TELKOMNIKA Telecommunication Computing Electronics and Control
Vol. 23, No. 6, December 2025, pp. 1495~1505
ISSN: 1693-6930, DOI: 10.12928/TELKOMNIKA.v23i6.27400

Retrieval-augmented generation for Arabic legal information: the family code case study

Jamal Hrimech, Mohammed Mghari, Youssef Zaz
Department of Computer Science, Faculty of Science, Abdelmalek Essaâdi University, Tetouan, Morocco

Article Info

Article history:

Received Jul 15, 2025
Revised Sep 19, 2025
Accepted Oct 19, 2025

Keywords:

Arabic-natural language processing
Large language model
Legal accessibility
Moroccan law
Retrieval-augmented generation
Semantic search

ABSTRACT

This document describes the implementation and evaluation of a retrieval-augmented generation (RAG) system to improve access to and understanding of Moroccan law, particularly the family code in Arabic. The research addresses the drawbacks of the widely used linguistic model applied to complex legal terminology in Arabic and aims to help citizens access crucial legal data. We built a new custom dataset with 2.5 k question-answer pairs while preprocessing and using the BGE-m3 embedding model in this experiment. Performance metrics, such as mean reciprocal rank (MRR), Recall@k, and F1-score, indicate that the RAG approach is effective compared to the use of standalone large language models (LLMs). Moreover, an evaluation on metrics such as the blue score, fidelity, response relevance, and contextual relevance indicated that the matching of meanings and context were well captured, which signifies a very good semantic understanding. The research highlights the need for language-specific model specialization in Arabic and presents its main challenges, such as dialectal variations and appropriate evaluation measures. The results indicate that well-developed RAG systems offer a promising approach to improving access to legal information in Arabic-speaking practice communities and to guiding future research and development in this field.

Corresponding Author:

Jamal Hrimech
Department of Computer Science, Faculty of Science, Abdelmalek Essaâdi University
P.O. Box. 2121, M'Hannech II, Tetuan, 93030, Morocco
Email: jamal.hrimech@etu.uae.ac.ma

1. INTRODUCTION

Retrieval-augmented generation (RAG) represents one such powerful and creative response to the limitations large pre-trained language models suffer from when applied to knowledge-rich natural language processing (NLP) tasks. Although these models have shown remarkable ability to memorize facts, they often have difficulty precisely querying that same knowledge, which manifests as "hallucinations" (the generation of false, unsupported facts) and opacity in their decision-making. The RAG models address this problem by combining parametric and non-parametric memory systems. In particular, RAG is the composition of a pre-trained generative model (parametric memory) and a retrievable knowledge base, which is usually a dense vector index of documents (non-parametric memory). Such a hybrid mechanism enables models to dynamically retrieve and utilize knowledge from the external source while generating text, which can make the generated response more factual, related and varied. This has a few important advantages: less hallucination, better performance on a variety of NLP tasks, easier world knowledge updates by replacing the non-parametric memory index, and more interpretability because we are pulling up human-readable documents. As a benefit, RAG has much fewer trainable parameters than the large parametric-only models and achieves more effective performance [1]. Additionally, RAG can significantly reduce graphics processing unit (GPU) and random access

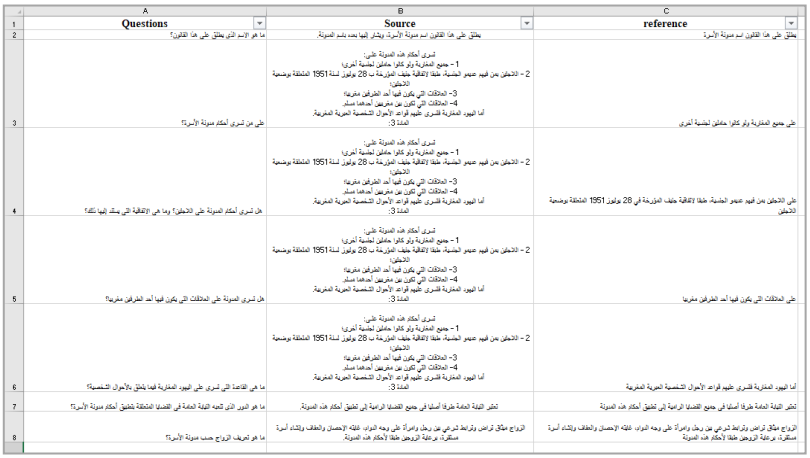
memory (RAM) costs when outsourcing knowledge to scalable data stores (as opposed to model weights), and robust scaling law and empirical results suggest that smaller models with large retrieval indices are competitive with or even outperform larger fully parametric models at a fraction of the GPU memory cost [2].

Furthermore, the RAG system can access and integrate related information in structured data, documents or databases, so that the information is not only suitable for the pre-setting context of the model, and the data source can be replaced flexibly according to the actual need for the conversion of the knowledge of the large language model (LLM) to a specific knowledge domain. From a RAG structure perspective, there are two fundamental components. First, the retriever is the process of finding and extracting relevant information from a large dataset or knowledge base by comparing the question vector and data vectors to find the closest ones in terms of meaning. In order to transform the data into a vector, a preliminary process called indexing is mandatory for storing the data in a vector database. This serves to optimize the search functionality so that retrieval is as fast and efficient as possible. Second, the generator is usually the LLM that leverages the information retrieved by the retriever and query to generate a coherent and factual answer [3].

Given the ability of RAG systems in improving the accuracy of LLMs, they have been applied in various fields, such as the banking sector, a RAG system named "UniAsk" intended for European bank employees accessing documentation related to policies, regulations and processes [4]. In the health and medicine sector, in particular for the purpose of interpreting recommendations, assisting with diagnosis, selecting eligibility for clinical trials, searching for clinical information and extracting information from scientific literature [5]. In the financial sector, a RAG pipeline has been developed for a comparative study of retrieval and prompting strategies in financial quality assurance tasks, informing production deployment recommendations, although the main focus is on the analysis of controlled components [6]. In the legal field, an implementation of the RAG system to assist content creators in disputes related to the fair use doctrine in American copyright law, which combines several elements (semantic search, judicial citation networks, and legal knowledge graphs), this work improves the legal relevance of the documents found, which is crucial for a solid fair use defense [7].

In this paper we will implement a RAG system in the Moroccan legal context, particularly the family code, which will facilitate citizens' understanding of the law. We will use multilingual embedding models (open source) proving their performance in the Arabic language for the recovery of relevant documents, and a large language model as a generator. The objective of this work is part of the right of access to information and the simplification of judicial procedures and processes. Despite its proven success, RAG has yet to be applied to the morpho syntactic and semantic intricacies of Arabic legal language. Challenges include handling legal terminology, maintaining the content validity of legal clauses reproduced from retrieval systems, and creating reliable evaluation sets, as well as the scarcity of annotated Arabic legal corpora. To fill this lacuna, this paper aims to answer the following research questions: what is the best way to chunk legal data in a way that acknowledges semantic representations? do standardized metrics properly measure the quality of the generated legaldocuments in Arabic? and which open-source multilingual embedding model achieves the best performance for the retrieval of the relevant articles of the family code?

To address these questions, the key contributions of this paper are as follows: to that end, we compile a new corpus of 2,500 arabic question-answer pairs focused on the Moroccan family code, obtained by a semi-automatic pipeline and manually checked. The structure of this dataset, which includes the question, source, and reference answer, is illustrated in Figure 1.



	A	B	C
1	Questions	Source	reference
2	ما هو الإسم الذي يطلق على هذا القانون؟	يطلق على هذا القانون اسم مدونة الأسرة، ويشار إليها بعده باسم المدونة.	يطلق على هذا القانون اسم مدونة الأسرة
3	على من تسري أحكام مدونة الأسرة؟	تسري أحكام هذه المدونة على: 1 - جميع المغاربة ولو كانوا يحاملون لجنسية أخرى؛ 2 - اللاجئين بمن فيهم عديمو الجنسية، طبقا لاتفاقية جنيف المؤرخة في 28 يوليوز لسنة 1951 المتعلقة بوضعية اللاجئين؛ 3- المحلات التي يكون فيها أحد المرتفقين مغربيا؛ 4- المحلات التي تكون بين مغربيين أحدهما مسلم. أما اليهود المغاربة فتسري عليهم قواعد الأحوال الشخصية المغربية. المادة 3:	على جميع المغاربة ولو كانوا يحاملون لجنسية أخرى
4	هل تسري أحكام المدونة على اللاجئين؟ وما هي الإتفاقية التي يستند إليها ذلك؟	تسري أحكام هذه المدونة على: 1 - جميع المغاربة ولو كانوا يحاملون لجنسية أخرى؛ 2 - اللاجئين بمن فيهم عديمو الجنسية، طبقا لاتفاقية جنيف المؤرخة في 28 يوليوز لسنة 1951 المتعلقة بوضعية اللاجئين؛ 3- المحلات التي يكون فيها أحد المرتفقين مغربيا؛ 4- المحلات التي تكون بين مغربيين أحدهما مسلم. أما اليهود المغاربة فتسري عليهم قواعد الأحوال الشخصية المغربية. المادة 3:	على اللاجئين بمن فيهم عديمو الجنسية، طبقا لاتفاقية جنيف المؤرخة في 28 يوليوز 1951 المتعلقة بوضعية اللاجئين
5	هل تسري المدونة على المحلات التي يكون فيها أحد المرتفقين مغربيا؟	تسري أحكام هذه المدونة على: 1 - جميع المغاربة ولو كانوا يحاملون لجنسية أخرى؛ 2 - اللاجئين بمن فيهم عديمو الجنسية، طبقا لاتفاقية جنيف المؤرخة في 28 يوليوز لسنة 1951 المتعلقة بوضعية اللاجئين؛ 3- المحلات التي يكون فيها أحد المرتفقين مغربيا؛ 4- المحلات التي تكون بين مغربيين أحدهما مسلم. أما اليهود المغاربة فتسري عليهم قواعد الأحوال الشخصية المغربية. المادة 3:	على المحلات التي يكون فيها أحد المرتفقين مغربيا
6	ما هي القاعدة التي تسري على اليهود المغاربة فيما يتعلق بالأحوال الشخصية؟	تسري أحكام هذه المدونة على: 1 - جميع المغاربة ولو كانوا يحاملون لجنسية أخرى؛ 2 - اللاجئين بمن فيهم عديمو الجنسية، طبقا لاتفاقية جنيف المؤرخة في 28 يوليوز لسنة 1951 المتعلقة بوضعية اللاجئين؛ 3- المحلات التي يكون فيها أحد المرتفقين مغربيا؛ 4- المحلات التي تكون بين مغربيين أحدهما مسلم. أما اليهود المغاربة فتسري عليهم قواعد الأحوال الشخصية المغربية. المادة 3:	أما اليهود المغاربة فتسري عليهم قواعد الأحوال الشخصية المغربية
7	ما هو الدور الذي تلعبه النيابة العامة في القضايا المتعلقة بتطبيق أحكام مدونة الأسرة؟	تعتبر النيابة العامة طرفا أصليا في جميع القضايا الرامية إلى تطبيق أحكام هذه المدونة.	تعتبر النيابة العامة طرفا أصليا في جميع القضايا الرامية إلى تطبيق أحكام هذه المدونة
8	ما هو تعريف الزواج حسب مدونة الأسرة؟	الزواج ميثاق تراض ورابط شرعي بين رجل وامرأة على وجه الدوام، غايته الإحصان والعفاف وإنشاء أسرة مستقرة، برعاية الزوجين طبقا لأحكام هذه المدونة.	الزواج ميثاق تراض ورابط شرعي بين رجل وامرأة على وجه الدوام، غايته الإحصان والعفاف وإنشاء أسرة مستقرة، برعاية الزوجين طبقا لأحكام هذه المدونة

Figure 1. Dataset structure

We propose a clause-based document division approach that is more semantically relevant than the standard character or token-based approach for legal texts. We present a comparative analysis of state-of-the-art multilingual embedding models on legal retrieval for the Arabic language and a discussion of evaluation metrics with focus on the challenges of the Arabic legal domain.

The rest of this paper follows a pattern: section 2 discusses related work. Section 3 adopts the methodology used in this paper. Section 4 presents the experiments and explains the results. Section 5 summarizes the conclusions and key ideas, as well as suggestions that future researchers could follow.

2. RELATED WORK

Recent years have seen the emergence of research applying augmented retrieval-generation (ARG) architectures systems that combine neural retrieval on semantic embedding's with large language model text generation for Arabic language tasks. El-Beltagy and Abdallah [8] present a comprehensive case study on the implementation and evaluation of RAG for Arabic texts. Their work explores the performance of different semantic embedding models (among them multilingual models like AraBert and Cohere 1) for retrieval, and a few long-term translation LLM models for generation. This work also investigates the difficulties of dialectal variations between documents and queries, and shows that despite the challenges, it was possible to compose RAG pipeline for Arabic by fusing existing semantic embeddings, and LLMs. Abdelazim *et al.* [9] the team concentrated on the core generation part of Arabic RAG systems, and compared ten state-of-the-art multilingual semantic embedding models to judge available Arabic search-generation outputs and performance. On the Arabic reading comprehension dataset (ARCD) testbed, they reported that the Microsoft E5sentence embedding model achieved the highest recall rate of 10, with above 90%. This work emphasizes the necessity of using Arabic-appropriate multilingual word embedding's for Arabic search tasks.

Alshammary *et al.* [10] they introduce the retrieval fact-checking prompt generation (RFPG) framework, a fact-aware RAG approach to question answering in Arabic, a low-resource language. The fact-checking is performed in an asymmetric multi-stage pipeline retrieval setup and the generation is conditioned on personalized prompts. The RFPG model demonstrated high accuracy (100%) in answering 123 Arabic questions, as well as a source citation accuracy of 98%, outperforming the standard RAG and recent LLMs such as GPT-4 and GPT-4o in their experiments. However, the scope of their evaluation was limited to a controlled question set rather than open or legal corpora.

Currently research that targets RAG in the legal field involves mostly retrieval rather than full generation pipelines. Jafar *et al.* [11] present a new approach to automate retrieval of Arabic legal texts using unsupervised topic modeling (Top2Vec) and density-based clustering (HDBSCAN). The authors address the morphological complexity and ambiguity of Arabic legal texts through preprocessing steps such as normalization and tokenization. But their system while achieving a document retrieval rate of 87% and a coverage rate for 80%, lacks the generative capability needed for full RAG.

Several more general legal technology studies have investigated or proposed RAG pipelines for legal texts, but generally not focused on Arabic. For example, Hindi *et al.* [12] provide a systematic survey of existing RAG architectures in the legal domain, discussing retrieval methods, evaluation metrics, and ethical considerations, but do not report direct experiments on Arabic legal corpora. Wahidur *et al.* [13] and Kalra *et al.* [14] introduce RAG frameworks tailored to the legal domain, with innovations such as recursive feedback or adaptive hybrid retrieval, but their work does not focus on Arabic or use primary Arabic legal documents.

Given the lack of previous studies on the application of RAG systems in the field of Moroccan law, a direct comparison is hardly possible. However, we can put our results in the broader context of legal artificial intelligence (AI) research. For example, work on the legal general language understanding evaluation (LexGLUE) benchmark has already noted the complexity of English legal language for standard language models [15]. On the other hand, the creation of specialized question-answering corpora from complex European regulations such as the general data protection regulation (GDPR) [16], is considered an important step for the development of reliable question and answering (Q&A) systems. Our approach is in line with systems such as ChatLaw [17] for Chinese law. Unsurprisingly, ChatLaw also confirms the need to integrate external knowledge to reduce factual errors made by LLMs in China. Finally, the performance of our research models can be compared to the reports of previous large information retrieval tests identified. This demonstrates the robustness of these models even when applied to a field as specialized as law.

To provide a clear comparison with prior art, Table 1 summarizes the most relevant studies discussed in the previous section. This Table 1 highlights the key differences in dataset size, language focus, evaluation metrics, and technical limitations, thereby positioning our contribution within the existing landscape of legal AI research.

Table 1. Comparative results of summarizing previous work

Aspect	LexGLUE (Chalkidis *et al*. [15])	Private International Law (PIL) dataset (Sovrano *et al*. [16])	ChatLaw (Cui *et al*. [17])
Dataset size	7 datasets, 200 K+ total instances	17 questions (9 new + 8 from previous work)	4 M samples across 10 major categories
Language	English only	English (EU regulations)	Chinese legal texts
Domain coverage	Multiple: European Convention on Human Rights (ECHR), US law, EU law, contracts	PIL only	Broad legal domains
Evaluation metrics	Micro/macro F1, multiple task-specific metrics	Top5-recall, Top5-precision, Top5-F1	Accuracy, F1, expert evaluation (4 dimensions: completeness, correctness, guidance, and authority)
Benchmarks used	7 standardized legal tasks (ECHR, Supreme Court of the United States (SCOTUS), and European Union Law Access Portal (EUR-LEX))	Custom PIL questions with expert validation	LawBench, legal professional exam
Human evaluation	Limited human evaluation baseline	Independent legal expert validation	Legal expert assessment on real cases
Performance results	Task-dependent: 50-95% F1	38.05% Top5-F1 overall	60.08% avg on LawBench (vs GPT-4: 52.35%)
Geographic scope	US/EU-focused legal systems	EU-focused (PIL)	China-focused legal system
Language limitations	English only	English only	Chinese primarily
Technical limitations	Missing human baselines, copyright restrictions	Low baseline performance, requires multi-hop reasoning	Privacy concerns, hallucination risks despite mitigation

3. METHOD

This work aims to exploit the RAG architecture in the Moroccan legal context and to this end we will evaluate the performance of multilingual embedding models (open weights) in the Arabic legal context, in particular family code terminology as well as the evaluation of the capacity of large multilingual language models for generating complete and precise responses. The method adopted in this paper is as follows: starting with the experimentation of multilingual embedding models in the legal context in Arabic language, then the experimentation of large language models based on one of the embedding models proving its high quality and speed ratio.

3.1. Dataset

In this subsection, we explore the dataset generation method used in this work as well as its preprocessing. The dataset used in this work contains 2.5 k questions and answers on Moroccan law was designed to evaluate our RAG system on the Arabic legal context. It was created with the assistance of a robust LLM for content generations, it was structured with three columns: {questions}, {source} and {reference}.
{Questions}: questions asked by the user;
{Source}: the document relating to the question asked;
{Reference}: the correct and complete answer to the questions.

The questions column contains all possible questions for each article in the Arabic legal document. These questions were generated using GEMINI 2.5, using a well-detailed and structured prompt Algorithm 1, in order to obtain the maximum number of question modalities that an individual or legal expert could ask. Answer generation was performed by another LLM (GPT-4o) using a specific prompt Algorithm 2, to answer questions generated by GEMINI 2.5. Diversifying LLMs in the questionanswer generation procedure overcomes the problem of propagated errors; that is, if an LLM makes an error in a question (ambiguity, misinterpretation), it will likely be repeated in the answer. However, we concede that generating the data using LLMs may introduce some bias. These biases can reveal themselves as an interpretive bias, where the model may exhibit a preference for a dominant or mainstream interpretation of a legal clause'in its vast though uncontrolled training data'missing out on key subtleties. There is a further danger of framing bias; the generated questions may serve to reduce complex legal issues to very simple faulty form or lead towards some particular kind of answer.

Of course, diversifying LLMs in question-answer generation can lead to inconsistency issues stemming from the different analysis and interpretation styles of each LLM, which can thus create question-answer misalignments. To address this potential for bias and inconsistency, we manually check the questions and their answers. This dataset has been checked and corrected by private law laureates, all the errors found in this dataset are generally related to the structuring of the questions, the answers and the source text (law clause),

in particular we find several answers generated in a way that is too concise, which causes ambiguity in the understanding of the context, thus we find a disorder at the level of mapping between questions and its answer and its source text.

Algorithm 1. Prompt used to generate questions of dataset

```
you are an expert in arabic law text, from this article in attached files, create questions
for any article with respecting the following instructions:
Output:
A TXT file with:
article_number: The article number (e.g 65).
questions: A list of question in Arabic:
Short articles (1-2 sentences): 1 questions
Longer articles (multiple sentences, clauses, or points): At least one question per distinct
point (e.g, clause, condition, exception, step), with a minimum of 3 questions and no upper
limit.
Guidelines:
1.Article Analysis:
       -For {short articles} (1-2 sentences), identify the core idea and generate one
question.
       -For {longer articles}, break the text into distinct points (e.g, each clause,
condition, exception, or procedural step) and generate at least one question per point.
Examples of distinct points:
       -A separate sentence or clause.
       -Each numbered condition or requirement (e.g, 1), 2), 3)).
       -Exceptions or special cases (e.g, if unable to speak).
1.Question Generation:
       -{Short articles}: Generate exactly one question.
       -{Longer articles}: Generate at least one question per distinct point identified. If
fewer than 3 points exist, pad with additional questions to reach 5. If more than 5 points
exist, generate questions for all points, ensuring more than 5 questions.
Use varied question types (e.g, direct, yes/no, hypothetical).
```

Algorithm 2. Prompt used to generate answers of dataset

```
you are an expert in arabic law text, from this question in attached files named, generate a
concise answer for the questions from the law texte in file "LAW" with respecting the following
instructions:
the answers should be arabic only;
each question associated with its item number and the answer must be generated from the text
of the same item number in the law text file;
output format:
article_1:
answers:
       1- "يطلق على هذا القانون اسم مدونة الأسرة"
article_2:
answers:
       1 - "على جميع المغاربة ولو كانوا حاملين الجنسية أخرى"
       2 - "على اللاجئين بمن فيهم عديمو الجنسية، طبقا لاتفاقية جنيف المؤرخة في 28 يوليوز 1951 المتعلقة بوضعية اللاجئين"
       3 - "على العلاقات التي يكون فيها أحد الطرفين مغربيا"
       4 - "أما اليهود المغاربة فتسري عليهم قواعد الأحوال الشخصية العبرية المغربية"
```

3.1.1. Preprocessing of legal text

Typically, legal documentation is managed by the relevant Moroccan ministry, which uses a common format for all legal documents, including the ministry's watermark, header, footer, and preamble. All these documents are published online in PDF format. To extract only the legal clauses from these documents, we use the Python library "PyMuPDF" for advanced PDF manipulation, as well as its powerful handling of Arabic characters and right-to-left (RTL) reading direction. This extraction should significantly improve retrieval accuracy, as it ensures that the vector database contains only text that is semantically relevant to potential legal queries, thus improving the signal-to-noise ratio. To facilitate the retrieval tool's exploitation of the dataset, we removed diacritical marks (tashkeel), which change the pronunciation but not the fundamental meaning. The expected impact is an increase in the retrieval rate, as this avoids inconsistencies due to the presence of diacritical marks in the source text, but absent from the user query.

We then standardized the Arabic characters by replacing the different forms of the Arabic letters (ة، إ، أ، آ) with a standard syllabary (ه، ا). This step standardizes common spelling variations in the Arabic script with the aim of further improving data retrieval, ensuring that relevant documents are not missed due to minor and semantically irrelevant character variations.

3.2. Embedding models

Embedding models are proving to be valuable tools for Arabic NLP, allowing to capture the semantic and syntactic nuances of words. There are different variants of these models, such as AraBERT, E5(Microsoft) multilingual and others that offer vector representations of Arabic words, thus facilitating information extraction [18]. In this work we use three different types of multilingual embedding model (open weights) having optimal dimension for similarity search task in Arabic legal context.

The BGE-m3 model proves its strengths were its multilingual capabilities, its ability to handle long and noisy texts, and its excellent performance in semantic similarity search, especially semantic similarities for the legal document retrieval task, it relatively a new model and provides meaningful representations in over 100 languages and at multiple depths: word, paragraph, and full text up to 8192 tokens, also it is well suited for a wide range of recovery tasks and supports multi-granularity text integrations [19]. Microsoft's E5 multilingual model also supports Arabic language and is optimized for the task of semantic similarity and passage retrieval, whose embedding dimension is 768 (E5-small) and 1024 (E5-large).

The choice of the embedding model is an important step, especially in the Arabic legal context, which has a specific characteristic such as long and complex sentences, the intensive use of the passive, the mixture between modern legal terminology and the concept of traditional Islamic law. This characteristic, known in legal texts, leads us to the evaluation of embedding models before integrating them into the RAG architecture. In this work, we use three embedding models in legal context retrieval in order to distinguish one among the three models that is the most efficient in terms of accuracy and light in use.

3.3. Indexing and splitting

Indexing is a preliminary in RAG systems, it facilitates the storage, retrieval, and similarity search of vector representations essential for the rapid search of relevant documents [20]. In this regard, efficient indexing allows RAG systems to adapt perfectly to large databases and return the most relevant information to user queries, thus enabling faster and more accurate responses. Splitting is an essential and distinctive aspect of the system to manage the considerable size of the information text and, consequently, to make document processing and rapid identification of useful information possible [21].

There are two main document-splitting techniques used in the RAG system: the recursive character splitter (RCS) and the token-based splitter (TTS), but in this work RCS and TTS method are not suitable for legal documents, we splitting by clauses to preserves the integrity of each complete clause, respects the logical structure of the legal document, avoids splitting a clause in the middle and maintains the legal context of each section.

3.4. Generator

In a RAG system architecture, the generator is based on an LLM that takes as input the user's query (question) and the context selected by the retrieval, i.e. the text potentially containing the elements of the response in order to produce an informative output that responds to the query entered by the user. For a generator to generate an informative answer to a question from a specific context, a structured prompt plays a vital role in helping the LLM determine the context and the question, as well as the guidelines to follow in their answers [22]. In this work and given the specificity of the legal field, in particular the accuracy of the information to be generated, we use an adapted prompt which directs the generator to follow the following directives:

```
1. "You are a specialized assistant in Moroccan family law. You must answer only based on
    the information contained in the excerpts of context provided below."
2. "If you do not find the information in these excerpts, state clearly that the information
    is not available in the provided documents."
3. "Never invent legal content not explicitly mentioned in the excerpts."
4. "Answer in Arabic only and use a maximum of three sentences and make the answer accurate
    and concise."
5. "Use the same terms contained in the reference text when possible."
Retrieval and Generation for a Legal Query {Input:} User query Q, Vector DB Vclauses, LLM
Mgen, embedding model Eemb {Output:} Answer A qemb ← Eemb(Q) Embed the user query Cretrieved
← SimilaritySearch (qemb, Vclauses, k = 5) Retrieve top-k clauses Pprompt ← Build Prompt
(Cretrieved, Q) Use prompt A ← Mgen (Prompt) Generate answer with JAIS {Return} A.
```

4. EXPERIMENTATION AND RESULTS

4.1. Retriever part

In this subsection, we assess the retriever part on the Moroccan legal document. The embedding models we use in this work are BGE-m3, E5-large, and E5-small to determine the best model among the three with the ability to handle the legal context in Arabic.

The assessment process begins by storing the legal documentation in the vector database after preprocessing and splitting the document. Then, we use the {question} and {source} columns of our dataset to calculate the necessary benchmarks to quantify the effectiveness of the retriever in a RAG system.

The assessment metrics used are MRR, which is widely used in information retrieval and recommendation systems. MRR evaluates how well a correct answer or relevant resource is classified after a query is made. MRR is an average based on the positions of the first correct reciprocal answer over a number of query questions and therefore provides a single score. Recall@k is used to evaluate for information retrieval systems. It investigates the fraction of relevant documents that are retrieved in the top k documents returned by a system. Therefore, it quantifies how effective a system is at surfacing items that are truly of interest to a user.

The F1 score is a performance metric that is primarily used to evaluate classification models, particularly in the case of imbalanced data. It is the mean harmonic of precision and recall, allowing us to obtain a combined score that is balanced for false positives and false negatives. Table 2 presents the comparative results of the three embedding models evaluated on these metrics.

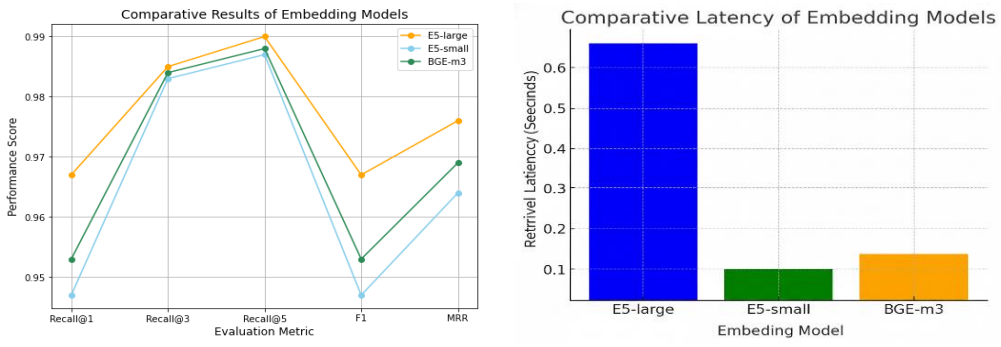
Table 2. Comparative results of embedding models E5-large, E5-small, BGE-m3

Metrics	E5-large	E5-small	BGE-m3	BGE-m3 vs E5-large	BGE-m3 vs E5-small
Recall@1	0.9670	0.9470	0.9530	-1.45%	+0.63%
Recall@3	0.9860	0.9840	0.9850	-0.10%	+0.10%
Recall@5	0.9900	0.9870	0.9880	-0.20%	+0.10%
F1	0.9670	0.9470	0.9530	-1.45%	+0.63%
MRR	0.9763	0.9645	0.9688	-0.77%	+0.45%
Retriever latency	0.657s	0.145s	0.178s	-0.479s	+0.033s

After analyzing the result according to Table 2, we see in the Recall@1 metric of the BGE-m3 model reaching 95.3%, which is less than 1.4% compared to E5-large and 0.6% more accurate than E5-small. Thus, in the Recall@5 metric of the same model, a value of 98.8%, which is practically equivalent to E5-large (a difference of 0.2%), and superior to E5-small. At the MRR metric level, the BGE-m3 model achieves 97% lower by 0.8% than E5-large and higher by 0.4% than E5-large, indicating a good overall positioning of relevant results. A consistency between Recall@1, F1@1 and MRR would mean that for each question, the correct law article is always the first result.

An excellent precision of relevant articles in the first position and a balance of performance/resources and near-perfect coverage, as well as excellent support for the Arabic language, was observed in the BGE-m3 embedding model. To provide a clearer visual representation of these findings, Figure 2 graphically summarizes the performance of the embedding models. Specifically, Figure 2(a) illustrates the comparative results across the key performance metrics discussed, visually confirming the competitive performance of BGE-m3 against the E5 models. In parallel, Figure 2(b) presents the retrieval latency for each model, highlighting the efficiency of BGE-m3, which offers a strong balance between speed and accuracy. BGE-m3 is an optimal choice for a production deployment of arabic legal retrieval system because it:

a) Offers excellent performance;
b) Offers a good balance between performance and resource consumption;
c) Benefits from a purpose-built retrieval architecture;
d) Demonstrates excellent Arabic language support.



(a) (b)

Figure 2. Comparison of embedding models; (a) metrics and (b) latency

4.2. Generator part

In this subsection, we assess the RAG system on generating answers to questions in the Moroccan legal context using the BGE-m3 embedding model. Generally, the generator relies on an LLM, prompt and retriever such that LLM receives the user's question and the retrieved document and then uses a guided prompt to generate a factual and non-hallucinatory answer, this is why the use of a robust LLM is essential in a RAG system.

According to Sengupta *et al*. [23], Jais Arabic instruction-based system (JAIS) is a state-of-the-art, Arabic-centric, instruction-optimized open generative language model with 13B parameters and built on the GPT-3 decoder architecture. JAIS being designed and trained from the ground up with a very large amount of Arabic data (alongside English) often gives it a better intrinsic understanding of nuances, morphology, syntax, and perhaps dialects (despite being primarily MSA). The JAIS model is evaluated across a wide range of Arabic NLP benchmarks, covering reasoning, knowledge, misinformation, and bias, and has been found to have significantly superior Arabic knowledge and reasoning capabilities compared to all existing open Arabic and multilingual models. In this assessment, we use six metrics to assess the LLM, namely cosine similarity, bilingual evaluation understudy score (BLEU) score, F1 score, faithfulness, answer relevance, and context relevance [24]. The performance results of our RAG system's generator, evaluated on the Moroccan law dataset using these six metrics, are presented in Table 3.

Table 3. Performance result of the generator on the Moroccan law dataset

	Cosine-similarity	BLEU score	BERTscore	ChrF++	F1 score	Faithfulness	Answer relevance	Context relevance
Jais-13B	69%	5.7%	73%	42,7%	21.9%	67%	79.09%	73.9%

After analyzing the performance result of the generator, we find a good capture of meaning and context of the answers which indicates an excellent semantic understanding, at the level of measuring the similarity between a generated answer and a reference answer. In the following, we can notice that the low value obtained for the legal text written in Arabic is normal, because, on the one hand, we find an extensive synonymy - that is, several synonymous words in a single language can express the same notion or idea, and, on the other hand, in the field of law, each term used has a legal and precise meaning, therefore, BLEU score does not differentiate between apparent synonyms that may have different legal implications. In addition, we recall that in general, the Arabic language has a rather complicated grammatical structure, with acceptable word order variations. This also influences the BLEU score, because, being based on n-grams, it can also penalize formulations that are correct, but simply differ [25].

Furthermore, in the context of NLP dealing with the Arabic corpus of legal texts, where the variability of the translated terminology and the complex morphosyntax pose major problems, the F1score suggests that the system is competent enough to partially capture the conception at stake by means of a diversity of subsumption while being translated by an adapted source lexicon even if it is different from the reference formulations.

we also deduce a strong correlation and good coherence between cosine metrics <-> Blue, Cosine <-> F1 and Blue <-> F1, which indicates respectively that the RAG system has an understanding that aids generation and an understanding and vocabulary related to lexical coherence.

```
PRACTICAL CASE: WHAT DO YOUR CORRELATIONS REVEAL?
Question: ما هي شروط صحة الزواج ؟
Reference: شُروطُ صِحَّةِ الزَّواج هي الوَلِيُّ والشَّاهِدانِ والصَّداقُ
Genrerated : يَتَطَلَّبُ الزَّواجُ الصَّحيحُ وُجودَ وَلِيّ وشُهودٍ ومَهرٍ
#Resulting metrics:
Cosinus: 0.85 → Excellent (same meaning, identical concepts)
BLEU: 0.15 → Low (different Word: ( ولي ≠ الولي , شُهودٍ≠ شّاهِدان , صَّداق≠ مَهر)
F1: 0.30 → Moderate (some common words: ( وُجودَ , الزَّواجُ
```

Then, after analyzing the results obtained on the "faithfulness, answer relevance, and context relevance metrics". In terms of measuring whether the generated response is faithful to the retrieved context, a moderate fidelity indicates that the model generally remains consistent with Arabic legal documents. For "answer relevance", we observe an excellent score of 79.09%, showing that the model generates relevant responses to the posed questions, in addition to a good ability of the system to identify and use appropriate contextual passages, suggesting an effective functioning of the retrieval module.

The deduced results then lead us to identify that BGE-m3 is the best, which means that development companies that create similar applications in the Arab world now have an open weight, efficient and proven

model, which accelerates the development of new legal technologies. Thus, the good functioning of our RAG system on Moroccan law proves that we have a reliable method that makes texts that are contextually complex (like the law) accessible to the general public, something that has a social impact helps people access legal information. Moreover, the low BLEU score calls on the research community to delve into this niche of developing evaluation methods for Arabic, especially in specialized fields like law. The experiments were run on a machine with the following specifications:

- GPU: T4×2
- GPU memory (VRAM): 16 GB
- System memory (RAM): 32 GB
- Framework: PyTorch

Table 4 provide qualitative examples to show the finer points of the model's performance. The incorrect generations illustrate deadly failure modes like returning a factually incorrect answer (hallucination) in Example 1 or providing the right but perilously incomplete answer in Example 2. Correct generations, however, indicate that the system is capable of generating a response which is semantically equivalent and factually correct for this question. Such a qualitative analysis is important to get an insight on real-world reliability of the model beside quantitative performance.

Table 4. Qualitative examples of correct and incorrect generations

Incorrect generations	Correct generations
Exemple 1 : Question : ما هو التصرف الذي يجب ألا يصدر من طالب الفسخ بعد علمه بتعذر الشفاء من العيب ليُقبل طلبه؟ Reference : يجب الا يصدر من طالب الفسخ ما يدل على الرضا بالعيب بعد العلم بتعذر شفائه Generated response : يفسخ الزواج الفاسد لعقده قبل البناء وبعده، وذلك في الحالات الآتية: - إذا كان الزواج في المرض المخوف لأحد الزوجين، إلا أن يشفى المريض بعد الزواج؛ - إذا قصد الزوج بالزواج تحليل المبتوتة لمن طلقها ثلاثا؛ - إذا كان الزواج بدون ولي في حالة وجوبه يعتد بالطلاق أو التطليق الواقع في الحالات المذكورة أعلاه، قبل صدور الحكم بالفسخ. Exemple 2 : Questions : ما هي الحقوق والواجبات التي تترتب على ثبوت الأبوة والبنوة؟ Reference : متى ثبتت بنوة ولد مجهول النسب ،يتبع أباه في نسبه ودينه، ويتوارثان وينتج عنه موانع الزواج، ويترتب عليه حقوق وواجبات الأبوة والبنوة. generated response: يثبت به حق التوارث	Exemple 1 : Questions: ما هو تعريف الخطبة Reference : الخطبة هي تعبير الطرفين بأي وسيلة متعارف عليها تفيد التواعد على الزواج Generated response : الخطبة هي تواعد المرأة والرجل على الزواج وهو عقد يلتزم بمقتضاه الخطيب أو الخطيبة بالامتناع عن الخطبة على غيره أو غيرها، Exemple 2 : Questions : ما هي القاعدة التي تسري على اليهود المغاربة فيما يتعلق بالأحوال الشخصية؟ Reference : اليهود المغاربة فتسري عليهم قواعد الأحوال الشخصية العبرية المغربية Generated response: تسري قواعد الأحوال الشخصية العبرية المغربية على اليهود المغاربة.

5. CONCLUSION

This research demonstrated that a properly configured RAG could significantly improve access to and understanding of Moroccan legal information, particularly the family code, an underexplored area, surpassing the capabilities of standalone LLMs. Specifically, our system achieved a retrieval recall of 98% for the relevant legal articles and a final answer relevance score of 79.09%, underscoring its effectiveness where baseline models often fail. Based on the quantitative results, the integrated BGE-m3 model proved to be the most suitable for this application, offering a good balance between performance, efficiency, and Arabic language. The identification of the BGE-m3 model as optimal provides a solid starting point for future work on semantic retrieval tasks in Arabic legal context. This work represents a step towards the real democratization of access to legal information for the Moroccan people. As these systems will be able to provide precise and understandable answers, they can help people better understand their rights and obligations and break down the barrier between citizens and the law and it highlighted the need to adapt a maximum length generator to Arabic in the BGE-m3 model and to optimize training techniques. Furthermore, this research resulted in the creation of a set of questions and answers, which could be refined and developed further.

Our immediate priority will be to improve the key components of the RAG system. This includes optimizing the BGE-m3 integration model on our specific legal corpus to further specialize its search capabilities. We also plan to explore more advanced RAG architectures, such as the integration of a reclassification model to refine the retrieved documents before their transmission to the generator. On the one hand, although broadening the scope is one of the guidelines for future work, it is essential to verify our methodology in other areas of law, such as criminal law, commercial law, and administrative law, as this will allow testing the system's ability to process more diverse and specialized legal terms and reasoning models, thus confirming the generalizability of our methodology. Furthermore, and in order to mitigate the biases inherent in the data generated by the LLM, we plan to develop a reference corpus, fully annotated and validated by legal experts. This expert-curated dataset will not only serve as a more reliable benchmark but will also enable a more systematic study of generation bias. Furthermore, we plan to explore the development of new

domain-specific evaluation measures that go beyond semantic similarity to assess the factual accuracy and logical consistency of generated legal responses.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their valuable feedback, which significantly improved the quality of this manuscript.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Jamal Hrimech	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Mohammed Mghari		✓		✓		✓		✓	✓	✓	✓	✓		
Youssef Zaz	✓		✓	✓	✓				✓		✓		✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study, specifically the 2,500 question-answer pairs generated for this research, are available from the corresponding author, J.H., upon reasonable request.

REFERENCES

[1] M. Sigala, A. Beer, L. Hodgson, and A. O'Connor, "Big data for measuring the impact of tourism economic development programmes: A process and quality criteria framework for using big data," *Big Data and Innovation in Tourism, Travel, and Hospitality. Springer Singapore*, pp. 57–73, 2019, doi: 10.1007/978-981-13-6339-9_4.

[2] G. Nguyen *et al.*, "Machine learning and deep learning frameworks and libraries for large-scale data mining: A survey," *Artificial Intelligence Review*, vol. 52, no. 1, pp. 77–124, Jun. 2019, doi: 10.1007/s10462-018-09679-z.

[3] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0197-0.

[4] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, "Deep learning approach for intelligent intrusion detection system," *IEEE Access*, vol. 7, pp. 41525–41550, 2019, doi: 10.1109/ACCESS.2019.2895334.

[5] K. Sivaraman, R. M. V. Krishnan, B. Sundarraj, and S. S. Gowthem, "Network failure detection and diagnosis by analyzing syslog and SNS data: Applying big data analysis to network operations," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 9S3, pp. 883–887, Aug. 2019, doi: 10.35940/ijitee.I3187.0789S319.

[6] A. D. Dwivedi, G. Srivastava, S. Dhar, and R. Singh, "A decentralized privacy-preserving healthcare blockchain for IoT," *Sensors*, vol. 19, no. 2, Jan. 2019, doi: 10.3390/s19020326.

[7] F. Al-Turjman, H. Zahmatkesh, and L. Mostarda, "Quantifying uncertainty in internet of medical things and big-data services using intelligence and deep learning," *IEEE Access*, vol. 7, pp. 115749–115759, 2019, doi: 10.1109/ACCESS.2019.2931637.

[8] S. R. El-Beltagy and M. A. Abdallah, "Exploring retrieval augmented generation in Arabic," *Procedia Computer Science*, vol. 244, pp. 296–307, 2024, doi: 10.1016/j.procs.2024.10.203.

[9] H. Abdelazim, M. Tharwat, and A. Mohamed, "Semantic embeddings for Arabic Retrieval Augmented Generation (ARAG)," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 14, no. 11, 2023, doi: 10.14569/ijacsa.2023.01411135.

[10] M. Alshammary, M. N. Uddin, and L. Khan, "RFPG: Question-answering from low-resource language (Arabic) texts using factually aware RAG," *2024 IEEE 10th International Conference on Collaboration and Internet Computing (CIC). IEEE*, pp. 107–116, Oct. 28, 2024, doi: 10.1109/cic62241.2024.00023.

[11] K. S. Jafar, A. A. Mohammad, A. A. Issa, and A. V. Panov, "Automating the search for legal information in Arabic: A novel approach to document retrieval," *Russian Technological Journal*, vol. 12, no. 5, pp. 7–16, Oct. 2024, doi: 10.32362/2500-316x-2024-12-5-7-1..

[12] M. Hindi, L. Mohammed, O. Maaz, and A. Alwarafy, "Enhancing the precision and interpretability of retrieval-augmented generation (RAG) in legal technology: A survey," *IEEE Access*, vol. 13, pp. 46171–46189, 2025, doi: 10.1109/access.2025.3550145.
[13] R. S. M. Wahidur, S. Kim, H. Choi, D. S. Bhatti, and H.-N. Lee, "Legal query RAG," *IEEE Access*, vol. 13, pp. 36978–36994, 2025, doi: 10.1109/access.2025.3542125..
[14] R. Kalra *et al*., "HyPA-RAG: A hybrid parameter adaptive retrieval-augmented generation system for AI legal and policy applications," *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U). Association for Computational Linguistics*, pp. 237–256, 2024, doi: 10.18653/v1/2024.customnlp4u-1.18.
[15] I. Chalkidis *et al*., "LexGLUE: A benchmark dataset for legal language understanding in English," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics*, 2022, doi: 10.18653/v1/2022.acl-long.297.
[16] F. Sovrano, M. Palmirani, B. Distefano, S. Sapienza, and F. Vitali, "A dataset for evaluating legal question answering on private international law," *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law. ACM*, pp. 230–234, June 21, 2021, doi: 10.1145/3462757.3466094.
[17] J. Cui *et al*., "Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model," 2023, *arXiv*, doi: 10.48550/ARXIV.2306.16092.
[18] G. Bourahouat, M. Abourezq, and N. Daoudi, "Word embedding as a semantic feature extraction technique in Arabic natural language processing: An overview," *The International Arab Journal of Information Technology (IAJIT)*, vol. 21, no. 2, 2024, doi: 10.34028/iajit/21/2/13.
[19] G. M. Csányi *et al*., "From fact drafts to operational systems: Semantic search in legal decisions using fact drafts," *Big Data and Cognitive Computing (BDCC)*, vol. 8, no. 12, p. 185, Dec. 2024, doi: 10.3390/bdcc8120185.
[20] Nishanth S. and Swetha S., "Enhancing RAG systems: A survey of optimization strategies for performance and scalability," *International Journal of Scientific Research in Engineering and Management (IJSREM)*, vol. 08, no. 06, pp. 1–5, June 2024, doi: 10.55041/ijsrem35402.
[21] E. Narimissa and D. Raithel, "Exploring information retrieval landscapes: An investigation of a novel evaluation techniques and comparative document splitting methods," *arXiv*, 2024, doi: 10.48550/ARXIV.2409.08479.
[22] P. Lewis *et al*., "Retrieval-augmented generation for knowledge-intensive NLP tasks," 2020, *arXiv*, doi: 10.48550/ARXIV.2005.11401.
[23] N. Sengupta *et al*., "Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models," 2023, *arXiv*, doi: 10.48550/ARXIV.2308.16149.
[24] S. Dara *et al*., "Context-aware auto-encoded graph neural model for dynamic question generation using NLP," *ACM Transactions on Asian and Low-Resource Language Information Processing*, Oct. 2023, doi: 10.1145/3626317..
[25] A. Elsaid, A. Mohammed, L. F. Ibrahim, and M. M. Sakre, "A comprehensive review of Arabic text summarization," *IEEE Access*, vol. 10, pp. 38012–38030, 2022, doi: 10.1109/access.2022.3163292.

BIOGRAPHIES OF AUTHORS

Jamal Hrimech received the Bachelor's degree in Network Administration and Information Systems Security from the University Abdelmalek Essaâdi, Tétouan, Morocco, in 2019, and the Master's degree in Intelligent Systems Processing from the University Mohammed V, Rabat, Morocco, in 2022. He is currently pursuing the Ph.D. degree at the Department of Computer Science at the Faculty of Sciences, University Abdelmalek Essaâdi, Tétouan, Morocco. His current research interests include generative AI and large language models (LLMs). He can be contacted at email: jamal.hrimech@etu.uae.ac.ma.

Mohammed Mghari received the Ph.D. degree in AI and Computer Science from the Abdelmalek Essaâdi University in 2024, and is currently an Assistant Professor at the Faculty of Sciences at Abdelmalek Essaâdi University, Tetouan. His current research interests include artificial intelligence, machine learning, deep learning, computer vision, and natural language processing. He can be contacted at email: mohammed.mghari@uae.ac.ma.

Youssef Zaz is a faculty member at the Université Abdelmalek Essaâdi, Faculty of Sciences of Tétouan, and an IEEE Senior Member. His research focuses on artificial intelligence, deep learning, smart grids, and renewable energy, with applications to sustainable energy systems. He is the author of numerous journal and conference publications, holder of three patents, and has supervised over a dozen Ph.D. students. He also served as President of the IEEE International Renewable and Sustainable Energy Conference (IRSEC) and International Conference on Multimedia Computing and Systems (ICMCS) conferences, contributing to the advancement of international research in renewable energy and computing. He can be contacted at email: youssef.zaz@uae.ac.ma.