

## Article

# CollectivIA: Two-Pipeline Multilingual Legal RAG for Moroccan Territorial Governance with LLM-Assisted and Regex-Based Chunking

Firiel Zouak , Omar El Beqqali and Jamal Riffi

Informatique, Innovation et Intelligence Artificielle Laboratory, Department of Informatics, Faculty of Sciences  
Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fes 30000, Morocco

\* Correspondence: zouak.firiel@usmba.ac.ma

## Abstract

Retrieval grounding is crucial for high-stakes administrative applications, since large language models remain prone to hallucinations when addressing legal questions. This problem is particularly relevant in Moroccan territorial governance, where official legislative PDFs have highly heterogeneous digital quality, user interactions often occur in Moroccan Darija, and the legal corpus is bilingual Arabic–French. This paper presents CollectivIA, a multilingual Retrieval-Augmented Generation system implemented for Moroccan territorial governance law. The system supports queries in French, Arabic, and Moroccan Darija and indexes 2272 article-level segments from sixteen official legislative documents. We compare two end-to-end retrieval pipelines: an LLM-assisted semantic chunking architecture using Gemini and ChromaDB and a regex-based chunking architecture using FAISS. Based on an expanded multilingual benchmark of 150 legal queries, with 50 queries per language group, the LLM-assisted pipeline achieved higher RAGAS scores than the regex-based pipeline, particularly improving Context Precision from 0.315 to 0.818. The multimodal Vision fallback successfully recovered 456 articles, which remained inaccessible under the regex-based pipeline. Overall, the LLM-assisted pipeline yielded legal boundaries with greater coherence and retrieved contexts with higher focus, while the regex-based design maintained a broader source diversity. These results suggest that LLM-assisted semantic chunking with multimodal fallback is a promising approach to enhance multilingual legal RAG over heterogeneous Moroccan legal corpora.

**Keywords:** retrieval-augmented generation; legal NLP; Arabic NLP; Moroccan law; Darija; document chunking; ChromaDB; FAISS; LLM evaluation; RAGAS; bilingual AI; digital governance



Academic Editors: Sardar Jaf,  
Basel Barakat and Yongqiang Cheng

Received: 12 April 2026

Revised: 3 June 2026

Accepted: 18 June 2026

Published: 29 June 2026

**Copyright:** © 2026 by the authors.  
Licensee MDPI, Basel, Switzerland.  
This article is an open access article  
distributed under the terms and  
conditions of the [Creative Commons  
Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

## 1. Introduction

The use of large language models (LLMs) for legal question answering has attracted growing research interest, as it has the potential to democratize access to legal information. This potential is, however, fundamentally limited by hallucination, the generation of factually incorrect content with unwarranted confidence. In the legal field, hallucination is not only a quality issue; it may jeopardize the reliability of legal information systems. Dahl et al. [1] measured hallucination rates between 69% and 88% for state-of-the-art general-purpose LLMs on legal queries, and Magesh et al. [2] found that even the best-performing commercial legal AI tool hallucinated in at least 17% of evaluated cases. Such rates are an unacceptable reliability risk in Moroccan administrative law, where decisions

should cite their legal basis precisely, and legal sources are subject to interpretation by administrative courts [3].

Lewis et al. [4] formalized Retrieval-Augmented Generation (RAG), which addresses hallucination by conditioning LLM generation on retrieved documentary evidence, anchoring responses in verifiable sources rather than relying only on the model's parametric memory. With applications including contract analysis [5], case law retrieval [6], and statutory interpretation [7], RAG has emerged as a major paradigm for legal AI systems [8,9]. However, the document chunking stage, namely the process by which lengthy legislative texts are divided into retrievable units, remains crucial to the quality of RAG outputs. This dependence has only recently been thoroughly described for legal corpora [10]. Regardless of the quality of the embedding model or the generation model, poor chunking can reduce retrieval precision [11].

To our knowledge, previous studies have not fully addressed the technical challenge presented by the Moroccan territorial governance context. The legislative corpus governing regions, prefectures, provinces, and communes is produced in both Arabic and French as a result of the 2011 constitutional reform [3]. Moroccan Darija, a dialectal Arabic variety combining Arabic morphology, French lexical borrowings, and Amazigh elements, is also frequently used by citizens when formulating administrative questions. This dialect lacks standardized orthographic conventions, which complicates traditional NLP preprocessing [12].

We present CollectivIA (<https://collectivia.ma>, accessed on 1 June 2026), a bilingual RAG system for Moroccan territorial governance law, and empirically compare two retrieval pipeline architectures evaluated with the RAGAS framework [13]. This study makes four contributions. We first characterize CollectivIA as a deployed research artifact from the observed behavior of 328 real user queries. Second, we compare two retrieval architectures on French, Modern Standard Arabic, and Moroccan Darija using an expanded benchmark of 150 legal queries, 50 queries per language group. Third, using four RAGAS metrics and complementary retrieval indicators, we show that Context Precision is one of the most informative retrieval-oriented metrics in our experimental setting. Fourth, we discuss the limitations of automated evaluation, including the potential for evaluator bias in RAGAS for low-resource dialectal inputs, and supplement the quantitative analysis with a limited human-in-the-loop validation on a representative subset. We also discuss implications of cloud-dependent LLM pipelines for digital sovereignty.

The paper is organized as follows. Section 2 reviews related work on legal RAG systems, document chunking strategies, and Arabic NLP. Section 3 presents the materials and methods. Section 4 reports the results. Section 5 discusses the findings and their limitations. Section 6 concludes with future research directions.

## 2. Related Work

### 2.1. Hallucination and Retrieval-Augmented Generation in Legal AI

Since the first neural models were applied to legal text, factual reliability has remained a recurring concern in legal AI. One of the first large-scale empirical studies of legal hallucinations was by Dahl et al. [1], who found hallucination rates between 69% and 88% on legal questions taken from U.S. case law for four frontier LLMs. Magesh et al. [2] further explored this analysis in the context of commercial legal AI tools, reporting that specialized tools reduce hallucinations but do not fully eliminate them, with a minimum hallucination rate of 17% observed. These results have prompted the widespread use of RAG as a mitigation strategy in legal AI deployments.

The RAG framework was introduced by Lewis et al. [4], who showed that retrieval-conditioned generation can reduce hallucination in knowledge-intensive NLP tasks. RAG

has since been adapted to legal applications with specific requirements for traceability, citation, and factual grounding. LegalBench-RAG [9] showed that popular embedding models are not as good on legal text as they are on general-domain benchmarks, and established text-level precision and recall as an important evaluation standard for legal retrieval systems. Keisha et al. [14] showed that open-source embedding models such as SBERT [15] and GTE-Large are cost-effective and can be comparable to proprietary embeddings for legal retrieval using adaptive query translation with Recall@k improvements of 30–95%.

Wiratunga et al. [16] proposed CBR-RAG, which combines dense retrieval with case-based reasoning to exploit the structure of legal precedent. Han et al. [17] introduced the AusLaw Citation Benchmark, comprising 55,000 Australian legal instances, and showed that instruction tuning with task-specific data outperforms domain-specific pre-training alone. A recent systematic review of RAG covering 128 studies from 2020 to 2025 reported that retrieval quality is often a stronger determinant of performance than generation model capability [18].

For non-English legal systems, Ni'Mah et al. [19] showed that structure-aware chunking performs better than fixed-size approaches for Indonesian maritime law. Ajay Mukund & Easwarakumar [20] applied Dynamic RAG to Indian legal summarization, while Reuter et al. [21] proposed an agentic approach for retrieving legislative definitions. To the best of our knowledge, no previous RAG study has specifically addressed Moroccan administrative law.

## 2.2. Document Chunking for Legal Retrieval

In RAG systems, document chunking has become an important research concern, especially for structured domains such as law. The most thorough analysis of chunking techniques for legal retrieval was carried out by Ferraris et al. [10], who showed that paragraph-level and fixed-size segmentation can compromise the normative integrity of legal provisions, a failure they call Document-Level Retrieval Mismatch (DRM). They found that article-level chunking with explicit structural markers reduces DRM across all top-k retrieval positions.

LLM-based semantic chunking outperformed recursive character text splitting (RCTS) by up to 9%, according to Wang et al.'s benchmarking of seven chunking strategies across five datasets [22]. Moura Júnior et al. [23] showed that chunking optimization alone increased answer correctness from 64% to 72% in a diverse corpus without changing the embedding or generation model. Retrieval mismatch is further reduced by prepending LLM-generated document summaries to chunks, as demonstrated by Anthropic's Contextual Retrieval [24]. These results motivate the comparison between regex-based boundary detection and LLM-assisted article extraction in this study, particularly for Arabic documents with less standardized typographic conventions.

## 2.3. Arabic and Multilingual NLP for Retrieval Systems

Arabic NLP poses specific challenges for retrieval systems. In their analysis of Arabic NLP difficulties, Darwish et al. [12] note that dialectal varieties, including Moroccan Darija, remain underrepresented in linguistic resources and evaluation benchmarks. Retrieval performance gaps for Arabic RAG compared with English baselines were reported by El-Beltagy and Abdallah [25], who attributed these gaps mainly to embedding-space imbalances in multilingual models trained on language-imbalanced corpora. Wu et al. [26] further showed that cross-lingual retrieval is a major bottleneck in multilingual RAG, particularly when the query language and the indexing language differ.

According to Samih and Maier [27], Moroccan Darija is a code-switching variety where Arabic morphology is often combined with borrowings from French and elements of Amazigh in a single utterance. The Open Arabic LLM Leaderboard [28], an important ref-

erence for Arabic retrieval evaluation, proposed ALRAGE, an Arabic-specific RAG benchmark with M3 embeddings [29]. Our work adds to this nascent literature by tackling a compound retrieval challenge in the Maghrebi legal domain, where dialectal input and bilingual legal corpora are combined, and current benchmarks do not exist.

#### 2.4. RAG Evaluation Frameworks

Methodological challenges in evaluating RAG systems have led to the development of specialized evaluation frameworks. RAGAS [13] provides four reference-free metrics: Faithfulness, Answer Relevancy, Context Precision, and Context Recall. These metrics are computed using an LLM judge and do not require gold-standard annotated passage–answer pairs. This makes RAGAS useful in domains where annotated QA resources are unavailable, as is currently the case for Moroccan legal question answering.

RAGAS has been used in legal RAG evaluation by Keisha et al. [14] and has been compared with human annotations in several studies [18]. A recent systematic review reported that RAGAS is among the most widely used RAG evaluation frameworks in the recent literature [18]. However, automated LLM-based evaluation can introduce evaluator bias, especially across languages and low-resource dialects. This is particularly the case for Moroccan Darija, where linguistic variation and non-standardized spelling might affect retrieval and evaluation.

In addition to automated evaluation, recent work on RAG evaluation has also focused on human-in-the-loop evaluation of dimensions such as context relevance, answer faithfulness, answer relevance, and citation correctness. We therefore rely on RAGAS as the core quantitative framework in this study and supplement it with a limited human-in-the-loop validation on a representative subset in order to verify whether the automatic trends are in agreement with manual evaluation.

### 3. Materials and Methods

#### 3.1. Legal Corpus

Sixteen official Moroccan legislative texts obtained from the public portal of the Secretariat General of Government (SGG) are included in the CollectivIA corpus [30]. These texts cover the 2011 Constitution [3], public finance and procurement legislation, including Decree 2-12-349 and Laws 45-08 and 47-06, the organic governance laws governing regions, prefectures/provinces, and communes, namely Laws 111-14, 112-14, and 113-14, as well as complementary administrative legislation, including Laws 55-19, 54-19, 12-90, 31-13, and Dahir 1.15.84. The full corpus composition for both architectures is presented in Table 1.

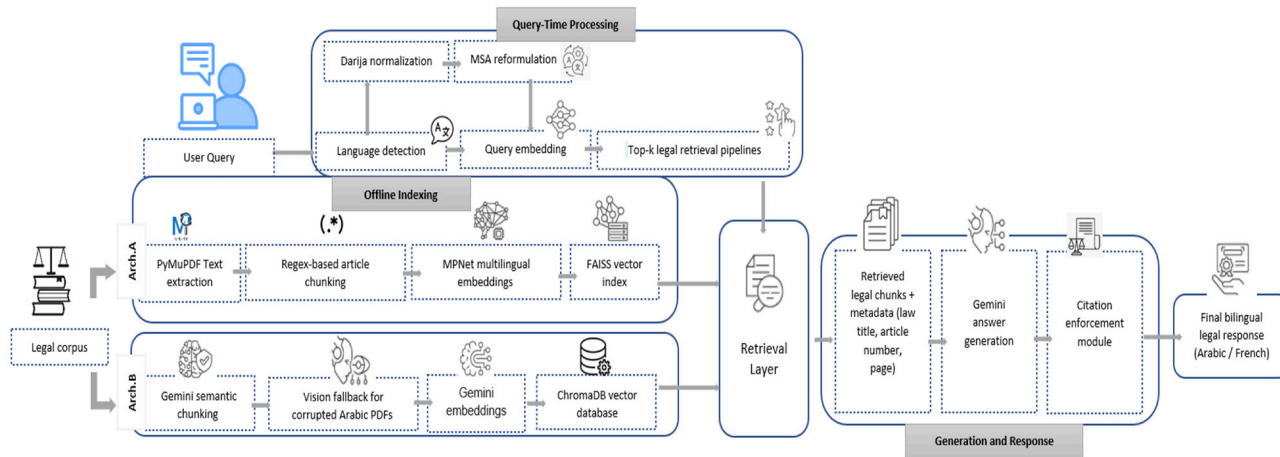
**Table 1.** Corpus composition and effectively indexed chunks produced by each architecture. Arch. A = Architecture A; Arch. B = Architecture B. (\*) indicates complete extraction failure under Architecture A due to encoding corruption. Chunk counts reflect the effective output of each extraction and indexing pipeline. Differences in indexed coverage are part of the practical evaluation of the two retrieval architectures.

| Legislative Text                    | Reference | Lang.   | Pages | Arch. A Chunks | Arch. B Chunks | Domain         |
|-------------------------------------|-----------|---------|-------|----------------|----------------|----------------|
| Loi organique—Régions               | 111-14    | FR      | 70    | 180            | 256            | Regional       |
| Loi organique—Préfectures/Prov.     | 112-14    | FR      | 70    | 150            | 209            | Provincial     |
| Loi organique—Communes              | 113-14    | FR      | 87    | 200            | 283            | Municipal      |
| Décret marchés publics              | 2-12-349  | FR      | 144   | 168            | 171            | Procurement    |
| Loi 45-08 finances CT               | 45-08     | FR      | 19    | 40             | 43             | Finance        |
| Loi 47-06 fiscalité (FR + AR)       | 47-06     | FR/AR   | 179   | 300            | 518            | Taxation       |
| 2011 المغربية المملكة دستور         | 2011      | AR      | 40    | 0 (*)          | 179            | Constitution   |
| Décrets d’application (3 texts)     | Various   | AR      | 484   | 0 (*)          | 456            | Multi-domain   |
| Other laws (55-19, 54-19, 12-90...) | Various   | FR/AR   | 85    | 163            | 157            | Administrative |
| TOTAL                               | 16 texts  | 9FR/7AR | 1178  | 1201           | 2272           | —              |

Raw text was extracted using PyMuPDF [31], followed by Unicode NFKC normalization [32], harakat removal (Unicode range U+064B–U+065F), and alif normalization, where U+0622, U+0623, and U+0625 were mapped to U+0627. These preprocessing steps are commonly used in Arabic NLP to reduce orthographic variation and prevent diacritical differences from producing distinct embedding representations for the same lexical item [12]. During corpus preparation, four categories of document quality issues were identified: (i) encoding corruption in Arabic PDFs, defined by a garbage character ratio exceeding 4% of all extracted characters and affecting four documents; (ii) typographic variation in article headers across publications; (iii) multi-column layouts interfering with sequential text extraction; and (iv) mixed Arabic–French script within single pages.

We evaluated the two architectures with their respective extraction and indexing workflows. This implies that the indexed corpora vary in the coverage of articles retrieved. Architecture A uses regex-based boundary detection on standard text extraction, and Architecture B combines LLM-assisted extraction with a multimodal Vision fallback for corrupted PDFs. Consequently, Architecture B was able to retrieve 456 Arabic articles that were not accessible in Architecture A. Such differences are considered for the interpretation of the comparative results, since they mirror the practical behavior of the entire retrieval pipelines when applied to heterogeneous official legal documents.

Figure 1 presents the overall architecture of CollectivIA, including the two offline indexing pipelines, the Darija-aware query normalization workflow, the unified retrieval layer, and the citation-constrained bilingual answer generation module. This end-to-end representation provides an overview of the proposed system before detailing its individual components. It also illustrates the methodological differences between Architecture A (Regex + FAISS) and Architecture B (Gemini + ChromaDB) while showing how both pipelines are integrated into the online question-answering workflow.



**Figure 1.** Global architecture of CollectivIA, illustrating the two offline indexing pipelines—Architecture A (Regex + FAISS) and Architecture B (Gemini + ChromaDB)—together with the Moroccan Arabic dialect Darija-aware query normalization workflow, the unified retrieval layer, and the citation-constrained bilingual answer generation module.

### 3.2. Architecture A: Regex-Based Chunking with FAISS

#### 3.2.1. Article Boundary Detection

Architecture A detects article boundaries using three compiled regular expression patterns applied to the reconstructed full text of each document. These patterns cover the three article-header conventions most frequently observed in the corpus:

- Pattern\_FR: `^\\s*Article\\s+ (premier|1er|\\d+)\\s*:? [re. IGNORECASE]`
- Pattern\_AR: `^\\s*(المادة|الفصل)\\s*(\\d+|[٩-٠]+) [re. UNICODE]`

- Pattern\_CO:  $\wedge s^*(\backslash d+)\backslash s + (\text{الفصل})\backslash s^*$  [2011 Constitution format]

For each regex match, the bold font flag (bit 4 of the span flags field) and a minimum font size threshold were verified using PyMuPDF typographic attributes. The threshold was set to  $\geq 10$  pt for French headers and  $\geq 11$  pt for Arabic headers. This dual criterion reduces false positives caused by in-body cross-references to articles. When fewer than two article boundaries were detected in a document, the system fell back to fixed-size chunking using an 800-character window with a 150-character overlap, which typically indicated encoding corruption or structural irregularity in the source PDF.

### 3.2.2. Embedding and Vector Indexing

Valid article chunks are encoded using the paraphrase-multilingual-mpnet-base-v2 model [15], a Siamese BERT network [33] that produces 768-dimensional dense embeddings optimized for semantic similarity across more than 50 languages. The embeddings are L2-normalized and indexed using FAISS IndexFlatIP [34], which performs an exact exhaustive inner-product search. Since both query and document embeddings are L2-normalized, the inner product is equivalent to cosine similarity:

$$\text{Sim}(q, d_i) = \frac{E(q) \cdot E(d_i)}{\|E(q)\| \cdot \|E(d_i)\|} \quad (1)$$

where  $q$  denotes the user query,  $d_i$  represents the  $i$ -th candidate document chunk,  $E(q)$  is the dense embedding vector of the query, and  $E(d_i)$  is the dense embedding vector of chunk  $d_i$ .

Because L2 normalization ensures that  $\|E(q)\| = \|E(d_i)\| = 1$ , the denominator in Equation (1) simplifies to 1. Consequently, FAISS IndexFlatIP computes similarity using the dot product:

$$\text{Sim}(q, d_i) = E(q) \cdot E(d_i) \quad (2)$$

The index and chunk metadata are serialized to disk and reloaded at server startup to avoid repeated HuggingFace model download and initialization overhead. In the production deployment, this reload process takes approximately 12 s.

## 3.3. Architecture B: LLM-Assisted Chunking with ChromaDB

### 3.3.1. Gemini-Powered Article Extraction

In Architecture B, Gemini 2.5 Flash [35] is used for article boundary detection and content extraction. The model processes the input text in sliding windows of five pages per API call. The chunking prompt (Appendix A) instructs the model to return a structured JSON array following the schema {article\_numero, article\_titre, contenu, langue} and to identify article boundaries using typographic, structural, and semantic cues. This method aims to manage typographic variation within the article headers by taking advantage of the model's multilingual knowledge of legal documents and article structure.

For documents with a garbage character ratio greater than 4%, pages are rasterized into PNG images using PyMuPDF (version 1.24.9) at twice the standard resolution ( $1190 \times 1684$  pixels) and submitted to Gemini through its multimodal Vision API. This Vision fallback bypasses the corrupted text extraction layer. Using this Vision fallback, 456 articles were recovered from four Arabic PDF documents for which Architecture A produced no valid chunks. To handle page-boundary splits, post-processing removes duplicate articles sharing the same article\_numero, cleans JSON formatting artifacts, and retains the longest extracted version.

### 3.3.2. Embedding and Persistent Storage

In Architecture B, chunks are embedded using gemini-embedding-001 [36], Google’s multilingual text embedding model. At the time of evaluation, this model ranked among the highest-performing multilingual embedding models on the MTEB Multilingual leaderboard, covering 117 languages. To exploit asymmetric retrieval optimization, `task_type = RETRIEVAL_DOCUMENT` is used during indexing, while `task_type = RETRIEVAL_QUERY` is used for query encoding. The model produces 3072-dimensional embeddings and supports Matryoshka Representation Learning (MRL) [37].

The resulting embeddings are stored in ChromaDB [38], a persistent vector database that uses Hierarchical Navigable Small World (HNSW) graphs [39] for approximate nearest-neighbor search. Retrieval is performed using cosine distance with  $ef_{search} = 100$ . ChromaDB’s built-in WHERE-clause filtering enables collectivity-type routing. For example, queries related to communes are filtered toward Law 113-14, while queries related to regions and prefectures/provinces are routed toward Laws 111-14 and 112-14, respectively.

### 3.4. Darija Normalization Module

Moroccan Darija queries are handled by a specialized preprocessing module before vector retrieval. Darija detection relies on three signals: (i) an Arabic character ratio greater than 0.20, indicating Arabic-script input; (ii) a 45-token dialectal marker lexicon including common Moroccan Darija terms such as *wash*, *kifash*, *bghit*, *dyal*, and similar tokens; and (iii) Arabizi character–digit mixing patterns. When Darija is detected, the query is forwarded to Gemini 2.5 Flash, using the normalization prompt as described in Appendix A, which requests the model to reformulate the input to MSA for retrieving legal information about Moroccan administrative law. The original dialectal form is preserved for response language detection, while the normalized query is used for vector retrieval.

This design is based on the cross-lingual query translation evidence reported in [26]. However, normalization can sometimes over-standardize or change dialectal nuances, especially in legal queries. This is discussed in more detail in Section 5.

### 3.5. Generation and Citation Enforcement

The retrieved chunks are displayed as numbered source blocks [SOURCE 1] to [SOURCE k] with the exact law title, article number and page number taken from the indexed metadata. We used Gemini 2.5 Flash for both French and Arabic outputs with `temperature = 0.1` and `max_tokens = 3000`. The generation prompts in Appendix A direct the model to only cite law names that appear verbatim in the retrieved source blocks and not to infer article numbers from its parametric knowledge.

For non-legal queries, a keyword-based out-of-scope classifier returns a standard refusal message without invoking an LLM generation. Following LegalBench-RAG [9] and the citation benchmark of Han et al. [17], we fixed the retrieval parameter `k` to 5 for all experiments.

### 3.6. Evaluation Metrics and Statistical Protocol

We evaluate both architectures using the RAGAS framework [13], which offers four reference-free RAG evaluation metrics automatically computed by an LLM judge without the need for manually annotated gold-standard passages. RAGAS has been used by Keisha et al. [14] in the context of legal RAG evaluation and has been compared to human annotations in several independent studies. However, the LLM-based evaluation might introduce evaluation bias, especially for low-resource dialectal inputs such as Moroccan Darija; thus, we interpret the RAGAS results as comparative markers rather than absolute measures of legal correctness.

The new evaluation benchmark contains 150 legal queries, including 50 French (FR), 50 Modern Standard Arabic (AR), and 50 Moroccan Darija (DA) questions per architecture. The queries cover the main domains of the corpus, including territorial governance, local public finance, procurement, administrative procedures, access to information, and urban planning. The following paragraphs define the four RAGAS metrics used in the evaluation.

**Faithfulness (Faith.)** measures the proportion of generated answer statements that are directly supported by the retrieved context:

$$Faith. = \frac{|\{s \in Answer : s \text{ is entailed by Context}\}|}{|\{s \in Answer\}|} \quad (3)$$

where

- $s$  denotes an elementary statement extracted from the generated answer.
- $Answer$  represents the set of all statements contained in the model-generated response.
- $Context$  corresponds to the set of document chunks retrieved during the retrieval phase.
- $|\cdot|$  denotes the set cardinality, i.e., the number of elements.

**Answer Relevancy (Relev.)** measures semantic alignment between the generated answer and the original question, computed as the mean cosine similarity between the original query embedding and embeddings of LLM-generated back-questions from the answer [13].

**Context Precision (CP)** measures whether the relevant chunks appear at the top ranks of the retrieved set. It rewards retrieval systems that place relevant documents higher, penalizing configurations that retrieve correct information but rank it low.

**Context Recall (CR)** measures the fraction of the ground-truth answer that can be attributed to the retrieved context chunks, indicating retrieval comprehensiveness.

In addition to the RAGAS metrics, we also include the average cosine similarity of the top 5 retrieved chunks for every query as a direct indicator of retrieval quality. For comparison of architectures, in general, across the 150-query benchmark, we use paired two-sided  $t$ -tests, and language-stratified comparisons are performed separately for the French, Arabic and Darija subsets. For the automated evaluation, a small human-in-the-loop validation was performed on a representative subset of 45 multilingual legal queries selected from the 150-query benchmark, including 15 French, 15 Modern Standard Arabic and 15 Moroccan Darija queries. This manual assessment was not intended to provide a statistically valid benchmark but to verify whether the trends revealed by the automatic RAGAS evaluation were in line with qualitative expert judgment. Manual evaluation was performed in four dimensions: Context Relevance, Answer Faithfulness, Answer Relevance, and Citation Correctness. Context Relevance measures the relevance of the retrieved passages to the user query. Answer Faithfulness measures whether the generated claims are supported by the retrieved context. Answer Relevance measures how well the answer responds to the original question. These dimensions are aligned with evaluation criteria used in RAG assessment frameworks such as RAGAS [13] and ARES [40]. Citation Correctness verifies whether the cited legal references are consistent with the retrieved sources and adopts the attribution-oriented evaluation approach proposed in recent citation-aware RAG studies [41]. We used this validation as a qualitative check of reliability and not as a replacement for the RAGAS-based quantitative evaluation.

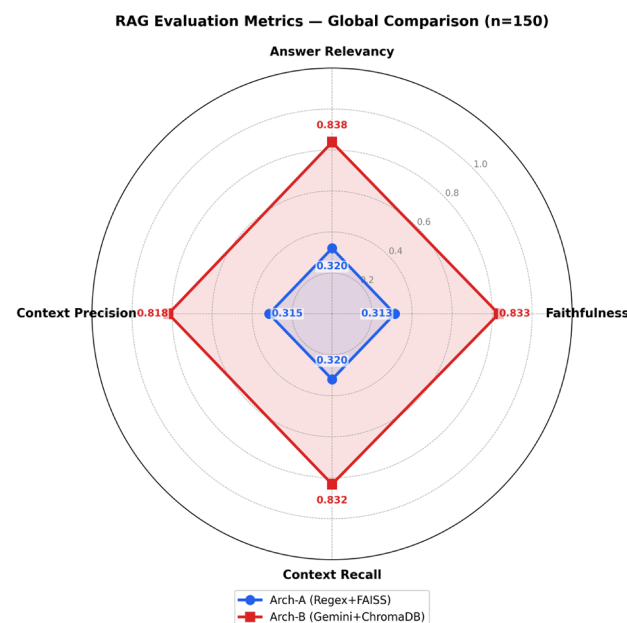
## 4. Results

Table 2 and Figures 2 and 3 present the global comparison of the two evaluated retrieval architectures across the four RAGAS metrics: Faithfulness, Answer Relevancy, Con-

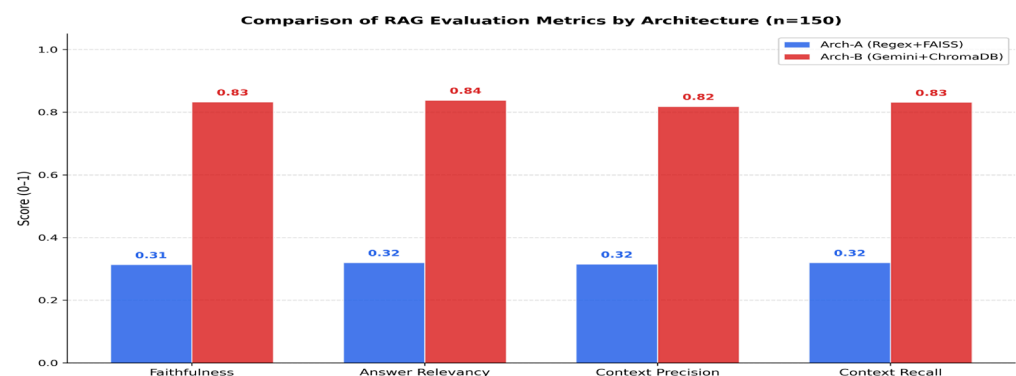
text Precision, and Context Recall. Source Diversity is also reported as an additional indicator of retrieval breadth.

**Table 2.** Evaluation results for Architecture A (Regex + FAISS) and Architecture B (Gemini + ChromaDB) across  $n = 150$  queries per architecture, including 50 French, 50 Modern Standard Arabic, and 50 Moroccan Darija queries. The table reports four RAGAS metrics, Source Diversity, and average retrieval cosine similarity. Welch’s two-sided  $t$ -tests were used for statistical comparisons. Cohen’s  $d$  is reported as an effect-size measure. Bold indicates the best value for each metric.

| Metric                  | Arch-A       | Arch-B       | Delta  | t-Stat. | p-Value | Cohen’s d |
|-------------------------|--------------|--------------|--------|---------|---------|-----------|
| Faithfulness            | 0.313        | <b>0.833</b> | +0.520 | −10.713 | <0.001  | +1.237    |
| Answer Relevancy        | 0.320        | <b>0.838</b> | +0.518 | −10.662 | <0.001  | +1.231    |
| Context Precision       | 0.315        | <b>0.818</b> | +0.503 | −10.332 | <0.001  | +1.193    |
| Context Recall          | 0.320        | <b>0.832</b> | +0.512 | −10.463 | <0.001  | +1.208    |
| Source Diversity (mean) | <b>4.135</b> | 1.842        | −2.293 | +11.992 | <0.001  | −1.385    |
| Avg. Retrieval Score    | <b>0.737</b> | 0.629        | −0.108 | +4.560  | <0.001  | −0.527    |



**Figure 2.** Global comparison of RAG evaluation metrics for Architecture A (Regex + FAISS) and Architecture B (Gemini + ChromaDB) across the expanded benchmark of 150 legal queries.



**Figure 3.** Global comparison of RAG evaluation metrics between Architecture A (Regex + FAISS) and Architecture B (Gemini + ChromaDB) across the expanded benchmark of 150 legal queries. Bars represent mean scores, and error bars indicate  $\pm 1$  standard deviation. All differences are statistically significant (Welch’s  $t$ -test,  $p < 0.001$ ).

At the global level, Architecture B (Gemini + ChromaDB) exceeds Architecture A (Regex + FAISS) on all four RAGAS metrics by a considerable margin. Architecture B has a Faithfulness of 0.833 to 0.313 for Architecture A and, more significantly, a Context Precision of 0.818 to 0.315. The most discriminative quantitative finding of this study is the gain in Context Precision ( $\Delta = +0.503$ ), indicating that the LLM-assisted pipeline retrieves more focused and legally coherent evidence than the regex-based architecture. This behavior is illustrated in Figure 2, where Architecture B consistently scores higher on all evaluation dimensions.

In contrast, Architecture A has a higher score in Source Diversity (4.135 vs. 1.842) and average retrieval cosine similarity (0.737 vs. 0.629), showing that the regex-based retrieval strategy maintains wider contextual neighborhoods and retrieves evidence from a larger set of legislative sources. This broader retrieval profile may account for the increased contextual breadth for some multilingual legal queries.

Context Recall also differs substantially between the two architectures (0.832 vs. 0.320), with Architecture B achieving a clear advantage. Combined with the large improvements observed in Faithfulness, Answer Relevancy, and Context Precision, this result suggests that the superiority of Architecture B is not limited to ranking quality alone but extends to the overall retrieval and grounding process.

Overall, these findings reveal a trade-off between retrieval breadth and retrieval focus. Architecture A prioritizes broader source coverage and higher source diversity, whereas Architecture B retrieves fewer but more coherent and citation-ready legal contexts, resulting in substantially higher end-to-end RAG performance.

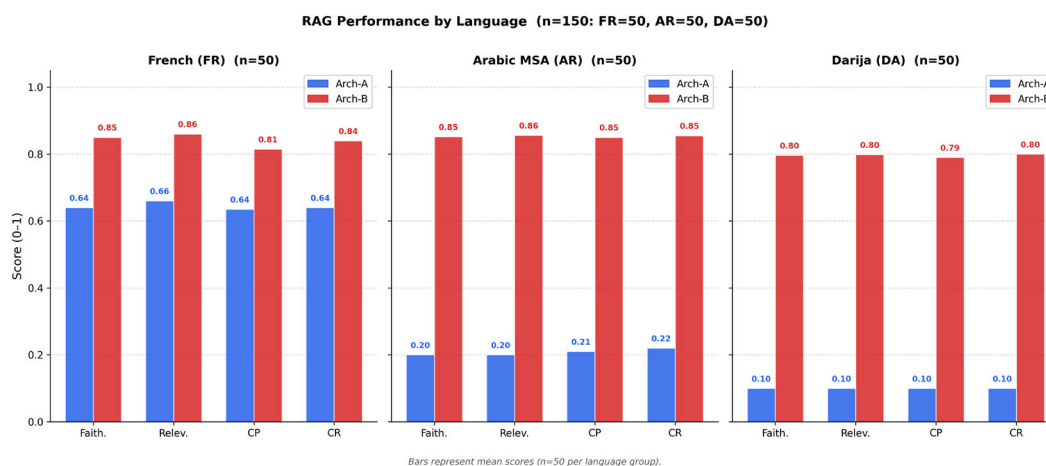
Beyond the automated RAGAS evaluation, we conducted a limited human-in-the-loop validation on a representative subset of multilingual legal queries. According to the methodology described in Section 3.6, four evaluation dimensions were evaluated: Context Relevance, Answer Faithfulness, Answer Relevance and Citation Correctness. The original three-level annotation scores (0 = not satisfied, 1 = partially satisfied, 2 = fully satisfied) were scaled to the interval [0, 1] to facilitate comparison with the RAGAS metrics.

As can be seen in Table 3, Architecture B received a higher score than Architecture A in all four evaluation dimensions. The largest differences were found in Answer Faithfulness, Answer Relevance and Citation Correctness, implying that the LLM-assisted pipeline produced answers that were more grounded in the retrieved legal context and more consistent with the cited legal sources. Overall, the human evaluation confirms the trends in the automatic RAGAS evaluation and further shows the robustness of Architecture B.

**Table 3.** Human-in-the-loop validation results on a representative subset of 45 multilingual legal queries (15 French, 15 Arabic, and 15 Moroccan Darija). Scores are normalized to the [0, 1] interval from the original three-level annotation scale (0–2).

| Metric               | Arch-A | Arch-B |
|----------------------|--------|--------|
| Context Relevance    | 0.267  | 0.700  |
| Answer Faithfulness  | 0.267  | 0.711  |
| Answer Relevance     | 0.289  | 0.711  |
| Citation Correctness | 0.267  | 0.711  |
| Mean Human Score     | 0.272  | 0.709  |

Figure 4 presents the multilingual evaluation stratified by French (FR), Modern Standard Arabic (AR), and Moroccan Darija (DA).



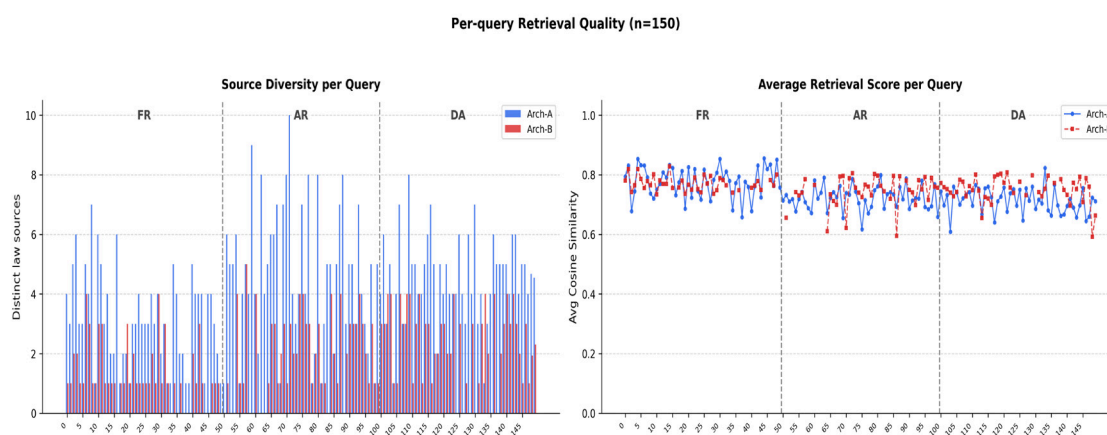
**Figure 4.** RAG performance by language group (FR = French, AR = Modern Standard Arabic, DA = Moroccan Darija) across the expanded benchmark of 150 legal queries (50 per language group). Bars represent mean scores, and error bars indicate  $\pm 1$  standard deviation.

Architecture B consistently performs better than Architecture A across all three language groups. For French legal queries, Architecture B achieves higher scores across all evaluation metrics, indicating that semantic chunking more effectively preserves article-level legal structure in formally organized legislative texts.

For Arabic queries, the performance gap is even larger. Architecture B is able to retrieve legal content with the multimodal Vision fallback, which Architecture A still cannot access. This results in more reliable retrieval and significantly better answer grounding.

Architecture B still has a clear advantage for Moroccan Darija queries. Dialect normalization and semantic chunking effectively improve the retrieval quality as they cope with the linguistic variability in Darija. The two architectures show a greater variability on dialectal inputs than on French queries, but Architecture B is more robust across all evaluation dimensions.

The language-stratified analysis demonstrates that Architecture B's advantage is not confined to a particular language group but is transferable across French, Arabic and Darija legal queries. Figure 5 further displays the retrieval mechanisms behind the observed performance differences.



**Figure 5.** Per-query retrieval quality across the expanded benchmark of 150 legal queries. (Left): Source Diversity per query, measured as the number of distinct legal sources retrieved in the top-k results. (Right): average retrieval cosine similarity per query. Vertical dashed lines separate the French (FR), Arabic MSA (AR), and Darija (DA) subsets.

To further examine the consistency of the human evaluation across languages, Table 4 reports the language-stratified annotation results.

**Table 4.** Language-stratified human-in-the-loop validation results ( $n = 15$  queries per language group). Scores are normalized to the  $[0, 1]$  interval from the original three-level annotation scale (0–2).

| Metric               | FR–A  | FR–B  | AR–A  | AR–B  | DA–A  | DA–B  |
|----------------------|-------|-------|-------|-------|-------|-------|
| Context Relevance    | 0.467 | 0.767 | 0.267 | 0.734 | 0.067 | 0.600 |
| Answer Faithfulness  | 0.467 | 0.800 | 0.267 | 0.734 | 0.067 | 0.600 |
| Answer Relevance     | 0.534 | 0.800 | 0.267 | 0.734 | 0.067 | 0.600 |
| Citation Correctness | 0.467 | 0.800 | 0.267 | 0.734 | 0.067 | 0.600 |
| Mean Human Score     | 0.484 | 0.792 | 0.267 | 0.734 | 0.067 | 0.600 |

Moreover, the language-stratified human evaluation further validates the robustness of Architecture B across all three language groups. For the queries in French, Arabic and Moroccan Darija, Architecture B always scores higher than Architecture A for all the evaluation dimensions.

The largest relative performance gap is observed for Moroccan Darija queries, which constitute the most difficult retrieval setting in terms of dialectal variation, code-switching, and the absence of standardized orthographic conventions. This observation is in line with the quantitative results of RAGAS and confirms that the combination of dialect normalization, semantic chunking and multimodal processing of documents improves the reliability of multilingual legal retrieval.

Since the human validation was performed on a subset of 15 queries per language group, these language-specific observations should be interpreted as qualitative evidence supporting the automated evaluation and not statistically conclusive comparisons.

Overall, the human evaluation corroborates the results of the automated assessment and shows that the performance gains of Architecture B generalize across the three languages represented in the corpus.

The Source Diversity analysis displays that within the three language groups, Architecture A gains information from more diverse legal sources than Architecture B. This confirms the global results shown in Table 2, where Architecture A obtains higher Source Diversity.

However, for most queries, the average retrieval cosine similarity is still quite close between the two architectures. This observation implies that downstream RAG performance cannot be explained by semantic similarity alone. The structural coherence and the legal completeness of the retrieved chunks seem to be equally important in legal retrieval.

Overall, Architecture A is biased towards broader retrieval coverage, while Architecture B is biased towards more focused, citation-ready retrieval. This retrieval behavior is consistent with Architecture B obtaining higher Faithfulness, Answer Relevancy, Context Precision and Context Recall.

## 5. Discussion

The main contribution of this study is that Architecture B performs better on all four RAGAS metrics. Architecture B improves Faithfulness, Answer Relevancy, Context Precision and Context Recall significantly over Architecture A. Among these metrics, precision shows the most significant improvement ( $\Delta = +0.503$ ), i.e., the LLM-assisted pipeline retrieves more focused and legally coherent evidence compared to the regex-based architecture. This finding is consistent with the observations of Ferraris et al. [10], who pointed out the importance of document segmentation for the retrieval quality of legal RAG systems.

A plausible explanation for this behavior can be found in the way the two architectures build retrieval units. Architecture A uses fixed-size fallback fragments, which may cross article boundaries when the regex boundary detection fails, especially in corrupted Arabic PDF documents. Therefore, the chunks retrieved may contain partially relevant information originating from several legal provisions. Architecture B, on the other hand, leverages LLM-assisted segmentation and multimodal document processing to generate more coherent article-level chunks, which leads to cleaner retrieval units and higher retrieval precision.

These findings, in addition to the specific implementation of CollectivIA, highlight the importance of the quality of retrieval units in multilingual legal RAG systems. However, the architectures being compared differ in several components such as the extraction, chunking, embedding and storage mechanisms. Hence, the observed performance differences should be interpreted at the full pipeline level, rather than only at the chunking level.

The inverse behavior of Source Diversity is also consistent with this interpretation. Architecture A retrieves evidence from a wider range of legal texts, while Architecture B retrieves evidence from fewer but more focused legal sources, consistent with its higher Context Precision (0.818 vs. 0.315). This wider retrieval profile might increase the coverage of context but also introduces a higher ratio of less relevant or partially related legal provisions, thus decreasing the retrieval precision.

The language-stratified evaluation shows the specific difficulties of retrieving Moroccan Darija. Both architectures are more variable on dialectal queries than on French queries. However, Architecture B always outperforms on all metrics of evaluation. These results indicate that the combination of dialect normalization, semantic chunking and multimodal document processing achieves more robust retrieval under dialectal linguistic variation. The results are further reinforced by the human-in-the-loop validation, which shows that Architecture B outperforms across all three language groups.

The results, divided by language, indicate that query normalization is still a significant component for multilingual legal retrieval. Specifically, we demonstrate that the combination of Darija normalization and semantic chunking contributes to improving the robustness of retrieval on dialectal queries. These results are consistent with findings by Wu et al. [26] on cross-lingual retrieval, who found that query reformulation can improve retrieval quality by reducing the linguistic variability between user queries and indexed documents.

While normalization increases the robustness of retrieval, it may at the same time sometimes diminish the dialect-specific nuances by rephrasing colloquial expressions in more standard legal language. Such a limitation of LLM-based normalization methods can impact the recovery of highly localized or region-specific dialectal expressions. Therefore, future work should explore normalization strategies that retain the benefits of query reformulation while preserving linguistic specificity.

The variations observed across language groups beyond retrieval performance pose larger questions of algorithmic fairness and access to public information. Retrieval failures for dialectal users may endanger equal access to legal information and public services in multilingual administrative settings. The performance improvement achieved with Darija normalization and multimodal document retrieval can therefore be interpreted not only as technical improvements but also as mechanisms for overcoming the linguistic barriers between citizens and administrative information. While this work does not directly focus on fairness metrics, the results show that robust multilingual retrieval can contribute to reducing digital inequalities in access to legal knowledge.

The agreement between the automated RAGAS evaluation and the human-in-the-loop validation supports the robustness of the observed results. Architecture B also

achieved better performance on the Context Relevance, Answer Faithfulness, Answer Relevance and Citation Correctness, indicating that the performance improvements in the quantitative evaluation are also reflected in the manual assessment.

Regarding evaluation metrics, the results demonstrate that Context Precision and Faithfulness are especially useful metrics for legal RAG systems, as they directly assess the quality of the retrieved legal evidence and its consistency with the generated answers. Answer Relevancy is still useful for the task of measuring the relationship between a question and its response but should be taken into account with retrieval-oriented metrics and human evaluation results when evaluating legal question-answering systems.

Three limitations warrant attention for future work.

#### **Darija coverage**

The current dialectal lexicon of 45 tokens has not been validated yet on regional variations from the Rif, Atlas or Souss-Massa regions, which are known to differ substantially [27]. The normalization module captures mainly the common urban Moroccan Darija. Geographic dialectal variation could be a bottleneck in retrieval and requires further study using larger and more diverse dialectal datasets.

#### **Sample size and benchmark availability**

The revised evaluation benchmark consists of 150 legal queries, significantly improving statistical robustness over the initial pilot evaluation, though the absence of a standardized Moroccan legal question-answering benchmark remains a limitation. We suggest future work to build a larger publicly available benchmark that allows for classical information retrieval metrics such as Recall@k, MRR and nDCG while enabling reproducible comparisons across legal RAG systems.

#### **Digital sovereignty**

Currently, Architecture B depends on external cloud-based services for chunking, embedding and generation. This reliance raises important questions on digital sovereignty in the context of the Moroccan strategy, Maroc Digital 2030 [42] and the recommendations of the CNDH [43] on the deployment of trustworthy AI. The future work should therefore focus on open-source options that can be deployed locally, such as models aligned with the Morocco–Mistral AI Memorandum of Understanding [44], to allow deployment at the institutional scale, while preserving data governance and operational autonomy.

## **6. Conclusions**

This paper introduced CollectivIA, a multilingual Arabic–French Retrieval-Augmented Generation (RAG) system for Moroccan territorial governance law, and systematically compared two retrieval pipeline architectures on a larger benchmark of 150 multilingual legal queries. The results demonstrate that the LLM-assisted pipeline (Architecture B) consistently outperforms the regex-based pipeline (Architecture A) across all four RAGAS metrics (Faithfulness, Answer Relevancy, Context Precision, and Context Recall). The most remarkable improvement was observed for Context Precision ( $\Delta = +0.503$ ), which demonstrates that Architecture B retrieves more focused and legally coherent evidence, leading to more reliable legal question answering.

The language-stratified evaluation revealed that the performance improvements of Architecture B generalize across queries in French, Modern Standard Arabic and Moroccan Darija. In particular, the combination of semantic chunking, multimodal document processing and dialect normalization turned out to be effective for dealing with heterogeneous legal documents and multilingual user interactions. These results were also validated by a human-in-the-loop validation showing that Architecture B was consistently preferred in terms of Context Relevance, Answer Faithfulness, Answer Relevance and Citation Correctness.

More broadly, the work emphasizes the importance of retrieval-unit quality and robustness of document processing in multilingual legal RAG systems. However, since the architectures we compare differ in more than one component, the difference in performance that we observe should be understood at the level of the full pipeline and not of a single architectural component.

Future work should focus on the construction of a standardized Moroccan legal question-answering benchmark supporting classical information retrieval metrics such as Recall@k, MRR, and nDCG. Additional research is also needed to improve coverage of regional Darija variations and to investigate locally deployable open-source alternatives that align with Morocco's digital sovereignty objectives.

**Author Contributions:** Conceptualization, F.Z., O.E.B. and J.R.; methodology, F.Z.; software, F.Z.; validation, F.Z., O.E.B. and J.R.; formal analysis, F.Z.; investigation, F.Z.; resources, F.Z.; data curation, F.Z.; writing—original draft preparation, F.Z.; writing—review and editing, F.Z., O.E.B. and J.R.; visualization, F.Z.; supervision, O.E.B. and J.R.; project administration, O.E.B. and J.R. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The 16 official Moroccan legislative texts forming the CollectivIA knowledge base are publicly available from the Secrétariat Général du Gouvernement of the Kingdom of Morocco at <https://www.sgg.gov.ma> (accessed on 17 June 2026). The evaluation query sets are available upon request from the corresponding author.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Appendix A. System Prompts

Every prompt is taken directly from the CollectivIA system (Architecture B) that has been deployed. Each five-page PDF batch is sent to Gemini 2.5 Flash with the chunking prompt. When Moroccan Darija is identified, the normalization prompt is used. For French and Arabic outputs, respectively, the generation prompts impose citation restrictions.

### *Appendix A.1. Chunking Prompt Structure (Gemini 2.5 Flash)*

The chunking prompt instructs the model to: (1) identify article boundaries from typographic and semantic cues covering French (Article N), Arabic (AlMadda N, AlFasl N), and the 2011 Constitution reversed format; (2) return exclusively a valid JSON array without preamble or explanation; (3) copy article content verbatim without summarization; (4) return an empty array for pages without articles such as tables of contents or annexes; (5) detect and label the language of each article as fr or ar. The JSON output schema specifies four fields: article\_numero (string, exact as written), article\_titre (string, empty if absent), contenu (full verbatim article text), and langue (fr or ar). For corrupted PDFs where the garbage character ratio exceeds 4 percent, pages are rasterized to PNG at 2× resolution and submitted via the Gemini multimodal Vision API with an identical prompt augmented by visual layout instructions.

### *Appendix A.2. Darija Normalization Prompt Structure*

The normalization prompt requests reformulation of the detected Darija query into Modern Standard Arabic appropriate for a legal information search about Moroccan administrative law, returning only the reformulated question without explanation or pream-

ble. The original dialectal form is preserved separately for response language detection. Darija detection uses three complementary signals: (i) Arabic character ratio exceeding 0.20; (ii) presence of tokens from a 45-item dialectal lexicon covering common Moroccan Darija vocabulary; and (iii) Arabizi character-digit mixing patterns such as the use of digits to represent Arabic letters.

### Appendix A.3. Generation Prompt Structure (French and Arabic)

Both French and Arabic generation prompts define CollectivIA as a specialized legal AI assistant for Moroccan administrative law with a scope strictly limited to territorial governance, local public finance, and general administrative law. The prompts enforce three citation rules: (i) cite only law names appearing verbatim in the retrieved source blocks labeled SOURCE 1 through SOURCE k; (ii) never invent or rephrase a law name or article number; (iii) if the relevant provision is not found in the retrieved context, explicitly state this without fabricating a citation. The prompts additionally instruct the model to propose two to three relevant follow-up questions and to maintain a formal administrative register appropriate to the Moroccan public administration context.

## References

1. Dahl, M.; Magesh, V.; Suzgun, M.; Ho, D.E. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *J. Leg. Anal.* **2024**, *16*, 64–93. [CrossRef]
2. Magesh, V.; Surani, F.; Dahl, M.; Suzgun, M.; Manning, C.D.; Ho, D.E. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. *J. Empir. Leg. Stud.* **2025**, *22*, 216–242. [CrossRef]
3. La Constitution Du Royaume Du Maroc—2011. Available online: <https://www.chambredesrepresentants.ma/fr/la-constitution-du-royaume-du-maroc-2011> (accessed on 30 May 2026).
4. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020*; Curran Associates Inc.: Red Hook, NY, USA, 2020; pp. 9459–9474.
5. Hendrycks, D.; Burns, C.; Chen, A.; Ball, S. CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. *arXiv* **2021**, arXiv:2103.06268.
6. Hou, A.B.; Weller, O.; Qin, G.; Yang, E.; Lawrie, D.; Holzenberger, N.; Blair-Stanek, A.; Van Durme, B. CLERC: A Dataset for U. S. Legal Case Retrieval and Retrieval-Augmented Analysis Generation. In *Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, NM, USA, 29 April–4 May 2025*; Chiruzzo, L., Ritter, A., Wang, L., Eds.; Association for Computational Linguistics: Albuquerque, NM, USA, 2025; pp. 7913–7928.
7. Guha, N.; Nyarko, J.; Ho, D.E.; Ré, C.; Chilton, A.; Narayana, A.; Chohlas-Wood, A.; Peters, A.; Waldon, B.; Rockmore, D.N.; et al. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. *arXiv* **2023**, arXiv:2308.11462.
8. Hindi, M.; Mohammed, L.; Maaz, O.; Alwarafy, A. Enhancing the Precision and Interpretability of Retrieval-Augmented Generation (RAG) in Legal Technology: A Survey. *IEEE Access* **2025**, *13*, 46171–46189. [CrossRef]
9. Pipitone, N.; Alami, G.H. LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain. *arXiv* **2024**, arXiv:2408.10343.
10. Ferraris, A.F.; Audrito, D.; Siragusa, G.; Piovano, A. Legal Chunking: Evaluating Methods for Effective Legal Text Retrieval. In *Legal Knowledge and Information Systems*; IOS Press: Amsterdam, The Netherlands, 2024; pp. 275–281.
11. Shi, F.; Chen, X.; Misra, K.; Scales, N.; Dohan, D.; Chi, E.; Schärli, N.; Zhou, D. Large Language Models Can Be Easily Distracted by Irrelevant Context. *arXiv* **2023**, arXiv:2302.00093.
12. Darwish, K.; Habash, N.; Abbas, M.; Al-Khalifa, H.; Al-Natsheh, H.T.; El-Beltagy, S.R.; Bouamor, H.; Bouzoubaa, K.; Cavalli-Sforza, V.; El-Hajj, W.; et al. A Panoramic Survey of Natural Language Processing in the Arab World. *arXiv* **2021**, arXiv:2011.12631.
13. Shahul, E.; Jithin, J.; Espinosa Anke, L.; Schockaert, S. 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, St. Julians, Malta, 21–22 March 2024*; Aletras, N., De Clercq, O., Eds.; Association for Computational Linguistics: St. Julians, Malta, 2024; pp. 150–158.

14. Keisha, F.; Singh, P.; Pallavi; Fernandes, D.; Manivannan, A.; Wicaksono, I.; Ahmad, F.; Rim, W.B. All for Law and Law for All: Adaptive RAG Pipeline for Legal Research. *arXiv* **2025**, arXiv:2508.13107.
15. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 3–7 November 2019; Inui, K., Jiang, J., Ng, V., Wan, X., Eds.; Association for Computational Linguistics: Hong Kong, China, 2019; pp. 3982–3992.
16. Wiratunga, N.; Abeyratne, R.; Jayawardena, L.; Martin, K.; Massie, S.; Nkisi-Orji, I.; Weerasinghe, R.; Liret, A.; Fleisch, B. CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering. In *Case-Based Reasoning Research and Development, Proceedings of the 33rd International Conference, ICCBR 2025, Biarritz, France, 30 June–3 July 2025*; Recio-Garcia, J.A., Orozco-del-Castillo, M.G., Bridge, D., Eds.; Springer Nature: Cham, Switzerland, 2024; pp. 445–460.
17. Han, J.; Burgess, P.; Shareghi, E. Evaluating LLM-Based Approaches to Legal Citation Prediction: Domain-Specific Pre-Training, Fine-Tuning, or RAG? A Benchmark and an Australian Law Case Study. *arXiv* **2025**, arXiv:2412.06272.
18. Brown, A.; Roman, M.; Devereux, B. A Systematic Literature Review of Retrieval-Augmented Generation: Techniques, Metrics, and Challenges. *Big Data Cogn. Comput.* **2025**, *9*, 320. [[CrossRef](#)]
19. Ni'Mah, I.; Aini, L.R.; Fajri, R.; Pebiana, S.; Hidayati, N.N.; Wijayanti, R. Evaluating Retrieval Augmented Generation (RAG) Chunking Strategy for Question Answering in Indonesian Law of The Sea. In *Proceedings of the 2025 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, Jakarta, Indonesia, 15–16 October 2025; pp. 190–195.
20. Ajay Mukund, S.; Easwarakumar, K.S. Optimizing Legal Text Summarization Through Dynamic Retrieval-Augmented Generation and Domain-Specific Adaptation. *Symmetry* **2025**, *17*, 633. [[CrossRef](#)]
21. Reuter, M.; Lingenberg, T.; Liepina, R.; Lagioia, F.; Lippi, M.; Sartor, G.; Passerini, A.; Sayin, B. Towards Reliable Retrieval in RAG Systems for Large Legal Datasets. In *Proceedings of the Natural Legal Language Processing Workshop 2025, Suzhou, China, 8 November 2025*; Aletras, N., Chalkidis, I., Barrett, L., Goanță, C., Preoțiuc-Pietro, D., Spanakis, G., Eds.; Association for Computational Linguistics: Suzhou, China, 2025; pp. 17–30.
22. Wang, X.; Wang, Z.; Gao, X.; Zhang, F.; Wu, Y.; Xu, Z.; Shi, T.; Wang, Z.; Li, S.; Qian, Q.; et al. Searching for Best Practices in Retrieval-Augmented Generation. *arXiv* **2024**, arXiv:2407.01219.
23. De Moura Júnior, P.R.; Lelong, J.; Blangero, A. Adaptive Chunking: Optimizing Chunking-Method Selection for RAG. *arXiv* **2026**, arXiv:2603.25333.
24. Contextual Retrieval in AI Systems. Available online: <https://www.anthropic.com/engineering/contextual-retrieval> (accessed on 31 May 2026).
25. El-Beltagy, S.R.; Abdallah, M.A. Exploring Retrieval Augmented Generation in Arabic. *Procedia Comput. Sci.* **2024**, *244*, 296–307. [[CrossRef](#)]
26. Wu, S.; Tang, J.; Yang, B.; Wang, A.; Jia, K.; Yu, J.; Yao, J.; Su, J. Not All Languages Are Equal: Insights into Multilingual Retrieval-Augmented Generation. *arXiv* **2024**, arXiv:2410.21970.
27. Samih, Y.; Maier, W. An Arabic-Moroccan Darija Code-Switched Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, 23–28 May 2016; Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., et al., Eds.; European Language Resources Association (ELRA): Portorož, Slovenia, 2016; pp. 4170–4175.
28. Open Arabic LLM Leaderboard—A Hugging Face Space by OALL. Available online: <https://huggingface.co/spaces/OALL/Open-Arabic-LLM-Leaderboard> (accessed on 1 June 2026).
29. Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; Liu, Z. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 11–16 August 2024*; Ku, L.-W., Martins, A., Srikumar, V., Eds.; Association for Computational Linguistics: Bangkok, Thailand, 2024; pp. 2318–2335.
30. Le SGG. Available online: <https://www.sgg.gov.ma/LeSGG.aspx> (accessed on 1 June 2026).
31. PyMuPDF Documentation. Available online: <https://pymupdf.readthedocs.io/en/latest/> (accessed on 1 June 2026).
32. UAX #15: Unicode Normalization Forms. Available online: <https://www.unicode.org/reports/tr15/> (accessed on 1 June 2026).
33. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
34. Johnson, J.; Douze, M.; Jégou, H. Billion-Scale Similarity Search with GPUs. *arXiv* **2017**, arXiv:1702.08734.
35. Comanici, G.; Bieber, E.; Schaeckermann, M.; Pasupat, I.; Sachdeva, N.; Dhillion, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv* **2025**, arXiv:2507.06261.
36. Gemini Embedding Now Generally Available in the Gemini API-Google Developers Blog. Available online: <https://developers.googleblog.com/gemini-embedding-available-gemini-api/> (accessed on 1 June 2026).

37. Kusupati, A.; Bhatt, G.; Rege, A.; Wallingford, M.; Sinha, A.; Ramanujan, V.; Howard-Snyder, W.; Chen, K.; Kakade, S.; Jain, P.; et al. Matryoshka Representation Learning. In *Advances in Neural Information Processing Systems*; Curran Associates Inc: Red Hook, NY, USA, 2022; Volume 35, pp. 20246–20258.
38. Chroma—Open-Source Search Infrastructure for AI. Available online: <https://www.trychroma.com> (accessed on 1 June 2026).
39. Malkov, Y.A.; Yashunin, D.A. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 824–836. [[CrossRef](#)] [[PubMed](#)]
40. Saad-Falcon, J.; Khattab, O.; Potts, C.; Zaharia, M. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico City, Mexico, 16–21 June 2024; Duh, K., Gomez, H., Bethard, S., Eds.; Association for Computational Linguistics: Mexico City, Mexico, 2024; pp. 338–354.
41. Gao, T.; Yen, H.; Yu, J.; Chen, D. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023*; Bouamor, H., Pino, J., Bali, K., Eds.; Association for Computational Linguistics: Singapore, 2023; pp. 6465–6488.
42. “Digital Morocco 2030” | Ministère de la Transition Numérique et de la Réforme de l’Administration. Available online: <https://www.mmmsp.gov.ma/fr/actualites/cliquez-ici-pour-d%C3%A9couvrir-la-strat%C3%A9gie-digital-morocco-2030> (accessed on 1 June 2026).
43. Rapport sur la Protection des Droits de l’Homme dans le Monde Numérique et les Systèmes d’Intelligence Artificielle. Available online: <https://cndh.ma/fr/rapport-sur-la-protection-des-droits-de-lhomme-dans-le-monde-numerique-et-les-systemes> (accessed on 1 June 2026).
44. Le MTNRA et Mistral AI Unissent Leurs Efforts Pour l’innovation Digitale | Ministère de la Transition Numérique et de la Réforme de l’Administration. Available online: <https://www.mmmsp.gov.ma/fr/actualites/le-mtnra-et-mistral-ai-unissent-leurs-efforts-pour-l%E2%80%99innovation-digitale> (accessed on 1 June 2026).

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.