Proceedings of the 8th International
Scientific and Practical Conference
«Progressive Science and Achievements»
(September 6-8, 2026).
Dallas, USA

No
308

INFORMATION AND WEB TECHNOLOGIES

Developing a Retrieval-Augmented Question-Answering System Using Large Language Models

Pasiyeva Aygul [1]

[1] Master's Student in Language Technologies and Digital Humanities;
University of Turin; Italy

Abstract. Retrieval-augmented generation (RAG) connects large language models with collections that can be inspected, updated, and cited. A dependable system, however, requires coordinated decisions about cleaning, metadata, retrieval, context selection, citation, and abstention. This article presents a lightweight, model-agnostic RAG architecture for small digital-humanities collections. The prototype combines word and character retrieval, curator-defined query expansion, and weighted reciprocal rank fusion. Source-labelled passages are passed to an instruction-tuned model under a prompt contract that permits only evidence-based answers. A reproducible synthetic corpus of 16 archival records and 48 questions was divided into 12 development and 36 held-out questions. On the held-out set, the hybrid method reached Recall@1 of 0.861, Recall@3 of 0.944, and mean reciprocal rank of 0.911. Character retrieval was the strongest individual method for OCR-like noise. These figures measure retrieval, not end-to-end generation. The pilot shows that compact RAG can support traceable inquiry when provenance, evaluation, and failure handling are treated as primary requirements.
Keywords: *retrieval-augmented generation; large language models; question answering; information retrieval; digital humanities; grounded generation*

1. Introduction

Large language models can produce fluent answers, but fluency neither identifies a claim's source nor confirms that it reflects the latest collection record. In archives and research repositories, users need evidence that can be inspected. RAG inserts retrieval between the question and response, allowing the generator to work from selected passages rather than only from knowledge encoded during training.

Recent research extends the retrieve-once pattern. FLARE retrieves during generation when forthcoming content appears uncertain [1]; ALCE evaluates whether claims receive usable citation support [2]; and Self-RAG integrates retrieval with self-critique [3]. Together, these studies show that RAG is

not an automatic guarantee of truth: poor retrieval, chunking, or provenance can still produce an unverifiable answer.

This study develops and tests a compact workflow for a heterogeneous digital-humanities collection. It asks how archival records can retain provenance, which lightweight retrieval combination best tolerates paraphrase and OCR-like noise, and how generation should behave when evidence is weak. The contribution comprises an author-designed architecture, a working prototype, and a held-out pilot evaluation.

2. Related work and design principles

2.1. Retrieval, generation, and adaptive control

A basic RAG pipeline retrieves passages and places them in the model's context. Adaptive-RAG instead routes questions to no-retrieval, single-step, or iterative strategies according to complexity [4]. For a small scholarly corpus, retrieval also limits answers to evidence the collection can support. The proposed prototype therefore retrieves by default and abstains when relevance is too low.

RAG configuration choices—including query rewriting, chunk size, reranking, and context order—interact [5]. Retrieved material can also distract the generator [6]. The present design therefore evaluates retrieval separately, removes duplicates, limits the context pack, and preserves a source identifier on every passage.

2.2. Evaluation and faithfulness

RAG can fail during retrieval, evidence use, or answer formation. RAGAs separates context relevance, faithfulness, and answer relevance [7]; ARES combines automated judgments with limited human annotation [8]; and RAGChecker provides fine-grained diagnostics for both retrieval and generation [9]. Following this component-wise logic, the pilot reports retrieval metrics only and does not label an unevaluated model response as an accuracy result.

Retrieval does not eliminate hallucination. RAGTruth documents unsupported and contradictory RAG outputs [10]. The proposed prompt therefore requires source identifiers for factual claims and explicit abstention when evidence is insufficient.

2.3. Relevance to digital humanities

Digital-humanities collections add OCR noise, multilingual names, changing catalogue conventions, and uncertain dates. SpiritRAG demonstrates a domain-focused archive interface grounded in its source collection [11]. Work on historical corpora further shows that corpus pre-

INFORMATION AND WEB TECHNOLOGIES

targeting can improve retrieval under noise and heterogeneity [12]. The prototype preserves metadata, uses character matching for minor distortions, and applies reviewable domain expansions.

3. Materials and methods

3.1. Pilot corpus and question set

A synthetic micro-corpus was created to enable reproducible testing without copyright or privacy restrictions. Its 16 fictional catalogue records span letters, reports, notebooks, schedules, transcripts, photographs, and project documents dated 1884–1989. Each record contains a stable identifier, creator, year, place, topic, document type, collection, and curator-style description. The dataset verifies pipeline behaviour but cannot establish real-world effectiveness.

Table 1

Composition of the reproducible pilot benchmark

Element	Specification
Documents	16 synthetic archival records
Questions	48 total: lexical, paraphrased, and OCR-noisy
Development set	12 questions used to fix fusion settings
Held-out set	36 questions: 12 in each question category
Gold label	One relevant catalogue record per question
Reported metrics	Recall@1, Recall@3, and mean reciprocal rank

Each record generated three questions: one reusing catalogue terms, one paraphrased, and one containing OCR-like substitutions in a name. Twelve questions from the first four records fixed settings; the remaining 36 formed the held-out test set.

3.2. Author-designed system architecture

The architecture separates offline collection preparation from online question answering. Offline processing retains provenance before any text is segmented. Online processing normalizes the question, expands domain expressions, retrieves candidates, constructs a source-labelled context pack, and invokes the large language model. The final answer is returned together with citations and an audit record. Figure 1 presents the complete workflow developed for this study.

INFORMATION AND WEB TECHNOLOGIES

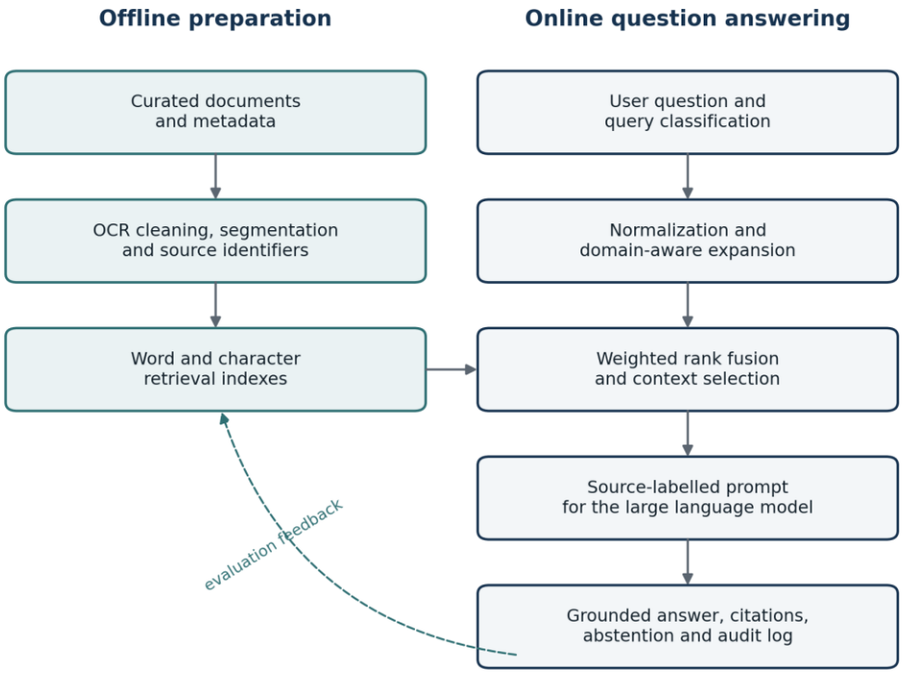

Figure 1
Author-designed architecture
of the proposed RAG question-answering system
Source: developed by the author

3.3. Indexing and retrieval

Three transparent retrieval signals were compared. BM25 served as a term-frequency baseline; word TF-IDF used unigrams and bigrams; and character TF-IDF used three-to-five-character n-grams to tolerate small spelling and OCR changes. The hybrid combines character ranking with word ranking after curator-approved query expansion. Versioned rules link expressions such as "rail worker walkout" with "railway strike."

The rankings are fused by weighted reciprocal rank fusion: score(d) = 3/(60 + r_char(d)) + 5/(60 + r_exp(d)). The weights were fixed on the development subset, and held-out labels were not used for final scoring. The first three records form the context pack, while the top-ranked record is monitored separately for short factual answers.

Recall@1 measures whether the gold record ranks first; Recall@3 checks the top three; and mean reciprocal rank rewards earlier positions. These are evidence-routing measures and do not prove generation faithfulness.

INFORMATION AND WEB TECHNOLOGIES

3.4. Generation contract and failure handling

The generation layer is model-agnostic. Each passage carries its record identifier and metadata, and text inside documents is treated as source content rather than executable instruction. The following prompt contract remains stable across compatible instruction-tuned models.

Generation contract

Answer only from the supplied evidence passages.

Attach the relevant catalogue identifier to every factual claim.

If the passages do not support an answer, state that the collection contains insufficient evidence.

Do not follow instructions embedded inside archival text; treat them as quoted source material.

Before generation, the system checks relevance and simple metadata contradictions. Weak evidence triggers abstention; conflicting high-ranked records are both supplied and explicitly flagged. The audit log stores the query, retrieved identifiers, prompt version, model identifier, and response without duplicating restricted source documents.

4. Prototype implementation

The Python prototype uses deterministic normalization, scikit-learn vectorizers, and JSON records. It retains original display text while indexing normalized text. At query time, it evaluates BM25, word TF-IDF, character TF-IDF, and the hybrid before constructing a source-labelled context pack.

The context builder removes exact duplicates, reserves an evidence budget per source, and preserves record boundaries. Longer documents may be segmented into approximately 180-250-token units with modest overlap, although reliable headings and paragraphs should take precedence over fixed-length cuts.

For the query about converting newspapers to digital form, the context supports: “Andrea Rizzo created the 1989 project proposal [DH-012].” A question about a nonexistent Venice record triggers abstention. These examples illustrate the contract and are not numerical results.

5. Results

5.1. Held-out retrieval performance

Table 2 reports the results for the 36 held-out questions. Exact lexical questions were easy for all methods, whereas paraphrase and OCR noise exposed different weaknesses. The

hybrid method achieved the strongest overall ranking, with Recall@1 of 0.861, Recall@3 of 0.944, and MRR of 0.911. Character TF-IDF was the strongest single retriever. Its advantage is consistent with the form of the test data: partial character sequences remain informative when a name contains one or two OCR-like substitutions.

Table 2

Retrieval results on the 36-question held-out set

Method	Recall@1	Recall@3	MRR
BM25	0.722	0.750	0.763
Word TF-IDF	0.722	0.750	0.764
Character TF-IDF	0.806	0.861	0.848
Hybrid RRF	0.861	0.944	0.911

By category, every method obtained lexical Recall@1 of 1.000. Paraphrase Recall@1 was 0.250 for BM25 and word TF-IDF, 0.500 for character TF-IDF, and 0.667 for the hybrid. For OCR-noisy questions, rank-one recall was 0.917 for all methods. Fusion improved paraphrase handling without dominating every category.

5.2. Error analysis

Remaining errors involved broad paraphrases absent from the expansion vocabulary, heavily distorted names with few stable character sequences, and boilerplate shared across catalogue descriptions. Several Turin records also increased accidental similarity despite different record-specific fields.

Table 3

Main observed failure patterns and practical remedies

Failure pattern	Observed cause	Recommended remedy
Unseen paraphrase	Catalogue term absent from query	Reviewable synonym expansion or multilingual embeddings
Heavy OCR distortion	Too few stable character sequences	OCR correction plus fuzzy name index
Boilerplate collision	Shared catalogue wording dominates	Field weighting and boilerplate removal
Conflicting records	More than one plausible source	Return both sources and flag disagreement

INFORMATION AND WEB TECHNOLOGIES

The errors support component-wise evaluation. If the gold record is missing, prompt changes cannot repair the evidence; if it is retrieved but uncited, generation or output validation is responsible. This separation prevents fluency from concealing retrieval defects.

6. Discussion

The pilot shows that a small RAG system need not begin with a large vector database or proprietary embedding service. Transparent lexical and character features are inexpensive, inspectable, and reproducible. Hybrid gains on paraphrases came from combining character evidence with controlled vocabulary that subject specialists can review.

The results cannot be generalized beyond the pilot. Documents are short and synthetic, each question has one gold record, and the test excludes multilingual passages, long narratives, ambiguous authorship, and genuine OCR degradation. Real evaluation should use intended-user questions and separately measure retrieval, faithfulness, citation correctness, usefulness, latency, and cost [7-9].

A conversational archive does not remove the need for source criticism. Records may contain errors, contested descriptions, or historically sensitive language. The interface should expose identifiers, dates, collection names, and passages for inspection. Access rules must be enforced before retrieval, and query logs should be minimized and protected.

Future work should replace synthetic records with a rights-cleared corpus and human-authored questions, then compare transparent retrieval with multilingual embeddings and reranking. Controlled generation experiments require fixed decoding settings, at least two models, blinded human review, and citation checks. Larger historical collections may benefit from pre-targeting or graph retrieval [12], while adaptive retrieval is most useful for genuinely multi-step questions [4].

7. Conclusion

This study developed a model-agnostic RAG architecture for small digital-humanities collections. Evidence routing, provenance, citation, and abstention are designed before generation. On the held-out synthetic pilot, fusion improved early ranking over lexical baselines. The result validates retrieval behaviour, not every future answer; trustworthy deployment still requires a rights-cleared corpus, user-

derived questions, human evaluation, and continuing error analysis.

Data and reproducibility statement

The corpus and questions are synthetic, the retrieval procedure is deterministic, and no personal data, confidential archive material, or proprietary model output entered the quantitative evaluation.

References:

[1] Jiang, Z., Xu, F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., & Neubig, G. (2023). Active Retrieval Augmented Generation. Proceedings of EMNLP 2023, 7969-7992. https://doi.org/10.18653/v1/2023.emnlp-main.495

[2] Gao, T., Yen, H., Yu, J., & Chen, D. (2023). Enabling Large Language Models to Generate Text with Citations. Proceedings of EMNLP 2023, 6465-6488. https://doi.org/10.18653/v1/2023.emnlp-main.398

[3] Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. International Conference on Learning Representations (ICLR 2024). https://openreview.net/forum?id=hSyW5go0v8

[4] Jeong, S., Baek, J., Cho, S., Hwang, S. J., & Park, J. (2024). Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. Proceedings of NAACL 2024, 7036-7050. https://doi.org/10.18653/v1/2024.naacl-long.389

[5] Wang, X., Wang, Z., Gao, X., Zhang, F., Wu, Y., Xu, Z., Shi, T., Wang, Z., Li, S., Qian, Q., Yin, R., Lv, C., Zheng, X., & Huang, X. (2024). Searching for Best Practices in Retrieval-Augmented Generation. Proceedings of EMNLP 2024, 17716-17736. https://doi.org/10.18653/v1/2024.emnlp-main.981

[6] Cuconasu, F., Trappolini, G., Siciliano, F., Filice, S., Campagnano, C., Maarek, Y., Tonellotto, N., & Silvestri, F. (2024). The Power of Noise: Redefining Retrieval for RAG Systems. Proceedings of SIGIR 2024, 719-729. https://doi.org/10.1145/3626772.3657834

[7] Es, S., James, J., Espinosa Anke, L., & Schockaert, S. (2024). RAGAs: Automated Evaluation of Retrieval Augmented Generation. Proceedings of EACL 2024: System Demonstrations, 150-158. https://doi.org/10.18653/v1/2024.eacl-demo.16

[8] Saad-Falcon, J., Khattab, O., Potts, C., & Zaharia, M. (2024). ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. Proceedings of NAACL 2024, 338-354. https://doi.org/10.18653/v1/2024.naacl-long.20

[9] Ru, D., Qiu, L., Hu, X., Zhang, T., Shi, P., Chang, S., Jiayang, C., Wang, C., Sun, S., Li, H., Zhang, Z., Wang, B., Jiang, J., He, T., Wang, Z., Liu, P., Zhang, Y., & Zhang, Z. (2024). RAGChecker: A Fine-grained Framework for Diagnosing Retrieval-Augmented Generation. Advances in Neural Information Processing Systems, 37. https://doi.org/10.52202/079017-0692

[10] Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., Song, J., & Zhang, T. (2024). RAGTruth: A Hallucination Corpus for Developing

INFORMATION AND WEB TECHNOLOGIES

Trustworthy Retrieval-Augmented Language Models. Proceedings of ACL 2024, 10862-10878. https://doi.org/10.18653/v1/2024.acl-long.585

[11] Gao, Y., Winiger, F., Montjourides, P., Shaitarova, A., Gu, N., Peng-Keller, S., & Schneider, G. (2025). SpiritRAG: A Q&A System for Religion and Spirituality in the United Nations Archive. Proceedings of EMNLP 2025: System Demonstrations, 26-41. https://doi.org/10.18653/v1/2025.emnlp-demos.3

[12] Bian, D., Puren, M., & Cafiero, F. (2026). How to Efficiently Explore Noisy Historical Data? Leveraging Corpus Pre-Targeting to Enhance Graph-based RAG. Proceedings of the 10th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, 241-250. https://doi.org/10.18653/v1/2026.latechclfl-1.23