

A Hybrid System with Automatic Classification and RAG for Moroccan Court Rulings

Sara IBRIHICH¹ , Ahmed OUSSOUS² , Ouafaa IBRIHICH³ , Mustapha ESGHIR⁴ 

Faculty of Sciences, Mohammed V University in Rabat, Morocco¹

Faculty of Sciences and Technologies of Mohammedia (FSTM), Hassan II University, Morocco²

Mohammed V University in Rabat, Morocco^{3,4}

Abstract—The Accessibility of Moroccan court judgments and decisions is significantly limited due to the lack of a specialized Arabic Judicial information retrieval architecture. Current NLP technologies focus on Modern Standard Arabic and fail to capture the specialized procedural language of judicial rulings from the courts, resulting in inadequate search and query answering assistance for legal professionals. This study introduces AraLex-RAG, a system that combines document extraction, automated categorization, and Retrieval-Augmented Generation (RAG) for Moroccan Arabic court texts. Neither the retrieval nor the classification components need a GPU; however, the RAG module has the possibility to use an external LLM API for generation. This study presents AraLex-RAG, a benchmark dataset of 2855 unstructured Moroccan judicial judgments in 26 different legal classes. It has been created using an automated workflow that combines specific Arabic text normalization, optical character recognition, and balanced class sampling. The information retrieval system is a combination of similarity-based semantic search and keyword-based search, and then uses Reciprocal Rank Fusion to fuse their results. The document categorization is based on a character-level text analysis approach mixed with a linear classifier, while the question-answering feature creates Arabic replies based on retrieved court rulings using a broad language structure. An accuracy of 92.87 percent is achieved (std 1.33 percent) and 76.15 percent macro F1 across 26 categories using five-fold stratified cross-validation on 2853 documents, outperforming the previous best on Moroccan court judgments (80.51 percent on a six-class subset) on a challenge 4 times tougher. Using 3,936 document query pairings, dense semantic retrieval performs better than the search-based keyword in all four assessment measures. On 45-stratified pairs of questions and answers, the query-answering component obtains a cosine similarity value of 0.845, which is similar to the most advanced Arabic query-answering systems on structured fields.

Keywords—Legal information retrieval; BM25; FAISS; Arabic NLP; TF-IDF; text classification; LinearSVC; Moroccan legal judgments

I. INTRODUCTION

Legal information retrieval (LIR) involves identifying pertinent statutes, law cases, and court rulings in answer to a legal search query [22]. Efficient information extraction systems significantly diminish the time of researching and enhance the uniformity of court judgments by guaranteeing that pertinent legal precedents are not ignored or overlooked. LIR systems have attained stability and maturity for the English language via platforms such as LexisNexis and

Westlaw [21], as well as studies like COLIEE [1], and even Chinese language is starting to gain focus and interest [2]. On the other hand, the Arabic legal corpora lack similar architecture. This difference and void are most pronounced in Morocco, where court rulings are articulated in a specialized form of formal and official legal Arabic that draws on traditional structures, field-specific lexicon, and administrative norms, which are not well represented by broader-purpose Arabic NLP tools.

Arabic has significant NLP obstacles: its morphologically complex architecture generates several forms of words from the same root; diacritic structures that clarify context are often absent in written text; and right-to-left directional display imposes software limitations. In the Moroccan Judicial field, these difficulties are aggravated by the lack of any annotated legal corpus, the particularity of Moroccan procedure-based Arabic, and the extraction of most court judgments from paper archives that are not available in a machine-readable format. Researchers are unable to expand on previous studies or methodically assess methodologies since the field does not presently provide a standard for Moroccan court judgments, a retrieval mechanism, or a validated question-answering pipeline. This study concurrently covers all three different unexplored areas.

This study is organized as follows: 1) AraLex-RAG: One of the golden standard corpora for the Moroccan Arabic juridical information extraction system, consisting of 2855 unstructured court judgments and rulings spanning 26 legal areas. Without a common dataset, reliable studies on Moroccan judicial Artificial intelligence are unattainable; this corpus establishes the basis for any future endeavors in the field. 2) Prove that deep learning is not as effective as traditional approaches when it comes to limited resources for Arabic legal categorization. Our classifiers overcome the preceding best on Moroccan judicial judgments, which only tackled six categories, while LinearSVC and TF-IDF obtain state-of-the-art accuracy over 26 categories. 3) The quantified extraction assessment on Moroccan Arabic court rulings demonstrates that intensive semantic search consistently surpasses keyword-based retrieval in this field. Level category research indicates that the difficulty of retrieval is associated with file similarity within a class, a discovery that informs further corpora development. 4) An exhaustive query-responding system assessed on Arabic legal text, using a major language approach based on recovered court rulings. Outcomes are similar to leading-edge Arabic query-responding systems in structured

areas, indicating that outstanding retrieval systems may be attained using freely available open-source frameworks. 5) A comprehensive framework that integrates OCR, information retrieval, categorization, and query answering into a single program operable on normal hardware without the need for a GPU. AraLex-RAG is a full tool for Moroccan legal experts and specialists.

The subsequent sections of this work are structured as follows. Section II examines pertinent background and related studies on international juridical data extraction, natural language processing in Arabic legal, current Arabic legal baseline, hybrid extraction techniques, and extracted augmented generation for legal query answering. Section III outlines the dataset creation process, system design, and assessment procedure. Section IV defines quantitative outcomes for categorization, extraction, and query responding. Section V examines the results, limits, and real-world consequences for legal experts in Morocco. Section VI concludes and highlights avenues for further research.

II. RELATED WORK

A. International Legal Information Retrieval Landscape

Judicial information retrieval has attained significant development in various highly resourced languages, establishing robust baselines and offering valuable insights for developing fields like Moroccan Arabic. The COLIEE sharing task sets [1] are the foremost English legal information retrieval reference, necessitating technologies to extract pertinent law cases and execute legal entailment logic. The highest-performing systems of 2023 integrate BM25 for preliminary candidate retrieval with using BERT-based neural re-ranking [1], illustrating the continued significance of hybrid linguistic semantic methodologies despite the prevalence of huge pre-trained systems. LEGAL-BERT [2], a specialized variant of BERT trained on judicial rulings, legislation, and legal agreements, demonstrates that pre-training language models on domain-specific legal content consistently outperforms generic-purpose systems. In a Recent study, Santosh and Upadhyay [3] established a cross-lingual section extraction standard utilizing European court rulings on human rights and discovered that multilingual models that have been pretrained continually perform poorly compared to monolingual peers in judicial documents, underscoring that linguistic scope compromises field depth.

In French, Authors [4] developed a binary classification approach to identify the existence of a cost allocation clause referred to as Article 700 in 1994 decisions from the French court of appeals. By combining TF-IDF vectorization with conventional classifiers, they surpass the domain-specific Judicial CamemBERT transformer's accuracy of 0.96 on targeted judgment sections. This discovery is enlightening: for well-specified classification assignments on organized law data sets, frequency-based approaches may rival transformer-based model designs, a conclusion that directly and profoundly impacts the architecture of AraLex's categorization unit.

The CAIL2018 dataset [5] in China, including 2.6 million criminal case judgments, has established itself as the standard norm for charge prediction and sentencing quantification.

Traditional SVM classification methods attain 94 % accuracy in charge prediction, surpassing most neural methodologies on this extensive, well balanced data. The translation of these results into Arabic, where corpora are much smaller and class borders are semantically and linguistically ambiguous, demands rigorous methodological adjustment.

B. Natural Language Processing in Moroccan and Arabic Legal Field

Zahir et al. [6] constructed an initial corpus of Court of Cassation judgments and trained CNN and GloVe embedding models on a six-category judgment prediction task, achieving an accuracy of 80.51%. The effort is authentic and beneficial, although the scope is limited: six general categories, just categorization, without data extraction or quality assurance. Word-level embedding techniques encounter difficulties with Arabic syntax when the training data is insufficient. Another study [7] indicates superior performance (94.58%, F1 = 94.45%) on a seven-class test using TF-IDF with SVM, specifically with legislative codes, documents from the government characterized by uniform formatting rather than court rulings. Court rulings are more complex and challenging. Neither research establishes a public baseline nor assesses extraction.

Arabic typology poses a further difficulty: extensive suffixation, prefixation, and cliticization result in significant surface form diversity, making vocabulary matching risky without exhaustive planning [20]. Systems pre-trained on Egyptian or Gulf Arabic have challenges when adapting to the Moroccan judicial records, which integrate French civil law terminology with Maliki jurisprudence in manners that generic embeddings fail to adequately represent.

C. Legal Arabic Standard: Moroccan Gap and Saudi Focus

The most advanced Arabic legal AI standards focus on Saudi law. ArabLegalEval [8] evaluates LLMs using multiple-choice based and open-ended inquiries, according on Saudi Ministry of Justice laws and Board of Experts norms. Even GPT-4o attains just around 77% on zero-shot multiple-choice question tasks, serving as a pertinent reminder that robust general approaches continue to have difficulties with domain-specific judicial Arabic. However, these benchmarks assess a model's pre-existing knowledge; they are not retrieval techniques and were not intended to locate content within a corpus. They also address Saudi law, which functions under a distinct legal system compared to Morocco. Moroccan law integrates French civil law with Maliki jurisprudence, resulting in a juridical Arabic that diverges from Gulf dialects in terminology, syntax, and procedural framework. There is no baseline or retrieval mechanism for Moroccan court rulings.

D. Reciprocal Rank Fusion and Hybrid Extraction

BM25 and dense retrieval provide distinct advantages, a fact that is well-known. BM25 is dependable for exact matching, particularly article numerals, party nomenclature, and procedural terminology. Dense retrieval, as shown by Karpukhin et al. [9] using Dense Passage Retrieval, identifies relevant documents even when the query and document do not share any vocabulary, by analyzing trained vector representations. In open-domain QA assessments, dense

retrieval demonstrates superior recall at shallow depths, although BM25 performs comparably when keyword matching is paramount.

Reciprocal Rank Fusion [10] integrates ranking lists by assigning a score of $1/(k + \text{rank})$ to each document from each system and aggregating the scores. The constant $k = 60$ prevents the highest-ranked document from dominating when an extraction algorithm exhibits excessive confidence. RRF requires neither training data nor score normalization, offering practical benefits in low-resource environments. Arabic legal inquiries combine precise terminology (case IDs, article numbers) with more general concepts (legal norms, procedure classifications), making the hybrid method particularly suitable.

E. Generation of Legal Questions-Answering Using Retrieval-Augmented Services

Lewis et al. [11] formalized Retrieval-Augmented Generation: the model obtains relevant materials prior to formulating responses, rather than relying only on parametric memory. This is of greater significance in legal applications than in most other fields. A memory-based system may create a seemingly authoritative legal assertion without any supporting documentation [12]. A citation referencing the exact court ruling may be verified. English legal QA systems like LEGAL-BERT [2] demonstrate enhanced response fidelity when employing RAG compared to its absence.

The study by Al-Rasheed et al. [13] presents Arabic retrieval-augmented generation that assesses GPT-3.5, GPT-4, and other specific Arabic versions on a dictionary question-answering challenge. The extracted context is a dictionary entry: concise, well-structured, and immediately relevant. GPT-4 achieves a Token F1 score of 0.90 and a Cosine Similarity of 0.87. While it serves as a valuable reference, dictionary entries provide the most concise form of retrieval settings. Unstructured court judgments are extended, more convoluted, and need the approach to extract pertinent information concealed inside extensive legal text.

III. METHODS AND MATERIALS

A. Dataset

1) *Data collection*: The AraLex-RAG corpus consists of 2855 Moroccan judicial rulings sourced from tribunal archives in both digital (DOCX, PDF) and scanned picture forms. The papers consist only of unstructured, informal Arabic text devoid of markup, annotations, or predefined categories. The document length ranges from 150 to more than 2,000 tokens, with an average of around 420 tokens. Each record is classified into one of 26 judicial classifications, including family, property, civil, administrative law, and labour, as determined by the tribunal department code of the document.

2) *Preprocessing pipeline*: All documents undergo a uniform four-phase (Fig. 1) pipeline, implemented consistently throughout both corpus construction and query execution to guarantee uniformity.

Phase 1 OCR: Scanned picture documents are transformed into text through the application of Tesseract with the Arabic language model (ara).

Phase 2 Quality Filtering: Records producing fewer than 30 characters post-OCR extraction are eliminated as retrieval failures.

Phase 3 Arabic normalization: a regex-based pipeline eliminates diacritics (tashkeel), converts all Alef variations (ا, آ, إ, ئ) to the plain Alef (ا), and removes non-Arabic letters and extraneous whitespace. No morphological analyzer or stemmer is employed; this is an intentional design decision to eliminate reliance on external Arabic NLP toolkits.

Phase 4: Class-balanced sample: stratified sampling yields the final corpus of 2855 documents.

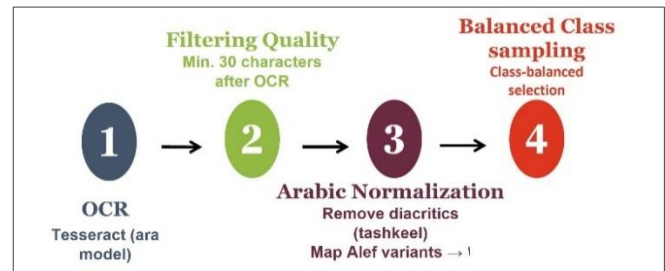


Fig. 1. AraLex-RAG preprocessing pipeline.

TABLE I. STATISTICS FROM THE ARALEX-RAG CORPUS

Property	Value
Total documents	2855
Judicial classifications	26
Document nature	Free-form, unstructured Arabic text
Input formats	PDF, DOCX, scanned image (JPEG/PNG)
Mean document length	~420 tokens
Median documents per category	71
Min / Max per category	1 / 602
Categories with < 30 docs	2 out of 26 (Single document)
Train/test split	80% / 20% (stratified)
Cross-validation	5-fold stratified (K=5, seed=42)
Storage backend	SQLite relational database

1) *Distribution of classes and imbalance*: Table I represents statistical data of the corpus, whereas Table II specifies how it is distributed by class. The dataset exhibits significant imbalance: groups range from a minimum of 1 to a high of 602. Two classifications (1619, 1608) each comprise a single record and are omitted from classifier assessment. The predominant classes, Judicial divorce "التطليق" (602 docs) and النفقة (552 docs), overshadow the corpus and enhance accuracy while diminishing macro F1, as elaborated within the discussion Section.

This disparity is intrinsic to the actual distribution of case categories in Moroccan courts and cannot be rectified via sampling without excluding the already limited minority-class instances. For assessment, an 80/20 stratified division maintains class ratios in both segments; 2853 records that meet quality standards are used for classification and retrieval trials.

TABLE II. ARALEX-RAG SYSTEM DISTRIBUTION OF DOCUMENTS BY CATEGORY

Code	Category	Number
Civil		
1201	Miscellaneous civil المتنوع المدني	86
1202	Tort liability المسؤولية التقصيرية	75
Family		
1601	Civil status amendment الإصحاح والتغيير	161
1602	Birth registration	50
1603	Death registration تسجيل الوفاة	88
1606	Alimony النفقة	552
1607	Judicial divorce التطلاق	602
1608	Unilateral divorce الطلاق	1
1609	Child custody الحضانة	56
1610	Spousal return الرجوع لبيت الزوجية	75
1611	Proof of marriage ثبوت الزوجية	70
1612	Right of family access صلة الرحم	65
1613	Lineage / paternity النسب	80
1614	Family enforcement order التذليل بالصيغة التنفيذية (أسرة)	65
1616	Guardianship of minors نواب القاصرين	55
1617	Guardianship of neglected children كفالة الأطفال المهملين	60
1618	Permission for polygamy الإذن بالتعدد	50
1619	Personal status الأحوال الشخصية	1
1620	Other personal status cases قضايا الأحوال الشخصية الأخرى	62
1623	Legal incapacitation التحجير	71
Labour		
1501	Labour disputes نزاعات الشغل	57
1502	Work accidents حوادث الشغل	90
Real Estate		
1401	Standard real estate العقار العادي	80
1402	Registered real estate العقار المحفوظ	101
1403	Real estate in registration العقار في طور التحفظ	102
Rental		
1303	Rent payment and eviction الأداء والإفراغ	100

B. Overview of System Architecture

AraLex-RAG system consists of five operational components: OCR, retrieval, categorization, generation, and interface, supported by a dual-layer storage backend composed of a SQLite relational database that holds all the textual content and associated information of the documents (fields: id, code, label, filename, filepath, file_number, year, decision_number, clean_text, full_text), alongside an FAISS vector index that retains the dimensional dense vector embeddings. Fig. 2 illustrates the whole design and the interconnections among these aspects.

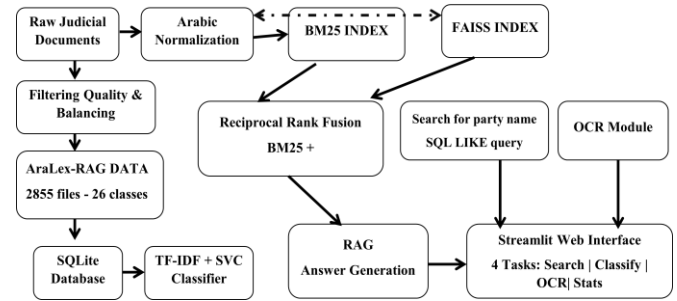


Fig. 2. Overview of AraLex-RAG system architecture.

Corpus indexing constitutes a singular offline procedure. Unstructured files, PDFs, DOCX formats, and scanned pictures are processed using the OCR module (Tesseract utilizing the Arabic language models) and undergo the four-phase normalization process outlined in Section A.2. Every standardized text is thereafter portrayed in two complementary formats: the BM25 engine constructs a token-level inverted index for sparse extraction, whilst the multilingual-e5-small encoder generates a dense embedding preserved in FAISS. Both visualizations come from the same standardized text, guaranteeing that the two retrieval methods function on equivalent input. All textual content, metadata, and embeddings are retained in the storage backend and used for each query.

During inquiry time, the user asks a question using natural language via the interface. The inquiry is subjected to the same normalization as the corpus, thereafter proceeding via a five-stage pipeline shown in Fig. 2. Initially, it is sent concurrently to the BM25 engine and the FAISS index, with BM25 ranking based on word frequency adjusted by inverse document frequency, while FAISS ranking relies on inner product similarity among embeddings. Both ranked classification lists are then integrated by Reciprocal Rank Fusion, which merges the distinct advantages of sparse and dense retrieval without necessitating score adjustment. The categorization module then utilizes the TF-IDF + LinearSVC model to forecast the legal category among the 26 options to which the question applies, hence offering a field labeling that contextualizes the obtained findings. Ultimately, when the generation unit is engaged, the first five extracted files are compiled into a coherent Arabic prompt and presented to LLaMA-3.3-70B via the Groq API, which generates a natural-language response based only on the collected choices. The whole local pipeline, including indexing, data extraction, and categorizing, does not need a GPU; only the RAG phase employs distant inference, allowing the system to remain operable on conventional office equipment.

C. Information Retrieval Algorithms

1) BM25 sparse information retrieval: The BM25 component employs an inverted index based on normalized document tokens. The relevance score is determined by the query $Q = \{q_1, \dots, q_m\}$ and the document D .

$$\text{Score}(D, Q) = \sum_{i=1}^n \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})} \quad (1)$$

Q: The query including terms.

D: This is the document that is being evaluated.

$f(q_i, D)$: Term frequency. The frequency of the query phrase q_i in document D.

$|D|$: Document D length (aggregate word count).

avgdl: Average length of documents throughout the whole corpus.

k_1 : A parameter for term frequency saturation. Augmenting this value amplifies the influence of phrase frequency. It's set here to 1.5 ($k_1 = 1.5$). (Standard Okapi BM25 defaults [14]).

b_1 : The parameter associated with length normalization. It manages the extent to which the length of a document impacts the score. It's set here to 0.75 ($k_1 = 0.75$). (standard Okapi BM25 defaults [14]).

$IDF(q_i)$: Inverse Document Frequency. Terms are weighted according to how rare they are throughout the corpus. Usually, it is calculated as:

$$IDF(q_i) = \ln \left[\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1 \right] \quad (2)$$

In this context, N represents the aggregate count of documents inside the corpus, whereas $n(q_i)$ denotes the quantity of documents that include the phrase q_i .

2) Dense semantic retrieval: FAISS: Dense retrieval utilizes the intfloat/multilingual-e5-small sentence transformer (386 dimensions) [15, 16] to encode both files and queries. The compact version was chosen for resource-efficient implementation: with 2855 documents, the whole FAISS index occupies about 4.3 MB of RAM, allowing sub-second search on conventional hardware without the need for a GPU. Semantic similarity is represented by cosine similarity, which corresponds to the inner product of L2-normalized vectors.

$$\text{sim}(u, v) = \frac{u \cdot v}{\|u\| \cdot \|v\|} = \hat{u} \cdot \hat{v} \quad (3)$$

where,

$u \cdot v$: represents the scalar product of the two vectors.

$\|u\|$ and $\|v\|$: the Euclidean norms (magnitudes) of the respective vectors.

Cosine similarity quantifies the angle between two vector representations inside the embedding space: a score of 1 signifies semantically equivalent texts, while a score of 0 denotes not related content. Normalizing each vector corresponding to unit length (L2 normalization) prior to storage guarantees that document length does not enlarge the score, allowing a concise judgment that is entirely relevant to take precedence over an extensive ruling that just briefly references the same words. This normalization also minimizes inner-product searches in FAISS to a cosine similarity query without any additional calculations at the time of querying.

Document vectors are maintained in an FAISS IndexFlatIP [17,19], which does precise inner-product searches with $O(N \cdot d)$ time complexity, with N being the count of indexed documents and d denoting the dimensionality of the embeddings. The index functions on L2-normalized vectors,

making inner-product search theoretically synonymous with cosine similarity ranking, therefore necessitating no further transformation during query execution.

3) Reciprocal rank fusion: BM25 and dense retrieval generate relevance ratings on dissimilar scales: BM25 ratings are unrestricted integers weighted by phrase frequency, while cosine similarity scores are confined to the range of [0, 1]. Integrating them via a weighted linear sum necessitates manual scale calibration on reserved data, an unfeasible limitation for a corpus that changes as fresh judgments are included. Reciprocal Rank Fusion (RRF) [10] bypasses this issue by ignoring score magnitudes and combining ranked lists only based on their ordinal positions.

$$RRF(d) = \sum_i \frac{1}{k + \text{rank}_i(d)} \quad (4)$$

where, $\text{rank}_i(d)$ is the placement of document d in the ordered list generated by retriever i, and $k = 60$ represents the smoothing constant from the initial formulation [10]. The parameter k controls the rate at which the score diminishes across ranks: with $k = 60$, a document ranked first contributes a score of $1/61 \approx 0.016$, whereas a document at rank 10 contributes $1/70 \approx 0.014$, resulting in a carefully moderated decline that ensures the entire top-10 range is treated almost equally, thus preventing the scenario where a single retriever excessively favors a limited selection of documents with extremely high scores.

In AraLex, after every query, two lists are merged: the BM25 ranked list derived from the inverted index and the FAISS cosine similarity ranked list. For any document included in either list, its RRF score aggregates $1/(60 + \text{rank}_i)$ from each retriever in which it is found; documents not included in a retriever's list do not contribute from that retriever. Documents that get high rankings from both systems, characterized by lexical and semantic relevance, combine contributions from both lists and hence increase to the highest level of the combined ranking. Fig. 3 shows the whole five-stage query process.

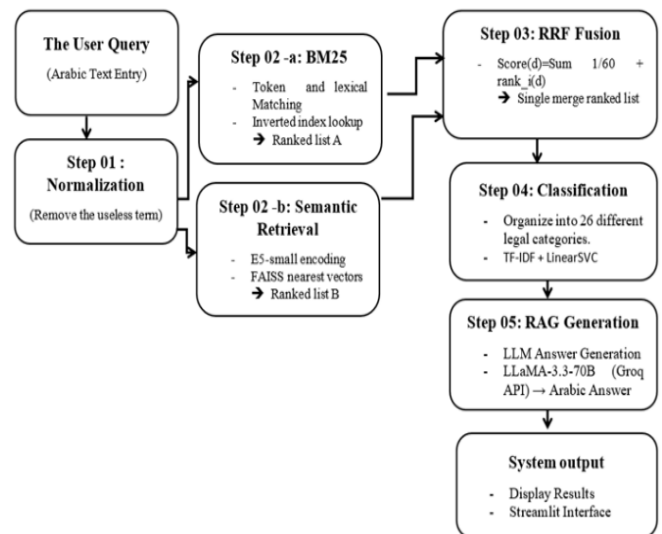


Fig. 3. AraLex-RAG pipeline procedure for query preprocessing.

D. Categorisation Algorithms

Document categorization is executed via a TF-IDF vectorizer in conjunction with a LinearSVC (Linear Support Vector Classifier). The representation of features is constructed using character n-grams instead of words. Arabic has a high degree of morphological complexity: a single root may produce several surface forms by prefixation, suffixation, and cliticization, resulting in word-level characteristics that regard morphological variations as completely distinct. Character n-grams ranging from lengths 2 to 4, meaning $\text{ngram_range}=(2,4)$, automatically identify common substrings across these variations, eliminating the need for a stemmer or morphological analysis. The 'char_wb' analyzer derives n-grams at word borders, augmenting each term with spaces to ensure that sequences of characters at the beginning and end of words are distinctly represented. Every term t within document d is assigned a TF-IDF [18] weight:

$$w(t, d) = tf(t, d) \times \log \left(\frac{N}{df(t)} \right) \quad (5)$$

where, $tf(t, d)$ indicates the occurrence of n-gram t inside document d , N represents the total count of documents in the corpus, and $df(t)$ indicates the quantity of documents that include t . The significance is elevated for n-grams that are prevalent in the present text but infrequent across the corpus, embodying the unique character traits of each legal classification. Each document is thus shown as a sparse vector in a high-dimensional space composed of these weights. LinearSVC was chosen as the classifier due to its exceptional compatibility with high-dimensional sparse feature spaces: it trains significantly faster than kernel-based SVMs while maintaining similar accuracy in text classification tasks. The LinearSVC optimizes the following hinge-loss objective:

$$\min_w \frac{1}{2} \|w\|^2 + C \cdot \sum_i \max(0, 1 - y_i(w^T x_i + b)) \quad (6)$$

where, w represents the weight vector that defines the decision limit, x_i stands for the TF-IDF feature vector for document i , $y_i \in \{-1, +1\}$ signifies its class label, and b is the bias term. The initial component, $\frac{1}{2} \|w\|^2$, serves as the regularization penalty that maintains a broad and generalizable decision boundary. The second component imposes penalties on documents that exceed the margin, with a weight of $C = 1.0$, the normal setting, which is effective in high-dimensional feature spaces when the feature count significantly surpasses the document count. Given that SVM is fundamentally binary, the 26-category issue is partitioned into 26 autonomous one-versus-rest (OvR) sub-problems: each classifier is tasked with differentiating one lawful category from the rest. During the prediction phase, all 26 classifiers operate concurrently, and the category with the greatest decision margin is selected for the document.

E. Retrieval-Augmented Generation

The RAG module bases LLaMA-3.3-70B replies on extracted Moroccan judicial rulings, according to the framework established by Lewis et al. [11]. During the inference phase, the process unfolds in three stages.

Initially, the user's request is analyzed by the hybrid RRF retrieval system, and the leading $k = 5$ docs are chosen. Secondly, these documents are compiled into a systematic Arabic prompt using the following standardized format: 1) a directive instructing the model to react only based on the supplied context and to reply in Arabic; 2) the five retrieved documents enumerated and concatenated; 3) the user's query. Third, the compiled prompt is sent to LLaMA-3.3-70B over the Groq cloud API with a temperature setting of $\tau = 0.1$. The reduced temperature setting decreases linguistic variability in outcomes and is suitable for factual legal inquiries where uniformity is prioritized above inventiveness. If the acquired records lack pertinent information on the inquiry, the model is directed to clearly indicate that the context fails to provide an adequate response, which exemplifies appropriate grounded behaviour and inhibits hallucination. The $k = 5$ context window was selected to strike a compromise between providing enough evidence and maintaining the prompt inside the model's contextual limitations. LLaMA-3.3-70B was chosen over smaller approaches because of its native support for Arabic and its 70 billion parameter scale, which offers enough capability for multi-document legal analysis. The Groq API delivers sub-second inference without the need for local GPU hardware, enabling deployment on conventional office computer systems. LLaMA-3.3-70B is an open-source model, hence avoiding the data privacy issues associated with transmitting sensitive legal documents to private API vendors.

F. Overview of Hyperparameter Settings

Table III summarizes the parameter configurations for each AraLex-RAG component, according to the sequence of the pipeline: sparse retrieval, dense retrieval, rank fusion, classification, and generation. All parameters were either derived from recognized defaults in the referenced literature or chosen according to the limitations of the intended deployment setting.

TABLE III. HYPERPARAMETER CONFIGURATIONS FOR ALL ARALEX-RAG COMPONENTS

Module	Parameter	Setting	Reasoning
BM25	k_1	1.5	Standard Okapi BM25 [14]
	b	0.75	
Dense encoder	Model	multilingual-e5-small	Arabic support and CPU-efficient
	Dim	386	Concise index (<3 MB)
FAISS	Index type	IndexFlatIP	Exact search; $N=2855$
RRF	k	60	Initial formulation [10]
TF-IDF	analyzer	char_wb	Within-word boundary n-grams
	ngram_range	(2, 4)	Morphological affixes in Arabic
LinearSVC	C	1.0	Standard regularization
	strategy	OvR (26 classes)	OvR multi-class
RAG	Model	LLaMA-3.3-70B	Groq API, temperature=0.1
	Context docs	5	Top-k retrieved

G. Evaluation Framework

The efficacy of classification was assessed using five-fold stratified cross-validation ($K=5$, $\text{random_state}=42$) on the 2853 documents categorized with a minimum of two samples. The remaining papers were not included since stratified splitting demands at least one sample per class for each fold, a requirement unattainable by single-instance classes. $K=5$ was chosen as the optimal balance between bias and variance in cross-validation: fewer folds (e.g., $K=3$) provide bigger test sets but increase variation in the estimation, while more folds (e.g., $K=10$) are resource-intensive and provide decreasing benefits on a dataset of this magnitude. Stratification guarantees that the class allocation in each fold mirrors the overall corpus distribution, which is essential due to the distinction within groups, with “التطبيق” providing 602 documents while multiple categories contribute less than 10. In the absence of stratification, a fold may intentionally lack instances of a rare category, causing the accuracy assessment to be unreliable.

The primary measures are mean accuracy and mean macro F1, precision, and recall over the five folds, shown with a standard deviation. Macro averaging allocates equivalent importance to each category, irrespective of their size, hence imposing the same penalties on the classifier for mistakes in both infrequent and common categories, serving as a more insightful metric than weighted F1 in scenarios of class imbalance. The standard deviation across folds measures the stability of results: a limited range indicates that the classifier generalizes reliably instead of overfitting to a specific data subset.

A distinct fixed 80/20 stratified division ($\text{random_state}=42$) was implemented to derive per-class recall, precision, and F1 metrics throughout all 26 categories. Cross-validation reports only total metrics; the reserved subset offers the detailed analysis necessary to figure out which legal classifications the classifier manages well and which need more data acquisition.

Extraction query and document pairings were generated automatically by the use of template-driven Arabic question formulation using the document metadata. For every document, a collection of Arabic queries was formed based on structured elements such as subject, judgment, year, case identifier, and the name of the court, with each template used only for papers that had the necessary field. This automated method removed a demand for human query labeling and generated 3936 query–document pairings inside the 2855-document corpus. The definitive pertinent document for each inquiry is the original document from which the question was derived, providing one affirmative per query.

The assessment of retrieval was undertaken on these 3936 pairings using Mean Reciprocal Rank [23] ($\text{MRR}@10$) and Normalised Discounted Cumulative Gain [24] ($\text{NDCG}@10$):

$$\text{MRR}@10 = \frac{1}{|Q|} \cdot \sum_i \frac{1}{\text{rank}_i} \quad (7)$$

$$\text{NDCG}@10 = \frac{\text{DCG}@10}{\text{IDCG}@10}, \text{DCG}@10 = \sum_n \frac{\text{rel}_n}{\log_2(n+1)} \quad (8)$$

where, $|Q| = 3,936$ is the total number of assessment queries, and rank_i refers to the place of the correct document in the list ranked for query i . A valid document that does not rank within the top 10 adds nothing to the total. $\text{MRR}@10$ consequently favors systems that position the accurate document as near to rank 1 as possible among the top 10 outcomes.

$\text{DCG}@10$ combines pertinence scores marked down by location: rel_n signifies the pertinence of the document at position n (1 if it is the accurate document, 0 if not), while the $\log_2(n+1)$ denominator punishes the value of pertinent documents located in lower positions. $\text{IDCG}@10$ is the optimal DCG, the highest attainable score when the accurate document occupies the top position. With one affirmative response per query, $\text{IDCG}@10$ equals $1/\log_2(2) = 1.0$; thus $\text{NDCG}@10$ simplifies to $1/\log_2(\text{rank}_i + 1)$ when the accurate document is located within the top 10, and 0 in all other cases. Dividing by IDCG aligns the score to the range of $[0, 1]$, facilitating comparability across searches.

Furthermore, Hit at 5 ($\text{Hit}@5$) evaluates the presence of the right document within the top 5 findings (1 for yes, 0 for no), averaged over any query. Precision at 1 ($\text{P}@1$) assesses whether the right document is positioned first (1 for yes, 0 for no), averaged across all of the searches. $\text{MRR}@10$ and $\text{NDCG}@10$ assess the quality of rankings, whereas $\text{Hit}@5$ and $\text{P}@1$ evaluate retrieval coverage and precision at the most stringent threshold. RAG generation underwent evaluation using a stratified sample of 45 question-and-answer pairs, chosen to include various sorts of legal queries. The number 45 was established due to the practical limitation of manual ground-truth retrieval, necessitating cautious examination of each source file to get precise reference responses, which were manually obtained from the relevant source materials, a conventional method in generative QA assessment, whereby human-authored references serve as the norm for comparing system outputs. Generation quality was evaluated using three complementary measures: Token F1 assesses the overlap of words between the generated output and the reference reply; ROUGE-L measures the longest common subsequence [25], rewarding responses that retain the correct order of words; and Cosine Similarity determines semantic similarity through E5 embeddings, ensuring that paraphrases of the correct answer receive high scores even with differing phrasing. All trials were executed on a typical laptop lacking a GPU, indicating that AraLex is operable without dedicated hardware.

IV. RESULTS

A. Classification Outcomes

Table IV displays the outcomes of cross-validation classification. The classifier attains an accuracy of $92.87\% \pm 1.33\%$ and a macro F1 score of $76.15\% \pm 3.89\%$. This difference between accuracy (92.87%) and macro F1 (76.15%) indicates class imbalance: accuracy is influenced by well-supported predominant classes, whereas macro F1 treats all 26 categories roughly the same, resulting in minority classes with fewer test samples adversely affecting the average.

TABLE IV. CLASSIFICATION OUTCOMES: FIVE-FOLD STRATIFIED CROSS-VALIDATION

Metrics	Accuracy	F1-score macro	Precision macro	Recall macro
Mean	92.87%	76.15%	80.27%	75.08%
Std Dev	±1.33%	±3.89%	—	—
Minimum Fold	90.86% (F5)	—	—	—
Maximum Fold	94.20% (F4)	—	—	—

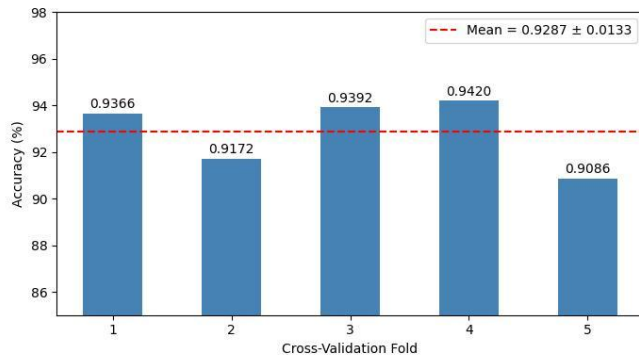


Fig. 4. Fold-specific accuracy throughout five cross-validation folds

Fig. 4 shows a small fold-to-fold range (3.34 percentage points), indicating that the classification algorithm maintains stability and does not exhibit overfitting to any specific data subset. The minimum fold (F5: 90.86%) and the highest (F4: 94.20%) vary by under one standard deviation.

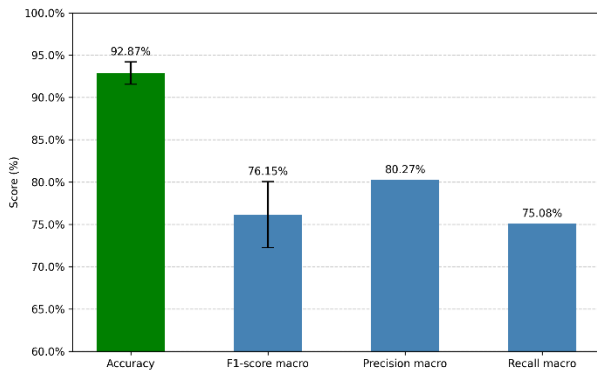


Fig. 5. System classification performance in 26 different legal areas.

Fig. 5 illustrates the four classification measures derived from 5-fold stratified cross-validation. The precision and recall macros are derived from the fixed 80/20 held-out split, whereas the accuracy and F1 macros include standard deviation error bars across folds.

B. Retrieval Outcomes

Table V presents a comparison of BM25-only, semantic-only, and hybrid RRF retrieval over 3936 query-document pairings. Semantic search frequently surpasses BM25 on all four parameters, whereas hybrid RRF attains moderate performance.

The aggregate metric numbers indicate the fundamental challenge of this assessment framework: the 3936 queries use template-driven queries (for instance, “What is the court’s ruling?” means “ما هو حكم المحكمة؟”) utilized across all sources. Due to the prevalence of analogous language across several legal categories, precise document retrieval using a general query is much more challenging than retrieval via document-specific questions.

As shown in Table V and Fig. 6, the semantic retriever provides the most significant relative enhancement compared to the BM25 baseline, suggesting that semantic representations greatly strengthen retrieval efficacy in the adversarial template-query context. Embracing this enhanced foundation, the hybrid RRF methodology significantly elevates retrieval efficacy, attaining the highest absolute scores across every assessment metric. The supplementary improvements over the semantic retriever are moderate (7–17%), although they remain constant across all assessed measures.

TABLE V. RETRIEVAL EVALUATION

Technique	BM25 only	Semantic only (E5)	Hybrid RRF
P@1	0.002	0.006	0.007
MRR@10	0.005	0.009	0.010
NDCG@10	0.008	0.014	0.015
Hit@5	0.006	0.011	0.012

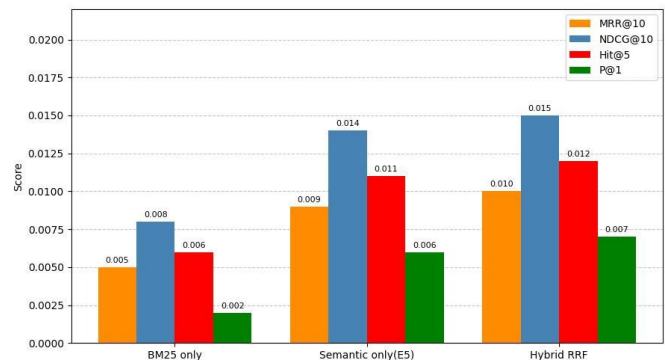


Fig. 6. Retrieval techniques comparison on AraLex corpus.

Category-specific examination reveals significant disparities in retrieval challenges. Minor categories like Permission for polygamy (n=50) attain well, due to the limited pool of available records. Extensive classifications like alimony (n=552) and judicial divorce (n=602) attain lower scores, since several records within these groups use basically the same terms, which complicates the identification of a particular document from a standard template inquiry.

C. Retrieval-Augmented Generation Outcomes

Table VI displays the assessment outcomes of RAG creation for 45 question-and-answer pairs. LLaMA-3.3-70B was accessed using the Groq API for every pair, with the five most relevant files given as context.

TABLE VI. RETRIEVAL-AUGMENTED GENERATION EVALUATION

Measure	Token F1 (SQuAD-style)	ROUGE-L	Cosine Similarity
Parameter	0.261	0.191	0.845
± Std Dev	0.091	0.065	0.062
Analysis	Conceptual overlap vs. fundamental truth	Subsequence with the longest in common	Semantic proximity (E5-embeddings)

Table VI presents the quality of RAG generation evaluated by three complementary measures. Cosine Similarity (0.845) demonstrates a robust semantic correspondence between the produced and reference responses, showing that LLaMA-3.3-70B reliably comprehends the essence of the accurate answer

TABLE VII. PLACEMENT OF ARALEX-RAG SYSTEM WITHIN RELATED WORKS

System	Technique	Experiment	Country / Field	Result
Our System (AraLex-RAG)	BM25 + FAISS + RRF	Retrieval + Classification RAG Question-Answering	Morocco 26 classes	Acc=92.87%
	TF-IDF + LinearSVC		Legal decisions	F1=76.15%
	LLaMA-3.3-70B (RAG)		(Table II)	Cosine(RAG)=0.845
[1]	BM25 + BERT re-rank	Legal Case Retrieval (Task1)	Canada Case Law	F1=0.3001 (THUIR)
[5]	TF-IDF + SVM	Prediction Charge	China - CAIL criminal cases	Acc=94.00%
[6]	GloVe + CNN	Judgment prediction	Morocco Six classes Court of Cassation	Acc=80.51%
[7]	TF-IDF + SVM	Legislative Docs classification	Morocco Seven legislative data fields Document types	F1=94.45% Acc=94.58%,
[13]	GPT-4 (RAG)	Arabic question answering (RAG)	Arabic dictionary (structured)	Cosine=0.87 F1=0.90

The systems included in Table VII cover three languages for different nations, illustrating the increasing global attention to legal NLP across various legal frameworks. Each pertains to a distinct objective: case retrieval, judgment forecasting, legislation categorization, or inquiry resolution, under varying resource constraints and assessment methods. AraLex-RAG enhances this domain by focusing on Moroccan Arabic judicial rulings, a context that combines limited resources with the unique language and legal attributes of the Moroccan judicial court.

V. DISCUSSION

TF-IDF combined with LinearSVC achieves 92.87% accuracy without doing any morphological analysis. The justification is clear: Moroccan legal classifications employ remarkably uniform procedural terminology, and character n-grams (n=2,4) inherently detect these subword structures, encapsulating morphological affixes without the need for a stemmer and maintaining resilience when identical roots manifest in various surface forms. Macro F1 (76.15%) is inferior to accuracy since it assigns identical weight to all 26 classes, including those with fewer test samples. The disparity between recall and precision (75.08% vs 80.27%) indicates that the classifier prioritizes caution, overlooking some instances

despite variations in phrasing. The values of Token F1 (0.261) and ROUGE-L (0.191) are diminished, which is anticipated for Arabic generative outputs. The model often chooses to paraphrase instead of directly replicating the reference response, adversely impacting word-overlap metrics while maintaining semantic similarity.

D. Benchmarking with Prior Work

Table VII positions AraLex-RAG within the context of comparable legal NLP systems. No direct quantitative comparison can be made among these rows, since each system functions inside a distinct language, domain, class count, and assessment process. The table is designed to demonstrate the range of each work rather than to evaluate performance.

rather than incorrectly labelling them, which is the appropriate compromise for a legal framework.

The retrieval statistics need a contextual framework. The assessment prompts are uniform template inquiries used across all papers, posing the query “What is the court’s ruling?” and requiring the system to pinpoint a singular document among hundreds that exhibit the identical procedural language, a process that is fundamentally hard. Under this restricted framework, semantic search exceeds BM25 by 75% on MRR@10, indicating that multilingual-E5-small locates files via semantic relevance rather than mere word correlation, which is crucial for Arabic legal texts because the same idea is expressed in many morphological ways. RAG has a distinct divergence in metrics: Cosine Similarity (0.845) is elevated, indicating that the produced responses are semantically accurate Arabic legal language; however, Token F1 (0.261) and ROUGE-L (0.191) are lower due to the model's tendency to paraphrase instead of replicating the reference response precisely.

The following three limitations should be mentioned properly. The corpus's size suggests that some categories contain fewer than 50 instances, which restricts the classifier and introduces noise into retrieval evaluations; still, with several thousand files per class, targeted document searches

would be practicable, enabling a more earnest exploration of neural techniques. The assessment of retrieval depends on computer-created template questions rather than genuine inquiries from legal professionals; a practitioner study would provide a more realistic depiction of operational performance. In conclusion, the present architecture employs the Groq cloud API for RAG generation, allowing for sub-second inference without reliance on local GPU hardware; a sensible advancement would be to replace it with a locally hosted open-source framework, thus making the system completely autonomous and fit for offline operation in judicial environments. The upcoming phase of action is enhancing the corpus with supplementary decision files that are already in possession.

VI. CONCLUSION

AraLex-RAG is the initial solution that simultaneously tackles information retrieval, categorization, and RAG question answering for court rulings in Moroccan Arabic. It is constructed upon AraLex-RAG, a set of 2855 unstructured judicial rulings across 26 law domains, and integrates BM25, FAISS dense retrieval, Reciprocal Rank Fusion, character n-gram TF-IDF classification, and LLaMA-3.3-70B creation into a singular program operable on a conventional laptop without a GPU. Assessment validates the feasibility of the methodology: the classifier attains 76.15% macro F1 and 92.87% cross-validation accuracy over 26 legal classifications, hybrid RRF retrieval surpasses both BM25 and semantic search separately, and RAG generation obtains a Cosine Similarity of 0.845, signifying robust semantic coherence among produced and reference responses. These findings indicate that lightweight, GPU-independent techniques may provide robust results on specialized Arabic legal documents, and that AraLex-RAG offers a feasible, implementable option for Moroccan legal practitioners and users without necessitating costly infrastructure.

ACKNOWLEDGMENT

This work was supported by the Ministry of Higher Education, Scientific Research and Innovation, the Digital Development Agency (DDA) and the CNRST of Morocco (Alkhawarizmi/2020/25).

REFERENCES

- [1] Kim, M.Y.; Xu, Y.; Goebel, R. (Eds.) Competition on Legal Information Extraction/Entailment (COLIEE) 2023. Lecture Notes in Computer Science; Springer: Berlin, Germany, 2023.
- [2] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The muppets straight out of law school," in Findings of EMNLP 2020, Online, 2020, pp. 2898–2904.
- [3] R. Upadhyaya and T. Y. S. S. Santosh, "LexCLiPR: Cross-lingual paragraph retrieval from legal judgments," in Proc. 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025), Vienna, Austria, 2025, pp. 13971–13993.
- [4] S. A. Mahmoudi, C. Condevaux, G. Zambrano, and S. Mussard, "Article 700 identification in judicial judgments: Comparing transformers and machine learning models," *Stats*, vol. 7, 2024, pp. 1421–1436.
- [5] C. Xiao, H. Zhong, Z. Guo, C. Tu, Z. Liu, M. Sun, et al., "CAIL2018: A large-scale legal dataset for judgment prediction," *arXiv preprint arXiv:1807.02478*, 2018.
- [6] J. Zahir, "Prediction of court decision from Arabic documents using deep learning," *Expert Syst.*, vol. 40, p. e13236, 2023, doi:10.1111/exsy.13236.
- [7] A. Bouhouche, M. Esghir, and M. Errachid, "Classification of Moroccan legal and legislative texts using machine learning models," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, pp. 1108–1114, 2024.
- [8] F. Hijazi et al., "ArabLegalEval: A multitask benchmark for assessing Arabic legal knowledge in large language models," in *Proc. ArabicNLP 2024*, Miami, FL, USA, 2024.
- [9] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. T. Yih, "Dense passage retrieval for open-domain question answering," in *Proc. EMNLP 2020*, Online, 2020, pp. 6769–6781.
- [10] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms Condorcet and individual rank learning methods," in *Proc. SIGIR 2009*, Boston, MA, USA, 2009, pp. 758–759.
- [11] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 9459–9474, 2020.
- [12] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, et al., "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, pp. 1–38, 2023, doi:10.1145/3571730.
- [13] R. Al-Rasheed et al., "Evaluating RAG pipelines for Arabic lexical information retrieval: A comparative study of embedding and generation models," in *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, 2025, pp. 155–164.
- [14] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Now Publishers Inc*, vol. 4, 2009.
- [15] L. Wang, N. Yang, X. Huang, et al., "Text embeddings by weakly-supervised contrastive pre-training," *arXiv preprint arXiv:2212.03533*, 2022.
- [16] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual E5 text embeddings: A technical report," *arXiv preprint arXiv:2402.05672*, 2024.
- [17] Johnson J, Douze M, and Jégou H. "Billion-scale similarity search with GPUs". *IEEE transactions on big data*. Vol.7, 535-547.
- [18] S. Ibrihich, A. Oussous, O. Ibrihich, and M. Esghir, "A Review on Recent Research in Information Retrieval," *Procedia Computer Science*, vol. 201, pp. 777–782, 2022.
- [19] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, "The Faiss Library," *IEEE Trans. Big Data*, 2025, doi:10.1109/TBDATA.2025.3618474. (early access)
- [20] N. Y. Habash, *Introduction to Arabic Natural Language Processing*. San Rafael, CA, USA: Morgan & Claypool, 2010.
- [21] Magesh, Surani, Dahl, Suzgun, Manning, Ho, "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools," *Journal of Empirical Legal Studies*, 2025, pp. 1–27.
- [22] S. Ceri, A. Bozzon, M. Brambilla, E. Della Valle, P. Fraternali, and S. Quarteroni, "The Information Retrieval Process," in *Web Information Retrieval*, Berlin, Germany: Springer, 2013, pp. 13–26.
- [23] N. Craswell, "Mean Reciprocal Rank," in *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds. Boston, MA, USA: Springer, 2009, p. 1703.
- [24] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, 2002.
- [25] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Proc. ACL-04 Workshop: Text Summarization Branches Out*, Barcelona, Spain, 2004, pp. 74–81.