How to Efficiently Explore Noisy Historical Data? Leveraging Corpus Pre-Targeting to Enhance Graph-based RAG

Donghan Bian[1,2] **Marie Puren**[2,1] **Florian Cafiero**[1]
[1]Centre Jean-Mabillon, École nationale des chartes - PSL, Paris, France
[2]EPITA Research Laboratory, EPITA, Paris, France
{donghan.bian, florian.cafiero}@chartes.psl.eu, marie.puren@epita.fr

Abstract

Graph-based Retrieval-Augmented Generation (RAG) is increasingly used to explore long, heterogeneous, and weakly structured corpora, including historical archives. However, in such settings, naive full-corpus indexing is often computationally costly and sensitive to OCR noise, document redundancy, and topical dispersion. In this paper, we investigate corpus pre-targeting strategies as an intermediate layer to improve the efficiency and effectiveness of graph-based RAG for historical research.

We evaluate a set of pre-targeting heuristics tailored to single-hop and multi-hop of historical questions on HistoriQA-ThirdRepublic, a French question-answering dataset derived from parliamentary debates and contemporary newspapers. Our results show that appropriate pre-targeting strategies can improve retrieval recall by 3–5% while reducing token consumption by 32–37% compared to full-corpus indexing, without degrading coverage of relevant documents.

Beyond performance gains, this work highlights the importance of corpus-level optimization for applying RAG to large-scale historical collections, and provides practical insights for adapting graph-based RAG pipelines to the specific constraints of digitized archives.

1 Introduction

Retrieval-Augmented Generation (RAG) has become a central paradigm for grounding large language models in external textual evidence, enabling more reliable generation in knowledge-intensive settings (Lewis et al., 2020; Gao et al., 2023; Poibeau, 2025). Recent work has shown that RAG can scale to long documents and complex queries, particularly when combined with graph-based representations that support multi-hop reasoning and global semantic structure (Larson and Truitt, 2024; Xiang et al., 2025). As a result, RAG-based systems are increasingly considered for exploratory access to large textual collections, including historical corpora. Beyond simple full-text search–often the default option for historians with limited digital experience–, RAG supports a more systematic, large-scale exploration of very large (and especially serial) sources. It also enables researchers to “dialogue” with their corpus, surfacing leads, aspects, and themes that were previously unknown, or less visible to domain experts, and that can subsequently be investigated through conventional methods of historical research.

At the same time, this promise comes with substantial practical challenges in archival settings. Historical archives constitute a particularly demanding testbed for such approaches. These documents are typically long, weakly structured, heterogeneous in genre, and are often affected by OCR noise (Piotrowski, 2012; Strange et al., 2014), a recurrent issue in DH corpora that often diverge from NLP benchmark assumptions (McGillivray et al., 2020). They also can exhibit strong redundancy and uneven information density, with large portions of text that are procedurally repetitive or only marginally relevant to a given research question. In this context, naive full-corpus indexing is not only computationally expensive, but can also amplify noise and degrade retrieval quality, especially in graph-based pipelines where indexing decisions directly affect graph construction costs.

While most existing work on RAG focuses on retrieval models, ranking strategies, or generation architectures (Gao et al., 2023), comparatively little attention has been paid to corpus-level decisions that precede retrieval, such as how much of a corpus should be indexed and under which constraints. This issue is particularly salient for historical research, where recent studies have shown both the potential and the limitations of applying RAG to digitized newspapers and parliamentary debates (Tran et al., 2024; Pellet et al., 2024). For large

Proceedings of SIGHUM 2026, pages 241–250
March 28-29, 2026 ©2026 Association for Computational Linguistics

historical collections, full-corpus indexing may be neither necessary nor desirable.

In this paper, we investigate corpus pre-targeting as an intermediate layer between raw historical archives and graph-based RAG pipelines. Pre-targeting consists in selectively reducing the indexed corpus based on query characteristics, with the goal of preserving relevant documents while filtering out material that is unlikely to contribute to answering a given class of questions. Building on recent proposals for pre-targeted RAG (Silvestre de Sacy et al., 2024), we hypothesize that different types of historical questions impose different retrieval constraints, and tailored pre-targeting strategies can improve the trade-off between retrieval effectiveness and computational cost.

We evaluate this hypothesis on HistoriQA-ThirdRepublic, a French-language question answering dataset derived from parliamentary debates and contemporary newspapers (Pellet et al., 2026). We distinguish between single-hop questions, which typically rely on localized factual information, and multi-hop questions, which require aggregating evidence across multiple documents. Our contributions are threefold: (1) a systematic evaluation of corpus pre-targeting strategies for graph-based RAG applied to historical archives; (2) empirical evidence that appropriate pre-targeting improves recall by 3–5% while reducing token consumption by 32–37%; and (3) practical insights for adapting RAG pipelines to the specific constraints of digitized historical collections.[1]

2 Related Work

Our work connects three research threads: (i) institutional and scholarly needs for source-grounded exploration of large historical archives, (ii) emerging Digital Humanities uses of RAG on noisy, heterogeneous collections, and (iii) graph-based RAG pipelines and corpus-reduction strategies that trade indexing cost for retrieval effectiveness.

Archive exploration, transparency, and accountability. Beyond academic use, access to institutional archives is increasingly framed in terms of transparency and accountability: computational methods can enable traceable, source-grounded claims over large documentary collections, provided that provenance and citation granularity remain explicit (Jo and Gebru, 2020; Cafiero, 2023; Cafiero et al., 2025). In that spirit, recent institutional initiatives experiment with question-answering interfaces over archival holdings, aiming to lower the barrier to exploration for non-specialists while preserving the ability to inspect underlying sources. For example, the National Library of Luxembourg has deployed a chatbot for exploring historical Luxembourgish newspapers (Bibliothèque nationale du Luxembourg, 2023), while the European Parliament Historical Archives provides an “Ask the EP archives” tool within a broader content-analysis environment for multilingual parliamentary material (European Parliament Historical Archives, 2024b,a).

RAG for historical research in Digital Humanities. RAG was initially introduced to ground generation on retrieved evidence for knowledge-intensive tasks (Lewis et al., 2020), and later work systematized a broad design space of retrieval, reranking, and generation modules (Gao et al., 2023). In Digital Humanities, RAG has been explored as a pragmatic approach to query long, heterogeneous, and OCR-noisy corpora where traditional search and close reading alone do not scale. As a result, RAG is increasingly being evaluated for examining large collections of historical documents (Nandula and Shenoy, 2024; Fan et al., 2025; Lee et al., 2025). Representative case studies include, in particular, historical newspapers (Tran et al., 2024) and reflections on RAG for parliamentary debates, with explicit discussion of methodological benefits (rapid exploration, cross-source synthesis) and risks (retrieval failures, over-trusting generated summaries) (Pellet et al., 2024). A recurring conclusion in this literature is that corpus characteristics–redundancy, uneven information density, OCR artifacts, and document structure–often dominate downstream performance, sometimes more than the choice of generator model.

Graph-based RAG and corpus reduction via pre-targeting. Graph-augmented RAG methods aim to improve evidence aggregation and multi-hop retrieval by exploiting explicit structure (entities/relations, neighborhoods, traversals) rather than relying solely on local semantic similarity. Survey work clarifies the main families of GraphRAG–like pipelines–graph construction, graph-guided retrieval, and graph-enhanced generation—and discusses when graphs help compared to simpler RAG stacks (Peng et al., 2025; Zhang et al., 2025a; Xiang et al., 2025). Concrete peer-reviewed

[1] Scripts available at Github

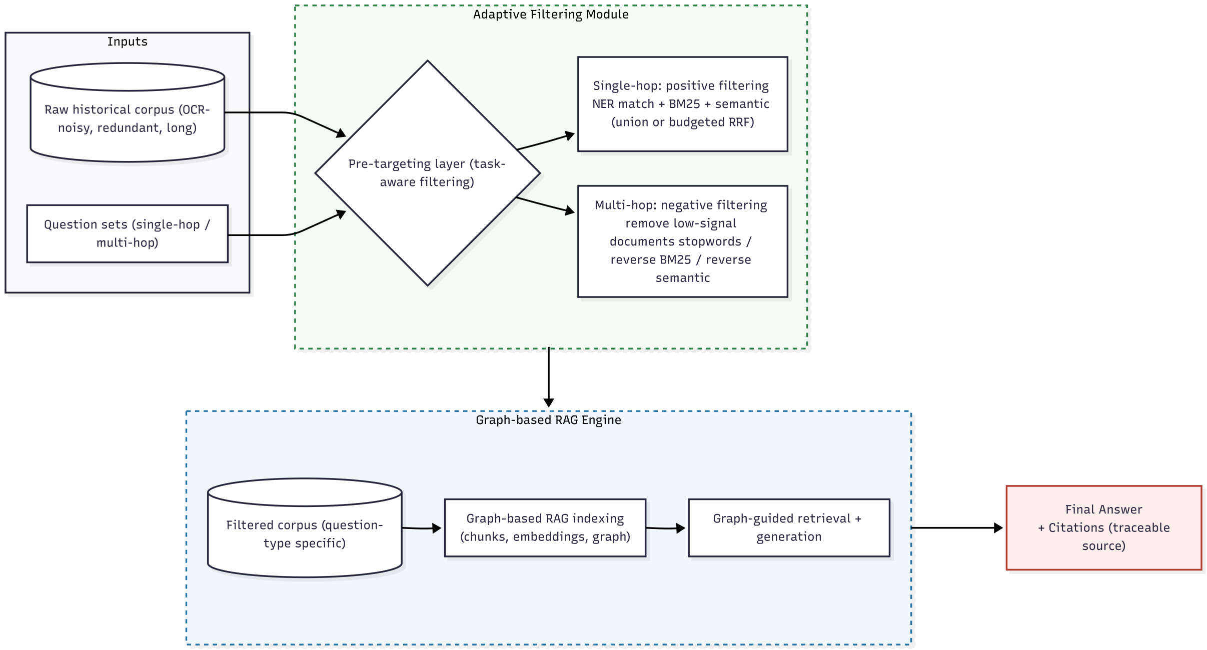

Inputs
Raw historical corpus (OCR-noisy, redundant, long)
Question sets (single-hop / multi-hop)
Adaptive Filtering Module
Pre-targeting layer (task-aware filtering)
Single-hop: positive filtering NER match + BM25 + semantic (union or budgeted RRF)
Multi-hop: negative filtering remove low-signal documents stopwords / reverse BM25 / reverse semantic
Graph-based RAG Engine
Filtered corpus (question-type specific)
Graph-based RAG indexing (chunks, embeddings, graph)
Graph-guided retrieval + generation
Final Answer + Citations (traceable source)

Figure 1: Overview of the pipeline. A task-aware pre-targeting layer reduces the indexed corpus differently for single-hop (positive filtering; union or budgeted Reciprocal Rank Fusion) and multi-hop questions (negative filtering). The filtered corpus feeds a shared Graph-based RAG indexing and QA stack, improving the recall–cost trade-off.

instantiations include HippoRAG, which combines LLM-driven graph induction with graph traversal mechanisms to improve multi-hop retrieval (Gutiérrez et al., 2024), and TRACE, which constructs knowledge-grounded reasoning chains over triples extracted from retrieved documents to integrate dispersed evidence (Fang et al., 2024).

However, graph construction can become the dominant cost on large archives, and indexing everything can amplify OCR noise and procedural redundancy. This motivates corpus reduction strategies that act before indexing, by filtering the corpus under query- or task-specific constraints while attempting to preserve gold evidence. Pre-targeted RAG formalizes this idea as a pre-filtering layer that reduces the indexed set prior to retrieval and generation (Silvestre de Sacy et al., 2024). We extend this line of work in a historical setting by (i) tailoring pre-targeting to question type (single-hop vs. multi-hop) and (ii) evaluating the resulting trade-offs specifically within a graph-based RAG pipeline on OCR-noisy parliamentary debates and newspapers.

3 Methodology

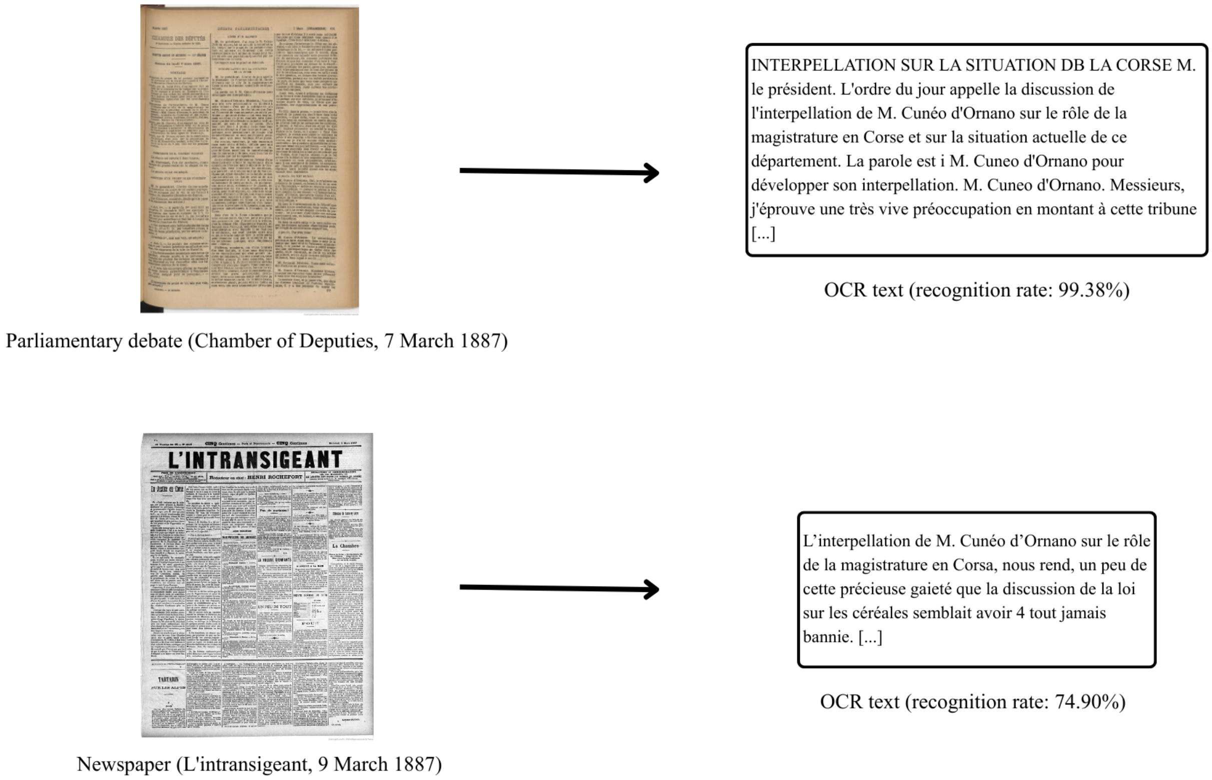
In this section, we outline our methodology. We begin by introducing the HistoriQA-ThirdRepublic dataset, and then present the strategies employed in the pre-targeting process along with the evaluation protocol used to assess their effectiveness.

3.1 HistoriQA-ThirdRepublic

HistoriQA-ThirdRepublic is a French-language question answering corpus derived from parliamentary debates and newspapers of the French Third Republic (1870-1940) (Pellet et al., 2026). The dataset contains historical questions designed to evaluate RAG systems in domain-specific contexts, specifically focusing on complex reasoning patterns typical of historical inquiry.

The dataset is constructed from three primary sources, focusing specifically on texts from 1887: parliamentary debates from the Chamber of Deputies, and two contemporary newspapers, *Le Gaulois* and *L'Intransigeant*. For the parliamentary debate texts, content not directly related to debates, such as agendas, procedural notes, and vote lists, was removed in advance. After applying a specific segmentation strategy, the final text corpus comprises 78 documents from *Le Gaulois*, 79 documents from *L'Intransigeant*, and 3,227 documents from parliamentary debates, as shown in figure 2.

Questions were then generated through iterative refinement using prompt engineering and validated by a historian to ensure factual and contextual coherence. Single-hop questions were generated ex-



INTERPELLATION SUR LA SITUATION DB LA CORSE M. le président. L'ordre du jour appelle la discussion de l'interpellation de M. Cunéo d'Ornano sur le rôle de la magistrature en Corse et sur la situation actuelle de ce département. La parole est i M. Cuneo d'Ornano pour développer son interpellation. M. Cuneo d'Ornano. Messieurs, j'éprouve une très vive préoccupation en montant à cette tribune [...]

OCR text (recognition rate: 99.38%)

Parliamentary debate (Chamber of Deputies, 7 March 1887)

L'interpellation de M. Cunéo d'Ornano sur le rôle de la magistrature en Corsa, nous rend, un peu de cette précieuse gaieté que la discussion de la loi sur les céréales semblait avoir 4 tout jamais bannie. [...]

OCR text (recognition rate: 74.90%)

Newspaper (L'intransigeant, 9 March 1887)

Figure 2: Corpus sources and OCR quality. The OCR texts are generated from scanned historical documents, and the recognition accuracy information is obtained from the corresponding Gallica pages.

clusively based on the content of parliamentary debates, whereas multi-hop questions were constructed following two strategies, newspaper-to-debate and newspaper-to-newspaper, resulting in three types of questions: follow-up questions (identifying opinions in debates and examining press reactions), comparative questions (contrasting differing viewpoints across sources), and bridge-entity questions (connecting sources through shared references).

To clarify the task distinction, we provide representative examples from HistoriQA-ThirdRepublic.

- **Single-hop example.** A question answerable from one debate document (localized factual evidence): *Quel est le rôle du bureau d'âge et comment est-il composé lors de l'ouverture de la session ordinaire de la Chambre des députés en 1887 ?* (What was the role of the Bureau of Age and how was it composed at the opening of the ordinary session of the Chamber of Deputies in 1887?)

- **Multi-hop example.** A question requiring evidence aggregation across at least two documents/sources (e.g., debate + newspaper, or newspaper + newspaper): *En 1887, comment le baron de Mackau critique-t-il la politique éducative du gouvernement, et quelle analyse l'article de presse en donne-t-il ?* (In 1887, how did Baron de Mackau criticize the government's educational policy, and what analysis did the newspaper article provide?)

These examples reflect the operational criterion used in our experiments: single-hop targets local evidence, whereas multi-hop requires cross-document synthesis.

The final QA dataset comprises 897 single-hop questions and 885 multi-hop questions, the indexing data (document ID) required to answer the questions is also integrated into the QA dataset.

The corpus contains OCR noise as shown in Figure 2. We adopted a fast and deliberately conservative heuristic to estimate the word error rate (WER), treating as errors words that exceed 20 characters in length, contain interleaved digits and letters, or include non-French characters. This approach yields an estimated WER of approximately 2.73%, which should be regarded as a lower bound, as the OCR error rates reported by Gallica, the digital library of the Bibliothèque nationale de France, for these sources are generally higher.[2]

3.2 Pre-targeting strategies regarding different type of questions

Overall, our approach is inspired by the methodology proposed by Silvestre de Sacy et al. (2024) The core objective is to define constraints for pre-filtering the corpus based on the characteristics of the query under investigation, thereby reducing the amount of text to be indexed as much as possible while preserving the target documents (gold documents) required to answer the questions. This strategy aims to lower the cost of graph construction and indexing time, partially mitigate the impact of OCR noise, and ultimately improve the performance of RAG systems during the question-answering stage.

With respect to question types, existing RAG approaches generally perform better on single-hop questions than on multi-hop questions (Larson and Truitt, 2024; Gutiérrez et al., 2025). This discrepancy stems from the fundamentally different nature of the two question types. Single-hop questions can be answered using information contained within a single document and are typically tied to concrete facts. In most cases, answers can be directly extracted from the original text. In contrast, multi-hop questions require aggregating and synthesizing information across multiple related documents, which places higher demands on the robustness of the retrieval stage as well as the long-context processing capabilities of large language models.

These observations directly motivate our experimental design. We evaluate the effectiveness of pre-targeting separately for single-hop and multi-hop questions in order to identify the most suitable strategies for each case. For each question type, we randomly sample ten sets of fifty questions from the HistoriQA-ThirdRepublic QA dataset without replacement and report the average performance of different pre-targeting strategies across the ten sets.

For single-hop questions, we assume that precise semantic matching is sufficient to rapidly localize the documents required to answer a given question. Accordingly, we adopt three complementary strategies. First, we apply GLiNER (Zaratiana et al., 2024), an advanced named entity recognition model, to detect four types of entities, including person, location, date, and organization, in the query, and subsequently search for these entities in the text corpus to retrieve the corresponding documents. Second, after tokenizing the query, we perform a bag-of-words retrieval over the entire corpus using BM25 (Robertson et al., 2009) and retain the top 500 ranked documents. Third, we segment the corpus into chunks of 1,000 tokens with an overlap of 200 tokens, compute the cosine similarity between the query and each chunk using the Qwen3-Embedding-0.6B model (Zhang et al., 2025b), and treat the document containing a chunk as a candidate document whenever the similarity exceeds 0.7. We further evaluate the union of the results produced by these three strategies to assess whether their combination yields additional improvements in retrieval performance.

In contrast to single-hop questions, multi-hop questions exhibit weaker direct semantic alignment with individual documents. Consequently, the objective of pre-targeting shifts from directly locating relevant texts to filtering out documents with low information density or those potentially unrelated to the question-answering task. We again employ three strategies. First, we define a stop word list and consider a document to be low in informational content if stop words account for more than 20% of its text, in which case the document is filtered out.[3] Second, we rank all documents using BM25 for each question and exclude any document that appears among the bottom 500 documents for at least one question. Finally, we compute cosine similarity using the same embedding model, but with a different criterion: if a document's similarity score with respect to all questions does not exceed 0.6, it is considered as unrelated documents. These strategies yield three lists of excluded documents, and we further explore the union of these lists to identify the most effective combination of filtering strategies.

The thresholds and cutoffs used in pre-targeting (e.g., cosine thresholds, BM25 rank cutoffs, stop-word ratio) are operational hyperparameters selected for this corpus and task configuration, not universal constants. Our objective is to evaluate whether task-aware corpus reduction can improve the recall–cost trade-off under realistic archival constraints, rather than to claim corpus-independent optimal values. In practice, these parameters should be re-calibrated when corpus characteristics

[2] OCR error information is available on the document's Gallica page, here is an example.

[3] The stop-word list is available in our GitHub repository (block 19).

change (OCR quality, document length distribution, genre balance, or language), and we therefore interpret the reported gains as evidence of the approach's usefulness in this setting, with external transfer requiring additional validation.

3.3 Retrieval performance

Based on the selection of the optimal pre-targeting strategies, we choose three filtered corpora and compare their performance against full-corpus indexing on the corresponding question sets. We adopt HippoRAG 2, one of the state-of-the-art RAG frameworks (Gutiérrez et al., 2025; Xiang et al., 2025), for retrieval and generation. For embedding and generation models, we use Qwen3-Embedding-0.6B and DeepSeek-V3.2 (Liu et al., 2025), respectively. To support the former, we implement additional scripts to enable compatibility between HippoRAG 2 and the Qwen 3 model architecture, while the latter is accessed via an API. Finally, we report both the recall performance of the different configurations and their token consumption. Figure 3 illustrates an example of a knowledge graph generated under full-corpus indexing.

4 Results

4.1 Pre-targeting stage results

We first evaluate the performance of different filtering strategies in the pre-targeting stage. Table 1 summarizes the results using three key metrics: recall percentage, which measures the proportion of gold documents retained relative to the total set of gold documents; precision percentage, representing the proportion of gold documents in filtered corpus; document count, indicating the number of documents remaining after filtering; and document coverage, defined as the percentage of documents preserved from the original corpus.

For single-hop questions, individual filtering strategies showed varying performance characteristics. BM25 Filtering achieved the highest recall at 97.00% while maintaining relatively low document coverage (34.93%, 1182.8 documents). In contrast, NER Filtering demonstrated 84.00% recall but required substantially more documents (1891.7 documents, 55.87% coverage), and Cosine Similarity Filtering achieved only 62.80% recall with the smallest document set (120.5 documents, 3.56% coverage).

When combining strategies, BM25 + Cosine Similarity (Union) emerged as the optimal approach, achieving 97.20% recall with 1185.4 documents (35.01% coverage). The motivation for adopting the union of the results is to account for the complementary information captured by bag-of-words models and dense vector representations. In addition, we apply Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) to combine these two lists of results by weighting documents according to their positions in each list, while constraining the size of the fused list to the longer of the two inputs to prevent an excessive expansion of the candidate set, thereby assessing their relative importance under the two ranking schemes. We observe that RRF attains the same recall as the union strategy, suggesting that it successfully fuses the complementary features captured by the two retrieval strategies in an efficient manner. Although the union of all three strategies (NER + BM25 + Cosine) reached the highest recall at 98.80%, it came at the cost of substantially increased document coverage (64.75%), making it less efficient for practical deployment.

For multi-hop questions, which require reasoning across multiple documents, reverse filtering strategies proved efficient. Among individual strategies, Reverse Cosine Similarity Filtering achieved the best balance with 98.64% recall using 1828.6 documents (54.00% coverage), outperforming both Reverse BM25 Filtering (98.15% recall, 61.52% coverage) and Reverse Stopwords Filtering (84.95% recall, 72.15% coverage). Combined strategies further improved performance, with Reverse BM25 + Cosine (Union) reaching 99.39% recall at 74.96% coverage. The most comprehensive approach, combining all three reverse strategies, achieved near-perfect recall at 99.52% but required processing 98.74% of the document collection (3343.3 documents), offering minimal practical advantage over simpler combinations.

Based on these results, we selected BM25 + Cosine Similarity (RRF) for single-hop questions and Reverse Cosine Similarity Filtering for multi-hop questions as our optimal strategies for the subsequent RAG pipeline (Figure 1), as they provide the best balance between high recall and computational efficiency.

These results should be read as corpus- and setup-specific: they show that simple task-aware heuristics can be effective on HistoriQA-ThirdRepublic, but they do not imply that the same thresholds will transfer unchanged to other histori-

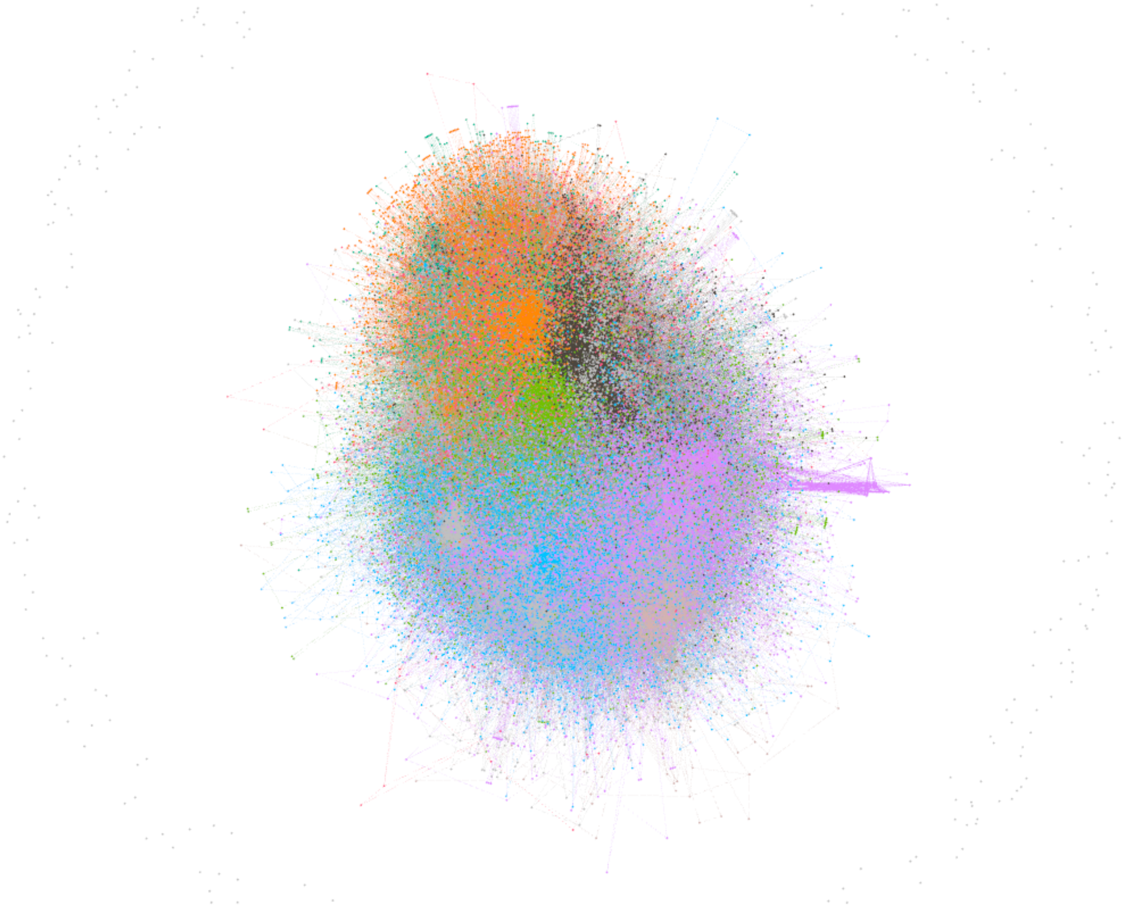


Figure 3: Knowledge graph generated from the HistoriQA-ThirdRepublic dataset under full-corpus indexing. The network comprises 30,514 nodes and 648,792 edges. We employed the Louvain algorithm (Blondel et al., 2008) for community detection, identifying 332 distinct communities with a modularity of 0.619, indicating a robust community structure. For visualization, the nodes are color-coded by their community membership, and the network layout is optimized using the Yifan Hu multilevel force-directed algorithm (Hu, 2005) to enhance interpretability.

cal collections. The main transferable claim is the principle of differentiated pre-targeting for single-hop versus multi-hop questions.

4.2 RAG performance

We then compared the performance of default full indexing against the pre-targeting approach across different retrieval depths (top-3, top-5, and top-10) for both question types. Table 2 presents the recall rates and token consumption for each strategy, with numbers in brackets representing the total tokens used across all three experimental runs.

For single-hop questions, the pre-targeting strategy demonstrated consistent improvements over full indexing across all retrieval depths. Pre-targeting achieved average recall rates of 55.33% at top-3, 62.00% at top-5, and 70.67% at top-10, representing improvements of 4.66, 2.67, and 2.00 percentage points respectively over full indexing (50.67%, 59.33%, and 68.67%). These improvements were achieved while using substantially fewer tokens on average: 12,156,845.7 tokens compared to 19,434,237 for full indexing, representing a 37.5% reduction in computational cost. The standard deviations across runs (±6.80, ±7.12, and ±6.60 for top-3, top-5, and top-10 respectively) indicate reasonable stability in performance.

The advantages of pre-targeting were more pronounced for multi-hop questions. Pre-targeting achieved recall rates of 43.00%, 52.67%, and 62.67% at top-3, top-5, and top-10 respectively, outperforming full indexing by 5.00, 4.34, and 2.00 percentage points. The performance gains were accompanied by a 32.1% reduction in average token usage (13,194,585.3 tokens versus 19,434,237 for full indexing). The lower standard deviations for multi-hop questions (±5.66, ±4.71, and ±2.87) suggest more stable performance compared to single-hop questions, particularly at higher retrieval depths where the standard deviation

Strategy	Recall (%)	Precision (%)	Doc Count	Doc Coverage (%)
Single-Hop Questions				
Individual Strategies:				
NER Filtering	84.00	2.28	1891.7	55.87
BM25 Filtering	97.00	4.10	1182.8	34.93
Cosine Similarity Filtering	62.80	26.69	120.5	3.56
Combined Strategies:				
BM25 + Cosine Similarity (Union)	97.20	4.10	1185.4	35.01
BM25 + Cosine Similarity (RRF)	**97.20**	**4.11**	**1182.8**	**34.93**
NER + BM25 + Cosine (Union)	98.80	2.29	2192.3	64.75
Multi-Hop Questions				
Individual Strategies:				
Reverse Stopwords Filtering	84.95	2.79	2443.0	72.15
Reverse BM25 Filtering	98.15	3.79	2083.1	61.52
Reverse Cosine Similarity Filtering	**98.64**	**4.35**	**1828.6**	**54.00**
Combined Strategies:				
Reverse BM25 + Cosine (Union)	99.39	3.14	2538.3	74.96
Reverse Stopwords + BM25 + Cosine (Union)	99.52	2.39	3343.3	98.74

Table 1: Average Performance of Individual and Combined Filtering Strategies (10 sets). The bold rows indicate our selected optimal strategies for the subsequent RAG pipeline. The low precision is a result of sparsity of gold documents relative to the scale of the corpus.

Strategy	Recall@3 (%)	Recall@5 (%)	Recall@10 (%)	Token Consumption (total)
Single-Hop				
Full Indexing	50.67	59.33	68.67	19,434,237
Pre-targeting (std.)	55.33 (±6.80)	62.00 (±7.12)	70.67 (±6.60)	12,156,845.7 (36,470,537)
Multi-Hop				
Full Indexing	38.00	48.33	60.67	19,434,237
Pre-targeting (std.)	43.00 (±5.66)	52.67 (±4.71)	62.67 (±2.87)	13,194,585.3 (39,583,756)

Table 2: Comparison of Full Indexing and Pre-targeting Performance. The information of token consumption is obtained from the official API platform.

dropped to ±2.87. Additionally, it is noteworthy that despite the substantial disparity in the total number of documents between the single-hop and multi-hop text corpora, their token consumption remains relatively close. This phenomenon can be attributed to the significant variation in token counts across texts from different sources in the original dataset (Pellet et al., 2026).

The pre-targeting approach demonstrated substantial computational efficiency gains while improving retrieval performance across both question types. For single-hop questions, the strategy reduced token usage by 37.5% while consistently improving recall rates. For multi-hop questions, despite requiring more tokens than single-hop pre-targeting due to the inherent complexity of multi-hop reasoning and also the different filtering strategies, pre-targeting still achieved a 32.1% reduction in computational cost compared to full indexing. These results indicate that pre-targeting not only enhances retrieval quality but also provides meaningful computational benefits, making it a more practical solution for large-scale RAG applications.

5 Discussion

Our results demonstrate that corpus pre-targeting is a viable strategy for enhancing graph-based RAG, achieving 3–5% higher recall while reducing token consumption by 32–37% in our experimental setting. This suggests that optimizing the document preprocessing pipeline can complement ongoing improvements in model architectures and retrieval algorithms, particularly for domain-specific applications where computational resources are limited. It should be noted, however, that practical deployment requires corpus-specific calibration (e.g., threshold tuning and rank cutoffs), since the heuristics evaluated here are not expected to generalize unchanged across collections with different OCR quality, textual genres, or language distributions.

Our approach assumes effective document identification through lexical-semantic similarity, which may not generalize to all query types. The upfront filtering cost (12-13M tokens) remains substantial for single queries, though this amortizes across multiple queries on the same corpus. Additionally, our evaluation on a single historical dataset requires validation across diverse domains and languages.

Promising directions include developing lightweight query routers for automatic strategy selection, adaptive filtering with dynamic thresholds, iterative pre-targeting to recover initially missed documents, and extending to multi-lingual scenarios. Incorporating OCR noise evaluation and the impact of the level of noise on performance could help refine our approaches depending on the corpus treated.

Acknowledgements

We would like to express our sincere gratitude to Aurélien Pellet for his invaluable support and assistance with the dataset. This work has been supported by PSL's CultureLab project.

References

Bibliothèque nationale du Luxembourg. 2023. L'intelligence artificielle au service du patrimoine imprimé luxembourgeois. Press release on the launch of the chatbot.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Florian Cafiero. 2023. Datafying diplomacy: How to enable the computational analysis and support of international negotiations. *Journal of Computational Science*, 71:102056.

Florian Cafiero, Jean-Philippe Cointet, and Grégoire Mallard. 2025. Digital accountability can relegitimate multilateralism. Preprint / working paper, HAL.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759.

European Parliament Historical Archives. 2024a. Archives content analysis dashboard and "ask the ep archives" (project page). Project web page (no peer-reviewed publication identified for this citation key).

European Parliament Historical Archives. 2024b. The ep archives unit launches its first generative ai tool. Web page describing the "Ask the EP archives" tool within the Archives Content Analysis Dashboard.

Yang Fan, Zhang Qi, Xing Wenqian, Liu Chang, and Liu Liu. 2025. Research on graph-retrieval augmented generation based on historical text knowledge graphs. *Preprint*, arXiv:2506.15241.

Jinyuan Fang, Zaiqiao Meng, and Craig MacDonald. 2024. Trace the evidence: Constructing knowledge-grounded reasoning chains for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8472–8494. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: neurobiologically inspired long-term memory for large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. From RAG to memory: Non-parametric continual learning for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 21497–21515. PMLR.

Yifan Hu. 2005. Efficient, high-quality force-directed graph drawing. *Math J*, 10:37.

Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: strategies for collecting sociocultural data in machine learning. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 306–316. ACM.

Jonathan Larson and Steven Truitt. 2024. Graphrag: Unlocking llm discovery on narrative private data.

Jeong Ha Lee, Ghazanfar Ali, and Jae-In Hwang. 2025. A retrieval-augmented generation system for accurate and contextual historical analysis: Ai-agent for the annals of the joseon dynasty. *Computer Animation and Virtual Worlds*, 36(4):e70048.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. DeepSeek-V3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Barbara McGillivray, Thierry Poibeau, and Pablo Ruiz. 2020. Digital Humanities and Natural Language Processing: "Je t'aime... Moi non plus". *Digital Humanities Quarterly*, 14(2).

Srianusha Nandula and Saachi Shenoy. 2024. Enhancing historical understanding with retrieval augmented generation. Retrieved 2026-01-05.

Aurélien Pellet, Julien Perez, and Marie Puren. 2024. Generative artificial intelligence and historical research: Challenges, potentials, and limitations. application of RAG to french parliamentary debates of the third republic (1881–1940). In *A Conversation between AI and the Humanities*.

Aurélien Pellet, Marie Puren, and Julien Perez. 2026. HistoriQA-ThirdRepublic: Multi-Hop Question Answering Corpus for Historical Research, Parliamentary Debates from the. Working paper or preprint.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2025. Graph retrieval-augmented generation: A survey. *ACM Trans. Inf. Syst.*, 44(2).

Michael Piotrowski. 2012. *Natural Language Processing for Historical Texts*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Thierry Poibeau. 2025. *Understanding Conversational AI*. Ubiquity Press, London.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and trends® in information retrieval*, 3(4):333–389.

Antoine Silvestre de Sacy, Adam Faci, Stéphane Pouyllau, and Léa Maronet. 2024. Pre-targeted-RAG: Retrieval Augmented Generation sur des groupes préciblés de communautés d'articles de recherche.

Carolyn Strange, Daniel McNamara, Josh Wodak, and Ian Wood. 2014. Mining for the meanings of a murder: The impact of ocr quality on the use of digitized historical newspapers. *Digital Humanities Quarterly*, 8(1).

The Trung Tran, Carlos-Emiliano González-Gallardo, and Antoine Doucet. 2024. Retrieval augmented generation for historical newspapers. In *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, pages 1–5.

Zhishang Xiang, Chuanjie Wu, Qinggang Zhang, Shengyuan Chen, Zijin Hong, Xiao Huang, and Jinsong Su. 2025. When to use graphs in RAG: A comprehensive analysis for graph retrieval-augmented generation. *arXiv preprint arXiv:2506.05690*.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. Gliner: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376.

Qinggang Zhang, Chuanjie Wu, Zhishang Xiang, and 1 others. 2025a. A survey of graph retrieval-augmented generation for customized large language models. *arXiv preprint arXiv:2501.13958*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025b. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.