ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Department of Classical Philology and Italian Studies

Second Cycle Degree in
Digital Humanities and Digital Knowledge

From Documents to Dialogue: Design, Implementation and Evaluation of a Question-Answering System for Geoportale Nazionale Archeologia

Dissertation in
Machine Learning for the Arts and Humanities

Supervisor
Prof. Giovanni Colavizza

Co-Supervisors
Prof. Paolo Bonora
Mario Caruso and Simone Persiani, BUP Solutions

Defended by
Lucrezia Pograri

Graduation Session II
Academic Year 2024/2025

Contents

Abstract

This dissertation presents the design and development of a question-answering (QA) system tailored to the Geoportale Nazionale Archeologia (GNA). The research addresses the challenge of extracting relevant information from archaeological documentation using Retrieval-augmented generation (RAG) amd natural language processing (NLP) techniques. The methodology combines transformer-based language models with domain-specific information extraction to enable intuitive, natural language querying of technical documentation related to archaeological data.

Keywords: Digital Humanities · Information Retrieval · Question-Answering Systems · Retrieval-Augmented Generation · Natural Language Processing · AI · Cultural Heritage.

List of Figures

List of Tables

Chapter 1

Introduction

At the swiftly evolving intersection of artificial intelligence (AI) and digital humanities (DH), computational methods have profoundly transformed access to and interpretation of cultural heritage resources. Question-answering systems (QAS), driven by advances in natural language processing (NLP) and retrieval-augmented generation (RAG), have become increasingly significant tools, offering new possibilities for engaging with extensive documentation and complex cultural repositories. This thesis emerges directly from an applied research experience conducted during an internship at BUP Solutions, aimed at exploring the realistic feasibility and effectiveness of these AI technologies in the context of cultural heritage. Specifically, the project focused on developing a specialized question-answering (QA) system for the Geoportale Nazionale Archeologia (GNA), Italy's primary repository for archaeological data.

The motivation for this study initially arose from a concrete, practical challenge: enabling efficient, intuitive, and contextually accurate access to the extensive and often fragmented archaeological documentation hosted by GNA. Archaeologists, heritage managers, and scholars regularly face difficulties in navigating vast volumes of technical reports, field notes, operational procedures, and geospatial data. In response, this project experimented with applying cutting-edge NLP and machine learning (ML) techniques – primarily Transformer-based language models and advanced information retrieval methods – to dynamically retrieve and synthesize relevant information based on user queries expressed in natural language.

Central to the chosen methodology is retrieval-augmented generation (RAG), an approach that significantly enhances traditional QA systems through the dynamic retrieval of domain-specific content, which augments the generative capabilities of language models. Instead of relying solely on internal model knowledge, RAG-based systems integrate external document retrieval with generative text production, resulting in greater reliability and contextually grounded responses – crucial qualities for scholarly and professional uses. While this approach inherently promises increased accuracy and reduced hallucinations compared to purely generative methods, it also involves several complexities and uncertainties, which were encountered firsthand during the development and evaluation phases, as will be discussed in the following chapters.

Rather than adopting a narrowly theoretical or idealized perspective, this study reflects the exploratory and evolving nature of hands-on experimentation, shaped by iterative cycles of trial-and-error, heuristic adjustments, and pragmatic responses to practical constraints such as computational limits, the absence of standardized evaluation benchmarks, and the linguistic complexity of the domain. This process brought to light the persistent tension between the ambitions of AI-driven solutions and the realities of applying them in intricate cultural contexts. In systems like the GNA's AI assistant, the focus necessarily shifts from abstract notions of "understanding" to measurable outcomes: the true test is not whether the system comprehends archaeology in any human sense, but whether it efficiently retrieves relevant information, navigates domain-specific nuances, and supports users in making informed decisions. In describing system capabilities, I am mindful of McDermott's famous warning against *wishful mnemonics* in AI, reminding us to resist the temptation to label what our systems do with grand terms like "understand" and instead to critically assess and communicate the functional scope and limits of their actual achievement (McDermott, 1976).

In light of this reality, this study deliberately avoids overstating the system's semantic or interpretive capabilities. Instead, it foregrounds the project's exploratory nature, acknowledging both methodological achievements and encountered limitations. The outcome represents a pragmatic yet innovative step toward applying AI in the digital humanities, offering insights into the real-world challenges and possibilities of using retrieval-augmented generation in cultural heritage contexts.

Ultimately, this project remains fundamentally hopeful. It demonstrates that even with inherent methodological challenges, AI-driven tools such as RAG-based QAS hold substantial promise for enhancing access to cultural heritage information. Through a transparent presentation of both the strengths and shortcomings discovered during this internship experiment, this thesis aims to contribute realistically yet optimistically to the ongoing dialogue between artificial intelligence and humanistic inquiry, offering a balanced vision of AI's evolving role in supporting cultural heritage scholarship.

Chapter 2

The Evolution of Question Answering Systems

This chapter introduces the foundations of question answering (QA) as both a computer science discipline and an applied technology. Before the emergence of large language models (LLMs)[1], Transformers[2], and modern generative AI[3], question-answering systems (QAS) progressed through distinct paradigms : from symbolic and rule-based architectures to traditional information retrieval (IR) models and early neural networks approaches (Jurafsky and Martin, 2024; Antoniou and Bassiliades, 2022). Early systems depended on domain-specific models, manually curated knowledge bases, keyword retrieval, or engineered features. In recent years, pre-Transformer language models such as BERT and GPT-2/3 have answered questions by extracting or generating responses from static training data alone (Caballero, 2021). While efficient and easy to deploy, these models exhibit factual inaccuracies, shallow contextual understanding, and limited adaptability to new or evolving information (Alanazi et al., 2021). They also frequently hallucinate or generate outdated responses, constrained by their static training corpora.

[1]Large Language Models (LLMs) are advanced AI systems trained on massive text datasets to generate and understand human language. For an accessible overview, see *A Very Gentle Introduction to Large Language Models without the Hype* (Riedl, 2023).

[2]The Transformer is a neural network architecture introduced in 2017 that efficiently models sequential data using a self-attention mechanism. The original paper, *Attention Is All You Need* by Vaswani et al. (2017), provides a foundational introduction.

[3]Generative AI refers to systems capable of producing new content, such as text, images, or audio, based on learned patterns. For more, see the *Stanford AI Index 2025 Report* (Maslej et al., 2025).

Models	QA Approach	Examples / Results
Symbolic / Rule-based (1960s–1980s)	Rule-based, domain-specific, handcrafted knowledge base	BASEBALL, LUNAR, SHRDLU
Early IR Approaches (1990s–mid-2010s)	Keyword retrieval, TF-IDF, BM25, open-domain ranking	TREC QA
Statistical / Seq2Seq (2000s–2018)	N-gram, embeddings, RNN/LSTM, statistical IR	Early neural QA, Reading comprehension in 2010s
Transformer-based	Pre-training, fine-tuning, self-attention	BERT (93% F1 on SQuAD), XLNet
Generative LLMs and agents	Prompting, retrieval-augmented generation, agentic reasoning	GPT-3, RAG pipelines

Table 2: Evolution of QA systems.

2.1 Pre-Transformer Era: Symbolic and Statistical Systems

The development of QAS prior to the rise of Transformers was shaped by several key methodological shifts and technological milestones. The earliest efforts prioritized manually curated knowledge bases and rules-based systems for precise but limited question matching. As the scope of QA expanded, techniques evolved to incorporate large-scale information retrieval methods, statistical modeling, and increasingly complex approaches to feature engineering and answer extraction. This trajectory ultimately set the stage for early neural models that leveraged word embeddings and sequence modeling, gradually moving the discipline toward data-driven architectures and deeper semantic representation. The following sections trace these major trends, illustrating how each contributed to the capabilities and limitations of pre-Transformer QA systems.

2.1.1 Rule-Based Systems (1960s–1980s)

Early QAS relied on highly constrained, domain-specific approaches built around manually constructed knowledge bases. These systems operated within carefully delineated boundaries, matching user questions to a limited set of predefined templates and answer patterns. While this design enabled highly precise responses in their target domains, it also rendered the systems brittle and inflexible – minor variations in user queries or topics outside the encoded scope often resulted in failure to provide meaningful answers.

Expert systems from this era encoded explicit inference rules and logical representations of knowledge, enabling a form of automated reasoning that was fundamentally deterministic. However, these approaches struggled to address ambiguity or generalize beyond the hand-curated domain, and could not scale to larger, more dynamic information environments (*Question answering,* 2025; Jurafsky and Martin, 2024).

Seminal examples of early domain-specific QA systems include:

- **BASEBALL** (1960s): Hand-coded rules and database logic for Major League Baseball[4] questions (Green et al., 1961).
- **LUNAR** (1971): Pattern matching and restricted knowledge base for geological questions about Moon rocks (Woods et al., 1972).
- **SHRDLU**[5] (late 1960s): Symbolic reasoning for a blocks-world robot in a toy domain (Winograd, 1971).
- **Unix Consultant (UC)**[6] and **LILOG**[7] (1980s): Domain-specific QA via linguistic rules and expert knowledge; though both projects remained at the demonstration stage, they contributed to advanced research in computational linguistics.

These early QA systems demonstrated the potential of automated question answering but highlighted the central challenge of balancing precision with generality and scalability. Their evolution would motivate the subsequent shift toward statistical and data-driven approaches (Jurafsky and Martin, 2024; Antoniou and Bassiliades, 2022).

[4]Major League Baseball (MLB) is the leading professional baseball league in North America. It is regarded as the world's premier baseball competition.

[5]SHRDLU was developed at the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) between 1968–70. The software allowed users to interact conversationally with a program that could manipulate, describe, and answer questions about objects in a virtual ̈blocks world ̧ a simplified environment containing various movable blocks. Read more about SHRDLU program here: https://hci.stanford.edu/winograd/shrdlu/.

[6]UC (QA) system, created at U.C. Berkeley (CA), answered queries about the Unix operating system using a hand-crafted knowledge base and could tailor responses to different user types (Wilensky et al., 1988).

[7]LILOG project was as a text-understanding system designed for tourism information in a German city (*Question answering,* 2025).

2.1.2 Classic Information Retrieval Strategies (1990s–mid-2010s)

As the volume of unstructured web data grew, QA moved toward ranking text passages with IR techniques like TF-IDF[8] and BM25,[9] to locate relevant content within large text collections. Open-domain QA systems – such as those in TREC QA[10] (Hirschman and Gaizauskas, 2001) – shifted the focus from structured fact retrieval to returning ranked sentences or extracting answer spans from retrieved passages. These approaches made it possible to scale QA to a broad range of topics and data sources, yet they also introduced notable challenges. Lacking deep understanding of natural language, IR-based QA systems often failed to interpret nuances, synonyms, or complex phrasing, and frequently missed correct answers that did not explicitly match the user's query terms (Antoniou and Bassiliades, 2022; Caballero, 2021).

2.1.3 Statistical Models and Feature Engineering (2000s–2018)

During the 2000s and 2010s, the adoption of n-gram models and statistical IR approaches (cf. TF-IDF, BM25, probabilistic models[11]) enabled reasoning over large corpora, moving beyond hand-crafted rules and enabling automated extraction of candidate answers from vast, unstructured datasets (Manning et al., 2008). The introduction of word embeddings (Word2Vec, GloVe) marked a significant advancement by capturing semantic similarities between words, thereby allowing models to generalize beyond simple keyword matching. These dense vector representations supported the emergence of recurrent neural networks (RNNs) and long short-term memory networks (LSTMs), which facilitated more accurate modeling of sequence and context in reading comprehension and retrieval-based QA tasks (Jurafsky and Martin, 2024).

A major milestone in this era was IBM's *Watson* system, which achieved notable success by winning the *Jeopardy!* challenge in 2011.[12] Watson's *DeepQA* architecture integrated hundreds

[8]TF-IDF (Term Frequency–Inverse Document Frequency) is a statistical method for ranking how important a word is to a document in a collection.

[9]BM25 is a ranking function that improves information retrieval by considering term frequency, document length, and saturation effects.
For more details on TF-IDF and BM25, read *Introduction to Information Retrieval* (Manning et al., 2008).

[10]TREC QA refers to the Question Answering track of the Text REtrieval Conference (TREC), a long-running evaluation series that has set benchmarks for open-domain QA research since 1999. See https://trec.nist.gov/data/qa.html

[11]Language Models for IR (LMIR) – such as n-gram models – estimate the probability of a query being generated by a document's language model. They capture local word dependencies and were widely used in early QA, speech recognition, and spelling correction, (Ponte and Croft, 1998) but were later outperformed by models like RNNs, LSTMs, and Transformers due to their limited handling of long-range context.

[12]The *Jeopardy! challenge* was a high-profile test where IBM *Watson* competed on the American television quiz show *Jeopardy!* against two of the show's greatest human champions. Watson's victory demonstrated significant progress in machine comprehension and open-domain question answering (Wikipedia IBM Watson). In February 2013, IBM announced that *Watson*'s first commercial deployment would assist with utilization

of NLP, IR and ranking components, employing sophisticated pipelines to analyze and combine evidence from diverse sources (Ferrucci et al., 2011). However, despite its advanced design, *Watson* relied on non-generative methods; it synthesized and ranked candidate answers but did not generate free-form responses from scratch.

Simultaneously, semantic QA systems began to emerge, mapping natural language questions to structured queries (e.g. using SPARQL language) executed over knowledge bases like Freebase and DBpedia. These systems required advanced components for entity recognition, relation extraction, and reasoning over symbolic representations. Typical architectures included steps like question analysis, sentence mapping, disambiguation, and query building, enabling automatic translation of natural language into formal queries over RDF data sources. Thanks to the usage of ontology-mapping and linguistic resources (e.g., WordNet), these approaches further bridged the gap between unstructured text and structured knowledge bases (Franco et al., 2020).

Throughout this period, feature engineering played a central role. Techniques such as conditional random fields (CRFs) and support vector machines (SVMs) enabled models to exploit hand-crafted features – including lexical overlap, question type, and answer patterns—to enhance answer extraction from retrieved texts. Hybrid QA systems appeared, combining keywords-based information retrieval methods for unstructured sources with knowledge-base querying for fact-based answers, thereby improving both coverage and precision (Antoniou and Bassiliades, 2022).

This period laid essential groundwork for the deep learning and neural approaches that would soon transform the QA landscape, highlighting the importance of both statistical modeling and intelligent feature design.

2.1.4 Early Neural and Generative Models (Late 2010s)

The late 2010s witnessed the adoption of neural architectures in question answering, building upon the foundational use of word embeddings and recurrent neural networks (RNNs). Embedding methods such as Word2Vec and GloVe allowed systems to capture deeper semantic relationships between words, providing a richer representation of both questions and candidate answers. RNNs, and their improved variants like long short-term memory networks (LSTMs) and gated recurrent units (GRUs), facilitated sequential modeling of language, enabling sys-

management decisions for lung cancer treatment at Memorial Sloan Kettering Cancer Center in New York City, in partnership with WellPoint (now Elevance Health) (Upbin, 2013).

tems to better process and compare question and answer pairs based on their context within a sentence or passage.

Despite these advancements, early neural QA models still faced significant limitations. The reliance on RNNs restricted their ability to effectively model long-range dependencies in text, often resulting in incomplete understanding when questions required reasoning across multiple sentences or broader contexts. While neural models improved matching between questions and answers, their performance remained constrained by the size and variety of the training data.

Around this time, encoder-decoder architectures began to appear in QA research, drawing inspiration from their success in machine translation. These generative models aimed to produce answers by generating sequences of text, rather than simply extracting passages from a source document. However, early generative QA systems often struggled with factual consistency: they had a tendency to copy or paraphrase the input text rather than synthesizing novel, precise answers. Additionally, these models sometimes hallucinated information or failed to maintain logical coherence in their generated responses, limiting their reliability in open-domain settings (Caballero, 2021).

These developments set the stage for the subsequent breakthroughs brought about by attention mechanisms and transformer-based architectures, which dramatically improved the handling of context and factuality in generative QA.

2.2 Blind Spots and Bottlenecks: The Shortcomings of Early Approaches

Earlier approaches to question answering were hindered by several fundamental limitations. Most notably, symbolic and rule-based systems suffered from severe domain restrictions, as their performance relied on hand-crafted knowledge bases and rigid rules that did not generalize well to new or broader topics (Alqifari, 2019). The brittleness of these systems was further exposed by their heavy dependence on template matching, which frequently led to failures when users phrased questions in unanticipated ways or employed linguistic variations (Hirschman and Gaizauskas, 2001). Information retrieval (IR) and statistical models, while more scalable, continued to struggle with true semantic understanding and contextual reasoning, often retrieving only superficially relevant snippets rather than synthesizing comprehensive or contextually rich answers (Alanazi et al., 2021; Diefenbach et al., 2018). The answers these systems produced were typically shallow, extracted verbatim from source texts rather than generated or adapted to the user's specific information need (Hirschman and Gaizauskas, 2001; Alqifari, 2019).

Substantial manual effort was required to design, maintain, and update rules, features, and parsers, creating significant bottlenecks and making adaptation to new domains costly and time-consuming (Alanazi et al., 2021). In addition, IR and knowledge base approaches frequently exhibited incomplete coverage, missing relevant answers due to differences in phrasing or limitations in their underlying datasets (Diefenbach et al., 2018). Early neural models, despite improvements, were generally confined to handling short text spans and struggled with complex or multi-sentence reasoning tasks. Finally, all these methods exhibited a strong dependence on the quantity and quality of available training data and engineered features, resulting in inconsistent performance across different domains and question types (Liu et al., 2022; Alanazi et al., 2021; Alqifari, 2019; Diefenbach et al., 2018; Hirschman and Gaizauskas, 2001).

These cumulative factors left pre-generative QA systems largely inflexible and brittle, with limited ability to provide context-aware, nuanced, or creative responses to user queries.

2.3 Deep Learning Breakthroughs

The advent of the Transformer architecture fundamentally reshaped the field of deep learning and revolutionized neural QAS. Introduced by Vaswani et al. in 2017, Transformers replaced RNNs and LSTMs with a self-attention mechanism that could model relationships between words regardless of their distance in the input sequence. This innovation allowed for efficient parallelization during training and inference, dramatically improving the scalability and performance of language models on a range of natural language processing tasks, including QA.

One of the earliest and most influential Transformer-based models was BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019). BERT employs a bidirectional attention mechanism and is pre-trained using a masked language modeling objective, allowing it to capture complex context from both directions in a sentence. When fine-tuned for question-answering benchmarks such as SQuAD (Rajpurkar et al., 2016), BERT achieved unprecedented accuracy – reaching Exact Match and F1 scores above 85% and 87% respectively on the SQuAD 2.0 leaderboard –, surpassing previous neural models and establishing a new standard for QA (Li and Zhang, 2024).

Building on this foundation, subsequent models explored variations and enhancements of the Transformer paradigm. XLNet, for example, employed a permutation-based language modeling objective, enabling it to better capture bidirectional context and achieve state-of-the-art results on several QA benchmarks (Z. Yang et al., 2020). In specialized domains, models such as BioBERT extended the BERT architecture with additional pre-training on biomedical texts, achieving top performance on domain-specific challenges like the BioASQ question answering competition (Yoon et al., 2019). Parallel research into model architectures also produced frameworks such as Dynamic Coattention Networks (DCN), which fused question and context representations through attention mechanisms and iterative decoding, further improving accuracy on reading comprehension tasks (Xiong et al., 2018).

Surveys in literature underscore a pronounced move toward both extractive and generative QA pipelines, with each stage – chunking, embedding, retrieval, and answer generation – now being explicitly modeled and systematically optimized (Farea and Emmert-Streib, 2025). At the same time, interest in conversational and multi-turn QA has grown rapidly, as Transformer-based models demonstrate substantial ability to manage dialogue context and maintain coherent, context-aware interactions with users (Yue, 2025; Antoniou and Bassiliades, 2022). Together, these advances have laid the foundation for generative AI systems and retrieval-augmented approaches that now dominate state-of-the-art QA research.

2.4 Large Language Models, Agents and Modular Pipelines

A clear distinction now emerges between "traditional" QA systems, primarily built upon general-purpose pre-trained language models – such as GPT, BERT, T5, etc. – and the new wave of modular approaches that dynamically retrieve external information sources. Traditional QA encompasses both extractive and generative paradigms, each defined by how they use the model's internal knowledge. Extractive QA models are designed to identify and extract exact answer spans directly from a provided text or document, making them highly effective for fact-based questions and reading comprehension tasks. Generative QA models, in contrast, use natural language generation to produce answers, often synthesizing or paraphrasing responses in ways that may not appear verbatim in the original text. However, despite their success, both of these paradigms are fundamentally limited by the static nature of their training data: they may struggle with rare, rapidly changing, or domain-specific queries, and are prone to hallucinations and outdated information (Farea and Emmert-Streib, 2025).

The latest advances in question-answering are characterized by the emergence of generative LLMs and retrieval-augmented generation (RAG) pipelines. In these systems, a retriever component dynamically searches external and continually updated knowledge bases, while a generator synthesizes fluent, grounded answers by conditioning on the retrieved information. This hybrid approach addresses many of the shortcomings of earlier Transformer-based QA models by significantly enhancing factual accuracy, contextual relevance, and system adaptability. Generative LLMs within the RAG pipeline are able to incorporate real-time knowledge, thereby reducing hallucinated content and providing up-to-date responses, even as external data sources evolve (Yue, 2025; Lewis et al., 2020).

Furthermore, RAG-based QA systems offer practical advantages for scalability. Rather than requiring full model retraining to accommodate new information, they can simply update or expand the external document index or knowledge base. This design allows for the integration of vast and dynamic data resources, enabling high coverage across domains and rapid adaptation to new information needs. At the same time, these benefits come with trade-offs. RAG architectures require more complex infrastructure, including document indexing and retrieval pipelines, which increase computational overhead and system latency compared to traditional, static QA models. As a result, deploying and maintaining RAG-based systems can be more challenging, especially at scale.

Recent research also explores QA agents that operate as multi-stage pipelines and modular systems, capable of orchestrating question understanding, document retrieval, reasoning, and answer synthesis in a coordinated workflow. This evolution illustrates a profound transformation in how question answering is conceptualized, implemented, and applied across a wide range of domains.

Tab. 4 summarizes the functional differences between traditional and RAG-based QA systems, highlighting the shift toward dynamic, retrieval-augmented, and generative approaches that characterize the current state of the discipline.

Feature	**Traditional QAS** *(e.g., BERT, GPT-2/3)*	**RAG QAS** *(Retriever + Generator)*
Knowledge source	Fixed (training data)	Dynamic (external docs/databases)
Answer type	Extracted or generated	Retrieved + generated (synthesized)
Factual accuracy	Limited (can hallucinate or be outdated)	High (grounded in retrieved, up-to-date information)
Contextual depth	Limited	Comprehensive, nuanced
Scalability	Moderate	High (can update external data sources)
Computational cost	Lower	Higher (due to retrieval/generation)
Latency	Lower (faster for simple queries)	Higher (retrieval step adds time)
Complexity of setup	Simpler	More complex to maintain
Adaptability	Less adaptable to new domains	Highly adaptable via updated document index

Table 4: Comparison of traditional vs. RAG question-answering systems. Adapted from https://www.geeksforgeeks.org/nlp/rag-vs-traditional-qa/.

Chapter 3

State of the Art

In the rapidly evolving landscape of artificial intelligence (AI), large language models (LLMs) have demonstrated remarkable results in text generation and understanding. Yet, when applied to real-world tasks such as question answering, these models still face significant limitations. As detailed in the previous chapter, LLMs are prone to hallucinations[1], rely on static and often outdated training data, and offer limited transparency or traceability in their outputs. Additionally, they may struggle to incorporate domain-specific context or organizational knowledge (Vaibhav, 2025). These factors pose challenges for domains, like cultural heritage, GLAM and archaeology, where reliability, provenance, and interpretive rigor are fundamental requirements (Di Marcantonio, 2024).

To address these concerns, retrieval-augmented generation (RAG)[2] has emerged as a crucial methodological advance. It improves the factual grounding and contextual relevance of generated answers, thorugh the integration of external and verifiable knowledge at inference time, thereby reducing the risk of generating fabricated or distorted information (Martineau, 2023). As discussed in Chap. 2, this approach represents a significant step beyond both traditional information retrieval and earlier neural QA models, which were often brittle, domain-dependent, or struggled to adapt to evolving information needs. The adoption of RAG in question-answering reflects a broader evolution within the field: from early symbolic and rule-based systems, through statistical and information retrieval approaches, to today's transformer-based, generative architectures. This shift has transformed not only the technical capabilities of QA systems but also their applicability to complex, heterogeneous knowledge domains. Although initially developed for open-domain question answering and enterprise search (Akkiraju et al., 2024; Jiang et al., 2024; Packowski et al., 2024; R. Yang et al., 2025;

[1]In the context of LLMs, hallucinations refer to outputs that are plausible-sounding but factually incorrect, fabricated, or unsupported by the underlying data or external sources (Harsh and Shobha, 2024).

[2]For more information about RAG technique, see https://en.wikipedia.org/wiki/Retrieval-augmented_generation.

Zhou et al., 2025), RAG pipelines are increasingly adopted in the humanities and cultural heritage contexts. In these sectors, where interpretive rigor, provenance, and information reliability are critical, RAG-based tools support scholars and professionals in navigating vast, fragmented knowledge repositories. While some initiatives employ RAG to analyze sensitive historical materials (Callaghan and Vieira, 2025; Ciletti, 2025; Sergeev et al., 2025; Fan et al., 2025), this thesis explores a distinct application: improving access to procedural and technical documentation, where clarity, consistency, and actionable guidance are the primary objectives.

This chapter therefore provides a comprehensive overview of the state of the art in retrieval-augmented generation, situates RAG within the current research landscape, outlines its core mechanisms, and examines its recent application in the digital humanities.

3.1 Core Mechanisms and Foundations of Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) is a hybrid approach designed to address several critical limitations of traditional LLMs, such as knowledge staleness, limited context awareness, and insufficient output traceability (Vaibhav, 2025; Gao et al., 2024; Gupta et al., 2024). While LLMs excel at producing fluent, human-like text, they often falter when facing domain-specific queries or requests for information beyond their training cutoff. RAG directly addresses these challenges by integrating external information retrieval within the generation process, allowing outputs to be more factual, up-to-date, and grounded in verifiable sources (X. Wang et al., 2024).

At its core, a typical RAG workflow consists of two main stages: **retrieval** and **generation** (ODSC-Community, 2024). The process begins with preprocessing and indexing, where raw data is cleaned, extracted, segmented into manageable "chunks", and encoded into vector representations. These embeddings are then stored in a vector database (e.g., Milvus, Faiss, Qdrant)[3] to facilitate efficient similarity searches. When a user submits a query, it is encoded in the same vector space, and the system retrieves the top-k most relevant chunks from the indexed knowledge base. In the subsequent stage, these retrieved documents are passed as context to a generative language model – often based on Transformer architectures (Vaswani

[3]Cf. glossary in Appendix B.

et al., 2017) – which synthesizes a response that blends the original query with external evidence, producing answers that are both coherent and contextually appropriate (Arslan et al., 2024).

This modular design (Fig. 1) enables the continuous incorporation of domain-specific and current information, overcoming the constraints of static model parameters. Recent contributions have helped to formally systematize the RAG pipeline's, with frameworks delineating specific interdependent modules such as query classification, retrieval, re-ranking, and generation (X. Wang et al., 2024; Gao et al., 2024).

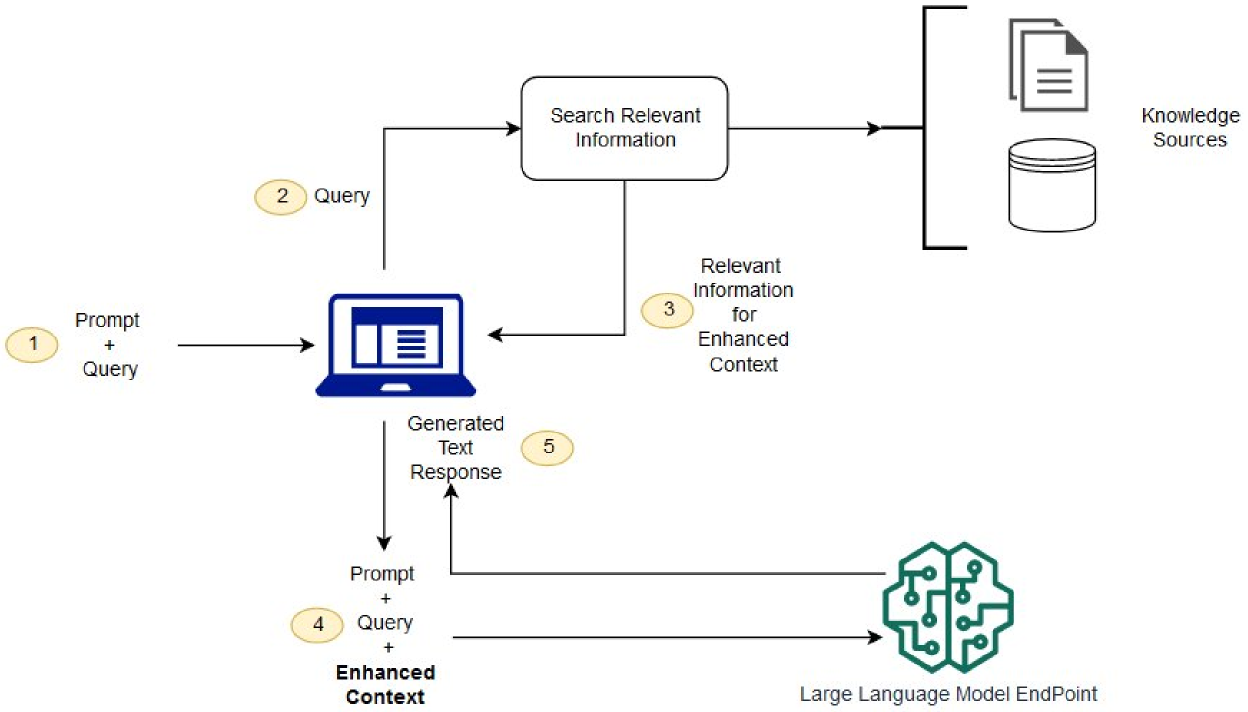

Figure 1: Typical RAG workflow.
Source: https://aws.amazon.com/de/what-is/retrieval-augmented-generation/.

3.1.1 Pipeline Design and Common Practices

The design of RAG pipelines can vary significantly based on the specific use case, domain, and available resources. However, several common practices have emerged that reflect the current state of the art in retrieval-augmented generation (Arslan et al., 2024; Gupta et al., 2024).

The standard workflow includes the following components:

- **Query understanding and classification:** Not all queries require retrieval from external sources. Advanced systems first analyse and classify incoming queries to determine whether retrieval is necessary or if the LLM alone suffices. This step leverages natural language understanding (NLU) techniques to extract key entities, relationships, and user intent, improving efficiency and reducing unnecessary retrieval latency.

- **Document indexing and chunking:** Raw data from source documents is preprocessed: cleaned, segmented into manageable "chunks" at token, sentence, or semantic level, and converted into dense vector representations (embeddings). Recent studies recommend dynamic or semantic chunking over simple fixed-size splitting, as it better preserves context and improves retrieval quality – especially in heterogeneous domains.

- **Embedding and Vector Database:** Both document chunks and user queries are embedded into a shared vector space using models fine-tuned for semantic similarity (e.g., BAAI/bge, LLM-Embedder, intfloat/e5). These vectors are stored in efficient vector databases (e.g., Milvus, Faiss, Qdrant), selected based on scalability, indexing strategies, and support for hybrid (vector plus keyword) search capabilities.

- **Retrieval and query transformation:** Upon receiving a user query, the system encodes it into a vector and retrieves the top-k most relevant chunks from the indexed knowledge base (KB) using similarity search. Robustness is enhanced through hybrid retrieval, which combines dense (vector-based, e.g., DPR, Contriever) and sparse (lexical, e.g., BM25) methods. Advanced query transformation techniques – including query rewriting, decomposition, or hypothetical document generation (e.g., HyDE)– can further improve retrieval effectiveness.

- **Re-ranking:** Initially retrieved candidates are often re-ranked based on relevance to the original query, using additional models (DLM-based) – e.g., cross-encoders like monoT5, monoBERT, or RankLLaMA, which jointly consider the query and each candidate – or more sophisticated algorithms through heuristics. This contextualization ensures that the most pertinent information is prioritized for the generative model.
  See Tab. 6 for a summary on re-ranking techniques.

- **Repacking and summarization:** In some cases, retrieved passages may be reorganized or summarized to distill key information, especially when dealing with lengthy corpora. This step can involve extractive summarization or abstractive (e.g., with Pegasus or T5) techniques to condense information and fit within the context window of the generator model.

- **Generation:** The generative model – usually a transformer-based LLM such as T5, BART, or GPT – synthesizes a response conditioned on both the original query and the retrieved context, integrating intrinsic model knowledge with external evidence to produce a coherent, accurate, and contextually grounded answer.

(Vaibhav, 2025; X. Wang et al., 2024; Gao et al., 2024; Gupta et al., 2024).

Algorithm	Rationale
Cross-Encoders	Joint encoding of query and document for fine-grained relevance scoring.
TILDE (Zhuang and Zuccon, 2021)	Token-level likelihoods for queries across a collection, allowing fast re-ranking by summing the probabilities of query tokens given each candidate passage.
Learning-to-Rank (LTR (Gupta et al., 2024))	Traditional machine learning ranking approaches: **a) Pointwise:** predicts relevance score for each document independently; **b) Pairwise:** compares pairs of documents to learn relative relevance; **c) Listwise:** considers the entire ranked list at once.
Hybrid sparse + dense scoring	Blends scores from dense retrievers (semantic similarity – e.g., DPR, Contriever) and sparse methods (lexical overlap – e.g., BM25, TF-IDF) for robust ranking. Sometimes uses learnable weighting (X. Wang et al., 2024).
Graph-based (Han et al., 2025)	Constructs a graph of candidates (nodes) based on relationships (semantic, citation, or knowledge graph edges), then uses graph algorithms (e.g., PageRank, label propagation) to identify central passages.
Self-RAG (LLM-enhanced reranking) (Asai et al., 2023)	Uses LLMs directly to score or select the most relevant passages, sometimes via few-shot prompting or chain-of-thought reasoning.

Table 6: Algorithms for document re-ranking in RAG pipelines.

3.1.2 Evaluation and Benchmarking

Evaluating RAG systems poses unique challenges, as traditional metrics like BLEU or ROUGE may not fully capture the quality of generated responses, particularly regarding factual accuracy and contextual relevance. To address these shortcomings, researchers have introduced specialized benchmarks and frameworks. For example, the Retrieval-Augmented Generation Assessment System (RAGAS) specifically targets the evaluation of answer faithfulness and contextual alignment, offering a more nuanced perspective on RAG system performance (Es et al., 2023). Human-in-the-loop evaluations have also gained prominence, with expert judges assessing generated responses on criteria such as coherence, factual accuracy, and relevance to original queries – providing richer and more reliable quality assessments than automated metrics alone (Gupta et al., 2024).

Continuous evaluation against established benchmarks like BEIR (Thakur et al., 2021) and TREC (Voorhees et al., 2005) remains essential for tracking performance across diverse retrieval and generation tasks. Regular benchmarking ensures that systems maintain high standards for accuracy, relevance, and user satisfaction. Even after deployment, proactive monitoring is critical. Tracking system performance, user interactions, and error patterns allows teams to address issues quickly and keep RAG solutions reliable and effective as data and usage patterns evolve (Amershi et al., 2019).

Ethical considerations must be embedded throughout the lifecycle of RAG systems. Prioritizing data privacy, addressing algorithmic bias, and complying with regulations such as GDPR are fundamental. This involves designing systems for transparency, explainability, and responsible AI use, as well as ongoing audits to ensure ethical standards are met (Jobin et al., 2019).

The trajectory of RAG systems points toward increasingly sophisticated applications that are deeply integrated into the workflows of research, industry, and cultural institutions. Innovations in evaluation frameworks, user interaction, and system scalability are steadily pushing the boundaries of what these models can achieve. As these technologies continue to mature, success will depend on the ability to combine robust benchmarking, user-centered feedback mechanisms, and adaptive optimization strategies. Addressing challenges related to factuality, scalability, and responsible deployment will be essential for building trustworthy systems capable of delivering high-quality, contextually relevant information. With continued progress, RAG systems are set to play a pivotal role in shaping the future of digital knowledge access and discovery. (X. Wang et al., 2024; Gao et al., 2024).

3.2 New Frontiers Applications

RAG systems are increasingly deployed across diverse domains – spanning academia, enterprise, and product environments – to enhance data accessibility, support decision-making, and facilitate natural language interaction with complex knowledge bases. Recent surveys and empirical studies document a growing array of scholarly applications of RAG, including:

- Automated literature review tools and citation management – e.g., LitLLM; (Agarwal et al., 2025), KNIMEZoBot; (Alshammari et al., 2023);
- Generation of summaries for large corpora of academic papers;
- Field-specific knowledge extraction, including biomedical and legal research support.

In one experiment, a RAG system was developed to assist data scientists through a combination of GROBID library[4] for structured bibliographic extraction, fine-tuned embeddings, semantic chunking, and an abstract-first retrieval strategy. The system's performance, assessed using the Retrieval-augmented generation Assessment System (RAGAS), demonstrated improved faithfulness and context relevance in response generation (Aytar et al., 2024). A similar approach was explored in the context of academic library systems, where RAG was applied to improve contextual retrieval through semantic indexing of structured metadata (e.g., MARC/RDA standards) and multimodal resources. Additionally, the framework introduced conversational querying via a natural language interface, supporting complex interdisciplinary searches and significantly improving document discoverability by synthesizing citation-backed responses from diverse scholarly sources – including journals, datasets, and videos. This solution also addressed challenges such as copyright compliance and ethical AI transparency (Bevara et al., 2025). Collectively, these studies affirm RAG systems' efficacy in alleviating information overload and improving research workflow discoverability.

In parallel, the work of (Soman and Roychowdhury, 2024) provides further critical insights into the design of RAG systems for domain-specific and technical content, closely aligning with the methodological framework adopted in the GNA question-answering system. Using IEEE telecommunications engineering corpora (i.e., wireless LAN specifications and battery glossaries) as testbeds, their analysis highlights key factors influencing retrieval quality, which include chunk size, sentence-level similarity, and the strategic placement of domain-specific terms. These aspects are similarly addressed in the GNA RAG pipeline (Pograri, 2025), which

[4]GROBID is a machine learning library designed to extract, parse, and convert raw documents, like PDFs, structured XML/TEI encoded documents (*GROBID*, 2008–2025).

applies customized tailored chunking, semantic preprocessing, and contextual embedding strategies. Both studies advocate for more nuanced, context-aware approaches to enhance precision in technical and highly structured domains.

Numerous recent graduate-level research projects have provided substantive input into the implementation and evaluation of RAG systems:

- Antolini (2025) developed a custom RAG system for open-domain question answering using both traditional (BM25, PRF) and advanced retrieval strategies, integrated with local LLMs. A novel Parametric RAG (PRAG) approach was also explored, embedding context into model parameters for performance gains.
- Caramanna (2024) investigated conversational agent architectures, comparing various LLM types and retrieval configurations.
- Florio (2024) implemented a LangChain-based RAG chatbot for corporate documentation, evaluating multiple vector database technologies.
- Salcuni (2025) applied RAG to the medical domain, improving LLM responses in hypertension care. The study used RAGAS to assess quality and relevance, focusing on personalization and accuracy.
- Nicoletti (2025) developed Essence Coach, a chatbot that integrates LLMs with the Essence software engineering standard. This system significantly outperformed generic LLMs like GPT-4o in domain-specific reasoning tasks.

3.3 RAG in the Digital Humanities

A growing body of research is exploring RAG applications within the digital humanities. One such example is the *iREAL* project, which applied RAG to interpret archival records from Aboriginal schools in Australia, demonstrating a careful balance between cultural sensitivity and historical accuracy (Callaghan and Vieira, 2025). Another initiative, *ValuesRAG*, focuses on cultural alignment in LLMs by integrating societal and demographic knowledge through retrieval-augmented contextual learning, experimenting with the *World Values Survey* dataset (Seo et al., 2025). In another case, the *Foggia Occupator Dataset* project applied a RAG model to post-WWII Italian periodicals, extracting information on political figures and stylistic traits (Ciletti, 2025).

Among the technical approaches explored in recent experiments on generative AI for digital scholarly editions, RAG emerges as a promising method for addressing challenges such as entity linking (EL) and the integration of external knowledge sources. Notably, RAG is recognized for its ability to mitigate hallucinations in named entity recognition (NER) and to enable the enrichment of text with information from structured databases or knowledge graphs (Pollin et al., 2025). For example, the *Regesta Imperii* project demonstrates how knowledge bases, including Neo4j graph databases, are leveraged within RAG pipelines to improve accuracy in information extraction, entity normalization, and semantic annotation (Kuczera and Armbruster, 2024). Similarly, the editorial workflow developed for the Hugo Schuchardt Archive outlines a process that combines prompt engineering, human-in-the-loop oversight, and RAG-enabled toolchains to enhance the generation of TEI-compliant XML, supporting more explainable and modular processing pipelines (Pollin et al., 2023). These and other experiments underscore the need for standardized workflows, robust evaluation protocols, and systematic research into both the strengths and weaknesses of LLMs and related tools in the editorial process, while also advocating for thoughtful engagement with advanced computational methods in the humanities (Pollin et al., 2024). As digital editions become increasingly complex and interconnected with broader knowledge infrastructures, the relevance and application of AI technologies – such as RAG – are both expected and desirable to grow accordingly.

RAG methodologies are being adopted within the GLAM sector (Galleries, Libraries, Archives and Museums) as well. In archival contexts, a smart assistant developed for querying the *Prozhito* digital archive of personal diaries combines text-to-SQL filtering, hybrid search, and automatic query reformulation, proving especially effective for historians and anthropologists without prior knowledge of database query languages (Sergeev et al., 2025). Meanwhile, in museum settings, a comparative evaluation of RAG systems versus large-context LLMs for answering multimodal questions about artworks demonstrated that the RAG approach offers superior precision and explainability (Ramos-Varela et al., 2025).

Innovations in graph-based retrieval are also gaining momentum. Techniques combining structured supervision and chain-of-thought prompting have been used to map character relationships in early modern English historiography, thereby reducing the manual workload typically associated with historical data annotation (Fan et al., 2025). Related directions are being explored within cultural heritage institutions, as seen in the *CAT-IA* initiative (Barbato, 2025), which integrates ArCo knowledge graph (Carriero et al., 2019) within a RAG system for provenance tracking, AI explainability (XAI), and structured metadata extraction. Designed to streamline and enrich user interactions with the General Catalogue of Cultural Heritage

(Catalogo generale dei beni culturali), *CAT-IA* marks a notable stride in applying advanced digital technologies to promote accessibility and valorization of cultural assets.

A complementary, conceptual perspective emerges in a critical mapping of the theoretical contours of RAG within the broader landscape of archives, libraries, and cultural heritage – articulating not only the potential for RAG-augmented LLMs to enhance the precision, accessibility, and contextualization of information retrieval, but also foregrounding the social and infrastructural challenges inherent in such integration. This analysis is accompanied by a caution against viewing RAG as a substitute for established cataloguing practices. This viewpoint encourages the field to reflect on both the affordances and the epistemic and ethical complexities introduced by RAG systems in digital humanities contexts (Di Marcantonio, 2024).

Finally, efforts to advance access to fragmented digital repositories – such as web archives – have increasingly adopted RAG methodologies. An illustrative bespoke prototype (Davis, 2025) transforms keyword-based search into semantically guided question answering, sharing architectural parallels with the GNA QA system presented in the context of this thesis. Both systems prioritize semantic retrieval over lexical matching using dense embeddings – e.g., *E5* variants (L. Wang et al., 2024) – to interpret queries in context, employ structured text processing pipelines to reduce noise in source materials, and apply optimized chunking strategies for retrieval accuracy. Crucially, these studies highlight RAG's potential to transform scattered and heterogeneous resources – whether web archives or catalographic procedures – into coherent, accessible knowledge through context-aware synthesis.

3.4 Future Directions

Ongoing research is rapidly pushing the frontiers of RAG, opening up new avenues that extend well beyond traditional information retrieval into domains such as scientific research and the digital humanities. Among the most promising innovations is the use of synthetic corpora to bolster the robustness and generalizability of RAG systems, particularly in low-resource or specialized domains where annotated data is scarce (Bor-Woei, 2024). This strategy not only improves retrieval accuracy but also addresses longstanding issues of bias, coverage, and representativity in humanities corpora.

RAG is also at the core of a new wave of applications that automate and enhance scholarly practices. In scientific research, advanced RAG frameworks – including agentic systems like PaperQA (Lála et al., 2023)– are being leveraged to conduct systematic literature reviews, automate evidence synthesis, summarize emerging trends, and provide transparent citation recommendations. These multi-stage agentic architectures enable recursive reasoning, dynamic tool use, and context-aware synthesis, often surpassing human-level performance in both retrieval and summary tasks. Their capacity for rigorous, scalable analysis offers substantial promise for the digital humanities as well, facilitating the synthesis of dispersed cultural resources and supporting complex, interpretive scholarly inquiries (Skarlinski et al., 2024).

Despite these advances, several critical research challenges remain. There is an urgent need to develop domain-adapted and multilingual LLMs that can process not just text, but also multimodal data such as images, tables, and audiovisual materials—a key requirement for both scientific and cultural heritage applications. Future RAG systems should be able to retrieve and reason over heterogeneous, cross-domain sources, necessitating robust mechanisms for source evaluation, multimodal fusion, and trust calibration. The ongoing development of benchmarks and evaluation datasets, tailored to the nuanced needs of fields such as the digital humanities, is essential to guide progress and ensure methodological rigor (Yue, 2025).

Another major direction is the semantic enrichment of RAG pipelines through the integration of ontologies and knowledge graphs. Ontologies, as formal domain knowledge models, provide structured frameworks that enable more precise and explainable retrieval semantic coherence, and the inclusion of ethical dimensions in generative AI. Complementing this, knowledge graphs capture complex relationships and support context-aware multi-hop reasoning, improving accuracy, explainability, and cultural sensitivity of outputs. Current research and practical applications span a range of initiatives – from ontology-guided entity typing to the grounding of AI in explicit ethical and procedural knowledge, demonstrating that these semantic tools are essential for creating robust, context-aware, and transparent RAG systems, addressing challenges in fields as diverse as healthcare, engineering, scientific discovery, and enterprise knowledge management (Tiwari et al., 2025; Ludwig et al., 2025; Bran et al., 2024; Sharma et al., 2024; Xiao et al., 2024; Park et al., 2024; DeBellis, 2024; Franco et al., 2020).

In the specific context of the digital humanities, the accelerated adoption of AI is shaping a transformative future for scholarship, curation, and access to cultural heritage. The diverse case studies and technical innovations, discussed in § 3.3, illustrate both the breadth of RAG's impact and the field's growing ambition. Across applications, from digital scholarly editions, to archival assistance and museum information systems, RAG has is emerging as a pivotal enabler for addressing the limitations of traditional search and annotation by supporting context-aware, semantically rich, and explainable information access.

Looking forward, several converging trends and open challenges will define the evolution of RAG in the digital humanities. First, technical advances such as the integration of knowledge graphs, graph-based retrieval, and multimodal pipelines are driving improvements in improving semantic linking and annotation of historical, literary, and artistic materials. Second, the increasing complexity of digital scholarly editions and GLAM infrastructures is catalyzing demand for standardized, reproducible workflows, robust evaluation protocols, and domain-adapted benchmarks, ensuring that RAG methods are critically assessed and tuned for the nuanced needs of humanistic research.

At the same time, as digital repositories become ever more fragmented, the promise of RAG lies in its ability to synthesize heterogeneous, dispersed data – transforming scattered web archives, periodicals, and catalogues into accessible, contextualized knowledge spaces. Yet, this evolution also foregrounds critical conceptual and ethical questions. As highlighted by recent critical perspectives, it is essential to position RAG as an augmentative technology: one that enhances, but does not replace, established cataloguing, metadata, and interpretive practices. human interpretive oversight, transparency, and cultural sensitivity must remain central, particularly as RAG systems are increasingly relied upon for knowledge production and mediation in complex social and historical domains (Di Marcantonio, 2024).

In sum, the next phase of RAG's development in the digital humanities will require sustained interdisciplinary collaboration and critical reflection. Researchers and practitioners must continue to experiment with new strategies, but also engage deeply with the epistemic, social, and infrastructural complexities of integrating advanced AI into cultural knowledge management. Ultimately, RAG applications stand poised not only to offer improved access to information, but they also invite a reimagining of the relationship between artificial intelligence and cultural knowledge production, fostering tools that augment – not displace – human creativity and understanding.

Chapter 4

Case Study: Question-Answering System for GNA

4.1 Geoportale Nazionale per l'Archeologia

The Geoportale Nazionale per l'Archeologia (GNA) (Mic, 2019) serves as the central online hub for collecting and sharing data resulting from archaeological investigations carried out across Italy. The project's primary goal is the creation of a dynamic archaeological map of the national territory, which is easily updatable over time, openly accessible, and designed for reuse and integration across multiple institutional and disciplinary contexts.

The inception of the GNA traces back to a 2014 *Memorandum of Understanding* signed by the Ministero dei Beni e delle Attività Culturali e del Turismo (MiBACT) – specifically the Segretariato Generale, the Direzione Generale per le Antichità (DG-Ant), and the Consiglio Nazionale delle Ricerche (CNR). This agreement laid the groundwork for a national geoportal aimed at safeguarding and enhancing cultural heritage through integrated digital infrastructure. However, it was the establishment of the Istituto Centrale per l'Archeologia (ICA) in 2016 that provided the structural and institutional foundation for the GNA. The ICA's mandate to define standards and promote digital archaeological databases gave renewed potential to the initiative, which culminated in the launch and formal presentation of the GNA at a ministerial venue in 2019 (Calandra, 2023).

4.1.1 Purpose and Scope

As the official repository for all research activities in archaeology and preventive archaeology – particularly those related to public infrastructure projects – the GNA platform was established to provide a unified national access point to essential archaeological data gathered across the country. This includes the interventions listed in Tab. 8, all conducted under the scientific supervision of the Italian Ministry of Culture (MiC).

Archaeological interventions	**Description**
Preventive archaeology reports	Data from excavations and surveys carried out ahead of construction projects (e.g., highways, railways, pipelines), often submitted by private firms or cultural heritage consultants.
Assisted scientific excavations records	Results from academic digs by universities or research institutions, including documentation of stratigraphy, finds, and site interpretation.
Accidental discoveries	Locations of fortuitous archaeological finds, such as during agricultural work or construction, reported to local heritage authorities. Typically include preliminary spatial data and descriptive reports.
Scheduled excavations	Long-term planned investigations, often at known heritage sites, including geospatial boundaries, uncovered structures, and findings.
Archaeological surveys	Surface survey data with GPS-tracked locations of finds, artifact scatters, and site features.
Cultural heritage GIS layers	External datasets from institutions (regional superintendencies, local governments, ICCD), e.g., maps of protected zones, risk maps, or site inventories.
Legacy data and digitized archives	Georeferenced digitizations of paper maps, notebooks, and archival records previously stored in non-digital formats, essential for integrating historical with current data.
Depository locations	Georeferenced storage locations of archaeological finds (museums, storerooms) associated with sites or interventions.
Remote sensing and aerial surveys	Drone imagery, LiDAR scans, or satellite data used to identify and map archaeological features not visible at ground level.
Paleontological sites	A specific level dedicated to paleontological sites is currently under study for future inclusion, aiming to protect this fragile heritage.

Table 8: Types of archaeological data sources integrated into the GNA.

These sources, once georeferenced and structured, are integrated into the GNA using standardized metadata and visualization protocols, to allow users to view, search, and analyze information in a spatially accurate and coherent manner (Boi, 2023; Acconcia, 2023).

4.1.2 Stakeholders and Intended Users

The development of the GNA saw significant acceleration during the COVID-19 pandemic, which provided both the urgency and institutional impetus toward the creation of a unified digital platform for managing archaeological data nationwide. This initiative built upon years of prior collaboration between key stakeholders, including the Istituto Centrale per l'Archeologia (ICA) and the Istituto Centrale per il Catalogo e la Documentazione (ICCD), who had already developed a cataloging structure to document archaeological assessments and identified sites within the Sistema Informativo Generale del Catalogo (SiGECweb) (Calandra, 2023; Boi, 2023). The pandemic underscored the limitations of purely textual cataloging and catalyzed a shift toward a more dynamic and geospatially grounded approach, leading to the adoption of a GIS-based framework better suited for preventive archaeology and territorial planning. The result was a consolidated national infrastructure designed not only to support compliance with cultural heritage protection regulations but also to enable data harmonization across previously fragmented practices (Acconcia, 2023).

The GNA is primarily intended for use by:

- Public administrators and government officials
- Professional archaeologists and cultural heritage consultants
- Stakeholders involved in public works, such as national infrastructure planners

For instance, major entities like TERNA (the national electricity grid operator), RFI (the Italian railway network), or the Milan Metro rely on the platform to assess archaeological constraints before launching construction projects. The platform helps them identify archaeological sites, deposits, and zones to avoid, ensuring the preservation of cultural heritage during the planning and development of public infrastructure.

Central to the system is a QGIS[1] template that standardizes data entry and visualization. This tool supports efficient integration of local information into the national infrastructure, offering users a unified territorial overview. It enables the comparison of diverse archaeological records, improves the quality of evaluations, and promotes transparency across institutional workflows. Thanks to its open-source foundation and modular structure, the GNA continues to evolve based on user feedback, maintaining a shared national standard while accommodating diverse local contributions (Calandra, 2023; Boi, 2023).

[1] QGIS is a free, open-source Geographic Information System (GIS) software used for creating, managing, and analyzing geospatial data.

4.1.3 User Manual and Operational Support

To guide users in correctly navigating the system, a collaboratively authored user manual (*manuale operativo*) is made available through a MediaWiki environment hosted on the GNA server (GNA, 2024). This living document offers structured instructions for data input and visualization within QGIS, including detailed documentation for using the GNA template plugin. These tools enable users to download and integrate standardized data layers – such as archaeological risk maps, identified sites, or project modules – directly into their GIS workflows. Complementing the manual, a Help Desk service is managed by Ada Gabucci,[2] who provides direct assistance to users facing technical difficulties or requiring clarification.

4.2 Proof of Concept

In response to the challenges users face when accessing and navigating the GNA operative manual, as well as the high volume of inquiries received by the Help Desk, a need emerged for a smarter and more efficient support solution. To address this, we developed an information system in the form of a question-answering chatbot designed to assist users directly and reduce the Help Desk's workload. Based on the current state of AI, machine learning, and digital humanities methodologies – as discussed in Chap. 3 and § 3.2 – retrieval-augmented generation (RAG) combined with natural language processing (NLP) was chosen as the most effective approach. This technology enables the chatbot to dynamically retrieve relevant information, which serves as an augmented knowledge base, allowing it to generate precise, context-aware, and up-to-date answers tailored to user queries.

The following chapter details the methodological framework and practical steps undertaken during the development of this chatbot system. It provides an in-depth explanation of the design choices, technical architecture, data preparation, implementation and evaluation processes.

[2]Ada Gabucci is a specialist in Roman-period archaeology, with expertise in stratigraphic methods, northern Italian material culture, and the structuring of archaeological data. She has over thirty years of experience consulting for public institutions, including the Italian Ministry of Culture (ICCD, ICA, DG-ABAP), its regional branches, the Veneto Region, and several universities, including Trieste, Venice, Verona, Bologna, Genova, and Pisa. Her work also encompasses cultural heritage cataloguing, ministerial regulations, and the design of complex Geographic Information Systems.
See: https://web.archive.org/web/20250724081422/https://conf24.garr.it/it/speaker/ada-gabucci.

Chapter 5

Methodology

5.1 Prototype

5.2 Full-scale implementation

5.3 Knowledge-Base Construction and Preprocessing

5.3.1 Sitemap Generation

- URL Discovery and Filtering - Crawling Policies and Content Parsing

5.3.2 Document Crawling and Chunking

- Content Extraction and Semantic Preservation - Chunking Strategy and Metadata Enrichment - Special Content Handling (text/tables/images)

5.3.3 Vector Embeddings and Storage

- Multilingual Embedding Model (multilingual-e5-large) - FAISS Index Optimization for Retrieval

5.4 Candidates Retrieval

- Top-k Similarity Search - Document Grouping by Provenance

5.5 Generation

- open source and proprietary LLMs (*Open Source vs. Proprietary LLMs: A Comprehensive Comparison*, 2025) - Mistral NeMo LLM Configuration - Response Generation and Validation

5.5.1 Prompt Engineering Techniques

- System Instructions and Constraints - Dynamic Citation Handling

5.6 User Interface

- Streamlit Application - Session State Management - Asynchronous Query Processing - Interaction Features - Inline Feedback Mechanism

5.7 Feedback Management

- Feedback Data Collection - streamlit local collection

- SQLite Database - Feedback Rating System (3-point Likert) - Automated GitHub sync - Conflict Resolution Protocols

5.8 Resource Optimization Strategies

5.8.1 Memory Management

- Garbage Collection Routines - Caching Mechanisms

5.8.2 Computational Constraints Mitigation

- CPU-Only Deployment Adaptations - Error Handling and Recovery

5.9 Evaluation

Chapter 6

Results

6.1 Evaluation

6.1.1 Qualitative analysis

6.1.2 Quantitave analysis

Chapter 7

Discussion

7.1 Further Development

specializzazione sul dominio archeologia

Chapter 8

Conclusion

Appendices

Appendix A

Implementation details

In this study, all the experiments have been performed on a system with an Intel Core i7-1185G7 CPU at 3.00 GHz, 16 GB of RAM, and integrated Intel Iris Xe Graphics with 128 MB of VRAM. Additionally, the following software packages have been used to implement the proposed approach: PyTorch, NumPy, Pandas, Matplotlib, SpaCy, HuggingFace...

Appendix B

Abbreviations and glossary

Table 9: Abbreviations and acronyms with their full forms and definitions.

Abbreviation / Term	Full form	Glossary definition
AI	Artificial Intelligence	The field of computer science dedicated to creating systems capable of performing tasks that typically require human intelligence, such as reasoning, learning, and problem-solving.
DH	Digital Humanities	An interdisciplinary field that applies computational methods and tools to humanities research, analysis, and dissemination.
QA	Question-Answering	A technology or task in which a system provides precise answers to questions posed in natural language.
QAS	Question-Answering System	A system designed to answer questions automatically by processing natural language input, often using methods from IR and NLP.
RAG	Retrieval-Augmented Generation	An approach combining information retrieval with generative models, allowing AI to reference external data sources when generating answers.
LLM	Large Language Model	A neural network trained on massive text corpora to generate or understand human language, such as GPT or BERT.
IR	Information Retrieval	The process of searching, retrieving, and ranking relevant documents or data from large collections based on user queries.
KB	Knowledge Base	A structured collection of information or data, often used to support reasoning, search, or retrieval in AI systems.
NL	Natural Language	The everyday language used by humans for communication, which NLP systems aim to understand and generate.
NLP	Natural Language Processing	The area of AI focused on enabling computers to understand, interpret, and generate human language.
NLG	Natural Language Generation	The process of automatically generating human-like text from structured data or models, often used in chatbots and content creation.
NL	Natural Language	The everyday language used by humans for communication, which NLP systems aim to understand and generate.

Continued on next page

Abbreviation / Term	Full form	Glossary definition
ML	Machine Learning	A subset of AI that involves training algorithms to recognize patterns and make decisions based on data.
TF-IDF	Term Frequency-Inverse Document Frequency	A statistical measure used in IR to evaluate how important a word is to a document relative to a corpus, balancing term frequency and document rarity.
BM25	Best Match 25	A ranking function used in IR to estimate the relevance of documents to a given search query, based on term frequency and document length normalization.
PRF	Precision-Recall-F1	Metrics used to evaluate the performance of classification models, where precision measures the accuracy of positive predictions, recall measures the ability to find all relevant instances, and F1 is the harmonic mean of precision and recall.
TREC	Text REtrieval Conference	An ongoing series of workshops and evaluations focused on advancing research in text retrieval and related tasks.
LMIR	Language Model Information Retrieval	A method of using language models to improve the effectiveness of information retrieval systems by leveraging their understanding of language and context.
RNN	Recurrent Neural Network	A type of neural network architecture designed to process sequential data by maintaining a form of memory of previous inputs.
LSTM	Long Short-Term Memory	A special kind of RNN capable of learning long-range dependencies, often used for tasks like language modeling or time series prediction.
CRF	Conditional Random Field	A probabilistic graphical model used for structured prediction, especially in NLP tasks such as sequence labeling.
SVM	Support Vector Machine	A supervised machine learning algorithm used for classification and regression, which finds the optimal boundary between classes in the feature space.
Word2Vec	Word to Vector	A technique for representing words as vectors in a continuous vector space, capturing semantic relationships between words based on their context in large text corpora.
GloVe	Global Vectors for Word Representation	An unsupervised learning algorithm for obtaining vector representations of words, which captures global statistical information from a corpus.
BERT	Bidirectional Encoder Representations from Transformers	A pre-trained language model that uses the Transformer architecture to understand the context of words in a sentence by considering both left and right contexts simultaneously.
GPT	Generative Pre-trained Transformer	A type of LLM that uses the Transformer architecture.
T5	Text-to-Text Transfer Transformer	A versatile LLM that treats all NLP tasks as text-to-text problems, allowing it to be fine-tuned for various applications.
E5	Embedding-based model for information retrieval	A family of models designed to generate high-quality embeddings for text, improving retrieval performance in IR tasks.
BGE	BAAI General Embedding	A family of models designed to generate high-quality embeddings for text, improving retrieval performance in IR tasks.

Continued on next page

Abbreviation / Term	Full form	Glossary definition
intfloat	intfloat/e5	A specific implementation of the E5 model, optimized for generating embeddings for information retrieval tasks.
LLM-Embedder	LLM-Embedder	A model designed to generate embeddings for text using large language models, enhancing the quality of semantic representations for retrieval tasks.
embeddings		Dense vector representations of text that capture semantic meaning, used in various NLP tasks including retrieval and classification.
chunking		The process of breaking down text into smaller, manageable pieces or "chunks" to facilitate processing and analysis in NLP tasks.
vector database		A specialized database designed to store and retrieve high-dimensional vectors efficiently, often used in RAG systems for managing embeddings.
retriever		A component of a system responsible for searching and retrieving relevant documents or information from a database or corpus based on user queries.
ranking function		A mathematical function used to score and order documents based on their relevance to a given query, often employed in IR systems.
XML	eXtensible Markup Language	A markup language used to encode documents in a format that is both human-readable and machine-readable, often used for data interchange.
TEI	Text Encoding Initiative	A set of guidelines for encoding literary and linguistic texts in XML, providing a standardized way to represent complex textual structures.
MARC/RDA	Machine-Readable Cataloging / Resource Description and Access	Standards for encoding bibliographic information in a machine-readable format, widely used in libraries and information systems.
GROBID	GeneRation Of BIbliographic Data	A machine learning library for extracting and structuring bibliographic information from scholarly documents, often used in academic publishing and research.
Milvus	Milvus Vector Database	An open-source vector database designed for efficient storage, indexing, and retrieval of high-dimensional vectors, commonly used in RAG systems.
Faiss	Facebook AI Similarity Search	A library for efficient similarity search and clustering of dense vectors, widely used in RAG systems for indexing and searching large datasets.
Qdrant	Qdrant Vector Database	An open-source vector database that provides efficient storage and retrieval of high-dimensional vectors, supporting hybrid search capabilities.
HyDE	Hybrid Document Embedding	A method that combines dense vector representations with traditional keyword-based indexing to improve retrieval performance in RAG systems.
Hybrid Search	Hybrid Search	A search approach that combines vector-based retrieval with traditional keyword search, allowing for more comprehensive and context-aware results in RAG systems.

Continued on next page

Abbreviation / Term	Full form	Glossary definition
TILDE	TILDE (TILDE is a framework for building RAG systems)	A framework designed to facilitate the development and deployment of RAG systems, providing tools for data preparation, indexing, and retrieval.
LTR	Learning-to-Rank	A machine learning approach used to optimize the ranking of search results based on user interactions and relevance feedback, improving the quality of retrieved documents in RAG systems.
self-rag	Self-Retrieval-Augmented Generation	A variant of RAG where the system retrieves relevant information from its own generated content, enhancing the context and accuracy of responses.
RAGAS	Retrieval-Augmented Generation with Adaptive Sampling	An advanced RAG approach that dynamically selects and retrieves the most relevant information based on the context of the query, improving the efficiency and accuracy of responses.

Acknowledgments

Bibliography

Acconcia, Valeria. 2023. "LA PUBBLICAZIONE DEI DATI NEL GEOPORTALE NAZIONALE PER L'ARCHEOLOGIA" [in ita]. In *GEOPORTALE NAZIONALE PER L'ARCHEOLOGIA,* 1. IT: Ministero della Cultura - Istituto Centrale per l'Archeologia, July. Accessed July 24, 2025. https://doi.org/10.60974/GNA_03.

Ada Gabucci. 2025, July. Accessed July 24, 2025. https://web.archive.org/web/20250724081422/https://conf24.garr.it/it/speaker/ada-gabucci.

Agarwal, Shubham, Gaurav Sahu, Abhay Puri, Issam H. Laradji, Krishnamurthy DJ Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. 2025. *LitLLM: A Toolkit for Scientific Literature Review.* ArXiv:2402.01788, March. Accessed July 23, 2025. DOI: 10.48550/arXiv.2402.01788. http://arxiv.org/abs/2402.01788.

Akkiraju, Rama, Anbang Xu, Deepak Bora, Tan Yu, Lu An, Vishal Seth, Aaditya Shukla, et al. 2024. *FACTS About Building Retrieval Augmented Generation-based Chatbots.* ArXiv:2407.07858, July. Accessed July 23, 2025. DOI: 10.48550/arXiv.2407.07858. http://arxiv.org/abs/2407.07858.

Alanazi, Sarah Saad, Nazar Elfadil, Mutsam Jarajreh, and Saad Algarni. 2021. "Question Answering Systems: A Systematic Literature Review" [in en]. *International Journal of Advanced Computer Science and Applications* 12 (3). ISSN: 21565570, 2158107X, accessed July 23, 2025. DOI: 10.14569/IJACSA.2021.0120359. http://thesai.org/Publications/ViewPaper?Volume=12&Issue=3&Code=IJACSA&SerialNo=59.

Alqifari, Reem. 2019. "Question Answering Systems Approaches and Challenges." In *Proceedings of the Student Research Workshop Associated with RANLP 2019,* edited by Venelin Kovatchev, Irina Temnikova, Branislava Šandrih, and Ivelina Nikolova, 69–75. Varna, Bulgaria: INCOMA Ltd., September. Accessed July 26, 2025. DOI: 10.26615/issn.2603-2821.2019_011. https://aclanthology.org/R19-2011/.

Alshammari, Suad, Lama Basalelah, Walaa Abu Rukbah, Ali Alsuhibani, and Dayanjan S. Wijesinghe. 2023. *KNIMEZoBot: Enhancing Literature Review with Zotero and KNIME OpenAI Integration using Retrieval-Augmented Generation.* ArXiv:2311.04310, November. Accessed July 21, 2025. DOI: 10.48550/arXiv.2311.04310. http://arxiv.org/abs/2311.04310.

Was ist RAG? – Retrieval Augmented Generation erklärt – AWS [in de-DE]. Accessed July 24, 2025. https://aws.amazon.com/de/what-is/retrieval-augmented-generation/.

Amershi, Saleema, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. "Software Engineering for Machine Learning: A Case Study." In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP),* 291–300. Montreal, QC, Canada: IEEE, May. ISBN: 9781728117607, accessed July 28, 2025. DOI: 10.1109/ICSE-SEIP.2019.00042. https://ieeexplore.ieee.org/document/8804457/.

Antolini, Gianluca. 2025. “Experimental Study on Retrieval-Augmented Generation: Engineering and Evaluation of a Custom RAG system for Open-Domain QA” [in en]. Master’s thesis, Computer Engineering, University of Padova, July. Accessed July 21, 2025. https://thesis.unipd.it/handle/20.500.12608/86949.

Antoniou, Christina, and Nick Bassiliades. 2022. “A survey on semantic question answering systems” [in en]. *The Knowledge Engineering Review* 37:e2. ISSN: 0269-8889, 1469-8005, accessed July 26, 2025. DOI: 10.1017/S0269888921000138. https://www.cambridge.org/core/product/identifier/S0269888921000138/type/journal_article.

Arslan, Muhammad, Hussam Ghanem, Saba Munawar, and Christophe Cruz. 2024. “A Survey on RAG with LLMs” [in en]. *Procedia Computer Science* 246:3781–3790. ISSN: 18770509, accessed June 19, 2025. DOI: 10.1016/j.procs.2024.09.178. https://linkinghub.elsevier.com/retrieve/pii/S1877050924021860.

Asai, Akari, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. *Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.* ArXiv:2310.11511, October. Accessed July 27, 2025. DOI: 10.48550/arXiv.2310.11511. http://arxiv.org/abs/2310.11511.

Aytar, Ahmet Yasin, Kemal Kilic, and Kamer Kaya. 2024. *A Retrieval-Augmented Generation Framework for Academic Literature Navigation in Data Science.* ArXiv:2412.15404, December. Accessed July 21, 2025. DOI: 10.48550/arXiv.2412.15404. http://arxiv.org/abs/2412.15404.

Barbato, Fabrizio. 2025. *Nasce Cat-IA, l’agente conversazionale che semplifica la consultazione del Catalogo generale dei beni culturali* [in it-IT], April. Accessed July 21, 2025. https://digitallibrary.cultura.gov.it/notizie/nasce-cat-ia/.

Bevara, Ravi Varma Kumar, Brady D. Lund, Nishith Reddy Mannuru, Sai Pranathi Karedla, Yara Mohammed, Sai Tulasi Kolapudi, and Aashrith Mannuru. 2025. “Prospects of Retrieval Augmented Generation (RAG) for Academic Library Search and Retrieval.” *Information Technology and Libraries* 44, no. 2 (June). ISSN: 2163-5226, 0730-9295, accessed July 19, 2025. DOI: 10.5860/ital.v44i2.17361. https://ital.corejournals.org/index.php/ital/article/view/17361.

Boi, Valeria. 2023. “IL GEOPORTALE NAZIONALE PER L’ARCHEOLOGIA: STANDARDIZZAZIONE E APERTURA DEI DATI” [in ita]. In *GEOPORTALE NAZIONALE PER L’ARCHEOLOGIA,* 1. IT: Ministero della Cultura - Istituto Centrale per l’Archeologia, July. Accessed July 24, 2025. https://doi.org/10.60974/GNA_02.

Bor-Woei, Huang. 2024. *Generative Large Language Models Augmented Hybrid Retrieval System for Biomedical Question Answering.* Technical report. Grenoble, France: University of Padova, Italy. https://ceur-ws.org/Vol-3740/paper-12.pdf.

Bran, Andres M., Alexandru Oarga, Matthew Hart, Magdalena Lederbauer, and Philippe Schwaller. 2024. “Ontology-Retrieval Augmented Generation for Scientific Discovery” [in en] (October). Accessed July 23, 2025. https://openreview.net/forum?id=DbZDbg2z9q.

BUP Solutions. *BUP Solutions.* Accessed July 25, 2025. https://www.bupsolutions.com/en/home_en/.

Caballero, Michael. 2021. “A Brief Survey of Question Answering Systems.” *International Journal of Artificial Intelligence & Applications* 12, no. 5 (September): 01–07. ISSN: 09762191, accessed July 26, 2025. DOI: 10.5121/ijaia.2021.12501. https://aircconline.com/ijaia/V12N5/12521ijaia01.pdf.

Calandra, Elena. 2023. "IL GEOPORTALE NAZIONALE PER L'ARCHEOLOGIA (GNA). UN'INTRODUZIONE" [in ita]. In *GEOPORTALE NAZIONALE PER L'ARCHEOLOGIA,* 1. IT: Ministero della Cultura - Istituto Centrale per l'Archeologia, July. Accessed July 24, 2025. https://doi.org/10.60974/GNA_01.

Callaghan, Samantha, and Miguel Vieira. 2025. *Prototyping a RAG System for Digital Humanities: Exploring AI/ML with Indigenous Data | King's Digital Lab* [in en], January. Accessed July 19, 2025. https://kdl.kcl.ac.uk/blog/ireal-rag/.

Caramanna, Gianluigi. 2024. "Progettazione e sviluppo di un chatbot basato su tecniche di Intelligenza Artificiale Generativa" [in it]. Master's thesis, Ingegneria Informatica, Università di Bologna, October. Accessed July 21, 2025. https://amslaurea.unibo.it/id/eprint/32820/.

Carriero, Valentina Anita, Aldo Gangemi, Maria Letizia Mancinelli, Ludovica Marinucci, Andrea Giovanni Nuzzolese, Valentina Presutti, and Chiara Veninata. 2019. *ArCo: the Italian Cultural Heritage Knowledge Graph.* ArXiv:1905.02840, May. Accessed July 23, 2025. DOI: 10.48550/arXiv.1905.02840. http://arxiv.org/abs/1905.02840.

Ciletti, Michele. 2025. "Retrieval-Augmented Generation systems for enhanced access to digital archives" [in en]. In *Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale, Proceedings del XIV Convegno Annuale AIUCD2025,* 663. Quaderni di Umanistica Digitale. Verona: AIUCD, June. ISBN: 978-88-942535-9-7. DOI: 10.6092/unibo/amsacta/8380. https://amsacta.unibo.it/id/eprint/8380/.

Davis, Corey. 2025. "Unlocking web archives: LLMs, RAG, and the future of digital preservation" [in en] (February). Accessed July 19, 2025. https://hdl.handle.net/1828/21379.

DeBellis, Michael. 2024. *Integrating Large Language Models (LLM) and ontologies to Implement Retrieval Augmented Generation* [in en], July. Accessed July 23, 2025. https://www.michaeldebellis.com/post/integrating-llms-and-ontologies.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* ArXiv:1810.04805, May. Accessed July 26, 2025. DOI: 10.48550/arXiv.1810.04805. http://arxiv.org/abs/1810.04805.

Di Marcantonio, Giorgia. 2024. "Intelligenza artificiale, Large Language Models (LLMs) e Retrieval-Augmented Generation (RAG). Nuovi strumenti per l'accesso alle risorse archivistiche e bibliografiche" [in it]. *Bibliothecae.it* 13, no. 1 (July): 146–173. Accessed July 28, 2025. DOI: 10.6092/ISSN.2283-9364/19982. https://bibliothecae.unibo.it/article/view/19982.

Diefenbach, Dennis, Vanessa Lopez, Kamal Singh, and Pierre Maret. 2018. "Core techniques of question answering systems over knowledge bases: a survey" [in en]. *Knowledge and Information Systems* 55, no. 3 (June): 529–569. ISSN: 0219-1377, 0219-3116, accessed July 26, 2025. DOI: 10.1007/s10115-017-1100-y. http://link.springer.com/10.1007/s10115-017-1100-y.

Es, Shahul, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. *Ragas: Automated Evaluation of Retrieval Augmented Generation.* Accessed June 19, 2025. DOI: 10.48550/ARXIV.2309.15217. https://arxiv.org/abs/2309.15217.

Fan, Yang, Zhang Qi, Xing Wenqian, Liu Chang, and Liu Liu. 2025. *Research on Graph-Retrieval Augmented Generation Based on Historical Text Knowledge Graphs.* ArXiv:2506.15241, June. Accessed July 19, 2025. DOI: 10.48550/arXiv.2506.15241. http://arxiv.org/abs/2506.15241.

Farea, Amer, and Frank Emmert-Streib. 2025. "Understanding question-answering systems: Evolution, applications, trends, and challenges" [in en]. *Engineering Applications of Artificial Intelligence* 156 (September): 110997. ISSN: 09521976, accessed July 26, 2025. DOI: 10.1016/j.engappai.2025.110997. https://linkinghub.elsevier.com/retrieve/pii/S0952197625009972.

Ferrucci, David, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, et al. 2011. "Building watson: An overview of the deepQA project" [in en-US]. *AI Magazine* (January). ISSN: 07384602, accessed July 27, 2025. DOI: 10.1609/aimag.v31i3.2303. https://research.ibm.com/publications/building-watson-an-overview-of-the-deepqa-project.

Florio, Michelangelo. 2024. "Progettazione e implementazione di un chatbot intelligente tramite piattaforma LangChain: studio e valutazione dei VectorDB" [in it]. Master's thesis, August. Accessed July 21, 2025. https://amslaurea.unibo.it/id/eprint/32282/.

Franco, Wellington, Caio Viktor, Artur Oliveira, Gilvan Maia, Angelo Brayner, V. Vidal, Fernando Carvalho, and V. Pequeno. 2020. "Ontology-based Question Answering Systems over Knowledge Bases: A Survey." In *Proceedings of the 22nd International Conference on Enterprise Information Systems,* 532–539. Prague, Czech Republic: SCITEPRESS - Science / Technology Publications. ISBN: 9789897584237, accessed July 26, 2025. DOI: 10.5220/0009392205320539. http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0009392205320539.

Gao, Yunfan, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. *Retrieval-Augmented Generation for Large Language Models: A Survey.* ArXiv:2312.10997. Accessed July 10, 2025. DOI: 10.48550/arXiv.2312.10997. http://arxiv.org/abs/2312.10997.

GNA, MiC. 2024. *Wiki GNA Manuale Operativo v.1.4,* February. Accessed July 24, 2025. https://gna.cultura.gov.it/wiki/index.php/Pagina_principale.

Green, Bert F., Alice K. Wolf, Carol Chomsky, and Kenneth Laughery. 1961. "Baseball: an automatic question-answerer" [in en]. In *Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference on - IRE-AIEE-ACM '61 (Western),* 219. Los Angeles, California: ACM Press. Accessed July 27, 2025. DOI: 10.1145/1460690.1460714. http://portal.acm.org/citation.cfm?doid=1460690.1460714.

GROBID. 2008–2025. https://github.com/kermitt2/grobid. swh: 1:dir:dab86b296e3c3216e2241968f0d63b68e8209d3c.

Gupta, Shailja, Rajesh Ranjan, and Surya Narayan Singh. 2024. *A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions.* ArXiv:2410.12837. Accessed June 19, 2025. DOI: 10.48550/arXiv.2410.12837. http://arxiv.org/abs/2410.12837.

Han, Haoyu, Yu Wang, Harry Shomer, Kai Guo, Jiayuan Ding, Yongjia Lei, Mahantesh Halappanavar, et al. 2025. *Retrieval-Augmented Generation with Graphs (GraphRAG).* ArXiv:2501.00309, January. Accessed July 27, 2025. DOI: 10.48550/arXiv.2501.00309. http://arxiv.org/abs/2501.00309.

Harsh and T. Shobha. 2024. “Comprehending and Reducing LLM Hallucinations” [in en]. *International Journal of Innovative Science and Research Technology (IJISRT),* 1222–1227. ISSN: 2456-2165, accessed June 19, 2025. DOI: 10.38124/ijisrt/IJISRT24JUL882. https://www.ijisrt.com/comprehending-and-reducing-llm-hallucinations.

Hirschman, L., and R. Gaizauskas. 2001. “Natural language question answering: the view from here” [in en]. *Natural Language Engineering* 7, no. 4 (December): 275–300. ISSN: 1351-3249, 1469-8110, accessed July 26, 2025. DOI: 10.1017/S1351324901002807. https://www.cambridge.org/core/product/identifier/S1351324901002807/type/journal_article.

Open Source vs. Proprietary LLMs: A Comprehensive Comparison. 2025, January. Accessed June 19, 2025. https://www.hostcomm.co.uk/blogs/open-source-vs-proprietary-llms-a-comprehensive-comparison.

Jiang, Zhuoxuan, Tianyang Zhang, Shengguang Bai, Lin Lin, Haotian Zhang, Yinong Xun, Jiawei Ren, Wen Si, and Shaohua Zhang. 2024. *Towards Enterprise-Specific Question-Answering for it Operations and Maintenance Based on Retrieval-Augmented Generation Mechanism.* Accessed July 23, 2025. DOI: 10.2139/ssrn.5069318. https://www.ssrn.com/abstract=5069318.

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. “The global landscape of AI ethics guidelines” [in en]. *Nature Machine Intelligence* 1, no. 9 (September): 389–399. ISSN: 2522-5839, accessed July 28, 2025. DOI: 10.1038/s42256-019-0088-2. https://www.nature.com/articles/s42256-019-0088-2.

Jurafsky, Daniel, and James H. Martin. 2024. “Chapter 14: Question Answering, Information Retrieval, and RetrievalAugmented Generation.” In *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models,* 2nd ed. Accessed July 26, 2025. https://web.stanford.edu/~jurafsky/slp3/.

Kuczera, Andreas, and Stephan Armbruster. 2024. *ChatGPT und generative KI in der mediävistischen Grundlagenforschung | H-Soz-Kult. Kommunikation und Fachinformation für die Geschichtswissenschaften | Geschichte im Netz | History in the web* [in de], July. Accessed July 28, 2025. https://www.hsozkult.de/event/id/event-142284.

Lála, Jakub, Odhran O’Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodriques, and Andrew D. White. 2023. *PaperQA: Retrieval-Augmented Generative Agent for Scientific Research.* ArXiv:2312.07559, December. Accessed July 25, 2025. DOI: 10.48550/arXiv.2312.07559. http://arxiv.org/abs/2312.07559.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, et al. 2020. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.* Accessed June 19, 2025. DOI: 10.48550/ARXIV.2005.11401. https://arxiv.org/abs/2005.11401.

Li, Jiawei, and Yue Zhang. 2024. *The Death of Feature Engineering? BERT with Linguistic Features on SQuAD 2.0.* ArXiv:2404.03184, April. Accessed July 28, 2025. DOI: 10.48550/arXiv.2404.03184. http://arxiv.org/abs/2404.03184.

Liu, Linqing, Patrick Lewis, Sebastian Riedel, and Pontus Stenetorp. 2022. *Challenges in Generalization in Open Domain Question Answering.* ArXiv:2109.01156, May. Accessed July 26, 2025. DOI: 10.48550/arXiv.2109.01156. http://arxiv.org/abs/2109.01156.

Ludwig, Heiner, Thorsten Schmidt, and Mathias Kühn. 2025. "An ontology-based retrieval augmented generation procedure for a voice-controlled maintenance assistant" [in en]. *Computers in Industry* 169 (August): 104289. ISSN: 01663615, accessed July 23, 2025. DOI: 10.1016/j.compind.2025.104289. https://linkinghub.elsevier.com/retrieve/pii/S0166361525000545.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval.* New York: Cambridge University Press. ISBN: 9780521865715.

Martineau, Kim. 2023. *What is retrieval-augmented generation (RAG)?* [In en-US]. Accessed July 19, 2025. https://research.ibm.com/blog/retrieval-augmented-generation-RAG.

Maslej, Nestor, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, et al. 2025. *Artificial Intelligence Index Report 2025.* ArXiv:2504.07139, July. Accessed July 26, 2025. DOI: 10.48550/arXiv.2504.07139. http://arxiv.org/abs/2504.07139.

McDermott, Drew. 1976. "Artificial intelligence meets natural stupidity" [in en]. *ACM SIGART Bulletin,* no. 57 (April): 4–9. ISSN: 0163-5719, accessed July 25, 2025. DOI: 10.1145/1045339.1045340. https://dl.acm.org/doi/10.1145/1045339.1045340.

Mic, GNA. 2019. *MiC GNA Geoportale Nazionale Archeologia* [in it], September. Accessed July 24, 2025. https://gna.cultura.gov.it.

Nicoletti, Sonia. 2025. "LLMs and Essence: Developing a Chatbot to Support Software Engineering Practices" [in en]. Master's thesis, Computer Science, University of Bologna, February. Accessed July 21, 2025. https://amslaurea.unibo.it/id/eprint/34197/.

ODSC-Community. 2024. *Retrieval-Augmented Generation (RAG): A Synergistic Approach to NLU and NLG* [in en-US], July. Accessed July 19, 2025. https://opendatascience.com/retrieval-augmented-generation-rag-a-synergistic-approach-to-nlu-and-nlg/.

Packowski, Sarah, Inge Halilovic, Jenifer Schlotfeldt, and Trish Smith. 2024. "Optimizing and Evaluating Enterprise Retrieval-Augmented Generation (RAG): A Content Design Perspective" [in en]. In *Proceedings of the 2024 8th International Conference on Advances in Artificial Intelligence,* 162–167. London United Kingdom: ACM, October. ISBN: 9798400718014, accessed July 23, 2025. DOI: 10.1145/3704137.3704181. https://dl.acm.org/doi/10.1145/3704137.3704181.

Park, Yeun, Paul Witherell, Nowrin Akter Surovi, and Hyunbo Cho. 2024. "Ontology-based Retrieval Augmented Generation (RAG) for GenAI-supported Additive Manufacturing" [in en]. *NIST* (August). Accessed July 23, 2025. https://www.nist.gov/publications/ontology-based-retrieval-augmented-generation-rag-genai-supported-additive.

Pograri, Lucrezia. 2025. *Question-Answering AI Assistant for Geoportale Nazionale Archeologia (GNA),* July. Accessed July 21, 2025. DOI: 10.5281/zenodo.16259759. https://zenodo.org/records/16259759.

Pollin, Christopher, Franz Fischer, Patrick Sahle, Martina Scholger, and Georg Vogeler. 2025. "When it was 2024 – Generative AI in the field of digital scholarly editions" [in en]. *Zeitschrift für digitale Geisteswissenschaften* 10. ISSN: 2510-1358, accessed July 28, 2025. DOI: 10.17175/2025_008. https://zfdg.de/2025_008.

Pollin, Christopher, Martina Scholger, Patrick Sahle, Georg Vogeler, Torsten Schaßan, Stefan Dumont, Franz Fischer, et al. 2024. "Workshop Generative KI, LLMs und GPT bei digitalen Editionen" [in deu] (March). Accessed July 28, 2025. https://zenodo.org/records/10893761.

Pollin, Christopher, Christian Steiner, and Constantin Zach. 2023. "New Ways of Creating Research Data: Conversion of Unstructured Text to TEI XML using GPT on the Correspondence of Hugo Schuchardt with a Web Prototype for Prompt Engineering. FORGE 2023. Tübingen" (October). Accessed July 28, 2025. DOI: 10.5281/ZENODO.8425162. https://zenodo.org/record/8425162.

Ponte, Jay M., and W. Bruce Croft. 1998. "A language modeling approach to information retrieval" [in en]. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval,* 275–281. Melbourne Australia: ACM, August. ISBN: 9781581130157, accessed July 27, 2025. DOI: 10.1145/290941.291008. https://dl.acm.org/doi/10.1145/290941.291008.

RAG vs Traditional QA - GeeksforGeeks. 2025, July. Accessed July 28, 2025. https://web.archive.org/web/20250728131605/https://www.geeksforgeeks.org/nlp/rag-vs-traditional-qa/.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. *SQuAD: 100,000+ Questions for Machine Comprehension of Text.* ArXiv:1606.05250, October. Accessed July 28, 2025. DOI: 10.48550/arXiv.1606.05250. http://arxiv.org/abs/1606.05250.

Ramos-Varela, Samuel, Jaime Bellver-Soler, Marcos Estecha-Garitagoitia, and Luis Fernando D'Haro. 2025. "Context or Retrieval? Evaluating RAG Methods for Art and Museum QA System." In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology,* edited by Maria Ines Torres, Yuki Matsuda, Zoraida Callejas, Arantza del Pozo, and Luis Fernando D'Haro, 129–136. Bilbao, Spain: Association for Computational Linguistics, May. ISBN: 9798891762480, accessed July 19, 2025. https://aclanthology.org/2025.iwsds-1.10/.

Riedl, Mark. 2023. *A Very Gentle Introduction to Large Language Models without the Hype* [in en], May. Accessed July 26, 2025. https://mark-riedl.medium.com/a-very-gentle-introduction-to-large-language-models-without-the-hype-5f67941fa59e.

Salcuni, Giuseppe Pio. 2025. "Utilizzo di tecniche RAG per la Valutazione e Comparazione dei Modelli LLM in ambito medico" [in it]. Master's thesis, Ingegneria Informatica, Università di Bologna, February. Accessed July 21, 2025. https://amslaurea.unibo.it/id/eprint/33387/.

Seo, Wonduk, Zonghao Yuan, and Yi Bu. 2025. *ValuesRAG: Enhancing Cultural Alignment Through Retrieval-Augmented Contextual Learning.* ArXiv:2501.01031, May. Accessed July 23, 2025. DOI: 10.48550/arXiv.2501.01031. http://arxiv.org/abs/2501.01031.

Sergeev, Alexander, Valeriya Goloviznina, Mikhail Melnichenko, and Evgeny Kotelnikov. 2025. *Talking to Data: Designing Smart Assistants for Humanities Databases.* ArXiv:2506.00986, June. Accessed July 19, 2025. DOI: 10.48550/arXiv.2506.00986. http://arxiv.org/abs/2506.00986.

Sharma, Kartik, Peeyush Kumar, and Yunqing Li. 2024. *OG-RAG: Ontology-Grounded Retrieval-Augmented Generation For Large Language Models.* ArXiv:2412.15235, December. Accessed July 23, 2025. DOI: 10.48550/arXiv.2412.15235. http://arxiv.org/abs/2412.15235.

Skarlinski, Michael D., Sam Cox, Jon M. Laurent, James D. Braza, Michaela Hinks, Michael J. Hammerling, Manvitha Ponnapati, Samuel G. Rodriques, and Andrew D. White. 2024. *Language agents achieve superhuman synthesis of scientific knowledge.* Accessed June 19, 2025. DOI: 10.48550/ARXIV.2409.13740. https://arxiv.org/abs/2409.13740.

Soman, Sumit, and Sujoy Roychowdhury. 2024. *Observations on Building RAG Systems for Technical Documents.* ArXiv:2404.00657, March. Accessed July 19, 2025. DOI: 10.48550/arXiv.2404.00657. http://arxiv.org/abs/2404.00657.

Thakur, Nandan, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. *BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models.* ArXiv:2104.08663, October. Accessed July 28, 2025. DOI: 10.48550/arXiv.2104.08663. http://arxiv.org/abs/2104.08663.

Tiwari, Yash, Owais Ahmad Lone, and Mayukha Pal. 2025. *OntoRAG: Enhancing Question-Answering through Automated Ontology Derivation from Unstructured Knowledge Bases.* Accessed July 23, 2025. DOI: 10.48550/ARXIV.2506.00664. https://arxiv.org/abs/2506.00664.

Text REtrieval Conference (TREC) QA Data. Accessed July 26, 2025. https://trec.nist.gov/data/qa.html.

Upbin, Bruce. 2013. *IBM's Watson Gets Its First Piece Of Business In Healthcare* [in en], February. Accessed July 27, 2025. https://www.forbes.com/sites/bruceupbin/2013/02/08/ibms-watson-gets-its-first-piece-of-business-in-healthcare/.

Vaibhav, Fanindra Mahajan. 2025. "Retrieval-augmented generation: The technical foundation of intelligent AI Chatbots." *World Journal of Advanced Research and Reviews* 26, no. 1 (April): 4093–4099. ISSN: 25819615, accessed July 23, 2025. DOI: 10.30574/wjarr.2025.26.1.1571. https://journalwjarr.com/node/1453.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. *Attention Is All You Need.* Accessed June 19, 2025. DOI: 10.48550/ARXIV.1706.03762. https://arxiv.org/abs/1706.03762.

Voorhees, E., D. K. Harman, and National Institute of Standards and Technology (U.S.), eds. 2005. *TREC: experiment and evaluation in information retrieval.* Digital libraries and electronic publishing. Cambridge, Mass: MIT Press. ISBN: 9780262220736.

Wang, Liang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024. *Text Embeddings by Weakly-Supervised Contrastive Pre-training.* ArXiv:2212.03533, February. Accessed July 25, 2025. DOI: 10.48550/arXiv.2212.03533. http://arxiv.org/abs/2212.03533.

Wang, Xiaohua, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, et al. 2024. *Searching for Best Practices in Retrieval-Augmented Generation.* ArXiv:2407.01219, July. Accessed July 18, 2025. DOI: 10.48550/arXiv.2407.01219. http://arxiv.org/abs/2407.01219.

Question answering [in en]. 2025. Page Version ID: 1293783575, June. Accessed July 26, 2025. https://en.wikipedia.org/w/index.php?title=Question_answering&oldid=1293783575.

Wilensky, Robert, David N. Chin, Marc Luria, James Martin, James Mayfield, and Dekai Wu. 1988. "The berkeley UNIX consultant project." *Comput. Linguist.* (Cambridge, MA, USA) 14, no. 4 (December): 35–84. ISSN: 0891-2017.

Woods, William, Ronald M. Kaplan, and Bonnie L. Webber. 1972. *The Lunar Sciences Natural Language Information System: Final Report* [in en]. Technical report 2378. Cambridge, Massachusetts: Bolt Beranek and Newman Inc., June. https://www.researchgate.net/publication/24285293_The_Lunar_Sciences_Natural_Language_Information_System.

Xiao, Jinfeng, Linyi Ding, James Barry, Mohab Elkaref, Geeth De Mel, and Jiawei Han. 2024. "ORAG: Ontology-Guided Retrieval-Augmented Generation for Theme-Specific Entity Typing" [in en]. August. Accessed July 23, 2025. https://openreview.net/forum?id=cKBmZ2PZ6c.

Xiong, Caiming, Victor Zhong, and Richard Socher. 2018. *Dynamic Coattention Networks For Question Answering.* ArXiv:1611.01604, March. Accessed July 26, 2025. DOI: 10.48550/arXiv.1611.01604. http://arxiv.org/abs/1611.01604.

Yang, Rui, Michael Fu, Chakkrit Tantithamthavorn, Chetan Arora, Lisa Vandenhurk, and Joey Chua. 2025. "RAGVA: Engineering retrieval augmented generation-based virtual assistants in practice" [in en]. *Journal of Systems and Software* 226 (August): 112436. ISSN: 01641212, accessed July 23, 2025. DOI: 10.1016/j.jss.2025.112436. https://linkinghub.elsevier.com/retrieve/pii/S0164121225001049.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. *XLNet: Generalized Autoregressive Pretraining for Language Understanding.* ArXiv:1906.08237, January. Accessed July 26, 2025. DOI: 10.48550/arXiv.1906.08237. http://arxiv.org/abs/1906.08237.

Yoon, Wonjin, Jinhyuk Lee, Donghyeon Kim, Minbyul Jeong, and Jaewoo Kang. 2019. *Pretrained Language Model for Biomedical Question Answering.* Accessed July 26, 2025. DOI: 10.48550/ARXIV.1909.08229. https://arxiv.org/abs/1909.08229.

Yue, Murong. 2025. *A Survey of Large Language Model Agents for Question Answering.* ArXiv:2503.19213, March. Accessed July 26, 2025. DOI: 10.48550/arXiv.2503.19213. http://arxiv.org/abs/2503.19213.

Zhou, Tianyu, Yuwei Wan, Ying Liu, and Maneesh Kumar. 2025. *Enabling interactive AI in industry 5.0 with RAG-enhanced GenAI Chatbots* [in en]. conference. Valencia, Spain. Accessed July 23, 2025. https://orca.cardiff.ac.uk/id/eprint/178617/.

Zhuang, Shengyao, and Guido Zuccon. 2021. "TILDE: Term Independent Likelihood moDEl for Passage Re-ranking" [in en]. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval,* 1483–1492. Virtual Event Canada: ACM, July. ISBN: 9781450380379, accessed July 27, 2025. DOI: 10.1145/3404835.3462922. https://dl.acm.org/doi/10.1145/3404835.3462922.