreted in an extended sense. While retaining their descriptive and identifying functions, they are integrated into a representation model that enhances their informational potential in relation to the textual content and semantic connections expressed in the KB. This results in modes of access and retrieval oriented toward thematic and contextual dimensions, enabling an information output consistent with the interpretative and enhancement purposes of a personal archive or collection. The described configuration should therefore be understood as a proposal still in the process of definition, oriented toward exploring in a controlled manner the possibilities offered by the integration of documentary representation models and systems of automatic language generation. The architecture was designed as an experimental environment, in which modes of access to content, the role of metadata and the logics of source selection can be observed, evaluated and progressively refined. Within this perspective, attention is focused less on the stability of the technical solution and more on the possibility of testing a methodological framework that, as will be discussed below, finds in the retrieval phase the open challenge to be addressed regarding the stability of RAG architectures (Huang and Huang 2024). It is in this phase, in fact, that the portion of bibliographic and documentary heritage that will actually be made available to generation is determined, thus conditioning informational quality, argumentative correctness and the enabling of citations. This process assumes not only technical but also epistemic relevance, constituting the core of the attempt to respond to critical issues such as limited reliability of responses and opacity of sources, well-known and extensively studied limitations of LLMs (Ji et al. 2023) and considered obstacles to their use in documentary mediation processes (Roncaglia 2023a). The platform architecture, articulated across three layers, React frontend, FastAPI backend, and PostgreSQL database extended with *pgvector*, integrates retrieval and generation within a single curatorial logic, as described in the project documentation.⁸ The *SmartAgent*⁹ interprets the user's query and decides whether to activate documentary retrieval through the two available tools: *retrieve_knowledge* for semantic querying of contents and *query_collection_metadata* for structured extraction of bibliographic and collection metadata. Semantic retrieval is based on a vector engine implemented with *pgvector* and on the *text-embedding-3-small* embeddings, which project both queries and previously segmented documents into a distributional space. Segmentation into chunks of approximately 1000 tokens with an overlap of 100 tokens makes it pos-

⁸ <https://github.com/EArtomRag/biblio-rag>

⁹ *SmartAgent* is the designation adopted in this work to indicate a software component designed to operate as an agent within the system. The term refers to the notion of *agent* in the classical artificial intelligence literature, understood as an entity capable of perceiving information from the environment and executing actions oriented toward the achievement of specific goals (Russell and Norvig 2022).

sible to maintain contextual coherence across segmentation thresholds, reducing informational dispersion, which is a known risk of text-splitting operations. During querying, the vectorized query is matched with the most relevant chunks via cosine similarity,¹⁰ according to an approach that is by now well established in the literature on semantic search (Reimers and Gurevych 2019; Karpukhin et al. 2020). In the standard configuration, up to ten chunks are selected, subsequently filtered by similarity and enriched with their respective metadata. The *query_collection_metadata* function extends this mechanism to the domain of structured data, allowing the *SmartAgent* to construct read-only SQL queries in order to respond to questions concerning the composition of the collection: counts, lists, chronological distributions, and other operations typical of bibliographic descriptive statistics. The operational flow unfolds according to a sequence that, while technically defined, reflects a precise methodological intention. The *SmartAgent* first classifies the input, distinguishing simple conversational exchanges from requests grounded in the contents of the collection. Only in the latter case the retrieval tools are activated: the query is transformed into an embedding, sent to the vector database and mapped back to a set of relevant chunks. These fragments, accompanied by their metadata, constitute the documentary context that will be transmitted to the generative model, together with a structured system of citations that allows users to verify precisely the sources employed. Every response that involves the use of the KB must include explicit references to the consulted documents, presented to the user as numbered citations and accompanied by a sources panel that makes it possible to trace back to metadata and, where available, to the textual passages used. The integration of a system of *guardrails*¹¹ further contributes to preventing the emergence of content unsupported by the corpus, mitigating the risk of hallucinations and constraining the production of outputs not compliant with the reference domain. The generation phase constitutes the final step of the RAG pipeline and, in the platform dedicated to the Artom collection, does not merely consist of a linguistic transformation of retrieved information. The agent is modelled as a genuine curatorial figure: a rigorous digital mediator, capable of combining expository clarity and accuracy with a tone described as “*scholarly but warm and engaging*”, designed to promote accessibility for diverse audiences. The construction of its communicative identity, a coherent set of style and expressive rules defined through an extended system prompt, responds to emerging practices of persona design in conversational systems, which show how the definition of a role and a personality contributes to orienting interaction quality and reliability (Pradhan and Lazar 2021). Generation is governed by the *Prompt Builder* (*builder.py*) and by the *Smart Engine*, which dynamically constructs the sequence of messages addressed to the model. When a response requires the use of documents, retrieved chunks are inserted as tool

¹⁰ Cosine similarity is a measure used to quantify the degree of similarity between two vectors in a multidimensional space. It is defined as the cosine of the angle between the vectors: the smaller the angle, the greater the similarity. The resulting value ranges between -1 and 1 (or more commonly between 0 and 1 when vectors are non-negative, as in the case of textual embeddings): values close to 1 indicate strong semantic proximity, while values close to 0 indicate the absence of a relationship. This method is widely used in distributed representations of knowledge to compare words, sentences, documents, or other cultural objects modelled through embeddings, as it captures similarity based on co-occurrence patterns or contextual relationships learned by language models.

¹¹ *Guardrails* are mechanisms external to the language model that filter, constrain, or correct inputs and outputs with the aim of ensuring safety, relevance, and reliability of responses, preventing undesired behaviours or outputs not compliant with the application domain (Bassani and Sanchez 2024).

messages, accompanied by the source category and metadata. This allows the model to operate within a structured and fully attributable context, reproducing in computational form the professional principle according to which a response must be accurate and grounded in verifiable sources, a criterion recognized in international standards for reference evaluation, from *IFLA Measuring Quality* (Poll 2007) to the information quality indicators included in the most recent versions of the ISO 11620:2023 standard (ISO 2023).

The evaluation framework

The development of a dedicated evaluation framework¹² constitutes one of the most methodologically significant elements of the project. This choice is situated within the broader debate on the evaluation of generative AI in documentary and bibliographic systems, a domain in which controlling informational quality and faithfulness to sources is considered indispensable in order to mitigate distortions and phenomena of algorithmic opacity (Sullutrone et al. 2024). The iterative nature of the evaluation adopted in this framework aligns with an orientation that is by now well established, emerging both from empirical analyses of the behaviour of generative models and from theoretical reflection on their use in complex contexts. Practices of cyclic examination of outputs, with subsequent revisions and renewed interaction, show how output quality requires continuous and structured verification (Natale et al. 2025). In parallel, studies on the dynamics of collaboration between humans and language models describe these systems as tools whose effective functioning depends on recursive processes of evaluation and interpretative refinement (Raisch and Fomina 2025). At a more general level, recent investigations on LLM evaluation highlight the limits of static benchmarks and point to the need for continuous, adaptive evaluative approaches rooted in the real tasks models are expected to address, emphasizing that retrieval and generation metrics must be considered jointly in order to capture the actual behaviour of such hybrid architectures (Yu et al. 2025). The objective of evaluation is not only to measure system accuracy but to understand how a generative model interacts with complex bibliographic and documentary information, testing not only the produced results but also the cognitive operations implicitly involved in their generation. From an operational standpoint, the framework is implemented as an application that orchestrates, in a structured manner, all phases of evaluation. Execution is started via the *run* command, based on a configuration suite that specifies the dataset, the RAG system to query and the set of metrics to be computed. A data-loading module reads a .jsonl file in which each line represents either a single query or a multi-turn conversation;¹³ for each of them, the application queries the RAG platform through a language model that simulates user behaviour, thus obtaining the system-generated response and allowing it to be evaluated under the same conditions as a real interaction. For each question–answer pair, a metrics manager then operates, applying in a coordinated way the different families of measurements included in the

¹² <https://github.com/EArtomRag/rag-evaluation-framework>

¹³ An interaction composed of multiple consecutive exchanges between user and model, in which each turn depends on the preceding ones. Unlike a single isolated query, a multi-turn dialogue evaluates the system's ability to maintain context, correctly interpret subsequent questions, manage anaphoric references (e.g., pronouns and synonyms), and ensure discursive coherence across the entire sequence.

framework: retrieval metrics (recall@k, mrr, ndcg@k), which assess the system's ability to rank relevant documents in the top positions, LLM-based RAG quality metrics (faithfulness, answer_relevancy, contextual_precision, contextual_recall) and a core set of operational metrics, such as cost and latency, which document the computational impact of different configurations. Within the second family fall the practices known as *LLM-as-a-judge*, today widely discussed in the literature on generative AI evaluation for their ability to provide judgments that are consistent and often well correlated with human ones (Gao et al. 2025). In this approach, a language model trained or instructed to act as an evaluator, analyzes the system's response, assigns a score according to the specified metric and, when required, provides a textual rationale for its judgment. Recent studies have shown that such evaluators can offer a reliable estimate of the semantic and contextual quality of responses, provided they are employed within adequately controlled evaluation pipelines (Huang et al. 2025). The results generated at this stage, including the system response, retrieved documents, metric scores and, where applicable, the explanations produced by the evaluator LLM, are then saved in structured form so that they can subsequently be analyzed both automatically and within the human review process. A reporting module produces a set of human-readable outputs: an interactive HTML report that displays responses and metrics in tabular form and CSV and JSON files that summarize aggregated results. It is from these artefacts that the *human-in-the-loop* (HIL)¹⁴ procedure is activated: through dedicated commands, the framework exports results to a CSV for human review and, once the evaluation columns have been completed by the involved experts, it analyzes the degree of alignment between manual judgments and those generated by the model, highlighting the cases of greatest disagreement and thereby providing systematic support for critical reflection on automatic metrics. A qualifying characteristic of the framework is, therefore, its hybrid nature: evaluations are conducted via *LLM-as-a-judge* practices, recognized as effective tools for providing scalable and consistent judgments but requiring expert supervision, and are systematically integrated with a *human-in-the-loop* process that ensures interpretative control (Amirizani et al. 2024). The system produces an initial judgment, while final validation is entrusted to a structured human intervention, activated in problematic cases or when the contextual, biographical, or bibliographic dimension requires specialized expertise. The human contribution is not limited to verifying the factual correctness of responses, but includes the assessment of narrative coherence, alignment with descriptive and bibliographic practices, and the contextual robustness of the interpretations provided. The figures involved in the HIL process play distinct and complementary roles. For the biographical dimension, the construction of questions related to documentation on the diaries and the verification of the system's responses in the most delicate cases are conducted in collaboration with Guri Schwartz, Associate Professor of Contemporary History at the University of Genoa and curator of the critical edition of Artom's diary texts, in order to ensure precise philological and historical control over generated reconstructions. For the bibliographic dimension, the definition of questions aimed at testing the platform's ability to transform a user prompt into SQL queries over the

¹⁴ *Human-in-the-loop* (HIL): an evaluation or control strategy in which humans intervene at key points in the process to verify, correct, or validate the results generated by the automated system. In the context of LLM evaluation, the expert's role consists in reviewing the judgments produced by the model, identifying interpretative or contextual errors, and providing a final validation that cannot be reduced to automated measurement alone.

database and to return formally correct representations of bibliographic information is carried out in dialogue with Maurizio Vivarelli, former Full Professor of Bibliography and Library Science at the University of Turin, who also contributes to validating the corresponding answers, verifying their coherence with bibliographic representation models. The *human-in-the-loop* thus operates at two decisive moments: in defining the benchmark and in verifying the judgments produced by the model-as-judge, ensuring an interpretative control that is indispensable in the most complex cases and keeping evaluation anchored to recognized disciplinary criteria. The construction of the benchmark of questions, the so-called golden standard, represents a central step of the framework. Unlike generic benchmarks used in AI, the one developed for this project is conceived as an epistemic device calibrated to the specificities of the Emanuele Artom collection, to actual research practices and to the cognitive operations that a documentary enhancement system is expected to support. The included question categories (factual, synthesis, comparison, SQL, bibliographic and out-of-scope) integrate three perspectives: the effective research practices on the Artom collection, the operations that a RAG system can concretely perform and models elaborated in the literature on semantic retrieval regarding the construction of evaluative scenarios. Each benchmark entry is defined in JSONL format and includes the question, the category, a manually drafted gold answer and a set of reference documents considered primary for generating the response. The construction of the golden standard is an interpretative work, grounded in the awareness that, in complex documentary contexts, a perfect answer does not exist: response evaluation is always situated and negotiated and what matters is not the uniqueness of the result but its justifiability with respect to the sources and the adopted relevance criteria. As shown both by the consolidated tradition of studies on informational relevance, which describe relevance judgment as an intrinsically variable, multidimensional, and context-dependent process (Barry and Schamber 1998; Saracevic 2007) and by more recent analyses that highlight the interpretative plurality of possible answers in retrieval and generation systems (Peikos and Pasi 2024), what can be asked of a gold standard is not the production of a definitive truth but the formulation of grounded and coherent answers with respect to the reference context. The framework does not only produce measurements but constitutes the engine of continuous platform improvement in which errors are classified as retrieval, generation or evaluation and then this taxonomy feeds iterative cycles of system revision. Evaluation outcomes thus guide prompt reformulation, optimization of search parameters, enrichment of the conceptual schema and, when necessary, redefinition of the benchmark categories themselves. The quantitative and qualitative results returned by automatic metrics and by comparison with human evaluation are not treated as definitive outcomes but as diagnostic elements that make it possible to interrogate system behaviour and to iteratively orient design decisions. The application of the architecture to a heterogeneous and stratified documentary corpus such as the Emanuele Artom collection makes this orientation particularly evident. In a context of such complexity, no single metric is able to exhaustively capture system functioning. Dimensions such as the accuracy and completeness of responses, their anchoring to sources and their adequacy with respect to the required cognitive operations emerge only if evaluative results are interpreted at a sufficiently granular level. For this reason, attention here is not placed on aggregated scores but on individual queries, considered as analytical units through which to observe the interaction between retrieval, generation and evaluation. It is at this step that the systematic comparison between automatic metrics and expert human judgment assumes a central role. Metrics provide

scalable and reproducible indicators, useful for identifying trends and recurring behaviours but tend to privilege formal aspects of coherence and semantic alignment. Human evaluation, instead, makes it possible to interpret these indicators in light of the cognitive operations required by the questions and the criteria specific to bibliographic and documentary mediation, making visible gaps and ambiguities that would otherwise remain opaque. It is precisely from the interaction between these two levels of observation that behavioural patterns attributable to specific architecture configurations emerge. The examples discussed here, collected in the table in Figure 1, are read, from this perspective, as diagnostic cases and therefore not as resolved problems or demonstrations of consolidated solutions but as situations from which concrete lines of work are being explored. The tuning hypotheses currently under exploration focus in particular on recalibrating retrieval strategies according to question types and on strengthening the recognition of complex documentary and bibliographic entities, alongside interventions on generation prompts and experimentation with evaluative criteria that are more sensitive to phenomena of incompleteness and to the explicit handling of not-knowing.

Tipo di domanda	Domanda	Risposta dell'LLM	Valutazione LLM-as-a-judge	Esito valutazione umana (Likert)
synthesis	Che cosa rivela la composizione della biblioteca di Artom riguardo ai suoi interessi intellettuali?	La composizione della biblioteca di Emanuele Artom riflette un interesse profondo per la storia antica, la cultura ebraica e la storia locale, in particolare valdostana. Laureato in lettere e studioso di storia, Artom orientò i suoi studi verso ambiti che coniugano ricerca storica e impegno civile. L'epigrafe posta nella biblioteca sottolinea come egli "alimentò dell'esperienza antica la sua speranza in una libera patria", suggerendo una connessione tra conoscenza del passato e ideali di libertà.	Faithfulness: 1.0 Answer relevancy: 1.0	Faithfulness: 3 - Parzialmente adeguata Answer relevancy: 4 - Adeguata
factual	Chi erano i membri della famiglia di Emanuele Artom?	Emanuele Artom era figlio di Emilio Artom e aveva un fratello di nome Ennio Artom. Il padre Emilio si sposò nel 1914 con la donna che sarebbe stata la madre di Emanuele. Emanuele nacque il 23 giugno 1915, mentre Ennio nacque il 25 aprile 1920 e morì il 29 luglio 1940. Non risultano menzionati altri fratelli o sorelle nelle fonti consultate.	Faithfulness: 1.0 Answer relevancy: ~0.56	Faithfulness: 3 - Parzialmente adeguata Answer relevancy: 3 - Parzialmente adeguata
comparison	Come varia la rappresentazione di Artom come intellettuale rispetto a quella come partigiano nelle diverse lettere di commemorazione?	Le commemorazioni intrecciano la figura di Artom come studioso e come partigiano. Da un lato viene ricordato come intellettuale raffinato, laureato in Lettere e studioso di storia e cultura ebraica; dall'altro come combattente della Resistenza, commissario politico del Partito d'Azione, esempio di coraggio e coerenza morale. Le due dimensioni non sono contrapposte, ma presentate come complementari: la cultura alimenta l'impegno civile e l'esperienza partigiana conferisce valore esemplare alla figura dello studioso.	Faithfulness: 1.0 Answer relevancy: 1.0	Faithfulness: 5 - Pienamente adeguata Answer relevancy: 4 - Adeguata
out_of_scope	Quale fu il primo libro che Artom lesse da bambino?	Dai documenti attualmente disponibili nella collezione, non emerge un riferimento esplicito al primo libro che Emanuele Artom lesse da bambino. Nei testi biografici e nei suoi scritti sono presenti descrizioni della sua infanzia e del suo carattere precoce, ma non viene menzionato il titolo del primo libro letto.	Faithfulness: 1.0 Answer relevancy: 0.5	Faithfulness: 5 - Pienamente adeguata Answer relevancy: 5 - Pienamente adeguata

Figure 1. The table reports examples of the evaluation outcome for some questions.

When a question requires a synthesis operation based on the composition of Emanuele Artom's personal library, system behaviour highlights some significant tensions. The response generated by the system appears, on first reading, solid and well constructed: the *LLM-as-a-judge* assigns it maximum scores for the analyzed metrics, on the basis of full coherence with retrieved contexts. It is the human judgment that makes visible a gap that automatic metrics alone do not capture with the same clarity. While recognizing the overall plausibility of the response and the correctness of the recalled information, the human reviewer assesses faithfulness as only partially adequate and answer relevancy as adequate but not fully centered. The critical point does not concern what is asserted but rather the argumentative path through which the response is built. The analysis of Artom's intellectual interests is in fact based primarily on a biographical and symbolic reconstruction (the education, scholarly activity, the value of the commemorative epigraph) whereas the ques-

tion presupposes a different kind of synthesis operation, grounded in the composition of the collection, in the works actually present and in the represented disciplinary areas. In the corpus, such information is available, also in the form of thematic summaries produced during knowledge base modelling but it is not activated with priority, thus leaving space for a generation phase that tends to compensate for this gap through highly plausible narrative inferences. This case made visible a structural tension between discursive coherence and documentary anchoring, orienting a set of intervention hypotheses that are currently being explored. In particular, work is underway on the possibility of strengthening the recognition of the library as an autonomous documentary entity within the knowledge base, emphasizing its distinction from biographical and paratextual materials. In parallel, differentiated retrieval strategies are being tested according to question type, with the aim of privileging, for queries of this kind, descriptive documents about the works and thematic summaries. On the generation side, this example also suggested the need to refine prompts in order to make the required level of analysis more explicit and to constrain recourse to inferences not directly supported by activated sources. The objective is not to limit the model's interpretative capacity but to orient it in a more controlled way, making recognizable the distinction between what is deduced from the documentary composition of the collection and what instead derives from a more general reconstruction of the subject's profile. The gap between automatic evaluation and human evaluation thus assumes an operational value, since it makes visible a system configuration that requires a targeted refinement of the architecture. If in the previous case the criticality emerged from an excess of narrative inference, the analysis of a factual question makes it possible to observe a different tension, related to the handling of informational completeness. When asked to identify members of Emanuele Artom's family, the system returns a response that, at the formal level, appears correct and coherent with activated contexts. The *LLM-as-a-judge* assigns a maximum score to faithfulness, recognizing the absence of explicit contradictions, while penalizing answer relevancy due to the presence of accessory information not strictly necessary with respect to the question. Human evaluation operates instead on a different level, shifting attention from the formal coherence of the response to its overall informational coverage. While detecting no factual errors, the reviewer assigns both faithfulness and answer relevancy a judgment that is only partially adequate, highlighting a lack of exhaustiveness with respect to the information actually available in the corpus, particularly in relation to the omission of Artom's mother, Amalia, who is extensively documented in the collection. From a documentary perspective, this case highlights a critical passage already glimpsed in the first example but articulated differently: a retrieval limit is not made explicit but is instead absorbed into the response in the form of an epistemic conclusion. The response asserts nothing false but produces an effect of completeness that does not reflect the real state of information available in the corpus. This type of behaviour escapes metrics oriented toward detecting contradictions such as faithfulness but emerges clearly in human evaluation, which is more sensitive to the requirements of exhaustiveness that characterize factual questions. In continuity with the previous case but shifting the axis from inference to coverage, this example shows how the quality of factual responses depends not only on formal correctness but on the system's ability to represent transparently what it knows and what, at that moment, it is not able to reconstruct. Comparison between automatic evaluation and human judgment makes this tension observable and traces it back to a system configuration that requires a targeted intervention on retrieval. The analysis of the next case makes it possible to

observe a different situation, showing a strong alignment between automatic evaluation and human judgment. The request to compare the representation of Emanuele Artom as an intellectual and as a partisan in commemorative letters triggers a response that the *LLM-as-a-judge* evaluates as fully adequate. The reviewer confirms the result, recognizing in the response a high degree of documentary anchoring, made explicit by the citation system that directly refers to retrieved contexts, and a correct interpretation of the cognitive operation required by the question. The comparison between the two dimensions of Artom's figure is not resolved into a schematic opposition but returns the intertwining of cultural formation and political action that characterizes commemorative sources. The observed alignment depends on the comparative strategy adopted in the response, which privileges a thematic synthesis rather than a point-by-point comparison document by document. Human evaluation recognizes this choice as legitimate and coherent with the question, but implicitly lets a margin of interpretative ambiguity emerge: the request for "variation" could also be read as prompting a more analytical comparison between individual letters, in addition to a transversal synthesis. Precisely for this reason, the case assumes a relevant methodological function, offering a reference point for calibrating system behaviour in the presence of complex comparative operations. On the one hand, it confirms that current retrieval and generation strategies are able to support articulated comparisons when relevant documents are correctly activated and the level of synthesis is adequate; on the other, it is orienting reflection on how to make the adopted comparative strategy more explicit at the generation stage. In this sense, alignment between automatic evaluation and human judgment does not represent an endpoint but a tool for verifying the overall coherence of the architecture and a basis for experimenting with controlled refinements. System behaviour can also be observed in the presence of questions that fall outside the informational domain of the corpus, as in the case of the request to identify the first book read by Emanuele Artom as a child. In this situation, the generated response does not attempt to fill the informational gap through inferences or conjectures but explicitly declares the absence of the datum in the available documents, distinguishing between what is attested in the corpus and what does not appear documented. Automatic metrics return a differentiated outcome, with high faithfulness scores and a penalization of answer relevancy, linked to the absence of a positive informational content. In a purely extractive reading, relevance is thus associated with the ability to provide positive content, regardless of the nature of the question. Human evaluation interprets the same behaviour in light of different criteria, linked to the documentary nature of the question, judging relevance as fully adequate as well, showing how, in the presence of an out-of-domain question, the most correct response consists in the transparent communication of the system's knowledge limit. From the human reviewer's standpoint, the response is relevant precisely because it informs the user about the boundaries of the corpus and the documentary nature of the available information, avoiding the introduction of speculative elements that would compromise the system's overall reliability. This example makes particularly evident a conceptual gap between certain automatic metrics and human evaluative criteria. Metrics such as answer relevancy, if interpreted in a non-contextualized way, can penalize negative answers or answers that explicitly state the absence of information, even when such answers represent the epistemically most appropriate behaviour. Comparison with human evaluation instead makes it possible to reformulate the very meaning of relevance according to question type, including the system's ability to recognize and make explicit its own limits as a positive indicator of quality. This case

does not highlight a criticality of the RAG architecture in the strict sense but rather a configuration in which system behaviour is epistemically adequate, while putting into tension the criteria through which such behaviour is evaluated. The interest of the example therefore lies in the possibility of interrogating not only generation strategies, but also the functioning of the metrics themselves and, more generally, of the evaluation process. In this perspective, ways are being explored to refine automatic metrics, distinguishing more explicitly between incomplete answers and correctly out of scope answers, for example through dedicated labels or evaluative criteria differentiated according to question type. The case thus shows how the refinement of the RAG architecture and that of the evaluation framework proceed in an interdependent way: it is not only the system that is evaluated, but the evaluative process itself that is continuously recalibrated in light of observed behaviours, making explicit the complementary nature of the two levels.

Conclusions

The work presented describes an ongoing experimentation process aimed at prototyping a Retrieval-Augmented Generation architecture in the library domain. At this stage, the objective is to explore, through a circumscribed and controlled case study, the methodological conditions that make it possible to meaningfully observe the behaviour of a generative system applied to a complex bibliographic and documentary heritage. The adoption of RAG is assumed from the outset as a problem of informational and communicative mediation, rather than as an extension of information retrieval models and techniques. This entails a shift in focus from the apparent quality of answers to the relationship they establish with sources, representational structures and the cognitive operations implicit in the questions. The construction of the knowledge base, the design of retrieval and the configuration of generation are not treated as independent components but as elements of a single methodological space, in which each design choice shapes the kind of knowledge the system is able to produce and make intelligible. The main contribution of the research, therefore, lies in the way evaluation is integrated into the development process itself, operating not as an *ex post* verification but as an analytical instrument that makes the system's structural tensions visible. The systematic comparison between automatic metrics and human judgment is not intended to determine which of the two approaches is more reliable but to use their misalignment as an interpretative key for understanding the architecture's behaviour and guiding its refinement. What emerges is an idea of RAG as an intrinsically unstable experimental environment, whose functioning cannot be adequately described through aggregated scores or generic benchmarks. The relevance of individual answers depends on the type of documentary operation elicited, the role assigned to sources and the system's ability to make explicit what is effectively supported by the available documents. Non-answers or explicitly limited answers thus also assume a positive epistemic value, challenging evaluation metrics that may tend to identify quality with the presence of explicit informational content. The architecture therefore appears less as a tool to be "applied" to libraries and more as a research object that enables reflection on consolidated categories of documentary mediation (relevance, exhaustiveness, citability, transparency) and, in particular, on their epistemological and technical articulation in the context of interaction with generative AI. The Artom collection case shows how such categories cannot simply be translated into computa-

tional parameters but require a process of methodological negotiation between automatic models and disciplinary expertise. It is within this negotiation, more than in the specific results of the experimentation, that the contribution of this study is located: in the attempt to define a framework of observation and evaluation capable of making the use of generative systems in the enhancement of the informational contents of bibliographic and documentary heritage analyzable, and therefore governable.

References*

- AIB (Associazione italiana biblioteche). 2019. 'Linee guida sul trattamento dei fondi personali.' <https://www.aib.it/documenti/linee-guida-sul-trattamento-dei-fondi-personali/>.
- Amirizani, Maryam, Jihan Yao, Adrian Lavergne, Elizabeth Snell Okada, Aman Chadha, Tan-ya Roosta, and Chirag Shah. 'LLMAuditor: A Framework for Auditing Large Language Models Using Human-in-the-Loop.' Preprint, submitted February 14, 2024. <https://doi.org/10.48550/arXiv.2402.09346>.
- Barry, Carol L., and Linda Schamber. 1998. 'Users' Criteria for Relevance Evaluation: A Cross-Situational Comparison.' *Information Processing & Management* 34 (2): 219–36. [https://doi.org/10.1016/S0306-4573\(97\)00078-2](https://doi.org/10.1016/S0306-4573(97)00078-2).
- Bassani, Elias, and Ignacio Sanchez. 2024. 'GuardBench: A Large-Scale Benchmark for Guardrail Models.' In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, edited by Yaser Al-Onaizan, Mohit Bansal and Yun-Nung Chen. Miami (USA): Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.1022>.
- Cassella, Maria. 2020. *Biblioteche Accademiche e Terza Missione*. Milano: Editrice Bibliografica.
- Chandra, Joydeep, Satyam Kumar Navneet, and Yong Zhang. 'The Hybrid Multimodal Graph Index (HMGI): A Comprehensive Framework for Integrated Relational and Vector Search.' Preprint, submitted October 11, 2025. <https://doi.org/10.48550/arXiv.2510.10123>.
- Chang, Yapei, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 'BoookScore: A Systematic Exploration of Book-Length Summarization in the Era of LLMs.' Preprint, submitted October 1, 2023. <https://doi.org/10.48550/arXiv.2310.00785>.
- Cox, Andrew M. 2024. 'Artificial Intelligence and the Academic Library.' *The Journal of Academic Librarianship* 50 (6): 102965. <https://doi.org/10.1016/j.acalib.2024.102965>.
- Cox, Andrew M., and Suvodeep Mazumdar. 2024. 'Defining Artificial Intelligence for Librarians.' *Journal of Librarianship and Information Science* 56 (2): 330–40. <https://doi.org/10.1177/09610006221142029>.
- Cox, Andrew M., Stephen Pinfield, and Sophie Rutter. 2018. 'The Intelligent Library: Thought Leaders. Views on the Likely Impact of Artificial Intelligence on Academic Libraries.' *Library Hi Tech* 37 (3): 418–35. <https://doi.org/10.1108/LHT-08-2018-0105>.
- Gao, Mingqi, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. 'LLM-Based NLG Evaluation: Current Status and Challenges.' *Computational Linguistics* 51 (2): 661–87. https://doi.org/10.1162/coli_a_00561.
- Gao, Yunfan, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, Haofen Wang. 2024. 'Retrieval-Augmented Generation for Large Language Models: A Survey'. arXiv:2312.10997. Preprint, submitted December 18, 2023. <https://doi.org/10.48550/arXiv.2312.10997>.

* Last date of access to the websites: February 25, 2026.

Harisanty, Dessy, Nove E. Variant Anna, Tesa Eranti Putri, Aji Akbar Firdaus, and Nurul Aida Noor Azizi. 2025. 'Is Adopting Artificial Intelligence in Libraries Urgency or a Buzzword? A Systematic Literature Review.' *Journal of Information Science* 51 (2): 511–22. <https://doi.org/10.1177/01655515221141034>.

Huang, Hui, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, Tiejun Zhao. 'An Empirical Study of Llm-as-a-Judge for Llm Evaluation: Fine-Tuned Judge Model Is Not a General Substitute for GPT-4'. Preprint, submitted March 5, 2024. <https://doi.org/10.48550/arXiv.2403.02839>.

Huang, Yizheng, and Jimmy Huang. 'A Survey on Retrieval-Augmented Text Generation for Large Language Models'. Preprint, submitted April 17, 2024. <https://doi.org/10.48550/arXiv.2404.10981>.

IFLA (International Federation of Library Associations and Institutions). 2020. 'IFLA Statement on Libraries and Artificial Intelligence.' <https://repository.ifla.org/handle/20.500.14598/1646>.

ISO (International Organization for Standardization). 2023. *ISO 11620:2023*. <https://www.iso.org/standard/83126.html>.

IPSOS. 2024. *Ipsos AI Monitor 2024: opinioni e atteggiamenti sull'Intelligenza Artificiale* | Ipsos. <https://www.ipsos.com/it-it/ipsos-ai-monitor-2024-opinioni-atteggiamenti-intelligenza-artificiale>.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Yu Tiezheng, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. 'Survey of Hallucination in Natural Language Generation.' *ACM Computing Surveys* 55 (12): 1–38. <https://doi.org/10.1145/3571730>.

Karpukhin, Vladimir, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, Wen-tau Yih. 2020. 'Dense Passage Retrieval for Open-Domain Question Answering.' In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, edited by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. <https://doi.org/10.18653/v1/2020.emnlp-main.550>.

La Gorga, Angelo, Roberto Testa, and Lorenzo Verna. 2025. 'LLM e Retrieval Augmented Generation (RAG) per le biblioteche. Sperimentazioni, prospettive e valorizzazione del patrimonio.' *DigitCult - Scientific Journal on Digital Cultures* 10 (1): 59–73. <https://doi.org/10.36158/97912566920714>.

Lana, Maurizio. 2023. 'Leggere l'IFLA Statement on Libraries and Artificial Intelligence al tempo di ChatGPT'. *Biblioteche Oggi Trends* 9 (1): 4–12. <https://doi.org/10.3302/2421-3810-202301-006-1>.

Ma, Rui, Kai Zhang, Zhenying He, Yinan Jing, X. Sean Wang, and Zhenqiang Chen. 'CHASE: A Native Relational Database for Hybrid Queries on Structured and Unstructured Data.' Preprint, submitted January 9, 2025. <https://doi.org/10.48550/arXiv.2501.05006>.

Marzal, Miguel Ángel, Angelo La Gorga, and Maurizio Vivarelli. 2025. 'Enriquecimiento del valor de los fondos personales en las bibliotecas universitarias: el fondo "Emanuele Artom" entre archivos, bibliotecas y repositorios digitales.' *Revista Española de Documentación Científica* 48 (1): 1653. <https://doi.org/10.3989/redc.2025.1.1653>.

Natale, Simone, Bruno Surace, Enrico Mensa, and Luca Befera. 2025. 'ChatGPT for Cultural Heritage and the Customization of Generative AI: A Talkthrough Analysis of the Luigi Einaudi Chatbot.' *New Media & Society*. <https://doi.org/10.1177/14614448251384258>.

Peikos, Georgios, and Gabriella Pasi. 2024. 'A Systematic Review of Multidimensional Relevance Estimation in Information Retrieval.' *WIREs Data Mining and Knowledge Discovery* 14 (5). <https://doi.org/10.1002/widm.1541>.

Poll, Roswitha. 2007. *Measuring Quality: Performance Measurement in Libraries*. Munich: K.G. Saur. <https://doi.org/10.1515/9783598440281>.

Pradhan, Alisha, and Amanda Lazar. 2021. 'Hey Google, Do You Have a Personality? Designing Personality and Personas for Conversational Agents.' In *CUI '21: Proceedings of the 3rd Conference on Conversational User Interfaces, New York, USA, July 27: 1–4*. <https://doi.org/10.1145/3469595.3469607>.

Prandi, Elena. 2014. *Emanuele Artom e i suoi libri. Analisi bibliografica del fondo conservato presso la Biblioteca 'Arturo Graf' di Torino*. <https://unitesi.unito.it/handle/20.500.14240/60978>.

Raisch, Sebastian, and Kateryna Fomina. 2025. 'Hybrid Problem-Solving with Large Language Models: A Reply to "Iterative Alternative Evaluation" and "an Assemblage Perspective".' *Academy of Management Review* 50 (2): 482–84. <https://doi.org/10.5465/amr.2024.0300>.

Reimers, Nils, and Iryna Gurevych. 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)*, edited by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. <https://doi.org/10.18653/v1/D19-1410>.

Roncaglia, Gino. 2023a. 'Intelligenze artificiali generative e mediazione informativa: una introduzione.' *Biblioteche oggi Trends* 9 (1): 13. <https://doi.org/10.3302/2421-3810-202301-013-1>.

Roncaglia, Gino. 2023b. *L'architetto e l'oracolo: Forme digitali del sapere da Wikipedia a ChatGPT*. Bari; Roma: Laterza.

Roncaglia, Gino. 2024. 'L'intelligenza artificiale generativa multimodale in ambito umanistico.' @ *DigitCult* 8 (2): 127-37. <https://dx.doi.org/10.36158/97888929589208>.

Russell, Stuart, and Peter Norvig. 2022. *Artificial Intelligence: A Modern Approach*. Hoboken: Pearson India Education Services Private Limited.

Sabba, Fiammetta. 2019. 'Third Mission, Communication, and Academic Libraries.' *Bibliothecae. It* 8 (2): 219–54. <https://doi.org/10.6092/ISSN.2283-9364/10368>.

Saracevic, Tefko. 2007. 'Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part II: Nature and Manifestations of Relevance.' *Journal of the American Society for Information Science and Technology* 58 (13): 1915–33. <https://doi.org/10.1002/asi.20682>.

Sullutrone, Giovanni, Riccardo Amerigo Vigliermo, Luca Sala, and Sonia Bergamaschi. 2024. 'Sensitive Topics Retrieval in Digital Libraries: A Case Study of Ḥadīṭ Collections.' In *Linking*

Theory and Practice of Digital Libraries, edited by Apostolos Antonacopoulos, Annika Hinze, Benjamin Piwowarski, Mickaël Coustaty, Giorgio Maria Di Nunzio, Francesco Gelati, and Nicholas Vanderschantz. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72440-4_5.

Tammaro, Anna Maria. 2024. 'Intelligenza artificiale e ruolo dei bibliotecari: riflessioni su due seminari di Jesus Lau.' *Bibelot: notizie dalle biblioteche toscane* 30 (1). <https://riviste.aib.it/bibelot/article/view/14035>.

Valacchi, Federico. 2022. 'The Parts and the Whole. Integrate Knowledge.' *JLIS.It* 13 (3): 1–11. <https://doi.org/10.36253/jlis.it-477>.

Voorhees, Ellen M. 2000. 'Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness.' *Information Processing & Management* 36 (5): 697–716. [https://doi.org/10.1016/S0306-4573\(00\)00010-8](https://doi.org/10.1016/S0306-4573(00)00010-8).

Yu, Hao, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2025. 'Evaluation of Retrieval-Augmented Generation: A Survey.' In *Big Data*, edited by Wenwu Zhu, Hui Xiong, Xiuzhen Cheng, Lizhen Cui, Zhicheng Dou, Junyu Dong, Shanchen Pang, Li Wang, Lanju Kong, and Zhenxiang Chen. Cham: Springer Nature. https://doi.org/10.1007/978-981-96-1024-2_8.