RESEARCH ARTICLE

# The Medieval Text and the Modern Question: AI for Multilingual Semantic Retrieval across Centuries

**Antoni Mestre**[1,2,*]

[1] Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, 46022 Valencia, Spain
[2] Valencian Graduate School and Research Network of Artificial Intelligence (ValgrAI), 46022 Valencia, Spain

## Abstract

**Multilingual access to medieval literature requires retrieval systems to bridge historical and cross-lingual semantic variation. This article introduces HistRAG, a language-aware framework combining lexical retrieval, multilingual dense representations, historical adaptation, hybrid fusion, and cross-encoder reranking. HistRAG separates the effects of document modernization, query adaptation, and retrieval architecture within a unified experimental setting. The framework is evaluated on a heterogeneous collection of Medieval Catalan literature using equivalent queries in three contemporary languages. Results show that lexical retrieval is highly sensitive to query language, while multilingual dense representations provide more robust cross-lingual access. Fixed hybrid fusion offers no systematic advantage over dense retrieval, and historical adaptations produce heterogeneous, non-additive effects. Cross-encoder reranking achieves the strongest average effectiveness, although its contribution varies across languages and evaluation subsets. These findings support an architecture- and language-sensitive approach to semantic retrieval from historically distant collections and provide a reproducible basis for extending AI-enabled access to other medieval textual traditions.**

**Keywords**: natural language processing, language-aware machine intelligence, multilingual semantic retrieval, hybrid information retrieval, medieval literature.

## 1 Introduction

The digitization of medieval literature has made historically significant collections increasingly available in machine-readable form. Yet digital availability does not guarantee effective access. Users formulate information needs in contemporary language, whereas medieval texts preserve historical spelling, vocabulary, syntax, and conceptual conventions. Relevant evidence may remain difficult to retrieve even when it is present in the collection.

This problem involves two distinct forms of mismatch. Temporal mismatch separates modern queries from historically situated language, while cross-lingual mismatch arises when users search in a language different from that of the source texts. Both are particularly important for medieval literature, where ideas may be expressed through dialogue, allegory, exemplification, or extended argument rather than through concise statements containing predictable

**Submitted:** 29 July 2026
**Revised:** 29 September 2026
**Accepted:** 07 October 2026
**Published:** 23 October 2026

**Vol.** 2, **No.** 4, 2026.
10.62762/TMI.2026.633155

***Corresponding author:**
✉ Antoni Mestre
antoni.mestre@upv.es

**Citation**
Mestre, A. (2026). The Medieval Text and the Modern Question: AI for Multilingual Semantic Retrieval across Centuries. *ICCK Transactions on Machine Intelligence*, 2(4), 236–256.

© 2026 ICCK (Institute of Central Computation and Knowledge)

keywords. Retrieval systems must recognize related meanings across both languages and historical periods.

Historical language processing has traditionally addressed temporal variation through spelling normalization, lexical mapping, and document modernization [1]. These methods can increase surface correspondence with contemporary queries, but they may also remove historically informative forms or leave deeper semantic differences unresolved. Query-side adaptation offers a less intrusive alternative by introducing historically plausible lexical expressions at retrieval time. Multilingual dense models address a different aspect of the problem by representing queries and passages in a shared semantic space, reducing their dependence on exact lexical overlap. Hybrid retrieval and cross-encoder reranking may provide additional improvements, although their value under simultaneous historical and linguistic variation remains insufficiently understood.

Existing research generally treats historical retrieval and multilingual retrieval as separate problems. Historical studies often focus on monolingual normalization, whereas multilingual benchmarks predominantly contain contemporary documents. It is unclear whether multilingual representations are robust to medieval language, whether document and query adaptation address complementary forms of mismatch, and whether lexical and semantic evidence remain useful when combined across languages.

This article presents HistRAG, a language-aware artificial intelligence framework for multilingual semantic retrieval from medieval literature. HistRAG integrates lexical retrieval, multilingual dense representations, score-level fusion, controlled historical adaptation, and cross-encoder reranking within a common evaluation pipeline. The framework is designed to compare these components rather than to assume that their combination is universally beneficial.

The empirical study uses a collection of 6,443 passages and approximately 1.69 million words from nine Medieval Catalan works. The corpus includes philosophical, theological, didactic, chivalric, and narrative material, with works by Ramon Llull alongside non-Lullian texts. Equivalent information needs are expressed in Modern Catalan, Spanish, and English, allowing query-language effects to be examined while preserving the underlying semantic intent.

The experiments compare BM25, multilingual dense retrieval, lexical-dense hybrid fusion, and cross-encoder reranking. They also distinguish between document-side adaptation, which transforms selected medieval forms toward contemporary Catalan, and query-side adaptation, which introduces historically plausible lexical alternatives into modern queries. The study focuses on passage retrieval as the evidence-selection stage of intelligent access systems; generated answers and their faithfulness are outside the scope of the present evaluation.

Query language and retrieval architecture turn out to interact strongly. BM25 works reasonably well for Modern Catalan queries and almost fails for Spanish and English ones, whereas multilingual dense retrieval closes most of that gap. A fixed lexical-dense fusion adds little to dense retrieval alone. Historical adaptation has smaller and less regular effects: query-side adaptation gives the largest gain among non-reranked systems but weakens on held-out intents, document-side adaptation mainly helps same-language BM25, and the two do not add up. Cross-encoder reranking gives the best average scores, with gains that differ by language and subset.

The main contribution is a controlled analysis of medieval retrieval as a joint temporal and cross-lingual alignment problem, in which lexical, dense, hybrid, adaptation, and reranking strategies are compared on the same multilingual information needs. Negative results are reported alongside positive ones, notably the lack of a consistent benefit from fixed fusion and the instability of combined adaptation.

Medieval Catalan serves as a demanding case study rather than as evidence of universal behavior across all medieval traditions. The historical transformations are computational proxies rather than expert critical editions, and the relevance judgments identify the passages used to construct the queries rather than every potentially relevant passage in the collection. The study aims to characterize architecture-dependent retrieval behavior and provide a reproducible basis for broader evaluation.

The remainder of the article is organized as follows. Section 2 reviews historical language processing, multilingual retrieval, and intelligent access to cultural-heritage collections. Section 3 presents the HistRAG framework and evaluation testbed. Section 4 describes the experimental methodology. Sections 5 and 6 report the results, ablations, and error analysis. Section 7 discusses their implications, and Section 8

presents the limitations and conclusions.

## 2 State of the Art

### 2.1 Historical Language Variation and Diachronic Representation

Computational research on historical language has largely developed around the representation and measurement of linguistic change. Early work on diachronic word embeddings demonstrated that distributional representations trained on temporally segmented corpora can reveal systematic patterns of semantic change [2]. Subsequent surveys established diachronic embeddings as a major methodological family while also identifying persistent problems involving corpus comparability, representation alignment, frequency effects, and the distinction between genuine semantic change and variation caused by data composition [3, 4]. More recent reviews have extended this discussion to contextual language models and large language models, which allow word meaning to be represented as a function of individual usage contexts rather than as a single static vector [5].

Contextual embeddings have improved the analysis of historical meaning by enabling semantic change to be estimated from distributions of contextualized usages [6]. Other approaches have sought to reduce the dependence on explicit post-hoc alignment between independently trained temporal spaces [7]. These developments are relevant to historical retrieval because they show that temporal variation cannot be reduced to spelling differences alone. A modern and a historical passage may employ recognizable word forms while assigning them different senses, or they may express closely related concepts through lexical items whose contemporary meanings have shifted.

The development of diachronic datasets has also supported more systematic evaluation. The DWUG resource represents word usages as graded graphs across time periods and languages, providing an explicitly usage-based framework for studying lexical semantic change [8]. DUKweb similarly offers diachronic word representations derived from the United Kingdom Web Archive, illustrating the value of large temporally organized corpora for studying changes in language use [9]. These resources have primarily been designed for lexical semantic analysis rather than document retrieval, but they establish two important principles for historical information access: temporal variation should be evaluated explicitly, and the same linguistic form cannot be assumed to preserve a stable representation across periods.

Historical corpora for Iberian languages remain less abundant than resources for English and other high-resource languages. The Open Diachronic Corpus of Historical Spanish provides an important example of a structured historical language resource for studying linguistic variation over time [10]. At the document-image level, CATMuS Medieval addresses the preceding stage of historical text processing by assembling a multilingual and cross-century dataset for handwritten text recognition (HTR) in the Latin script [11]. Such resources highlight the fact that computational access to medieval collections depends both on reliable textual transcription and on models capable of handling historical variation after transcription. Nevertheless, neither diachronic semantic-change datasets nor handwritten-text-recognition corpora directly address the retrieval of relevant passages from medieval literature using contemporary multilingual queries.

### 2.2 Multilingual Semantic Retrieval

Multilingual retrieval has increasingly moved from translation-dependent pipelines toward shared representation models. Cross-lingual encoders such as XLM-R showed that large-scale multilingual pretraining can support transfer across languages without requiring task-specific parallel data for every language pair [12]. Knowledge distillation has likewise been used to transfer the geometry of strong monolingual sentence encoders into multilingual models, placing semantically equivalent sentences from different languages in a shared embedding space [13]. Contrastive approaches such as SimCSE further demonstrated that relatively simple objectives can improve sentence representations by explicitly organizing semantically related and unrelated examples in the embedding space [14].

In information retrieval, dense retrievers encode queries and documents independently and rank documents through vector similarity. Cross-lingual training allows these systems to retrieve documents in a language different from that of the query, reducing the need for exact lexical correspondence or explicit machine translation [15]. Knowledge-enhanced pretraining has been proposed as a means of improving this alignment by incorporating multilingual semantic relations into the training process [16]. Multilingual E5 extends this line of work through weakly supervised contrastive

pretraining over multilingual text pairs, producing general-purpose embeddings applicable to retrieval and semantic similarity tasks [17].

The development of broad evaluation suites has revealed both the capabilities and limitations of dense retrieval. BEIR demonstrated that retrieval models that perform strongly within their training domains may generalize unevenly across heterogeneous tasks and corpora [18]. MTEB expanded this perspective by evaluating text embeddings across multiple task families, languages, and datasets, showing that no single representation model dominates uniformly across all settings [19]. MIRACL provides a specifically multilingual retrieval benchmark covering 18 languages and has become an important resource for evaluating cross-lingual and language-specific retrieval effectiveness [20].

These benchmarks have substantially advanced multilingual retrieval evaluation, but they predominantly represent contemporary language. Their documents and queries generally do not differ by several centuries of orthographic, lexical, and conceptual development. Consequently, strong multilingual alignment on MIRACL, BEIR, or MTEB does not guarantee that the same model will align contemporary queries with medieval passages. Historical retrieval introduces a form of domain shift that differs from conventional cross-domain transfer because it affects the linguistic system itself rather than only document genre or subject matter.

This distinction is particularly relevant for low-resource historical varieties. A multilingual encoder may align Modern Catalan, Spanish, and English effectively while representing Medieval Catalan less reliably because historical forms are rare or absent from its pretraining data. Dense retrieval can reduce cross-lingual mismatch without fully resolving temporal mismatch. Conversely, a method designed to normalize historical spelling may improve retrieval within one language but provide limited assistance when the query is expressed in another. Evaluating both dimensions in the same controlled setting is necessary to determine which capabilities arise from multilingual representation learning and which require explicit historical adaptation.

### 2.3 Lexical Retrieval, Hybrid Fusion, and Reranking

Lexical retrieval remains a strong and interpretable baseline for information access. BM25 ranks documents using term frequency, inverse document frequency, and document-length normalization, and it continues to perform competitively across many retrieval domains [18, 21]. Its underlying principles and relationship to inverted-index retrieval are established components of modern information-retrieval systems [22]. In historical collections, lexical retrieval can be particularly effective for proper names, distinctive terminology, and rare expressions that retain similar forms across periods.

Its dependence on token overlap, however, creates an evident limitation when contemporary queries and medieval passages use different spellings or lexical choices. The problem is more severe in cross-lingual retrieval, where related queries and passages may share few surface forms. Dense retrieval is designed to address this limitation, but lexical and dense methods often capture different evidence. Hybrid retrieval combines their scores or rankings in an attempt to preserve exact matches while also recognizing broader semantic correspondence.

The effectiveness of hybrid retrieval depends on actual complementarity between the components. If the lexical model contributes useful names and historical terms, fusion may improve the candidate set. If lexical scores are nearly uninformative for Spanish or English queries over Medieval Catalan documents, however, fusion may add noise or merely reproduce the dense ranking. Score distributions may also differ across query languages and document transformations, making a single fusion weight unsuitable for all conditions. Hybrid retrieval should be treated as an empirical design choice rather than an inherently superior architecture.

Cross-encoder rerankers provide a later and more computationally intensive stage of retrieval. Unlike bi-encoders, which represent queries and passages separately, cross-encoders jointly process each query-passage pair and can model detailed token-level interactions [23]. Their principal limitation is computational cost, since the model must be evaluated independently for every candidate pair. Research on matrix-factorization approximations has explored how cross-encoder behavior can be used more efficiently for nearest-neighbor search [24]. More recent work on minimal-interaction cross-encoders similarly seeks to retain part of the interaction capacity of rerankers while reducing their computational requirements [25].

For medieval retrieval, reranking is potentially valuable because it can reassess passages retrieved through incomplete lexical or dense signals. Even so, a reranker cannot recover relevant passages that are absent from the initial candidate set, and a multilingual cross-encoder trained on contemporary data may itself exhibit language- or period-dependent behavior. Average improvements may conceal heterogeneous effects across query languages. This motivates evaluation not only of global ranking metrics but also of whether reranking redistributes performance between languages or source works.

### 2.4 Intelligent Access to Cultural-Heritage Collections

Digital libraries have long sought to broaden access to cultural heritage by aggregating digitized objects, metadata, and historical texts. Europeana established an influential model for cross-institutional cultural-heritage aggregation in Europe [26], while the World Digital Library pursued global access to culturally significant materials [27]. National and regional initiatives such as TELDAP have combined digitization, e-learning, and digital-archive infrastructures [28]. Systems including Manuscriptorium and Kramerius illustrate the continuing development of platforms designed to make written heritage, including medieval materials, accessible through digital-library interfaces [29].

These infrastructures solve essential problems of preservation, cataloging, and dissemination, but keyword search and metadata browsing provide only limited support for users who do not know the vocabulary, language, or organization of the original materials. Recent research has explored knowledge graphs, semantic retrieval, and retrieval-augmented generation (RAG) as interfaces to cultural-heritage collections. The Nanjing Yunjin question-answering system combines knowledge-graph representations with RAG to provide access to domain-specific cultural information [30]. Topic-RAG applies topic-based retrieval to historical newspapers, restricting retrieval to thematically relevant portions of a large archive before generating responses [31]. Recent work has also investigated real-time RAG-based access to historical archives, emphasizing the engineering requirements of interactive retrieval over large historical collections [32].

These studies demonstrate increasing interest in conversational and generative access to cultural heritage. However, end-to-end RAG performance depends heavily on the quality of the retrieval stage. A fluent generated answer may conceal weak evidence selection, and answer-level evaluation alone does not reveal whether a failure arose from retrieval, generation, or the underlying collection. This is particularly important for medieval literature, where interpretative ambiguity and historically specific language increase the risk of producing plausible but insufficiently grounded answers. Retrieval should be evaluated independently before it is embedded in a generative system.

Existing cultural-heritage RAG studies also tend to address collections in which the main challenge is scale, thematic organization, or specialized domain knowledge. They provide less evidence about retrieval when query and document differ simultaneously in language and historical period. Furthermore, the use of modern semantic models over historical materials is often treated as an implementation decision rather than as an object of controlled comparison. This leaves open the question of how lexical retrieval, multilingual embeddings, historical adaptation, hybrid fusion, and reranking behave within the same medieval collection.

### 2.5 Research Gap

The literature reviewed above provides strong foundations for individual components of historical semantic retrieval, but these components have largely been studied in isolation. Diachronic natural language processing (NLP) has developed methods and resources for detecting semantic change, aligning historical representations, and processing noisy historical documents. Multilingual retrieval has produced shared representation models and extensive contemporary benchmarks. Hybrid retrieval and cross-encoder reranking have improved ranking effectiveness across many modern domains. Digital-humanities and cultural-heritage research has begun to incorporate semantic search and RAG into historical archives.

What remains insufficiently examined is the interaction between temporal and cross-lingual mismatch at the passage-retrieval level. Existing diachronic resources do not generally evaluate contemporary multilingual queries against medieval literary passages, while multilingual benchmarks do not normally expose retrievers to substantial historical language variation. Cultural-heritage RAG systems, in turn, rarely isolate the contribution of document normalization, query adaptation, dense representation, fusion, and reranking under controlled linguistic conditions.

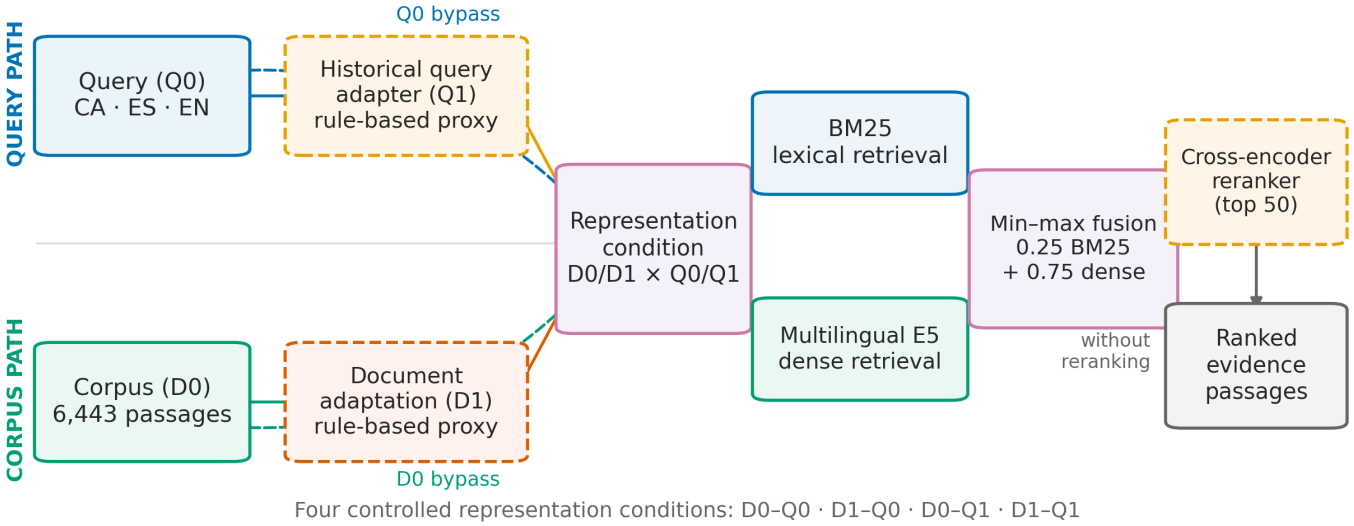


**Figure 1.** Overview of the HistRAG retrieval pipeline. Contemporary queries in Modern Catalan, Spanish, or English follow either the original $Q0$ path or the controlled historical-query adaptation $Q1$ path. Medieval passages are retrieved from either the original corpus $D0$ or its adapted representation $D1$. Each representation condition is evaluated using lexical and multilingual dense retrieval, score-level fusion, and optional cross-encoder reranking.

The present study addresses this gap by evaluating these components within a common language-aware framework. Rather than assuming that historical normalization and multilingual representation learning solve the same problem, it treats document modernization, historical query adaptation, lexical retrieval, dense retrieval, hybrid fusion, and reranking as separable interventions. Equivalent information needs in Modern Catalan, Spanish, and English allow the effect of query language to be examined while keeping semantic intent constant. Medieval Catalan is a useful test case: it is historically distant from its modern descendant, attracts readers who search in other languages, has few computational resources, and spans several literary forms.

## 3 HistRAG Framework and Evaluation Testbed

### 3.1 Framework Overview

HistRAG is a language-aware retrieval framework designed to study semantic access to medieval literature under simultaneous temporal and cross-lingual mismatch. The framework receives a contemporary-language query and retrieves ranked evidence passages from a historical collection. It combines lexical retrieval, multilingual dense representations, score-level fusion, controlled historical adaptation, and cross-encoder reranking within a common experimental pipeline.

The framework is evaluated through a case study based on Medieval Catalan literature. This setting is particularly demanding because the queries are formulated in contemporary languages, whereas the documents preserve historical spelling, vocabulary, syntax, and conceptual expression. The same underlying information need is expressed in Modern Catalan, Spanish, and English, allowing the effect of query language to be examined while keeping semantic intent constant.

HistRAG distinguishes between the original document and query representations and their historically adapted counterparts. Let $D0$ denote the original medieval corpus and $Q0$ the contemporary query. A document-side transformation produces $D1$, while a query-side historical adapter produces $Q1$. These alternatives define four controlled representation conditions:

$$D0\text{-}Q0, \qquad D1\text{-}Q0, \qquad D0\text{-}Q1, \qquad D1\text{-}Q1.$$

The first condition preserves both the original documents and the contemporary queries. The second modifies the document representation while leaving the query unchanged. The third adapts the query while retaining the original corpus, and the fourth applies both interventions simultaneously. This factorial design permits document modernization and historical query adaptation to be examined separately

rather than being treated as a single preprocessing operation. Figure 1 summarizes the complete pipeline.

### 3.2 Medieval Literary Corpus

The evaluation corpus contains 6,443 passages extracted from nine Medieval Catalan works, representing approximately 1.69 million words. The collection includes philosophical, theological, didactic, chivalric, and narrative material. A substantial proportion of the corpus consists of works by Ramon Llull, while the inclusion of non-Lullian texts provides greater variation in authorship, genre, and literary style.

The corpus was segmented into passage-level retrieval units intended to contain sufficient local context while remaining suitable for ranking. Passage identifiers preserve the source work and passage position, enabling results to be analyzed by text and author group. Exact duplicate passage identifiers and exact textual duplicates were removed during corpus preparation. The resulting collection contains no passages shorter than the predefined minimum length, while two indivisible sentences slightly exceed the nominal upper segmentation threshold.

The collection is heterogeneous in textual quality. Most works are based on relatively clean digital editions, whereas part of the corpus contains noisier digitized text. This variation reflects realistic cultural-heritage conditions but also introduces a possible source of retrieval error. The present study treats the corpus as an operational historical-text collection rather than as a uniform critical edition.

The original corpus representation, $D0$, preserves the available medieval text. The adapted representation, $D1$, applies deterministic lexical substitutions intended to reduce selected orthographic and lexical differences between medieval and contemporary Catalan. This transformation is used as a controlled document-side intervention. It should not be interpreted as a philologically validated modernization or as a replacement for an expert edition.

### 3.3 Multilingual Query Set

The evaluation is organized around semantic intents linked to evidence passages in the corpus. Each intent is expressed in Modern Catalan, Spanish, and English. The three versions correspond to the same information need and are grouped during statistical analysis rather than counted as independent observations.

The available pilot set contains 20 semantic intents, of which 17 are answerable from the corpus and three are designed as unanswerable controls. This produces 60 query formulations in total and 51 answerable query-language combinations. The questions were constructed from identified evidence passages so that the intended answer could be traced to a specific source location. This evidence-first procedure reduces the risk of generating questions that are unrelated to the collection.

The contemporary query representation is denoted by $Q0$. The historical query adapter produces $Q1$ by introducing rule-based historical lexical alternatives intended to increase correspondence with the medieval corpus. The adapter operates at retrieval time and does not modify the underlying semantic intent. As with $D1$, $Q1$ is a controlled computational proxy rather than an expert reconstruction of historical language. The split into development and test intents is described in Section 4.1.

### 3.4 Lexical and Multilingual Dense Retrieval

The lexical component uses Best Matching 25 (BM25) retrieval. BM25 assigns scores according to term frequency, inverse document frequency, and passage-length normalization. It provides a strong and interpretable baseline and is expected to perform best when queries and passages share distinctive vocabulary or named entities. Its dependence on surface overlap, however, makes it vulnerable to both historical lexical variation and cross-lingual queries.

The dense component uses multilingual E5 representations. Queries and passages are encoded independently into a shared vector space, and retrieval is performed using cosine similarity. The multilingual encoder is intended to capture semantic correspondence when query and document vocabulary differ, including cases in which they are written in different languages. Since the model is primarily trained on contemporary data, its performance on medieval language remains an empirical question rather than an assumed capability.

Lexical and dense retrieval are executed independently for each of the four representation conditions. This produces comparable rankings for original and adapted documents and queries. The design makes it possible to determine whether a transformation benefits primarily lexical matching, dense semantic alignment, or both.

### 3.5 Score Normalization and Hybrid Fusion

The hybrid configuration combines the BM25 and dense rankings at score level. Because lexical and cosine-similarity scores are not directly comparable, both are normalized independently using min-max normalization within each query:

$$\widehat{s}_m(q,d) = \frac{s_m(q,d) - \min_{d'} s_m(q,d')}{\max_{d'} s_m(q,d') - \min_{d'} s_m(q,d')}, \quad (1)$$

where $s_m(q,d)$ is the score assigned by retrieval model $m$ to passage $d$ for query $q$.

The normalized scores are combined as

$$s_{\text{hybrid}}(q,d) = 0.25\,\widehat{s}_{\text{BM25}}(q,d) + 0.75\,\widehat{s}_{\text{dense}}(q,d). \quad (2)$$

The fusion weights were selected using the development partition and were subsequently fixed for evaluation. The greater weight assigned to the dense component reflects its stronger development performance under multilingual querying. The selected weights should not be interpreted as universally optimal beyond the present corpus and model configuration.

Hybrid retrieval is evaluated under $D0$-$Q0$, $D1$-$Q0$, $D0$-$Q1$, and $D1$-$Q1$. Comparing these conditions permits the effect of each historical adaptation strategy to be measured relative to the same lexical-dense architecture.

### 3.6 Cross-Encoder Reranking

The final retrieval stage applies a multilingual cross-encoder to the highest-ranked candidates. Unlike the dense retriever, which encodes queries and passages independently, the cross-encoder jointly processes each query-passage pair and estimates its relevance through token-level interaction.

For computational efficiency, only the top 50 first-stage candidates are reranked; the choice of first-stage configuration is given in Section 4.2.

The cross-encoder cannot retrieve passages that are absent from the initial candidate set. Its performance depends on both its pairwise relevance estimates and the recall of the first-stage retriever. This distinction is important when interpreting improvements in ranking metrics.

### 3.7 Relevance Judgments and Evaluation Scope

The available relevance judgments (qrels) link each answerable semantic intent to its identified evidence passage. These judgments provide a verified origin passage for evaluating whether the retrieval system recovers the evidence used to construct the query. They do not constitute exhaustive pooled relevance assessments across the entire corpus.

A passage retrieved from a different work may contain valid or partially relevant evidence without being represented in the current qrels. Retrieval failures are defined relative to the annotated origin evidence rather than as definitive judgments of semantic irrelevance. Work-level error categories are used as diagnostic indicators and are interpreted cautiously.

The present testbed evaluates passage retrieval, not generated answers. Measures of answer correctness, faithfulness, citation support, or historical interpretation are outside the scope of the current experiment. Retrieval effectiveness is assessed with Recall at ranks 5 and 10, mean reciprocal rank, and normalized discounted cumulative gain at rank 10 (Section 4.3). Statistical inference groups the three query languages by semantic intent to avoid treating translations of the same question as independent observations.

The testbed is intended to support a controlled comparison of retrieval architectures and representation conditions. Its current scale and relevance judgments permit the identification of language- and architecture-dependent patterns, but they do not support claims of exhaustive benchmark coverage or universal generalization across medieval literary traditions. Medieval Catalan is used as a demanding empirical case study within the broader problem of multilingual retrieval from historically distant literature.

## 4 Experimental Methodology

### 4.1 Experimental Design

The experiments compare nine retrieval configurations derived from the components described in Section 3. The design varies the retrieval architecture and the representation of the documents and queries. The original medieval corpus is denoted by $D0$, while $D1$ denotes its rule-based adapted representation. Contemporary queries are denoted by $Q0$, and their historically adapted counterparts by $Q1$.

The complete set of configurations is shown in

Table 1. BM25 and multilingual dense retrieval are evaluated on the original documents and on the document-side adaptation. Hybrid retrieval is additionally evaluated with query adaptation and with the simultaneous application of document- and query-side transformations. The final configuration reranks the candidates produced by the strongest development-selected non-reranked system.

**Table 1.** Retrieval configurations evaluated in the study. $D0$ and $Q0$ denote the original medieval documents and contemporary queries; $D1$ and $Q1$ denote their controlled rule-based adaptations.

| ID | Retrieval method | Documents | Queries |
|---|---|---|---|
| S1 | BM25 | $D0$ | $Q0$ |
| S2 | BM25 | $D1$ | $Q0$ |
| S3 | Multilingual dense | $D0$ | $Q0$ |
| S4 | Multilingual dense | $D1$ | $Q0$ |
| S5 | Hybrid | $D0$ | $Q0$ |
| S6 | Hybrid | $D1$ | $Q0$ |
| S7 | Hybrid | $D0$ | $Q1$ |
| S8 | Hybrid | $D1$ | $Q1$ |
| S9 | Cross-encoder reranking | $D0$ | $Q1$ |

The original partition contains five development intents and fifteen test intents. Twelve test intents are answerable and constitute the held-out retrieval evaluation. Development intents are used only to select the hybrid fusion weights and the first-stage configuration supplied to the reranker. Once selected, these settings are frozen before evaluation on the held-out test partition.

Because the pilot collection contains a limited number of information needs, we additionally report a pooled exploratory analysis over all 17 answerable intents. This secondary analysis increases descriptive resolution but includes the development partition and is not interpreted as an independent estimate of generalization. Results from the held-out test and pooled analyses are kept separate throughout the paper.

The three unanswerable intents are not included in the ranking metrics because they have no annotated relevant passage. They are retained in the released query set for future evaluation of rejection and abstention behavior, which is outside the scope of the present retrieval experiment.

### 4.2 Retrieval Models and Hyperparameters

The lexical baseline uses BM25 over passage-level retrieval units. No parameters are selected using the held-out test partition. The model ranks passages according to lexical overlap, term informativeness, and document-length normalization. BM25 is evaluated independently for $D0$ and $D1$, allowing the effect of document-side adaptation on lexical matching to be isolated.

Dense retrieval uses the publicly available `intfloat/multilingual-e5-small` checkpoint. The model produces 384-dimensional representations for queries and passages, which are ranked using cosine similarity. Input sequences are truncated to a maximum length of 256 tokens. The encoder is used without additional fine-tuning on the medieval corpus, so the experiment evaluates the transfer of a contemporary multilingual representation model to historically distant text.

Hybrid retrieval uses the per-query min-max normalization and the 0.25/0.75 weighting of Eqs. (1) and (2), fixed on the development partition before test evaluation.

The reranking stage uses `cross-encoder/mmarco-mMiniLMv2-L12-H384-v1`, a multilingual MiniLM cross-encoder trained on mMARCO [33]. The model jointly encodes each query-passage pair and reranks the top 50 candidates returned by the first-stage hybrid retriever. The selected candidate generator corresponds to the $D0$-$Q1$ condition, which obtained the strongest development performance among the non-reranked configurations. The reranker is likewise used without task-specific fine-tuning.

Neither the dense retriever nor the cross-encoder is trained on the test queries. The experiment examines zero-shot transfer to Medieval Catalan rather than adaptation through supervised learning. The only development-dependent choices are the hybrid fusion weights and the selection of the candidate-generating configuration for reranking.

### 4.3 Evaluation Metrics

Normalized discounted cumulative gain at rank 10 (nDCG@10) is used as the primary evaluation metric. It rewards systems that place the annotated evidence passage near the beginning of the ranking and assigns no gain when that passage is absent from the top ten results. For a ranked list of passages, discounted cumulative gain is computed as

$$\mathrm{DCG@}k = \sum_{i=1}^{k} \frac{2^{\mathrm{rel}_i} - 1}{\log_2(i+1)}, \tag{3}$$

where $\mathrm{rel}_i \in \{0, 1\}$ is the relevance of the passage at rank $i$ (binary in the present qrels). DCG@$k$ is normalized by the ideal value IDCG@$k$, obtained by ranking all relevant passages first:

$$\mathrm{nDCG@}k = \frac{\mathrm{DCG@}k}{\mathrm{IDCG@}k}. \tag{4}$$

The current qrels contain one origin-evidence passage for each answerable query variant. Consequently, nDCG@10 primarily captures the rank assigned to that passage rather than graded coverage of multiple relevant passages.

Recall at ranks 5 and 10 is reported to measure whether the annotated evidence occurs within the corresponding retrieval depth:

$$\mathrm{Recall@}k = \frac{|R_q \cap L_q^k|}{|R_q|}, \tag{5}$$

where $R_q$ is the set of annotated relevant passages for query $q$ and $L_q^k$ is the set of the top-$k$ retrieved passages. Mean reciprocal rank (MRR) is used as a complementary rank-sensitive measure:

$$\mathrm{MRR} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{1}{\mathrm{rank}_q}, \tag{6}$$

where $\mathcal{Q}$ is the set of evaluated queries and $\mathrm{rank}_q$ is the position of the first annotated relevant passage. Queries for which the passage is not retrieved receive a reciprocal rank of zero.

Metrics are calculated first at the query-language level. For overall comparisons, the Catalan, Spanish, and English formulations of the same information need are grouped under their shared semantic intent. The three languages are averaged within each intent before macro-averaging across intents. This procedure prevents translated versions of the same question from being treated as statistically independent observations.

Language-specific results are also reported separately. These analyses use the same set of semantic intents for Catalan, Spanish, and English and allow differences between languages to be interpreted under controlled semantic content.

### 4.4 Statistical Analysis

Uncertainty is estimated using paired non-parametric bootstrap resampling. The semantic intent, rather than the individual translated query, is used as the resampling unit. For overall analyses, all three language variants belonging to an intent are retained together during each resample. This preserves the dependence structure between translations and ensures that the effective sample size corresponds to the number of distinct information needs.

A total of 10,000 bootstrap samples are drawn with replacement using a fixed random seed of 20260729. Ninety-five percent confidence intervals are obtained from the 2.5th and 97.5th percentiles of the bootstrap distribution. For language-specific analyses, intents are resampled independently within each language while preserving the same bootstrap procedure.

Pairwise comparisons are based on within-intent differences in nDCG@10. Seven contrasts are defined before inspecting the final aggregate results:

1. BM25 $D1$-$Q0$ versus BM25 $D0$-$Q0$;
2. dense $D1$-$Q0$ versus dense $D0$-$Q0$;
3. hybrid $D1$-$Q0$ versus hybrid $D0$-$Q0$;
4. hybrid $D0$-$Q1$ versus hybrid $D0$-$Q0$;
5. hybrid $D0$-$Q1$ versus hybrid $D1$-$Q0$;
6. hybrid $D1$-$Q1$ versus hybrid $D0$-$Q1$;
7. cross-encoder reranking versus hybrid $D0$-$Q1$.

Two-sided bootstrap $p$-values are corrected jointly using the Holm procedure [34] to control the family-wise error rate at $\alpha = 0.05$. Both unadjusted and Holm-adjusted values are reported. Effect magnitude is additionally summarized using paired Cohen's $d_z$:

$$d_z = \frac{\overline{\Delta}}{s_\Delta}, \tag{7}$$

where $\overline{\Delta}$ is the mean within-intent difference and $s_\Delta$ is its standard deviation.

Statistical significance is not used as the sole criterion for interpreting the results. Confidence intervals, effect direction, magnitude, consistency across languages, and sensitivity to the development/test partition are considered jointly. In particular, results based on the pooled 17-intent analysis are treated as exploratory

**Table 2.** Retrieval effectiveness on the held-out test partition. Results are macro-averaged over the 12 answerable semantic intents, with the Modern Catalan, Spanish, and English formulations grouped within each intent. nDCG@10 is the primary metric.

| ID | Method | Condition | R@5 | R@10 | MRR | nDCG@10 | 95% CI |
|---|---|---|---|---|---|---|---|
| S1 | BM25 | $D0$-$Q0$ | 0.222 | 0.250 | 0.152 | 0.175 | [0.094, 0.266] |
| S2 | BM25 | $D1$-$Q0$ | 0.250 | 0.250 | 0.205 | 0.214 | [0.127, 0.303] |
| S3 | Multilingual dense | $D0$-$Q0$ | 0.278 | 0.306 | 0.274 | 0.277 | [0.081, 0.485] |
| S4 | Multilingual dense | $D1$-$Q0$ | 0.278 | 0.306 | 0.275 | 0.276 | [0.068, 0.498] |
| S5 | Hybrid | $D0$-$Q0$ | 0.278 | 0.306 | 0.272 | 0.271 | [0.064, 0.500] |
| S6 | Hybrid | $D1$-$Q0$ | 0.278 | 0.306 | 0.275 | 0.270 | [0.061, 0.485] |
| S7 | Hybrid | $D0$-$Q1$ | 0.278 | 0.333 | 0.278 | 0.280 | [0.074, 0.518] |
| S8 | Hybrid | $D1$-$Q1$ | 0.333 | 0.361 | 0.292 | <u>0.299</u> | [0.091, 0.543] |
| S9 | Cross-encoder | $D0$-$Q1$ | **0.389** | **0.444** | **0.304** | **0.328** | [0.148, 0.515] |

*Note.* D0 and Q0 denote the original medieval documents and contemporary queries; D1 and Q1 denote controlled document and query adaptations. Bold indicates the best result, and underline indicates the second-best nDCG@10.

even when their unadjusted confidence intervals exclude zero.

### 4.5 Language-Architecture Analysis

Because the principal empirical question concerns the interaction between query language and retrieval architecture, system performance is examined separately for Modern Catalan, Spanish, and English. The primary descriptive comparison contrasts the language profile of BM25 with that of multilingual dense retrieval.

For each architecture, the cross-lingual gap is defined as the difference between performance on Catalan and the mean performance on Spanish and English:

$$G_m = \mathrm{nDCG@10}_{m,\mathrm{CA}} - \frac{\mathrm{nDCG@10}_{m,\mathrm{ES}}}{2} - \frac{\mathrm{nDCG@10}_{m,\mathrm{EN}}}{2}, \tag{8}$$

where $m$ denotes the retrieval architecture. A reduction in $G_m$ when moving from BM25 to dense retrieval indicates that the dense architecture is less dependent on direct linguistic proximity to the Medieval Catalan documents.

The corresponding difference-in-differences is defined as

$$\Delta G = G_{\mathrm{dense}} - G_{\mathrm{BM25}}. \tag{9}$$

A negative value indicates that multilingual dense retrieval reduces the disparity between same-language and cross-lingual queries relative to lexical retrieval. This interaction is evaluated using paired bootstrap resampling at the semantic-intent level, following the same procedure described above. It directly tests the central architecture-dependent language pattern rather than relying only on separate system averages.

### 4.6 Reproducibility

The experimental package preserves the processed corpus inventory, passage identifiers, multilingual queries, relevance judgments, development/test partition, adapted representations, model configurations, per-query rankings, aggregate metrics, bootstrap outputs, tables, and figures. A hash-based manifest records the files included in the frozen pilot and the subsequent analysis.

All retrieval configurations are executed through scripted pipelines rather than manual ranking or post-hoc editing. The experimental configuration records the model checkpoints, fusion weights, candidate depth, embedding dimensionality, maximum input length, bootstrap seed, and number of resamples. Automated tests verify corpus and identifier consistency, the expected number of systems and query variants, metric ranges, and the generation of the principal result files.

Dense representations were computed on a central processing unit using 384-dimensional embeddings and a maximum sequence length of 256 tokens. No commercial application programming interface calls were required, and no generated results were substituted for missing model executions.

## 5 Results

### 5.1 Overall Retrieval Effectiveness

The principal evaluation was conducted on the 12 answerable semantic intents in the held-out test partition. Table 2 reports the results of the nine retrieval configurations. The cross-encoder configuration, S9, achieved the highest effectiveness, with an nDCG@10 of 0.328, a mean reciprocal rank of 0.304, and Recall@10 of 0.444. The strongest non-reranked configuration was the hybrid system combining document and query adaptation, S8, which obtained an nDCG@10 of 0.299. Historical query adaptation without document adaptation, S7, reached 0.280.

The multilingual dense baseline provided a substantial improvement over lexical retrieval. Dense retrieval over the original corpus, S3, achieved an nDCG@10 of 0.277, compared with 0.175 for BM25 under the same $D0$-$Q0$ representation condition. Document adaptation increased BM25 to 0.214, but produced almost no change for the dense retriever, whose nDCG@10 shifted from 0.277 to 0.276. The corresponding hybrid configurations, S5 and S6, reached 0.271 and 0.270, respectively.

These results indicate that the main improvement over the lexical baseline comes from multilingual semantic representation rather than from score fusion or document adaptation. Historical query adaptation and reranking provide additional gains, but their magnitude is comparatively modest on the held-out test set. The progression from BM25 $D0$-$Q0$ to the cross-encoder raises nDCG@10 by 0.153, although this difference combines changes in representation, first-stage retrieval, query adaptation, and reranking and should not be interpreted as the isolated effect of a single component.

The pooled exploratory analysis over all 17 answerable intents produces the same broad ordering but higher estimates. In this analysis, the cross-encoder reaches an nDCG@10 of 0.470, followed by Hybrid $D0$-$Q1$ at 0.393 and Hybrid $D1$-$Q1$ at 0.362. Dense and Hybrid $D0$-$Q0$ obtain nearly identical scores of 0.337 and 0.336, respectively. Because the pooled analysis includes the development intents used to select the fusion weights and first-stage reranking configuration, these values are treated as descriptive rather than as independent estimates of generalization.

### 5.2 Robustness across Query Languages

The most pronounced result is the interaction between query language and retrieval architecture. Figure 2 presents nDCG@10 separately for Modern Catalan, Spanish, and English in the pooled exploratory analysis. The same 17 semantic intents are used in each language, so the comparison holds the underlying information need constant.

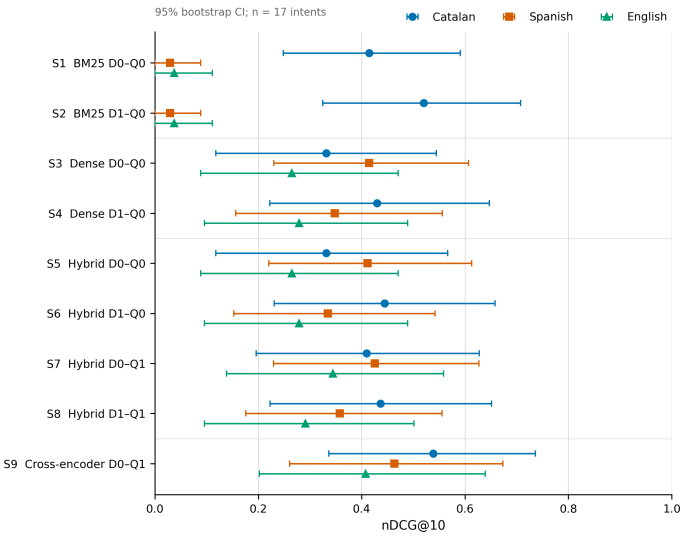


**Figure 2.** Language-specific nDCG@10 for the nine retrieval configurations in the pooled exploratory analysis. Points show mean effectiveness and horizontal bars show 95% bootstrap confidence intervals over 17 semantic intents. The three language formulations of each intent contain equivalent information needs.

BM25 is strongly dependent on linguistic proximity to the source collection. Under the original $D0$-$Q0$ condition, it reaches an nDCG@10 of 0.415 for Catalan, but only 0.029 for Spanish and 0.037 for English. The descriptive gap between Catalan and the mean of the two cross-lingual conditions is approximately 0.382. Document adaptation raises Catalan performance to 0.520, while the Spanish and English results remain unchanged. This pattern is consistent with the document-side intervention increasing lexical correspondence between Medieval and Modern Catalan without addressing the absence of shared vocabulary in cross-lingual queries.

Multilingual dense retrieval substantially changes this language profile. Dense $D0$-$Q0$ obtains an nDCG@10 of 0.331 in Catalan, 0.414 in Spanish, and 0.265 in English. The large same-language advantage observed for BM25 is reduced to approximately zero at the aggregate level. The Spanish point estimate is higher than the Catalan estimate, but their confidence intervals overlap widely; the result

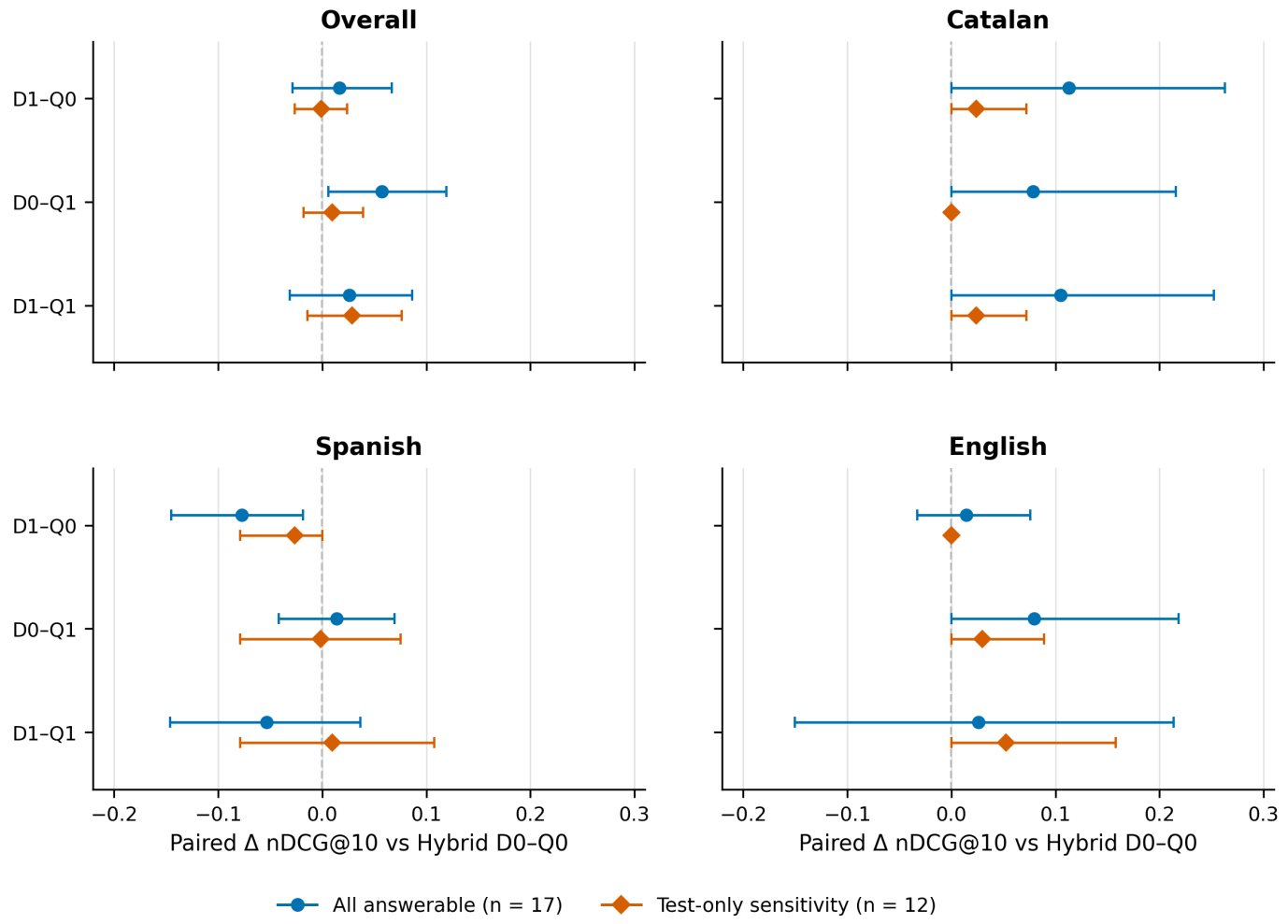


**Figure 3.** Paired changes in nDCG@10 relative to Hybrid $D0$-$Q0$. Blue circles show the pooled exploratory analysis over 17 answerable intents, and orange diamonds show the held-out test analysis over 12 answerable intents. Error bars represent 95% bootstrap confidence intervals.

should be interpreted as evidence of reduced language dependence rather than as evidence that Spanish is intrinsically easier than Catalan.

Hybrid retrieval closely reproduces the dense pattern. Under $D0$-$Q0$, the hybrid system obtains 0.331 for Catalan, 0.411 for Spanish, and 0.265 for English. Historical query adaptation produces a more balanced profile of 0.410, 0.425, and 0.344, respectively. Catalan and English both gain 0.079, while Spanish changes little $(+0.014)$.

The cross-encoder obtains the highest mean nDCG@10 in all three languages: 0.538 for Catalan, 0.463 for Spanish, and 0.408 for English. Still, the intervals remain broad because of the limited number of semantic intents. The language-specific results support the conclusion that multilingual dense representations and reranking are more robust than lexical retrieval under cross-lingual shift, but they do not establish a statistically ordered ranking among the three languages.

### 5.3 Effects of Historical Adaptation

Document-side and query-side adaptation produce different effects across retrieval architectures. In the pooled analysis, document adaptation increases BM25 nDCG@10 from 0.160 to 0.196. The paired difference is $+0.036$, with a 95% confidence interval of $[0.007, 0.069]$. The corresponding effects for dense and hybrid retrieval are smaller: $+0.016$ and $+0.017$, respectively, and both intervals include zero.

The stronger effect on BM25 is concentrated in Catalan. Document adaptation increases Catalan BM25 performance from 0.415 to 0.520, while producing no measurable change in Spanish or English. For dense and hybrid retrieval, the effect is heterogeneous: adaptation improves Catalan and English point estimates but reduces the Spanish

estimate. Document modernization appears to influence lexical correspondence more directly than multilingual semantic alignment.

Query-side adaptation produces the largest pooled improvement among the non-reranked configurations. Hybrid $D0$-$Q1$ improves over Hybrid $D0$-$Q0$ by $+0.057$ nDCG@10, with a 95% confidence interval of $[0.006, 0.119]$. A direct comparison between query adaptation and document adaptation favors $D0$-$Q1$ by $+0.041$, with an interval of $[0.009, 0.075]$. However, neither contrast remains statistically significant after Holm correction, with adjusted $p$-values of 0.131 and 0.080, respectively.

Applying document and query adaptation simultaneously does not produce an additive improvement in the pooled analysis. Hybrid $D1$-$Q1$ scores 0.362, compared with 0.393 for $D0$-$Q1$, corresponding to a paired difference of $-0.031$ with a confidence interval of $[-0.101, 0.018]$. The available evidence does not support treating the two transformations as reliably complementary.

### 5.4 Effect of Cross-Encoder Reranking

Cross-encoder reranking provides the highest average performance in both the held-out and pooled analyses. On the held-out test partition, reranking increases nDCG@10 from 0.280 for its first-stage Hybrid $D0$-$Q1$ retriever to 0.328. Recall@10 increases from 0.333 to 0.444, indicating that the reranking configuration is also associated with stronger recovery of the annotated evidence within the top ten positions.

In the pooled analysis, the increase is larger, from 0.393 to 0.470. The paired mean difference is $+0.077$, although its 95% confidence interval, $[-0.040, 0.202]$, includes zero and the Holm-adjusted $p$-value is 0.826. The result should be described as the strongest average performance rather than as a statistically confirmed superiority over the first-stage hybrid system.

The reranking effect also varies by language. In the pooled analysis, the changes relative to Hybrid $D0$-$Q1$ are $+0.129$ for Catalan, $+0.038$ for Spanish, and $+0.063$ for English. None of the corresponding confidence intervals excludes zero. In the held-out analysis, the estimated changes are positive for Catalan and Spanish but slightly negative for English. This variation indicates that reranking does not provide a uniform gain across all linguistic conditions.

### 5.5 Sensitivity to the Evaluation Partition

Figure 3 compares adaptation effects in the pooled 17-intent analysis with those obtained from the 12 answerable held-out test intents. The figure shows that the direction and magnitude of several effects depend on whether the development data are included.

The cross-encoder decreases from an nDCG@10 of 0.470 in the pooled analysis to 0.328 on the held-out test. Hybrid $D0$-$Q1$ decreases from 0.393 to 0.280, while Hybrid $D1$-$Q1$ decreases from 0.362 to 0.299. Hybrid $D0$-$Q0$ changes from 0.336 to 0.271. The ranking of the main systems is broadly preserved, but the advantages associated with query adaptation and reranking become smaller.

The interaction between document and query adaptation is especially sensitive to the split. In the pooled analysis, adding $D1$ to the query-adapted system reduces nDCG@10 by 0.031. On the held-out test, $D1$-$Q1$ exceeds $D0$-$Q1$ by approximately 0.019. This change of direction precludes a stable conclusion about complementarity. More generally, the split analysis indicates that the larger effects observed when development and test intents are pooled should not be interpreted as expected performance on unseen information needs.

## 6 Ablation and Error Analysis

### 6.1 Dense Retrieval versus Hybrid Fusion

The comparison between dense and hybrid retrieval provides little evidence that lexical fusion improves the unadapted multilingual system. In the pooled analysis, Dense $D0$-$Q0$ achieves an nDCG@10 of 0.337, while Hybrid $D0$-$Q0$ obtains 0.336. Their language-specific results are also almost identical: both score 0.331 in Catalan and 0.265 in English, while the Spanish score changes only from 0.414 to 0.411.

The same pattern occurs after document adaptation. Dense $D1$-$Q0$ and Hybrid $D1$-$Q0$ obtain nDCG@10 values of 0.352 and 0.353 in the pooled analysis. On the held-out test, dense retrieval reaches 0.276 and the corresponding hybrid system 0.270. The addition of the lexical component provides no stable advantage under either document condition.

This ablation suggests that BM25 contributes limited complementary evidence when a large proportion of the queries are cross-lingual. The selected fusion assigns only 25% of the normalized score to BM25, yet even this contribution does not improve the dense ranking consistently. The result does not imply that

lexical retrieval is generally unnecessary for historical collections. BM25 remains useful for Catalan queries and may recover distinctive names or historically stable expressions. In the present multilingual setting, however, its benefits are not sufficiently consistent to improve aggregate dense retrieval.

### 6.2 Document Adaptation, Query Adaptation, and Their Interaction

The adaptation ablation indicates that document-side and query-side transformations do not behave as interchangeable interventions. For hybrid retrieval, document adaptation increases pooled nDCG@10 by only 0.017 relative to $D0$-$Q0$, whereas query adaptation produces an increase of 0.057. The direct paired contrast favors query adaptation by 0.041. Although the unadjusted interval excludes zero, the Holm-adjusted $p$-value of 0.080 does not meet the predefined significance threshold.

The language-specific results help explain this difference. Document adaptation primarily benefits Catalan, where it increases hybrid nDCG@10 from 0.331 to 0.444. Its effect is negative in Spanish, where performance decreases from 0.411 to 0.334, and small in English, where it increases from 0.265 to 0.279. Query adaptation instead produces 0.410 in Catalan, 0.425 in Spanish, and 0.344 in English. Unlike document adaptation, it lowers none of the three languages: Catalan and English each gain 0.079 and Spanish 0.014.

Combining both interventions does not yield a stable interaction. In the pooled analysis, $D1$-$Q1$ underperforms $D0$-$Q1$, particularly in Spanish and English. In the held-out test, however, the combined configuration performs slightly better. The direction of the interaction changes across partitions. The appropriate conclusion is not that document and query adaptation are redundant, but that the present evidence does not demonstrate reliable complementarity.

The asymmetry between the interventions is also operationally relevant. Document adaptation transforms all 6,443 passages, whereas query adaptation is applied only when a query is submitted. Given the absence of a stable document-side gain and the lower processing requirements of query-side adaptation, the latter represents the more promising direction for further evaluation. This interpretation remains provisional because both transformations are controlled rule-based proxies rather than independently validated historical reconstructions.

### 6.3 Variation across Source Works and Author Groups

Retrieval effectiveness varies considerably across source works. Under the cross-encoder configuration, the pooled macro-average is 0.614 across the 12 intents linked to works by Ramon Llull, compared with 0.124 across the five non-Lullian intents. The same imbalance is visible before reranking: Hybrid $D0$-$Q0$ obtains 0.403 for Lullian intents and 0.175 for non-Lullian intents, while Hybrid $D0$-$Q1$ obtains 0.484 and 0.175, respectively.

These differences should not be interpreted as evidence of an inherent author preference. Author group is confounded with the number of intents, individual works, source quality, split allocation, passage style, and semantic content. The non-Lullian group contains only five answerable intents, while the Lullian group contains twelve. Moreover, the largest work in the corpus, *Llibre de contemplació*, contributes 4,104 passages but no relevant evidence intent in the current pilot, preventing its retrieval difficulty from being estimated directly.

At the individual-work level, cross-encoder nDCG@10 ranges from 0.000 for *Curial e Güelfa* and *Lo somni*, with two intents each, to 1.000 for *Llibre de proverbis*, also with two intents. Intermediate values include 0.272 for *Llibre d'amic e amat*, 0.547 for *Llibre de les bèsties*, 0.601 for *Llibre del gentil e dels tres savis*, 0.717 for *Llibre d'intenció*, and 0.621 for *Tirant lo Blanc*. The number of intents per work ranges from one to four, so these values are diagnostic rather than reliable work-level performance estimates.

The zero scores for *Curial e Güelfa* and *Lo somni* indicate that the annotated origin passages are not recovered in the top ten under the strongest configuration. They may reflect differences in style, narrative organization, vocabulary, or benchmark coverage. The available data do not permit these explanations to be separated, but the pattern shows that the aggregate performance is not uniformly distributed across the literary collection.

### 6.4 Retrieval Misses

Across the pooled analysis, 265 of the 459 system-query combinations fail to retrieve the annotated origin passage within the top ten, corresponding to an operational miss rate of 57.7%. The rate falls from 74.5% for BM25 $D0$-$Q0$ to 41.2% for the cross-encoder configuration, indicating that semantic

retrieval and reranking reduce, but do not eliminate, evidence-recovery failures.

The most frequent diagnostic pattern is a top-ranked passage from a different source work than the annotated evidence. This occurs in 197 of the 265 misses, or 74.3%. Because the available qrels identify only the passage used to construct the query, a different-work result cannot automatically be classified as semantically incorrect. It may represent unrelated retrieval, a thematically related but insufficient passage, or valid alternative evidence that is absent from the annotations. For this reason, the category is interpreted as different-work top-1 relative to the origin evidence, rather than as a definitive wrong-work error.

Among the 68 misses outside the different-work category, 11 are classified as cross-lingual lexical gaps corresponding to Spanish or English BM25 queries. Of the remaining 57, the origin passage appears between ranks 11 and 50 in 37 cases, indicating that the representation places the evidence among the candidates but not high enough for a top-ten interface; a further 15 occur between ranks 51 and 100, and five fall outside the top 100. Within this subset, insufficient ranking is far more common than complete non-recovery.

The cross-lingual lexical gap category is relatively small and should not be interpreted as the full extent of cross-lingual lexical failure, since many such cases are already included in the different-work category. The near-zero BM25 language results provide a clearer measure of the broader cross-lingual limitation.

The error analysis points to source discrimination and ranking, rather than complete absence from the candidate set, as the main observable problems. However, the incomplete qrels limit semantic interpretation. A larger pooled relevance assessment would be needed to distinguish genuinely irrelevant different-work passages from alternative evidence. The current taxonomy is used to characterize retrieval behavior rather than to assign definitive semantic error labels.

## 7 Discussion

The results show that retrieval from medieval literature cannot be understood as a uniform semantic-matching problem. Performance depends strongly on the relationship between query language, historical representation, and retrieval architecture. The clearest pattern is the contrast between lexical retrieval and multilingual dense retrieval. BM25 remains useful when contemporary queries preserve substantial lexical continuity with the source language, but its effectiveness deteriorates sharply when queries are expressed in Spanish or English. Multilingual dense retrieval substantially reduces this dependency, although it does not remove all variation across languages or source works.

This contrast can be interpreted as a distinction between lexical and multilingual semantic evidence. BM25 relies primarily on the recurrence of informative surface forms and benefits from names, distinctive terms, and vocabulary that has remained relatively stable over time. Such evidence is valuable in historical collections and explains why lexical retrieval remains competitive for Modern Catalan queries. However, the same mechanism becomes a limitation when a query differs from the medieval passage in both language and period. Spanish and English queries may express the appropriate concept without sharing tokens with the source text, leaving BM25 with little basis for ranking.

The dense retriever addresses part of this limitation by representing queries and passages in a shared multilingual space. Its comparatively balanced language profile suggests that contemporary multilingual encoders can transfer useful semantic structure to historically distant texts, even without task-specific fine-tuning. This transfer is not complete. Medieval forms remain underrepresented in the training data of general-purpose encoders, and English continues to be a relatively difficult query language in several configurations. Nevertheless, the transition from BM25 to dense retrieval changes the retrieval problem from one dominated by lexical proximity to one in which cross-lingual semantic correspondence becomes operationally useful.

The near-equivalence between dense and hybrid retrieval is also informative. Hybrid systems are often expected to combine complementary lexical and semantic signals, but this complementarity is not automatic. In the present setting, lexical evidence contributes strongly for Catalan queries and very weakly for cross-lingual queries. A single global fusion weight must combine signals whose usefulness changes according to the query language. The result is a hybrid system that largely reproduces the dense ranking rather than consistently improving it.

This finding does not imply that lexical retrieval is unnecessary for medieval collections. Exact matching remains valuable for proper names,

technical expressions, citations, and historically stable terminology. Instead, the results indicate that lexical contribution should probably be conditioned on the query and retrieval context. Future systems could estimate language proximity, lexical overlap, or uncertainty before deciding how strongly to weight BM25. A language-aware fusion mechanism may be more appropriate than a fixed score combination, particularly when the same collection is queried in both the source language and external languages.

Aggregate multilingual scores can obscure this language-dependent variation. BM25's respectable Catalan result masks near-zero scores in Spanish and English, while dense retrieval has a far more even profile; a macro-average alone would conceal the main finding of the study.

Multilingual robustness is not the same as language invariance. Point estimates differ across Catalan, Spanish, and English, and some differences change in the held-out analysis. Multilingual encoders may be influenced by the amount and quality of pretraining data available for each language, by translation choices in the query set, and by the historical relationship between the source and query languages. Catalan and Spanish are closely related Romance languages, while English introduces greater linguistic distance. However, the observed ordering is not sufficiently stable to support a general linguistic hierarchy. The stronger conclusion is that retrieval architecture changes the magnitude and structure of language dependence.

Historical adaptation produces a more heterogeneous picture. Document-side adaptation improves lexical retrieval primarily by increasing correspondence between Medieval and Modern Catalan forms. This effect is consistent with the assumptions underlying normalization-based historical retrieval. Its limited impact on Spanish and English is equally important: modernizing a Medieval Catalan passage toward contemporary Catalan does not directly align it with queries in another language. Document modernization and multilingual retrieval address different forms of mismatch.

The limited gains observed for dense retrieval after document adaptation suggest that multilingual representations already absorb some superficial orthographic variation or that the rule-based transformation does not modify the aspects of historical language that most affect semantic retrieval. Dense models may be relatively insensitive to minor spelling substitutions while remaining vulnerable to historically specific vocabulary, syntax, and conceptual expression. A stronger or linguistically validated modernization procedure might produce different results, but it could also introduce interpretative distortions. The present findings support treating document modernization as an architecture-dependent intervention rather than as a universally beneficial preprocessing step.

Query-side adaptation shows a more promising aggregate tendency. It can introduce historically plausible lexical cues while preserving the original corpus, and it incurs processing costs only when a query is issued. Its strongest practical advantage is not merely the observed increase in pooled effectiveness, but its operational flexibility. Query adaptation can be varied by language, user need, or collection without maintaining a second transformed version of the entire corpus.

At the same time, the held-out analysis shows that the magnitude of the query-adaptation effect is uncertain. The improvement is smaller on unseen intents, and it does not remain statistically significant after multiplicity correction. The result should be interpreted as evidence that query-side adaptation merits further study, not as confirmation that it is superior to document modernization. More sophisticated adapters could use historical lexicons, contextual language models, or constrained generation, but they would require explicit validation to ensure that the intended meaning is preserved.

The combined $D1$-$Q1$ condition does not provide stable evidence of complementarity. In the pooled analysis, adding document adaptation to the adapted-query condition reduces effectiveness, whereas the held-out test shows a small positive difference. This change of direction suggests that the two transformations sometimes reinforce one another and sometimes modify useful lexical or semantic signals in incompatible ways. Their interaction may depend on the individual query, source work, and language.

This result has methodological implications. Historical adaptation should not be evaluated only by comparing a fully transformed pipeline with an unadapted baseline. A factorial design is necessary to distinguish document-side, query-side, and interaction effects. Without this separation, an apparent improvement could be attributed to modernization even when it is driven primarily by query reformulation, or

a combined null result could conceal opposing component effects.

Cross-encoder reranking has the best mean scores, but the gain is not secure. The reranker benefits from joint query-passage interaction and can correct some ordering errors produced by the first-stage retriever. Its higher Recall@10 indicates that the complete reranked configuration more frequently places the annotated evidence within the visible result set. However, reranking cannot recover passages that are absent from the initial candidate pool, and the paired improvement over the first-stage hybrid system is not statistically conclusive.

The variation in reranking effects across languages suggests that contemporary multilingual cross-encoders may also exhibit language-conditioned behavior. The pooled results are positive in all three languages, but the held-out English effect is slightly negative. This does not demonstrate harm, given the wide confidence intervals and limited sample, but it shows that a stronger aggregate result does not guarantee uniform benefit. Rerankers intended for multilingual historical retrieval should be evaluated for both average effectiveness and distributional consistency across languages.

The source-level analysis reveals an additional form of heterogeneity. Performance is substantially higher for the available Lullian intents than for the small non-Lullian subset, while some individual works show no recovery of the annotated evidence within the top ten. These differences cannot be assigned directly to authorship because work, genre, passage style, textual quality, and split allocation are confounded. Nevertheless, they indicate that a retrieval system can perform well on average while failing systematically for particular literary traditions.

The predominance of different-work passages among top-ranked misses reinforces this concern. In some cases, the retrieved passage may be irrelevant or only thematically related. In others, it may constitute valid alternative evidence that is absent from the origin-based qrels. Historical and literary collections frequently discuss related concepts across multiple works, making single-passage relevance annotations especially restrictive. The error analysis exposes both retrieval failures and limitations in the current evaluation resource.

Four practical points follow for low-resource retrieval. First, useful multilingual access can be obtained through zero-shot dense models even when no supervised medieval retrieval dataset is available. Second, combining components does not necessarily improve performance; simple dense retrieval may be a stronger and more economical baseline than a fixed hybrid architecture. Third, lightweight query adaptation may be more practical than transforming an entire collection, but it must be evaluated independently by language and on held-out intents. Finally, system development should prioritize coverage across works and genres, not only improvements in overall ranking metrics.

More broadly, the study suggests that intelligent access to medieval literature requires modular and language-aware evaluation. Temporal adaptation, multilingual representation, lexical matching, and reranking solve partially overlapping problems. Their usefulness changes according to the linguistic relationship between users and texts. A retrieval architecture for historical collections should be designed as a conditional system in which representations and fusion strategies can respond to query language, historical distance, and collection characteristics rather than as a fixed universal pipeline.

## 8 Limitations and Conclusion

The study has several limitations that constrain the interpretation of its results. The held-out evaluation contains only 12 answerable semantic intents, while the broader 17-intent analysis includes development data used to select the hybrid weights and reranking configuration. The held-out results are the more credible estimate of generalization, whereas the pooled analysis provides additional descriptive resolution but may produce optimistic effect estimates. The small number of intents also results in wide confidence intervals and limited power after correction for multiple comparisons.

The relevance judgments identify the origin passage used to construct each answerable query but do not exhaustively annotate all relevant passages in the corpus. A system can retrieve valid evidence from another passage or work and still be counted as unsuccessful. This issue is particularly important for literary and philosophical collections, in which similar concepts recur across texts. The reported metrics should be interpreted as recovery of the annotated origin evidence rather than complete semantic recall. Future evaluations should construct pooled graded qrels and adjudicate alternative evidence from multiple systems.

The document and query transformations are controlled rule-based proxies. They have not been validated as philologically complete modernizations or historical reconstructions, and their effects cannot be generalized to expert normalization procedures. Some substitutions may improve surface correspondence without addressing deeper semantic shift, while others may remove useful historical distinctions. A larger study should validate meaning preservation and historical plausibility on stratified samples before treating either intervention as a mature system component.

The corpus is also unevenly distributed. Works by Ramon Llull constitute most of the passages, and the *Llibre de contemplació* alone contributes a substantial proportion of the collection without contributing an answerable intent to the current pilot. The observed Lullian and non-Lullian differences are confounded by coverage, sample size, genre, and source quality. Broader evaluation requires a more balanced allocation of intents across works and explicit coverage of the largest and noisiest corpus components.

Only one multilingual dense checkpoint, one fixed hybrid weighting scheme, and one cross-encoder were examined. The results characterize these configurations rather than all dense, hybrid, or reranking architectures. Larger encoders, historically adapted language models, learned fusion mechanisms, and multilingual rerankers may behave differently. The use of fixed min-max fusion is especially restrictive because the relevance of lexical evidence varies substantially by query language.

The study focuses on Medieval Catalan as a case study. Although the joint problem of temporal and cross-lingual mismatch applies to other historical traditions, the empirical findings cannot be assumed to transfer directly to Latin, Old French, Middle English, or other medieval languages. Differences in linguistic continuity, digitization quality, orthographic standardization, and model pretraining coverage may produce different architecture-dependent effects.

Finally, the evaluation concerns passage retrieval rather than end-to-end retrieval-augmented generation. The name HistRAG reflects the intended role of the framework as an evidence-retrieval component, but no generated answers are assessed here. Stronger retrieval does not by itself guarantee historically accurate, faithful, or well-cited generation. Answer correctness and evidence support require a separate evaluation with blind judgments and explicit abstention criteria.

Within these limits, the main finding is that query language and retrieval architecture interact strongly. On the pooled analysis, BM25 drops from 0.415 nDCG@10 for Catalan queries to below 0.04 for Spanish and English, whereas multilingual dense retrieval keeps all three languages between 0.26 and 0.42; fixed fusion with BM25 leaves the dense ranking essentially unchanged. Historical adaptation has smaller and less stable effects: query-side adaptation helps most among non-reranked systems but does not survive Holm correction and shrinks on held-out intents, document modernization mainly helps same-language BM25, and the two are not additive. Cross-encoder reranking gives the highest averages (0.328 held-out, 0.470 pooled) without a statistically confirmed advantage over its first stage.

These findings support a language-aware and modular approach to historical retrieval. Medieval collections should not be treated simply as contemporary corpora with unusual spelling, nor should multilingual access be reduced to translation. Temporal and cross-lingual mismatch affect different components of the retrieval process, and methods should be evaluated according to the particular combination of language, historical distance, and architecture.

HistRAG provides a reproducible foundation for this evaluation. Future work should expand the held-out query set, introduce graded pooled relevance judgments, validate historical transformations, balance evidence across works and genres, and compare additional multilingual and historically adapted models. Extending the framework to other medieval languages would determine which findings are specific to Medieval Catalan and which reflect broader properties of retrieval across languages and centuries. A subsequent end-to-end study could then examine how retrieval quality influences answer correctness, faithfulness, citation support, and responsible abstention in generative access to historical literature.

## Data Availability Statement

Data will be made available on request.

## Funding

This work was supported without any funding.

## Conflicts of Interest

The author declares no conflicts of interest.

## AI Use Statement

The author declares that no generative AI was used in the preparation of this manuscript.

## Ethical Approval and Consent to Participate

Not applicable. This study did not involve human participants, personal data, or animals.

## References


[1] Bollmann, M. (2019). A large-scale comparison of historical text normalization systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 3885-3898). [CrossRef]

[2] Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1489-1501). [CrossRef]

[3] Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: A survey. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1384-1397). https://aclanthology.org/C18-1117/

[4] Tahmasebi, N., Borin, L., & Jatowt, A. (2021). Survey of computational approaches to lexical semantic change detection. *Computational approaches to semantic change, 6*(1). [CrossRef]

[5] Periti, F., & Montanelli, S. (2024). Lexical semantic change through large language models: a survey. *ACM Computing Surveys, 56*(11), 1-38. [CrossRef]

[6] Martinc, M., Kralj Novak, P., & Pollak, S. (2020). Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 4811-4819). https://aclanthology.org/2020.lrec-1.592/

[7] Xu, X., Deng, K., Hu, F., & Li, L. (2019). An improved historical embedding without alignment. *arXiv preprint arXiv:1910.08692*. [CrossRef]

[8] Schlechtweg, D., Tahmasebi, N., Hengchen, S., Dubossarsky, H., & McGillivray, B. (2021, November). DWUG: A large resource of diachronic word usage graphs in four languages. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 7079-7091). [CrossRef]

[9] Tsakalidis, A., Basile, P., Bazzi, M., Cucuringu, M., & McGillivray, B. (2021). DUKweb, diachronic word representations from the UK Web Archive corpus. *Scientific Data, 8*(1), 269. [CrossRef]

[10] Sánchez-Martínez, F., Martínez-Sempere, I., Ivars-Ribes, X., & Carrasco, R. C. (2013). An open diachronic corpus of historical Spanish. *Language resources and evaluation, 47*(4), 1327-1342. [CrossRef]

[11] Clérice, T., Pinche, A., Vlachou-Efstathiou, M., Chagué, A., Camps, J. B., Levenson, M. G., ... & Kiessling, B. (2024, August). CATMuS Medieval: A multilingual large-scale cross-century dataset in Latin script for handwritten text recognition and beyond. In *International Conference on Document Analysis and Recognition* (pp. 174-194). Cham: Springer Nature Switzerland. [CrossRef]

[12] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020, July). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8440-8451). [CrossRef]

[13] Reimers, N., & Gurevych, I. (2020, November). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 4512-4525). [CrossRef]

[14] Gao, T., Yao, X., & Chen, D. (2021, November). Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 6894-6910). [CrossRef]

[15] Shi, P., Zhang, R., Bai, R. H., & Lin, J. (2021, November). Cross-lingual training of dense retrievers for document retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning* (pp. 251-253). [CrossRef]

[16] Zhang, H., Gong, Y., Liu, D., Zhang, S., He, X., Lv, J., & Guo, J. (2024, May). Knowledge enhanced pre-training for cross-lingual dense retrieval. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 9810-9821). [CrossRef]

[17] Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). Multilingual E5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*. [CrossRef]

[18] Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*. [CrossRef]

[19] Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2023). MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 2014-2037). [CrossRef]

[20] Zhang, X., Thakur, N., Ogundepo, O., Kamalloo, E., Alfonso-Hermelo, D., Li, X., ... & Lin, J. (2023). MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics, 11*, 1114-1131. [CrossRef]

[21] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., & Gatford, M. (1995). Okapi at TREC-3. In D. K. Harman (Ed.), *Overview of the Third Text REtrieval Conference (TREC-3)*, NIST Special Publication 500-225 (pp. 109-126). Gaithersburg, MD: National Institute of Standards and Technology. https://trec.nist.gov/pubs/trec3/t3_proceedings.html

[22] Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press. [CrossRef]

[23] Nogueira, R., & Cho, K. (2019). Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*. [CrossRef]

[24] Yadav, N., Monath, N., Angell, R., Zaheer, M., & McCallum, A. (2022). Efficient nearest neighbor search for cross-encoder models using matrix factorization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 2171-2194). [CrossRef]

[25] Vast, M., Morand, V., van Cooten, B., Soulier, L., Mothe, J., & Piwowarski, B. (2026). MICE: Minimal interaction cross-encoders for efficient re-ranking. *arXiv preprint arXiv:2602.16299*. [CrossRef]

[26] Purday, J. (2009). Think culture: Europeana.eu from concept to construction. *The Electronic Library, 27*(6), 919-937. [CrossRef]

[27] Akhoon, I. u. H., & Khan, M. Y. (2022). Evaluating the World Digital Library: A digital framework for global cultural preservation. *ShodhKosh: Journal of Visual and Performing Arts, 3*(2), 1074-1081. [CrossRef]

[28] Lin, S. C., & Yen, E. (2012). An introduction to Taiwan e-Learning and Digital Archives Program (TELDAP). *International Journal of Humanities and Arts Computing, 6*(1-2), 1-7. [CrossRef]

[29] Babka, L. (2026). From the Middle Ages to the Present–Making the Written Heritage available in the Manuscriptorium, Kramerius and eSbirky. cz Digital Libraries. *Z Badań nad Książką i Księgozbiorami Historycznymi, 20*(1), 121-131. [CrossRef]

[30] Xu, L., Lu, L., Liu, M., Song, C., & Wu, L. (2024). Nanjing Yunjin intelligent question-answering system based on knowledge graphs and retrieval augmented generation technology. *Heritage Science, 12*(1), 118. [CrossRef]

[31] Murugaraj, K., Lamsiyah, S., During, M., & Theobald, M. (2025). Topic-RAG for historical newspapers: Enhancing information retrieval in humanities research through topic-based retrieval-augmented generation. *Computational Humanities Research, 1*, e15. [CrossRef]

[32] Konstantinidis, S., Lotfy, H., Erne, A., Zahiragic, F., Kan, M. Y., & Vlachos, M. (2026). Improving Access to Historical Archives with Real-time RAG-based Systems. *arXiv preprint arXiv:2607.03440*. [CrossRef]

[33] Bonifacio, L., Jeronymo, V., Abonizio, H. Q., Campiotti, I., Fadaee, M., Lotufo, R., & Nogueira, R. (2021). mMARCO: A multilingual version of the MS MARCO passage ranking dataset. *arXiv preprint arXiv:2108.13897*. [CrossRef]

[34] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65-70. https://www.jstor.org/stable/4615733

**Antoni Mestre** is an artificial intelligence researcher and university lecturer whose work lies at the intersection of hybrid intelligence, responsible AI, human oversight, explainability, and AI governance. He is an Associate Professor at the Universitat Oberta de Catalunya and has conducted research at the Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València. (Email: antoni.mestre@upv.es)