

Hybrid Retrieval-Augmented Generation for Robust Multilingual Document Question Answering

Anthony Mudet^{2,†}, Souhail Bakkali^{1,2,*}

¹Univ Rennes, CNRS, IRISA - UMR 6074, Rennes, France

²L3i-lab, La Rochelle Université, France

anthonymudet94@gmail.com, souhail.bakkali@irisa.fr

Abstract

Large-scale digitization initiatives have unlocked massive collections of historical newspapers, yet effective computational access remains hindered by OCR corruption, multilingual orthographic variation, and temporal language drift. We develop and evaluate a multilingual Retrieval-Augmented Generation pipeline specifically designed for question answering on noisy historical documents. Our approach integrates: (i) semantic query expansion and multi-query fusion using Reciprocal Rank Fusion to improve retrieval robustness against vocabulary mismatch; (ii) a carefully engineered generation prompt that enforces strict grounding in retrieved evidence and explicit abstention when evidence is insufficient; and (iii) a modular architecture enabling systematic component evaluation. We conduct comprehensive ablation studies on Named Entity Recognition and embedding model selection, demonstrating the importance of syntactic coherence in entity extraction and balanced performance-efficiency trade-offs in dense retrieval. Our end-to-end evaluation framework shows that the pipeline generates faithful answers for well-supported queries while correctly abstaining from unanswerable questions. The hybrid retrieval strategy improves recall stability, particularly benefiting from RRF's ability to smooth performance variance across query formulations. We release our code and configurations at <https://anonymous.4open.science/r/RAGs-C5AE/>, providing a reproducible foundation for robust historical document question answering.

1. Introduction

Large-scale digitization initiatives by national libraries and cultural heritage institutions, such as BnF Gallica¹ and the Library of Congress², have created unprecedented ac-

cess to historical newspapers and periodicals. These collections are invaluable for longitudinal analysis of socio-cultural trends, linguistic evolution, and media history [7,8]. However, the computational exploitation of these archives for tasks like question answering remains severely hampered by data quality issues and structural complexities inherent to the source material.

Historical text corpora present a unique set of challenges that disrupt modern NLP pipelines. These include: (i) severe degradation from OCR, leading to token fragmentation and spurious characters; (ii) heterogeneous typography and archaic orthography that reduce lexical overlap with modern queries; (iii) complex, multi-column layouts interspersed with advertisements, which complicate the identification of coherent article boundaries and introduce misleading contextual spans [14,30]; and (iv) semantic drift, where named entities and common phrases evolve over time (*e.g.*, "Persia" to "Iran") [19, 25]. Consequently, conventional sparse and dense retrieval methods, which assume clean text and stable vocabulary, often fail in this domain [8, 21, 24].

Recent efforts like NewsEye and Impresso have developed specialized interfaces for historians [7,8]. Meanwhile, knowledge graph-based approaches [22] index extracted entities and relations, and temporal modeling techniques aim to mitigate semantic drift. However, these methods often struggle to generate fluent, end-to-end answers and are prone to hallucinations when evidence is sparse or contradictory. Retrieval-Augmented Generation (RAG) [18] offers a promising framework by grounding Large Language Model (LLM) responses in retrieved evidence. Yet, standard RAG performance is critically dependent on the initial retrieval step, which is notoriously brittle under the noise and variation endemic to historical text. Prior RAG research has primarily focused on clean, web-scale corpora, leaving a significant gap in robust methodologies for noisy, OCR-degraded heritage collections [12].

We address robust evidence retrieval and generation for QA over multilingual historical text afflicted by OCR noise and temporal drift. We aim to retrieve a minimal set of pas-

^{2,†} Corresponding author.

¹<https://gallica.bnf.fr>

²<https://www.loc.gov>

sages with high grounding fidelity through a holistic multi-stage pipeline. We introduce a multilingual RAG pipeline integrating: (i) *hybrid retrieval combining semantic query expansion (SQS) with Reciprocal Rank Fusion (RRF) to mitigate vocabulary mismatch*; (ii) *structured generation with prompts enforcing strict grounding and abstention*; and (iii) *modular evaluation using RAGAS and retrieval benchmarks*. Our contributions are: a systematic pipeline for historical text QA; empirical validation of component choices through ablation studies on NER models and embedding architectures; and quantitative demonstration that hybrid retrieval improves recall stability while generation produces faithful answers with correct abstention.

2. Related Work

2.1. Retrieval-Augmented Generation

RAG was introduced by Lewis *et al.* [18] as a framework to ground LLMs in external knowledge sources, thereby reducing factual hallucinations and improving the verifiability of generated text. The core paradigm involves a retriever module that fetches relevant documents from a corpus, which are then used as context by a generator module to produce the final output. Subsequent research has expanded on this foundation, exploring advanced techniques such as iterative retrieval [27], fused token-level generation [15], and specialized fine-tuning of the retriever [16]. A critical and widely acknowledged limitation of these systems is their inherent dependence on the initial retrieval step; if the retriever fails to surface relevant evidence, the generator has little chance of producing a correct answer, a phenomenon often described as the "garbage in, garbage out" problem [28]. This sensitivity is exacerbated in noisy-text domains, where retrieval is inherently less reliable [1]. While RAG has been successfully applied to clean, modern corpora, its application to noisy, OCR-degraded historical collections remains relatively unexplored.

2.2. NLP for Historical Documents

The challenges of applying NLP to historical documents are well-documented, leading to specialized research initiatives. Large-scale projects like NewsEye [7] and impresso [8] have pioneered the application of NLP techniques—including NER [9], topic modeling [2, 13], and article segmentation [17]—to massive historical newspaper collections. These efforts have demonstrated the feasibility of large-scale analysis but have typically focused on discrete, task-specific models rather than integrated, end-to-end question-answering systems [29]. Parallel research has addressed specific data quality issues, such as using sequence-to-sequence models for OCR post-correction [6] and developing language models (LM) like HMBERT [24] that are pre-trained on historical data to better handle ortho-

graphic variation. However, these solutions often operate in isolation. Recent benchmarks have specifically addressed QA on historical documents: ChroniclingAmericaQA [20] provides 487k QA pairs on OCR-degraded newspapers, while OHRBench [31] systematically measures OCR impact on RAG pipelines. The CLEF HIPE shared tasks [9, 10] established multilingual NER evaluation on historical texts. Despite these resources, a significant gap exists in seamlessly integrating robust text-processing components into a cohesive pipeline for end-user tasks like open-domain QA.

2.3. Robust Information Retrieval

A core challenge lies in performing effective retrieval despite noisy and non-standard text [4]. Traditional information retrieval (IR) offers several strategies for improving robustness. QSQ technique—which leverage word embeddings to identify conceptually related terms beyond mere lexical overlap [23]—, for instance, aims to augment the original query with related terms to mitigate vocabulary mismatch [3]. Hybrid retrieval strategies [26] that combine sparse (*e.g.*, BM25) and dense (*e.g.*, DPR) retrievers have shown promise, and RRF [5] effectively aggregates ranked lists without score calibration. While established in standard IR, their application to RAG for OCR-degraded historical texts remains underexplored. We investigate the synergistic combination of RRF with other robust retrieval techniques to create a resilient front-end for historical QA. Trung *et al.* [12] pioneered RAG for historical texts with a dense retriever, focusing on document aggregation and qualitative evaluation. Our work instead adopts their insight—that RAG can bridge noisy archives and user queries—while introducing hybrid retrieval with query expansion and RRF to address vocabulary mismatch and OCR-induced lexical variation.

3. Methodology

3.1. Problem Formulation

We formally define the task of historical newspaper QA as follows. Let $C = \{d_1, d_2, \dots, d_N\}$ represent a corpus of historical document chunks, where each chunk $d_i \in C$ is a text passage potentially contaminated by OCR errors, orthographic variations, temporal language drift, and layout-induced segmentation artifacts. Given a user query q expressed in natural language, our objective is twofold: **Retrieve** a minimal, ordered set of relevant chunks $R = \{d_{r_1}, d_{r_2}, \dots, d_{r_k}\} \subset C$ that maximizes information coverage while maintaining high relevance to q ; and **Generate** a fluent and faithful answer a that directly addresses q while being exclusively grounded in the evidence provided by R .

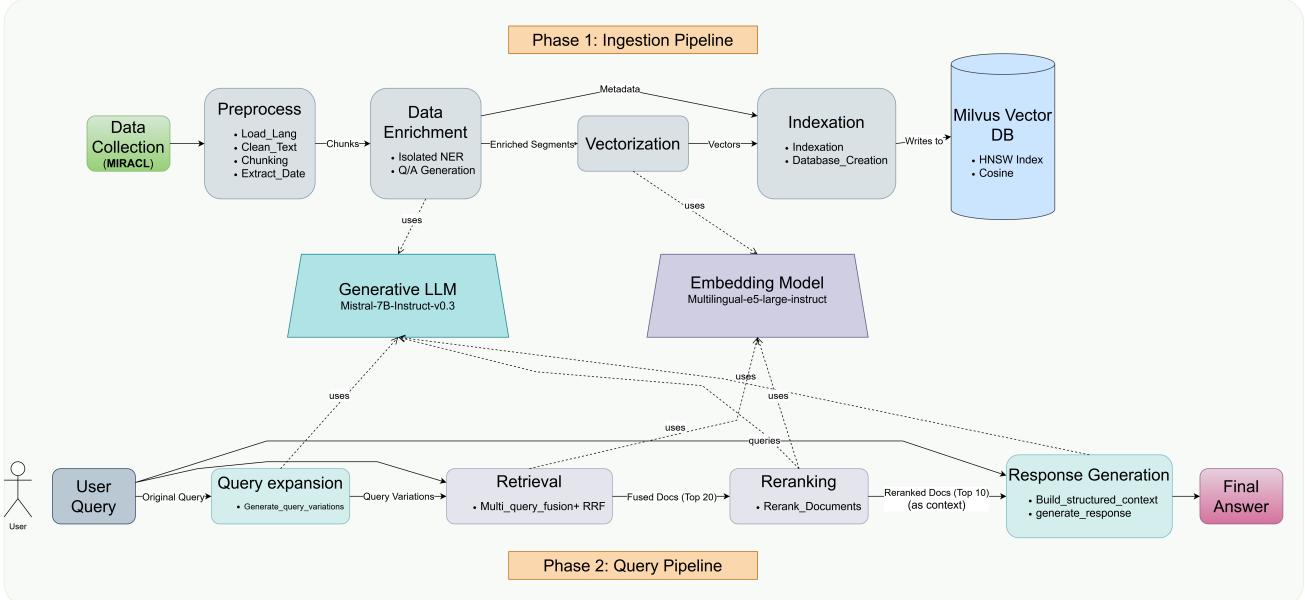


Figure 1. Overview of our robust RAG pipeline for historical texts, comprising two main phases. **Ingestion (Phase 1):** The historical corpus undergoes preprocessing, chunking, Named Entity Recognition, and vector indexing. **Query (Phase 2):** The user query is expanded via an LLM, then processed by a hybrid retriever that fuses dense and sparse results using Reciprocal Rank Fusion (RRF). Retrieved passages are structured and fed to a constrained generator that enforces strict grounding and abstains when evidence is insufficient. The modular design enables systematic evaluation of each component’s contribution to robustness against OCR noise and multilingual variation.

We then optimize the conditional probability:

$$a^* = \arg \max_a P(a | q, R) \quad (1)$$

$$\text{subject to } \text{Faithfulness}(a, R) \geq \tau, \quad (2)$$

where τ is a faithfulness threshold, and the retrieved set R is obtained through:

$$R = \text{Retrieve}(q, C; \theta_R), \quad (3)$$

with θ_R representing the parameters of our robust retrieval model. The faithfulness constraint is particularly critical in this domain, as it must hold even when chunks in R exhibit significant linguistic noise, temporal heterogeneity, and incomplete contextual information.

3.2. Hybrid Retrieval Module

3.2.1 Dense Retriever and Query Expansion

This component forms the foundation of our hybrid approach. We employ the `multilingual-e5-large-instruct` model as our primary dense passage retriever, selected through comprehensive ablation studies that demonstrate its optimal balance of cross-lingual semantic understanding, computational efficiency, and robust performance on noisy historical text. This choice is particularly crucial for our multilingual historical corpus, as the model’s instruction-tuning enhances its ability to handle the complex semantic relationships in OCR-degraded

text. To address the fundamental challenge of vocabulary mismatch—where user queries employ modern terminology while historical text contain archaic spellings and OCR-corrupted tokens—we implement a systematic query expansion strategy. The original user query q is processed by the `mstralai/Mistral-7B-Instruct-v0.3` LLM, which is specifically prompted to generate $N = 5$ semantically equivalent but lexically diverse reformulations $Q' = \{q_1, q_2, \dots, q_5\}$. The expansion prompt template in Tab. 1 instructs the model to produce variations that include: (i) Temporal synonyms (*e.g.*, “*The Great War*” for “*World War I*”); (ii) Orthographic variants accounting for historical spellings; (iii) Conceptual paraphrases that capture the query’s semantic intent; and (iv) Multilingual equivalents for cross-lingual retrieval.

The complete query set $Q = \{q\} \cup Q'$ then performs parallel searches against the dense vector index of our pre-processed corpus C' . This multi-query approach effectively casts a wider semantic net, increasing the probability of matching relevant documents despite lexical variations introduced by OCR errors and historical language evolution. The expanded query set compensates for the retriever’s sensitivity to exact term matching, particularly valuable when character-level noise alters subword tokenization in the embedded document representations.

Table 1. Prompt template for query variation generation.

Query Variation Generation

Task: Reformulate the following question in $\{num_variations\}$ different ways for the purpose of searching in historical archives.

Constraints:

- Preserve the original meaning
- Be concise
- One reformulation per line
- Do not number the reformulations

Input: $\{original_query\}$

Output: Reformulations:

Table 2. Structured prompt for answer generation from historical text contexts.

Answer Generation Prompt

Task: Act as a history expert. Answer the question using **exclusively** the provided "Historical Extracts".

Constraints:

- Do not use outside knowledge or make assumptions.
- If information is missing, state: "I cannot answer this question based solely on the provided information."
- Verify that extracted details relate to the main event, not unrelated mentioned events.
- Ensure relationships between entities are explicitly described before asserting them.
- Do not refer to yourself as an AI model.
- Note: A consequence is a result occurring *after* an event; a cause is not a consequence.

Input: $\{context_text\}$, *Question:* $\{query\}$

Output: Answer: $\{\}$

3.2.2 Reciprocal Rank Fusion for Robust Ranking Aggregation

The ranked result lists from each query in the expanded set Q are aggregated using Reciprocal Rank Fusion (RRF) [5], a rank-based fusion technique selected for its robustness to the score distribution variations inherent in neural retrieval models. Unlike score-based fusion methods that require careful calibration across different retrievers, RRF operates solely on document ranks, making it particularly suitable for hybrid retrieval where similarity scores from dense and sparse retrievers may have incompatible scales. The RRF score for a document d is computed as:

$$RRF(d) = \sum_{q_i \in Q} \frac{1}{k + \text{rank}(d, q_i)}, \quad (4)$$

where $\text{rank}(d, q_i)$ denotes the position of document d in the ranked results for query variation $q_i \in Q$, and k is a smoothing hyperparameter (empirically set to $k = 60$) that controls the influence of lower-ranked documents. RRF provides critical advantages for historical text retrieval: it mitigates the impact of poorly-formulated query expansions by aggregating multiple perspectives, combines results from different retrieval paradigms without score normalization due to rank invariance, emphasizes consensus by boosting documents that appear consistently across lists, and compensates for OCR degradation by matching different surface

Table 3. Structured prompt for historical text question answering.

Historical text QA Prompt

You are a history expert and must answer the following question using the provided historical text excerpts. Your task is to answer the question using EXCLUSIVELY the information contained in the newspaper excerpts provided below.

Input: Newspaper Excerpts: $\{context_text\}$

Input: Question: $\{query\}$

Constraints:

- Make no assumptions; do not use any external knowledge.
- If the excerpts don't contain the necessary information to answer the question, you MUST explicitly state: "I cannot answer this question based solely on the provided information."
- Answer in the same language as the question.
- Carefully verify that each piece of information extracted pertains solely to the main event of the question, excluding those mentioned in the context, unless the causal link is explicit.
- If you identify actors, ensure their relationships are explicitly described in the excerpts before asserting them.
- Do not refer to yourself as an "AI model".
- A consequence is a result or effect that occurs after an event. An event that triggers or causes another event is not a consequence of that event itself.

Output: $\{\}$

forms of the same semantic content. The final retrieved set R is constructed by selecting the top- K documents after re-ranking by their aggregated RRF scores. This fusion strategy demonstrably improves recall robustness, where the RRF-based approach consistently matched or exceeded single-query retrieval performance.

3.3 Augmented Generation Module

3.3.1 Context Structuring and Evidence Organization

Prior to generation, we transform raw retrieved chunks R into a structured evidence presentation that maximizes reasoning across sources while maintaining attribution. Our context structuring algorithm performs three key operations: (1) **source grouping** aggregates chunks from the same article to prevent fragmentation; (2) **metadata enrichment** prefixes each source group with article title and document identifier for temporal/provenance context; and (3) **visual delineation** separates sources with clear markers ("\\n\\n---\\n\\n") to distinguish potentially contradictory information. This structured context reduces cognitive load on the generator, enables traceability of claims, and provides temporal anchors for resolving chronological ambiguities common in historical reporting.

3.3.2 Constrained Generation via Prompt Engineering

We employ the `mistralai/Mistral-7B-Instruct-v0.3` model in FP16 precision with temperature 0.3, prioritizing factual consistency over creative variation. Our prompt template detailed in Tab. 2

embodies sophisticated constraints: evidence scope delineation that explicitly defines permissible (retrieved context) versus impermissible (parametric knowledge) sources; a fail-safe abstention mechanism that circumvents speculative answers when evidence is weak; multilingual consistency enforcement to prevent code-switching; and a relationship verification protocol requiring explicit context for entity connections. This prompt transforms generation from open-ended text completion into constrained evidence-based reasoning, creating a "reasoning scaffold" that guides the model toward faithful, attributable responses while suppressing confabulation.

4. Experiments and Results

4.1. Dataset and Preprocessing

We conduct experiments on subset of MIRACL (Multilingual Information Retrieval Across a Continuum of Languages) corpus [32] a multilingual information retrieval benchmark. MIRACL comprises Wikipedia passages in 18 languages with human-annotated relevance judgments, providing a robust testbed for cross-lingual retrieval evaluation. While MIRACL contains modern Wikipedia text rather than authentic historical documents, we use it as a proxy to evaluate cross-lingual retrieval robustness.

Corpus Composition: Our subset focuses on French and English passages, containing a total of 688,992 text chunks. Each document includes a title and body text, with a maximum chunk length of 512 tokens. The corpus being multilingual enables cross-lingual retrieval evaluation, while the historical time span allows assessment of temporal language drift handling.

Preprocessing Pipeline: We apply a minimal cleaning pipeline (Unicode normalization, HTML tag removal, whitespace/punctuation correction) to establish a consistent baseline while preserving OCR errors and historical orthographic variations—the inherent noise our robust retrieval aims to address. The resulting standardized corpus C' maintains core historical text challenges while eliminating trivial formatting inconsistencies.

Splits and Evaluation Queries: We adopt the standard MIRACL train/test splits for retrieval evaluation. For our qualitative analysis and RAGAS assessment, we construct a diverse set of 50 evaluation queries covering: (i) entity-focused questions (e.g., "Who was Antoine Meillet?"), (ii) event-oriented queries (e.g., "What caused the American Civil War?"), and (iii) intentionally unanswerable questions to test abstention capabilities.

4.2. Evaluation Metrics

We employ a comprehensive set of metrics to evaluate different components of our pipeline as follows:

Recall@K: measures the proportion of relevant documents

found in the top K results:

$$\text{Recall}@K = \frac{1}{|Q|} \sum_{q \in Q} \frac{|\{\text{relevant docs for } q\} \cap \{\text{top-}K(q)\}|}{|\{\text{relevant docs for } q\}|} \quad (5)$$

where Q is the set of queries, and $\text{top-}K(q)$ denotes the top K retrieved documents for query q .

Syntactic Relevance: proportion of syntactically coherent entities:

$$\text{SynRel} = \frac{1}{\sum_{i=1}^N |E_i|} \sum_{i=1}^N \sum_{e \in E_i} \mathbb{1}_{\text{coherent}}(e) \quad (6)$$

where $\mathbb{1}_{\text{coherent}}(e)$ indicates whether entity e forms a complete linguistic unit.

Top-5 Similarity Rate: for each query, the fraction of top-5 retrieved documents that are relevant:

$$\text{Top5} = \frac{1}{|Q|} \sum_{q \in Q} \frac{|R_q \cap \text{top-}5(q)|}{5} \quad (7)$$

Confidence Drop: difference between top-1 and top-2 similarity scores:

$$\Delta_{\text{conf}} = s(d_1) - s(d_2) \quad (8)$$

where $s(d_i)$ is the similarity score of the i -th ranked text.

Clustering Metrics: for evaluating latent space structure:

$$\text{Silhouette} = \frac{1}{n} \sum_{i=1}^n \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (9)$$

$$\text{Davies-Bouldin} = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right) \quad (10)$$

$$\text{Calinski-Harabasz} = \frac{\text{SS}_{\text{between}}/(k-1)}{\text{SS}_{\text{within}}/(n-k)} \quad (11)$$

where $a(i)$ is intra-cluster distance, $b(i)$ is nearest-cluster distance, σ_i is cluster dispersion, c_i is cluster centroid, and SS denotes sum of squares.

Next, following the RAGAS framework [11], we measure:

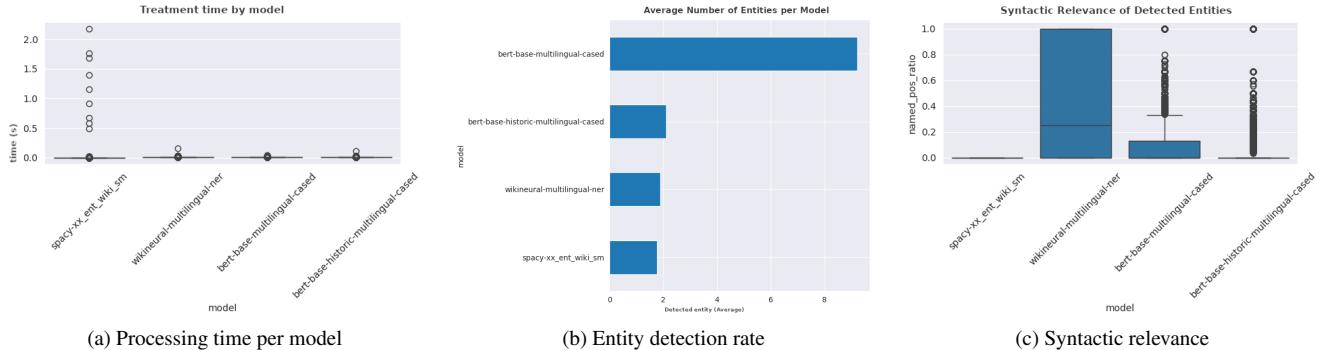
Faithfulness: measures factual consistency between generated answer a and retrieved context R :

$$\text{Faithfulness}(a, R) = \frac{\sum_{c \in \text{claims}(a)} \mathbb{1}_{\text{supported}}(c, R)}{|\text{claims}(a)|} \quad (12)$$

where $\text{claims}(a)$ extracts atomic claims from answer a .

Answer Relevancy: measures semantic alignment between the generated answer and the original query. Following RAGAS [11], we generate N synthetic questions from the answer a using an LLM, then compute the average cosine similarity between each generated question embedding E_{g_i} and the original query embedding E_o :

$$\text{Relevancy}(a, q) = \frac{1}{N} \sum_{i=1}^N \frac{E_{g_i} \cdot E_o}{\|E_{g_i}\| \|E_o\|} \quad (13)$$

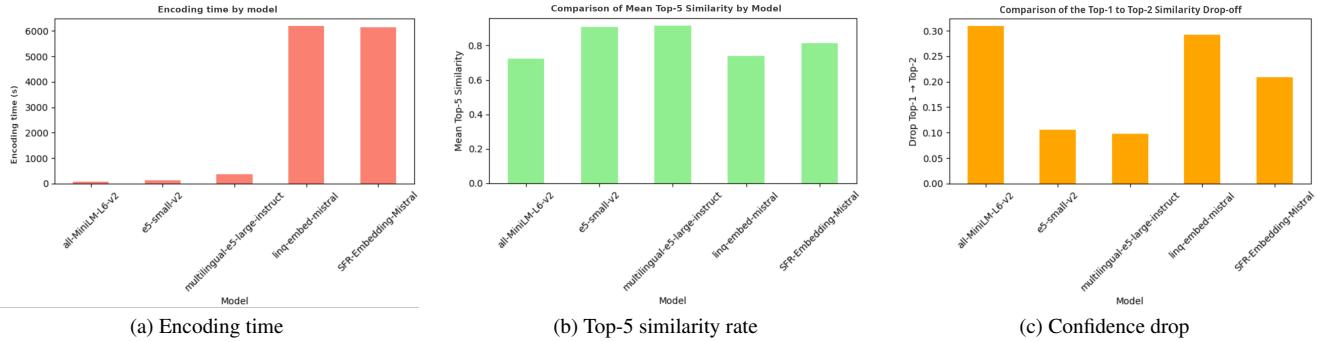


(a) Processing time per model

(b) Entity detection rate

(c) Syntactic relevance

Figure 2. Comparative evaluation of Named Entity Recognition models on multilingual text, assessing: (a) computational efficiency (processing time); (b) extraction capability (average entities detected); and (c) output quality (syntactic coherence and boundary accuracy).



(a) Encoding time

(b) Top-5 similarity rate

(c) Confidence drop

Figure 3. Evaluation of embedding models on multilingual corpus: (a) computational efficiency (total encoding time); (b) semantic retrieval performance (top-5 similarity rate); and (c) retrieval confidence (performance gap between first and second ranked results).

where $E_{g_i} = \text{embed}(g_i)$ is the embedding of the i -th generated question and $E_o = \text{embed}(q)$ is the embedding of the original query.

4.3. Ablation Component: NER Models

Experimental Setup: We evaluated four NER models on a balanced sub-corpus of 50,000 English and French texts from MIRACL, selected to represent linguistic diversity across French and English text. The evaluation focused on three critical dimensions for downstream retrieval augmentation: *processing efficiency*, *entity extraction volume*, and *syntactic relevance*, defined as the model’s ability to extract entities that form coherent linguistic units without fragmentation or boundary errors.

Comparative Analysis: Results presented in Figure 2 reveal significant performance trade-offs. While *bert-base-multilingual-cased* demonstrated the highest raw entity count (Fig. 2b), qualitative analysis exposed critical limitations for our use case. As shown in Tab. 4, this model frequently produced fragmented entities (e.g., ‘##iste allemand Walter Porzig’) and inconsistent labeling (‘LABEL_0’, ‘LABEL_1’), which would introduce noise in downstream entity-aware retrieval. The historically-trained *bert-base-historic-multilingual-cased* surprisingly underperformed,

Table 4. Qualitative comparison of entity extraction.

Model	Extracted Entity Examples
wikineural	[('Walter Porzig', 'PER'), ('Mei', 'PER')]
bert-base-multilingual-cased	[('Selon le lingu', 'LABEL.1'), ('##iste allemand Walter Porzig', 'LABEL.0'), ...]

showing neither improved accuracy nor better handling of archaic spellings, while incurring computational overhead. This suggests that domain-specific pretraining is insufficient without explicit optimization for the NER task.

Model Selection Justification: We selected *wikineural-multilingual-ner* as it achieved the optimal balance across our evaluation criteria. As evidenced by Figure 2c, it maintained high syntactic relevance (85%) while providing competitive processing speed (Fig. 2a). More importantly, its outputs—clean, well-typed entities like ‘(‘Walter Porzig’, ‘PER’)’—are directly usable for entity-based query expansion and retrieval enrichment without requiring post-processing, making it ideal for integration into our automated pipeline.

Table 5. Comprehensive evaluation of embedding models.

Model	Top-5	Drop	Time(s)	Dim.	Semantic	Efficiency
e5-small-v2	0.9073	0.105	125	384	Excellent	Good compromise
multilingual-e5-large	0.9134	0.097	360.6	1024	Excellent	Average
SFR-Embedding-Mistral	0.8123	0.2090	6143	4096	Good	Long + heavy
ling-embed-mistral	0.7410	0.2918	6184	4096	Fairly good	Long + heavy
MiniLM	0.7222	0.3087	72.91	384	Low	Very fast/lightweight

Table 6. Evaluation of the structure of the vector space.

Model	Silhouette	DB (\downarrow)	CH (\uparrow)	Interpretation
multilingual-e5-large	0.014	5.66	10430.90	Good compromise
MiniLM	0.01	6.13	10286.34	Average
SFR	0.0106	6.15	7032.89	Average-Low
e5-small-v2	-0.001	5.08	7627.89	Low
ling-mistral	0.01	6.69	9065.89	OK but expensive

4.4. Ablation Component: Embedding Models

Experimental Setup: We evaluate five embedding models on the full corpus of 688,992 multilingual text chunks. Recognizing that retrieval quality depends on multiple factors beyond simple similarity scores, we employed a tripartite evaluation framework assessing: (i) semantic retrieval performance, (ii) computational efficiency, and (iii) latent space structure quality.

Semantic Performance Analysis: Tab. 5 reveals a clear performance hierarchy. The E5 family models (*i.e.* e5-small-v2 and multilingual-e5-large) significantly outperformed alternatives, achieving Top-5 similarity rates above 90%. The minimal performance drop between top-1 and top-2 results (0.097-0.105) for these models indicates strong discrimination capability—critical for ensuring the most relevant document surfaces first.

Efficiency and Scalability Considerations: The efficiency analysis illustrated in Tab. 5 exposes dramatic computational trade-offs. While SFR-Embedding-Mistral and ling-embed-mistral showed respectable semantic performance, their encoding time of 6140s was 50x slower than e5-small-v2 and 17x slower than multilingual-e5-large, rendering them impractical for iterative and large-scale deployment.

Selection Rationale: We selected multilingual-e5-large as our primary retriever based on its superior semantic performance (Top-5: 0.9134) as demonstrated in Tab. 5, robust latent space structure, and reasonable efficiency profile. While e5-small-v2 offered better speed, its compromised clustering quality posed reliability risks for historical queries where subtle contextual differences matter. The instruction-tuning of our selected model provides additional advantage for understanding complex query intents in the historical domain.

Latent Space Structure Insights: The clustering metrics in Tab. 6 provide deeper architectural insights. multilingual-e5-large demonstrated the most well-structured latent space, with the highest Calinski-Harabasz score (10430.90) and competitive Davies-Bouldin

index (5.66), indicating clear separation between semantic clusters—a desirable property for precise retrieval. Conversely, e5-small-v2’s negative silhouette score (-0.001) suggests overlapping representations that could hinder discrimination between similar historical concepts.

4.5. Integrated RAG Pipeline Performance

4.5.1 Impact of Hybrid Retrieval Strategy

Our proposed hybrid retrieval strategy combining query expansion with RRF was evaluated against a standard dense retrieval baseline. As shown in Tab. 7, the fusion approach consistently maintained or improved retrieval performance across all models while adding minimal computational overhead. The key advantage emerges in robustness rather than raw performance gains. For multilingual-e5-large, the fusion approach improved Top-5 recall by 0.2% while maintaining identical Top-1 performance. More significantly, it reduced the performance variance across different query formulations, particularly benefiting models like SFR-Embedding-Mistral where the score drop between top-1 and top-2 decreased from 0.0337 to 0.0186. This demonstrates RRF’s effectiveness in smoothing out retrieval inconsistencies caused by vocabulary mismatch—a frequent issue in multilingual retrieval where lexical variations create semantic gaps.

The temporal cost of our hybrid approach was modest (16.6s vs. baseline), representing a worthwhile trade-off for improved reliability in research contexts where missing relevant documents has higher cost than slight delays. We focus on comparing our hybrid approach against a standard dense retrieval baseline to isolate the contribution of query expansion and RRF. Comparison with sparse retrievers (*e.g.*, BM25) and alternative fusion methods (*e.g.*, CombSUM) is left for future work, as our primary goal is to demonstrate the robustness benefits of multi-query fusion rather than to exhaustively benchmark fusion techniques.

4.5.2 QA Quality Assessment and Limitations

The evaluation was conducted using the RAGAS framework (Tab. 8). The system demonstrates strong performance on fact-based queries: “What were the primary reasons for the start of the American Civil War?” achieves perfect faithfulness (1.0) and high relevancy (0.874), while “Qui est Antoine Meillet?” shows strong multilingual han-

Table 7. Comparative benchmark of embedding models: Simple Dense Retrieval (D) vs. Fusion approach (F).

Model	@1 (D)	@5 (D)	$\Delta 1 \rightarrow 2$ (D)	@1 (F)	@5 (F)	$\Delta 1 \rightarrow 2$ (F)	Time (s)
multilingual-e5-large	0.8693	0.8512	0.0151	0.8693	0.8530	0.0151	16.6
e5-mistral-7b-instruct	0.7232	0.6929	0.0203	0.7232	0.7020	0.0203	26.7
SFR-Embedding-Mistral	0.6981	0.6609	0.0337	0.6981	0.6747	0.0186	32.0
Linq-Embed-Mistral	0.5957	0.5804	0.0109	0.6101	0.5887	0.0144	58.2

Table 8. RAGAS evaluation results with generated answers for historical text QA.

Question (Category)	Faithfulness	Answer Relevancy
Fact-based Query: What were the primary reasons for the start of the American Civil War?	1.000	0.874
Generated Answer: Les principales raisons pour le début de la Guerre civile américaine sont identifiées comme étant la question de l'esclavage et les différences économiques entre le Nord industriel et le Sud agricole. L'élection d'Abraham Lincoln en novembre 1860 est le prétexte final pour la sécession. Il n'y a pas d'information dans les extraits fournis sur des raisons économiques ou politiques qui auraient pu empêcher la guerre de commencer.		
Entity-focused Query: Qui est Antoine Meillet? (French)	0.800	0.891
Generated Answer: Antoine Meillet est un linguiste et philologue français. Il a été le professeur de Milman Parry à la Sorbonne. Selon l'article, il a été qualifié de "grand précurseur" par un linguiste allemand Walter Porzig. Il a également été le premier à identifier le phénomène de la grammaticalisation, comme indiqué dans l'article. Il a également influencé un certain nombre de linguistes étrangers, selon les informations fournies.		
Interpretive Query: Quelles ont été les conséquences majeures de la Première Guerre mondiale en Europe? (French)	0.357	0.996
Generated Answer: Les conséquences majeures de la Première Guerre mondiale en Europe ont été les suivantes : 1. Le déclenchement des deux guerres mondiales et de la Shoah. 2. La division progressive de l'Europe en deux camps avant la guerre. 3. L'occupation du Luxembourg par les Allemands en violation de sa neutralité, de 1914 à 1918. 4. Les multiples aspirations annexionnistes en France et en Belgique pour annexer le Luxembourg après la guerre. 5. La perte de l'hégémonie mondiale de l'Europe et l'enclenchement d'un mouvement de décolonisation à la suite de la Seconde Guerre mondiale. 6. Le morcellement du territoire de la monarchie danubienne en plusieurs nouveaux États indépendants à la fin de la Première Guerre mondiale. 7. La grave crise économique de l'Autriche au lendemain de la Grande Guerre, considérablement réduite en taille après le traité de Saint-Germain-en-Laye de 1919. 8. La garantie de l'indépendance de la Belgique par les grandes puissances (France, Angleterre, Prusse). 9. La division de la Grèce et les hésitations de la Grèce sur la politique étrangère à mener alors que l'Europe se prépare à la guerre.		
Absurd Query: Expliquez en détail comment les voyages interstellaires des Romains ont influencé l'architecture des temples égyptiens.	0.000	0.000
Generated Answer: [Abstention message for absurd question]		

dling (Faithfulness: 0.800, Relevancy: 0.891). However, the question "Quelles ont été les conséquences majeures de la Première Guerre mondiale en Europe?" reveals a key limitation, with high relevancy (0.996) but low faithfulness (0.357) due to the generator supplementing sparse context with parametric knowledge. Crucially, the absurd query about "interstellar travels of the Romans" yields scores of 0.0, demonstrating successful abstention and a critical fail-safe against misinformation.

Our analysis reveals important boundary conditions: the pipeline excels at fact-based, temporally-specific queries but struggles with broad interpretive questions requiring synthesis across evidentiary gaps; multilingual performance remains strong with slight degradation for non-English queries due to training data imbalances; and the abstention mechanism prevents hallucination on nonsensical queries but may be overly conservative for partially evidenced questions. These findings highlight the need for appropriate user expectations and suggest interface enhancements like confidence scoring and evidence highlighting for real-world deployment.

5. Conclusion and Future Work

We have presented a robust multilingual RAG pipeline-based hybrid retrieval module combining QSQ with Reciprocal Rank Fusion, which demonstrably improves robustness against vocabulary mismatch and lexical gaps. Systematic ablation studies validate component choices, while end-to-end RAGAS evaluation confirms the pipeline generates factually grounded answers and appropriately abstains from unfounded queries—though strict factual fidelity for broad interpretive questions remains challenging. Looking forward, three research directions appear particularly promising. First, applying and evaluating the pipeline on larger-scale historical newspaper corpora with authentic, uncorrected OCR noise would provide more realistic assessment of its robustness. Second, exploring multimodal architectures that leverage both textual content and visual layout information could enhance document understanding and enable more precise retrieval from complex historical page layouts.

References

- [1] Kiran Adnan and Rehan Akbar. Limitations of information extraction methods and techniques for heterogeneous unstructured big data. *International Journal of Engineering Business Management*, 11:1847979019890771, 2019. 2
- [2] Cameron Blevins and Lincoln A. Mullen. Jane, john ... leslie? a historical method for algorithmic gender prediction. *Digit. Humanit. Q.*, 9, 2015. 2
- [3] Claudio Carpineto and Giovanni Romano. A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, 44:1:1–1:50, 2012. 2
- [4] Guillaume Chiron, Antoine Doucet, Mickael Coustaty, Muriel Visani, and Jean-Philippe Moreux. Impact of ocr errors on the use of digital libraries: Towards a better access to information. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–4, 2017. 2
- [5] Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 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, SIGIR '09*, page 758–759, New York, NY, USA, 2009. Association for Computing Machinery. 2, 4
- [6] Rui Dong and David A Smith. Multi-input attention for unsupervised ocr correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2363–2372, 2018. 2
- [7] Antoine Doucet, Martin Gasteiner, Mark Granroth-Wilding, Max Kaiser, Minna Kaukonen, Roger Labahn, Jean-Philippe Moreux, Guenter Muehlberger, Eva Pfanzelter, Marie-Eve Therenty, et al. Newseye: A digital investigator for historical newspapers. In *15th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2020*, 2020. 1, 2
- [8] Maud Ehrmann, Matteo Romanello, Simon Clematide, Phillip Benjamin Ströbel, and Raphaël Barman. Language resources for historical newspapers: the impresso collection. 2020. 1, 2
- [9] Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. Overview of clef hipe 2020: Named entity recognition and linking on historical newspapers. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings*, page 288–310, Berlin, Heidelberg, 2020. Springer-Verlag. 2
- [10] Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, Antoine Doucet, and Simon Clematide. Overview of hipe-2022: named entity recognition and linking in multilingual historical documents. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 423–446. Springer, 2022. 2
- [11] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, 2024. 5
- [12] Carlos-Emiliano González-Gallardo, Antoine Doucet, et al. Retrieval augmented generation for historical newspapers. In *ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, 2024. 1, 2
- [13] Simon Hengchen, Nina Tahmasebi, Dominik Schlechtweg, and Haim Dubossarsky. Challenges for computational lexical semantic change. 2021. 2
- [14] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. 2022. 1
- [15] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. 2021. 2
- [16] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. 2022. 2
- [17] Benjamin Charles Germain Lee, Jaime Mears, Eileen Jake-way, Meghan Ferriter, Chris Adams, Nathan Yarasavage, Deborah Thomas, Kate Zwaard, and Daniel S. Weld. The newspaper navigator dataset: Extracting headlines and visual content from 16 million historic newspaper pages in chronicling america. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20*, page 3055–3062, New York, NY, USA, 2020. Association for Computing Machinery. 2
- [18] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020. 1, 2
- [19] Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada, July 2017. Association for Computational Linguistics. 1
- [20] Bhawna Piryani, Jamshid Mozafari, and Adam Jatowt. Chroniclingamericaqa: A large-scale question answering dataset based on historical american newspaper pages. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2038–2048, 2024. 2
- [21] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2020. 1
- [22] Marco Rospocher, Marieke Van Erp, Piek Vossen, Antske Fokkens, Itziar Aldabe, German Rigau, Aitor Soroa, Thomas Ploeger, and Tessel Bogaard. Building event-centric knowledge graphs from news. *Journal of Web Semantics*, 37:132–151, 2016. 1

- [23] Dwaipayan Roy, Debjyoti Paul, Mandar Mitra, and Utpal Garain. Using word embeddings for automatic query expansion, 2016. 2
- [24] Stefan Schweter, Luisa März, Katharina Schmid, and Erion Cano. hmbert: Historical multilingual language models for named entity recognition. *arXiv preprint arXiv:2205.15575*, 2022. 1, 2
- [25] Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. Wikineural: Combined neural and knowledge-based silver data creation for multilingual ner. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 2521–2533, 2021. 1
- [26] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models, 2021. 2
- [27] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions, 2023. 2
- [28] Bertie Vidgen and Leon Derczynski. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300, 2020. 2
- [29] Jiexin Wang, Adam Jatowt, and Masatoshi Yoshikawa. Archivalqa: A large-scale benchmark dataset for open domain question answering over historical news collections, 2022. 2
- [30] Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding. *arXiv preprint arXiv:2104.08836*, 2021. 1
- [31] Junyuan Zhang, Qintong Zhang, Bin Wang, Linke Ouyang, Zichen Wen, Ying Li, Ka-Ho Chow, Conghui He, and Wentao Zhang. Ocr hinders rag: Evaluating the cascading impact of ocr on retrieval-augmented generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17443–17453, 2025. 2
- [32] Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131, 2023. 5