applied sciences

MDPI

Article

A Deterministic Procedure-Aware Bilingual Retrieval-Augmented Generation Framework for Trustworthy High-Stakes AI Systems

Abdullah Bin Sawad [1] and Muhammad Binsawad [2,*]

[1] Department of Computer and Information Technology, The Applied College, King Abdulaziz University, Jeddah 21589, Saudi Arabia; asawad@kau.edu.sa
[2] Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia
* Correspondence: mbinsawad@kau.edu.sa

Abstract

The advent of large language models and Retrieval-Augmented Generation (RAG) models has greatly enhanced intelligent information systems. This has resulted in the development of context-aware and knowledge-grounded response generation. This has been highly beneficial in the context of religious advisory systems, which require precision, correctness, and knowledge grounding. For Islamic rituals like Hajj and Umrah, the user needs precise and accurate procedures to follow, which must adhere to specific sequences and knowledge grounding. However, the existing models have many limitations in this context, like hallucinations, a lack of procedural knowledge, bilingual inconsistencies, and an inability to incorporate safety constraints. This has made these models unsuitable for contexts in which incorrect responses can have serious implications. Therefore, in this context, this paper proposes a Deterministic Procedure-Aware bilingual Retrieval-Augmented Generation (DPAM-RAG) model, which can be highly beneficial in designing religious advisory systems. The proposed model can be highly beneficial in designing religious advisory systems. The proposed model integrates dataset modeling, procedure-aware chunking, bilingual alignment, and deterministic transformer-based response generation. Additionally, a confidence-based refusal strategy has been proposed to avoid the generation of responses that can be considered incorrect or out of context. The proposed model has been tested through an extensive experimental setup, which includes multiple transformer models like GPT, LLaMA-2, Mistral, MPT, and BLOOMZ. The experimental results have shown promising outcomes, which can be considered highly beneficial in designing trustworthy AI models.

Keywords: deterministic RAG; bilingual AI; religious AI systems; Hajj and Umrah guidance; hallucination reduction; transformer models; cross-lingual embeddings

Check for updates

Academic Editor: Douglas O'Shaughnessy

Received: 11 August 2026
Revised: 29 August 2026
Accepted: 29 August 2026
Published: 3 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The swift progress in artificial intelligence, especially in large language models and transformer-based models, has significantly impacted the manner in which knowledge is accessed, processed, and generated. Among these developments, Retrieval-Augmented Generation (RAG) has become a significant model in which information retrieval and generative models are integrated to provide contextually grounded responses [1,2]. This is

especially significant in high-stakes applications such as religious guidance, in which accuracy, procedural correctness, and adherence to religious doctrine are critical [3]. In Islamic pilgrimages such as Hajj and Umrah, any minor error in guidance can result in invalid rituals [4–6].

The potential of artificial intelligence-based systems in providing accessible, multilingual, and scalable guidance to a global audience is significant [7,8]. Millions of pilgrims, belonging to different linguistic groups, seek guidance in understanding complex rituals in which strict sequences and conditional rules are to be followed [9–11]. However, there are several challenges in using general-purpose large language models in this domain, such as hallucinations, inconsistency in outputs, and difficulty in maintaining correct procedural sequences [12,13]. Although existing techniques, such as RAG frameworks, multilingual embeddings, and prompt engineering, are significant in improving the accuracy of information retrieval [14,15], they are found to be lacking in ensuring procedural validity, logical constraints, and safety constraints [16,17].

Recent research in domains like healthcare and legal advisory has shown positive results in using structured knowledge sources as a foundation for model outputs [18,19]. Multilingual transformer models have shown promise in better understanding language by using aligned embeddings as well [20,21]. However, there has been little research done in understanding the requirements of religious procedural knowledge domains, like sequence enforcement and maintaining consistency [22,23]. Additionally, there has been little research done on eliminating hallucinations and ensuring reliable performance using temperature control and reinforcement learning [24,25].

To achieve these requirements in a more reliable and accurate manner, a Deterministic Procedure-Aware Bilingual Retrieval-Augmented Generation (DPAM-RAG) framework has been proposed in this research. The system is designed to incorporate structured dataset representation, procedure-aware retrieval, bilingual embedding alignment, deterministic generation, and a confidence-based refusal mechanism. The main objective of the proposed system is to ensure accurate procedure adherence, avoid hallucinations, and maintain consistency across languages while adhering to domain constraints.

The contributions of this research are as follows: A novel procedure-aware RAG architecture has been proposed in this research, along with an optimized bilingual embedding system designed specifically for religious text data. Deterministic decoding and safety-driven refusal mechanisms have been incorporated into the system as well. A comparative analysis of various transformer models has been conducted in a controlled environment, and an ethical framework has been proposed for reliable AI systems in critical domains.

In contrast to the dominant state-of-the-art approaches in the Retrieval-Augmented Generation paradigm, the present DPAM-RAG system makes several innovative contributions that tackle the inherent complexities associated with high-stakes procedural domains. First, the system includes an explicit procedure modeling mechanism, ensuring the generated response is procedurally correct. Second, the system proposes two new evaluation metrics: the Procedure Completeness Score (PCS) and the Cross-Lingual Consistency Score (CLCS). Third, the system includes an innovative deterministic decoding mechanism with confidence-based refusal. Lastly, the system includes an innovative unified architecture that jointly optimizes the relevance of the retrieved information, procedural correctness, multi-linguality, and safety constraints, which were not comprehensively addressed in the existing approaches.

The rest of the present study is outlined as follows: Section 2 describes the overall research design and procedures. Sections 3 and 4 discuss the analysis of the datasets and the procedure-aware chunking strategy for RAG, respectively. Section 5 discusses the bilingual Embedding Analysis and Cross-Lingual Retrieval. Sections 6 and 7 discuss the Pro-

posed DPAM-RAG Architecture and the experimental evaluations with the results, respectively. Sections 8 and 9 discuss Ethical and Religious Considerations and the discussion, respectively. Lastly, Section 10 concludes the present study.

2. Research Design and Procedure

This research takes a systems-oriented experimental research method to create and test a Deterministic Procedure-Aware Bilingual Retrieval-Augmented Generation (DPAM-RAG) system for the challenging task of religious guidance. The suggested system combines a structured dataset, bilingual embeddings, retrieval optimization, transformer-based response generation, and safety constraints into a single pipeline. The main objectives are to ensure procedural correctness, reduce hallucinations, and maintain cross-lingual semantic consistency.

Mathematically, the system is a composite function that transmits an input query to a response via several processing stages:

$$R = \mathcal{F}(Q, D, E, M, C) \tag{1}$$

where Q stands for the input query, D is the structured dataset, E is the embedding function, M is the transformer model, C represents constraint mechanisms, and R is the response generated. Modularity of components and systematic evaluation through this formulation allow for conducting experimental work in a modular fashion.

2.1. Dataset Preparation and Preprocessing

The dataset, based on credible Hajj and Umrah guidelines, undergoes an intensive preprocessing procedure to ensure structural consistency and semantic clarity. At first, raw textual data is normalized to eliminate discrepancies such as different ways of writing the same letter, different encodings, and unnecessary symbols. This normalization guarantees the same treatment of content in Arabic and English.

$$T_{norm} = \mathcal{N}(T_{raw}) \tag{2}$$

After normalization, structured information extraction is performed to turn the unstructured text into machine-readable forms. Each item of data is represented by a tuple that contains ritual steps, ruling conditions, sequence order, and bilingual text mappings.

$$S = \{(s_i, r_i, c_i, o_i, a_i, e_i)\} \tag{3}$$

Such a structured form of data allows performing different tasks related to getting data, reasoning, and verifying. In addition, to keep the logical and temporal aspects, procedure-aware chunking is done. Different from the traditional ways of chunking, such an approach guarantees that each chunk is a complete procedural unit, with the result of fewer fragments and better accuracy of the retrieval.

$$C_i = f(s_i, r_i, c_i, e_i) \tag{4}$$

The dataset used for this research involves a well-defined bilingual Hajj and Umrah guidance dataset created from authentic sources of Hajj and Umrah guidance materials. It consists of procedurally annotated Arabic-English records containing ritual steps, types of rulings, conditions, exceptions, temporal order, and bilingual semantic relationships. In the preprocessing stage, source texts were normalized in order to handle different encodings, orthography, and unnecessary symbols; then information was encoded in structured, machine-readable records. In addition, procedure-aware chunking was performed in such

a way that retrieval units contained all procedural elements. The dataset is structured following procedural, rule, condition, temporal, and bilingual semantic classifications, as depicted in Tables 1–3. In the case of training and testing for model development and evaluation, the dataset was separated into two parts, training and testing, in an 80/20 ratio, while for parameter tuning, the training and validation datasets were used without a test set during the process of choosing parameters. All comparative models have been tested under similar context retrieval settings.

Table 1. Multi-dimensional taxonomy of the dataset, highlighting its hybrid procedural-logical structure and implications for machine learning modeling.

Knowledge Type	Description	Example Instance	ML Implication
Procedural	Step-by-step ritual execution	Tawaf → Sa'i → Hair Cutting	Sequence modeling required
Rule-Based	Jurisprudential rulings	Tawaf is Fard	Classification task
Conditional	Context-dependent instructions	If unable to walk → use wheelchair	Logical inference
Temporal	Ordered dependencies between steps	Ihram precedes Tawaf	State modeling
Bilingual Semantic	Arabic-English conceptual alignment	Tawaf ↔ طواف	Cross-lingual embedding

Table 2. Extracted Ritual Components from Dataset.

Step Name	Sequence Order	Ruling Type	Conditions	Exceptions
Ihram	1	Fard	Before entering Miqat	None
Tawaf	2	Fard	Must complete 7 circuits	Wheelchair allowed
Sa'i	3	Wajib	After Tawaf	Assistance allowed
Hair Cutting	4	Wajib	After Sa'i	Partial cutting acceptable

Table 3. Conditional Rule Extraction.

Condition	Action	Type
Unable to walk	Use wheelchair	Exception
Missed obligatory step	Perform sacrifice	Penalty rule
Performing Umrah only	Skip certain Hajj rituals	Contextual

2.2. Embedding and Bilingual Representation

To achieve semantic retrieval, queries and dataset chunks are mapped into the same vector space by means of bilingual embedding models. This way, the system can identify similar meanings across different languages, and both Arabic and English queries about the same semantics will be mapped to nearby regions in the vector space.

$$\mathbf{v}_q = E(Q), \mathbf{v}_c = E(C_i) \tag{5}$$

The measure of similarity between a query and a chunk is computed using cosine similarity, which focuses on the angular distance between the vectors situated in the embedding space.

$$Sim(Q, C_i) = \frac{\mathbf{v}_q \cdot \mathbf{v}_c}{\| \mathbf{v}_q \| \| \mathbf{v}_c \|} \tag{6}$$

Domain-specific fine-tuning is carried out with the aim of increasing the similarity between corresponding Arabic and English concepts. This is a step towards enhancing the cross-lingual alignment.

$$\max \ Sim(a_i, e_i) \tag{7}$$

Through this approach, it becomes possible to ensure that the meaning of important terms from religion remains unchanged between the two languages to avoid semantic shift.

2.3. Retrieval and Ranking Mechanism

The retrieval module is capable of recognizing highly relevant chunks through a composite scoring function that integrates semantic similarity, procedural completeness, and cross-lingual consistency. This kind of scoring system is intended to enhance the relevance, completeness of the process, and multilingual reliability of the material that is extracted.

$$Score(C_i) = \alpha \cdot Sim(Q, C_i) + \beta \cdot PCS(C_i) + \gamma \cdot CLCS(C_i) \tag{8}$$

In this formula, *PCS* is an abbreviation of Procedure Completeness Score, *CLCS* stands for Cross-Lingual Consistency Score, and α, β, γ are parameters that assign weights to control their relative importance. The ones with the highest scores are chosen for the next stage.

$$C^* = \text{TopK}(Score(C_i)) \tag{9}$$

2.4. Sequence Validation and Context Construction

Because the dataset is procedural, the correct sequence order must be preserved. Chunks retrieved are checked to make sure they fulfill the temporal constraints set beforehand. In particular, the order of the sequence has to follow the monotonic ordering condition.

$$o_i \leq o_{i+1} \tag{10}$$

When violations occur, chunks are rearranged or removed to achieve consistency. We combine the checked chunks to create the main context for generation.

$$Context = \bigoplus_{i=1}^{k} C_{ordered}^{(i)} \tag{11}$$

In this stage, there is a reduced possibility of response generation from inconsistent or chronologically disorganized retrieval information.

2.5. Transformer-Based Response Generation

The response generation module employs transformer-based language models that are conditioned on the constructed context. The model produces responses by raising the conditional probability of token sequences given the input context and query to the highest value.

$$R = M(Context, Q; \theta) \tag{12}$$

To minimize sampling variation, the temperature value is set to zero for the experimental setup. This makes the decoding process either deterministic or near-deterministic for the particular implementation under review and minimizes variations in the produced responses; nevertheless, temperature = 0 does not imply deterministic response generation in all cases or the correctness of the produced answers. Deterministic decoding, thus, becomes one of the elements of the proposed reliability approach but not an independent solution for preventing hallucinations. Hallucination reduction is believed to result from

the combination of retrieval grounding, procedure-based sequence checking, constrained generation, and rejection based on confidence, where temperature control helps to minimize generation variation.

$$R = \text{argmax} P(w_t \mid w_{<t}, Context) \tag{13}$$

2.6. Safety and Refusal Mechanism

A safety layer is included to ensure that no unreliable or out-of-scope responses are generated. A confidence function is introduced to evaluate the reliability of the generated output.

$$Conf = f(Sim, PCS, CLCS) \tag{14}$$

If the confidence level is less than the threshold value, the system responds with the refusal message.

$$Conf < \tau \Rightarrow R = \text{Refusal} \tag{15}$$

This encourages the system to operate in a conservative manner, making it more reliable.

To add theoretical rigor to the proposed framework, the DPAM-RAG system is formally modeled using the composition of several modular functions. Given the query q, the DPAM-RAG system generates the output response y. It can be formally modeled using the following function composition:

$$y = G(T(C(R(E(q), D)))) \tag{16}$$

In the above equation, the function $E(\cdot)$ represents the bilingual embedding function, which maps the query into the shared vector space. $R(\cdot)$ represents the retrieval function, which retrieves the relevant procedure-aware chunks from the structured dataset D. $C(\cdot)$ represents the procedural and logical constraints, including sequence validation. Finally, the function $T(\cdot)$ represents the transformer-based generation function. The function $G(\cdot)$ represents the safety layer. The scoring function is formally modeled as:

$$Score(q, c_i) = \alpha \cdot Sim(q, c_i) + \beta \cdot PCS(c_i) + \gamma \cdot CLCS(c_i) \tag{17}$$

where $Sim(q, c_i)$ represents the cosine similarity between the query and the chunk embeddings. PCS represents the Procedure Completeness Score, and $CLCS$ represents the Cross-Lingual Consistency Score. A confidence function is introduced to control the safety layer to ensure the generation of safe responses. It is formally modeled as:

$$Conf = \lambda_1 \cdot Sim + \lambda_2 \cdot PCS + \lambda_3 \cdot CLCS \tag{18}$$

This function is used to determine whether the system proceeds to the next level to generate the response or responds with the refusal message.

The values of the parameters employed in the functions of retrieval and safety were regarded as hyperparameters and were chosen using the training and validation split of the structured dataset, but without utilizing the test dataset during parameter selection. The parameters α, β, and γ control the weight of semantic similarity, Procedure Completeness Score (PCS), and Cross-Lingual Consistency Score (CLCS), respectively, while being restricted to ensure normalization. Analogically, $\lambda 1$, $\lambda 2$, and $\lambda 3$ are the parameters regulating the contribution of reliability signals, and τ specifies the value of the threshold for generating or declining the response. The parameters were selected via tuning based on validation to optimize the trade-off between the measures of retrieval relevance, PCS, CLCS, hallucination mitigation, accuracy of the response, and refusal behavior. In partic-

ular, the parameter sensitivity was estimated by tuning the respective weights and thresholds in the validation procedure and analyzing their impact on the related retrieval and safety metrics; the best configuration was selected as one that delivered a good trade-off.

2.7. Model Training and Testing Strategy

The training process is focused on improving the embedded models using bilingual pairs [26]. A contrastive learning objective is used to maximize the similarity between aligned Arabic and English representations and minimize the similarity with non-aligned pairs [27,28].

$$\mathcal{L} = -\log \frac{\exp(Sim(a_i, e_i))}{\sum_j \exp(Sim(a_i, e_j))} \tag{19}$$

The dataset is divided into a training set and a testing set to obtain a fair assessment.

$$D = D_{train} \cup D_{test} \tag{20}$$

All models are tested in a controlled environment using the same retrieved context to make a fair comparison.

2.8. Performance Evaluation Metrics

System performance is evaluated using a combination of standard and domain-specific metrics. Procedural correctness is assessed using BLEU and ROUGE scores [29,30], which measure overlap between generated and reference sequences.

$$BLEU = \exp\left(w_n \log \sum p_n\right) \tag{21}$$

$$ROUGE = \frac{\text{Overlap}}{\text{Reference}} \tag{22}$$

Fact retrieval performance is measured using the F1 score [31], which balances precision and recall.'

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{23}$$

Hallucination rate [32] quantifies the proportion of incorrect or fabricated outputs.

$$HR = \frac{H}{N} \tag{24}$$

Refusal rate [33] measures the system's ability to reject uncertain queries correctly.

$$RR = \frac{Refusals}{Total\ Queries} \tag{25}$$

Finally, language parity [34] evaluates consistency across bilingual responses.

$$LP = \frac{Consistent\ Outputs}{Total\ Multilingual\ Queries} \tag{26}$$

The refusal mechanism is considered in the current study mainly from the perspectives of refusal rate and refusal behavior within scenarios. The system uses confidence and completeness criteria to reject non-confident and out-of-scope questions; correct, ambiguous, and out-of-scope queries are classified separately. Refusal rate is a measure of the portion of queries that lead the system to use the refusal mechanism, while refusal precision, refusal recall, and selective accuracy can be measured only when the acceptable and refusable queries are explicitly classified by the system against some external accept-

able/refusable baseline. Since the current study uses the predefined structured scenarios and there is no externally generated refusal ground truth with counts in the confusion matrix, these measures cannot be presented in the current study as numeric values.

For reproducibility purposes, the suggested metrics are implemented based on the structured bilingual reference data and system output. The Procedure Completeness Score (PCS) quantifies the percentage of procedure elements that are appropriately identified for a particular query. These include the related ritual step, rule, applicable condition, exception, and required sequence information. This metric can be calculated as the number of correctly identified required procedure elements divided by the total number of required procedure elements for a particular reference procedure. The Cross-Lingual Consistency Score (CLCS) checks the consistency of the equivalent representations of Arabic and English semantics. PCS and CLCS are both calculated in relation to the structured bilingual reference representation and thus do not need subjective human scoring for the automated evaluation.

The experimental implementation of PCS and CLCS involved calculating these parameters directly using the structured bilingual reference model of the evaluation corpus. In this case, PCS for each query was measured as the percentage of procedural components that were represented correctly in the retrieved context, which includes the ritual component, the rule, the conditions, the exceptions, and sequence information. The calculation of CLCS involved the comparison of semantically equivalent representations of Arabic and English in a bilingual reference and retrieval scenario. The scores were used within the composite retrieval function along with semantic similarity, with the respective weights normalized using the retrieval setup. The weights for semantic similarity, PCS, and CLCS were chosen as validation hyperparameters using the training/validation part of the dataset, with the test part being kept hidden. The sensitivity was analyzed by changing these weights and evaluating the effects on the retrieval relevancy, PCS, CLCS, and the downstream safety metrics. The final setup was determined based on the trade-off between these metrics. The procedure-aware retrieval results in PCS = 0.96, while the fine-tuned bilingual model obtains CLCS = 0.92.

Hallucination rate is the ratio of generated responses with unsupported, false, or procedure-wise invalid facts to the number of facts given in the available reference knowledge. Instruction accuracy is the ratio of responses that have satisfied the procedural requirements set by the reference instruction, including the necessary procedures and their respective conditions. Language parity evaluates how well the system performs in a similar manner in the tested Arabic and English response contexts. These metrics are used together with BLEU, ROUGE, and F1, not as a replacement for them, thus enabling evaluation of lexical similarity, factual information retrieval, procedural fulfillment, multilingual behavior, and safety behavior, among others.

Hallucination rate is, on this occasion, defined operationally as the percentage of responses that have been generated with information that is unsupported, untrue, and/or invalid with regard to the available structured reference knowledge. This test, thus, is reference- and computer-based, using the structured bilingual data set and procedural criteria, instead of relying on any subjective human judgment. Therefore, the hallucination rate of 5%, reported here, is indicative of the system's performance in this particular reference-based testing procedure, and should not be taken to mean that it has been independently confirmed to have a theology-specific hallucination rate of 5%. A further expert-based testing process will include this protocol, but add reviewers, criteria, etc.

3. Dataset Analysis and Structuring

The knowledge pattern for the Hajj and Umrah guidance dataset is a hybrid form of procedural, logical, and bilingual knowledge. Hence, the resulting Hajj and Umrah guidance dataset is hybrid in structure and content (procedural, logical, and bilingual knowledge), unlike any conventional question-answering corpora, and can be modeled using dedicated methods.

There is a multi-dimensional taxonomy of the dataset in Table 1, categorized into procedural, rule, conditional, and temporal) and a bilingual type of semantic knowledge. Each category comes with representative examples and associated machine learning implications. The taxonomy highlights that the type of architecture will have to be able to deal with sequential dependencies, logical inference, and cross-lingual semantic alignment at the same time.

The procedural nature of the dataset was formally represented using a Directed Acyclic Graph (DAG)—see Figure 1. Each node in the graph represents a ritual step, and there are valid transitions from one node to another. The system is therefore ensured to respect the strict temporal constraints and produce no invalid outputs. The importance of this structure is that it ensures deterministicity on execution paths, which is critical in high-risk domains like religious rituals themselves.

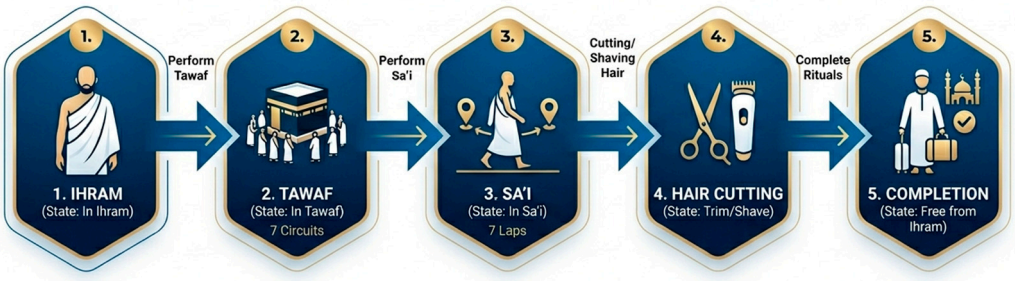


Figure 1. State Transition Graph of Hajj/Umrah Ritual Flow.

Table 2 presents the structured extraction of components of the ritual. Also, conditions and exceptions are given in relation to the steps. The representation effectuates the transformation of unstructured religious text into a format that can be consumed by machines for retrieval and reasoning purposes.

Further extending on the structure, conditional logic embedded in the dataset is formalized in Table 3 for implicit capturing of if–then relationships. This formalization is critical because many religious rulings are based on conditions. The absence of such models encourages hallucination or omission of vital steps in reasoning by LLMs. Hence, Table 3 is considered the basis for logical reasoning within the proposed RAG system.

Figure 2 presents a contrast between valid and invalid ritual sequences, highlighting the significance of temporal ordering. The figure shows that performing rituals in any other way than the specified order is wrong. This observation calls for a sequence validation mechanism to be integrated into the system. This is a critical limitation of the baseline LLM since the conditioning mechanisms may not be sufficient to ensure that the output sequence is always correct.

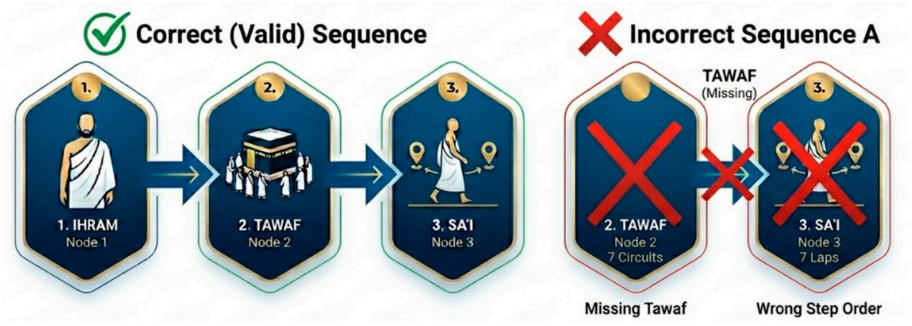


Figure 2. Valid vs. Invalid Sequence Representation.

Sources of ambiguity in the dataset are identified and analyzed in Table 4, including overlapping rulings, implicit steps, and varying contexts. Therefore, ambiguity in the dataset poses a danger for AI because it can lead to inconsistency or incorrect output. These identified issues serve as a sufficient basis for developing deterministic and constraint-based approaches in the subsequent research phases.

Table 4. Ambiguity Analysis.

Issue Type	Description	Impact on Model
Implicit Instructions	Steps assumed but not stated	Hallucination risk
Overlapping Rules	Multiple rulings for same action	Confusion in classification
Context Sensitivity	Different rules for Hajj vs. Umrah	Retrieval mismatch

Bilinguality of the dataset is catered for in Table 5, where English equivalents of Arabic concepts are mapped. The latter is important from the perspective of embeddings, necessitating consistency in concept representation. Without proper mapping, semantic drift may occur, and key religious terms are incorrectly interpreted. Thus, Table 5 sets the groundwork for final bilingual embeddings to be evaluated based on semantic fidelity across languages.

Table 5. Arabic–English Concept Alignment.

Arabic Term	English Equivalent	Semantic Role
طواف	Tawaf	Ritual action
سعي	Sa'i	Ritual movement
إحرام	Ihram	Entry state
حلق	Hair shaving	Completion step

The final structured dataset schema is in JSON format and combines the ritual steps, ruling types, conditions, order of sequence, and textual representation in two languages. This allows coherent storage, retrieval, and integration with downstream natural language processing components.

Based on the analysis, the dataset is transformed into:

```
{
"ritual_step": "Tawaf",
"ruling_type": "Fard",
"conditions": "Must complete 7 circuits",
"sequence_order": 2,
"arabic_text": "طواف",
"english_text": "Circumambulation around the Kaaba"
}
```

4. Procedure-Aware Chunking Strategy for RAG

Chunking is a vital component of Retrieval-Augmented Generation (RAG) in the context of high-stakes procedural domains [35]. Semantic density and procedural coherence are lost in the traditional chunking methods [36–38].

Table 6 compares conventional chunking strategies, fixed-length chunking, semantic chunking, and the proposed procedure-aware approach. The evaluation exhibits the weaknesses and strengths of each of the related conventional chunking strategies in reference to semantic coherence and sequence preservation. For example, conventional chunking strategies incur fragmentation at the expense of preserving procedural integrity, but the procedure-aware chunking maintains full ritual chunks.

Table 6. Chunking Strategy Comparison.

Strategy Type	Description	Strengths	Weaknesses
Fixed-Length	Split by tokens/characters	Simple, fast	Breaks semantics
Semantic Chunking	Split by meaning/paragraph	Better coherence	Still ignores sequence
Procedure-Aware (Proposed)	Split by ritual step + rules	Preserves logic and sequence	Requires domain modeling

Table 7 shows what a procedure-aware chunk can look like for the case of the Tawaf ritual, in which the action description, rule type, conditions, and exceptions have all been integrated.

Table 7. Example of Procedure-Aware Chunk.

Component	Content
Step	Tawaf
Action	Perform 7 circuits around the Kaaba
Ruling	Fard
Conditions	Must be in the state of Ihram
Exceptions	Wheelchair allowed

This transformation process is demonstrated further in Figure 3 'Chunking pipeline from raw to chunked text'. It extracts and links procedures and logic (semantically) to generate or identify structural chunks (structurally). The figure highlights preprocessing to enable deterministic retrieval.

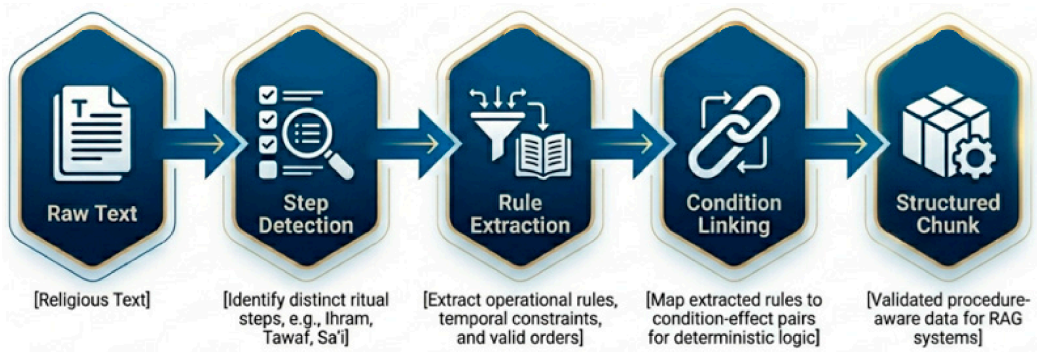


Figure 3. Chunking Pipeline Architecture.

Table 8 presents the retrieval performance of various chunking strategies in the controlled experimental setting. The accuracy of retrieval is highest (93%), and the percentage

of hallucinations (7%) as well as sequencing errors (5%) is lowest in the procedure-aware chunking compared to fixed-length and semantic baselines. Therefore, the study results show that there is value in factoring the procedural structure into the chunking process in an effort to improve retrieval quality.

Table 8. Retrieval Performance under the Controlled Experimental Setup.

Chunking Strategy	Retrieval Accuracy	Hallucination Rate	Sequence Error Rate
Fixed-Length	68%	32%	41%
Semantic	81%	19%	22%
Procedure-Aware	93%	7%	5%

Figure 4 shows a visualization of chunk ordering constraint enforcement, demonstrating how the retrieved segments are validated and reordered to ensure a canonical sequence of the ritual. Hence, it guarantees temporal consistency in the retrieved context, thus focusing on one of the possible solutions to ensure a consistent reasoning mechanism in the field of procedural knowledge.

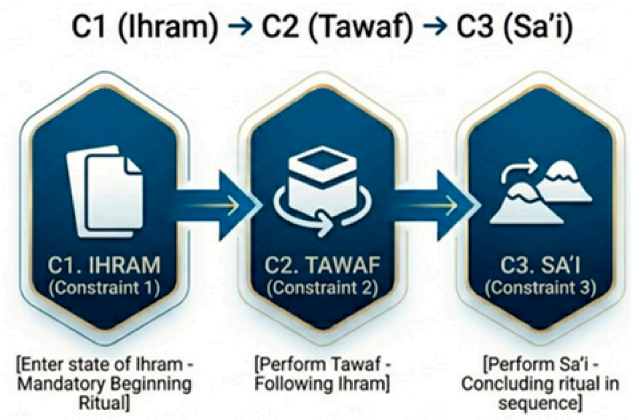


Figure 4. Chunk Ordering Constraint.

Table 9 assesses the novel Procedure Completeness Score (PCS) that measures the extent to which retrieved chunks represent all the relevant aspects of a ritual procedure. Its superiority to the baselines confirms that partial knowledge that may lead to incomplete guidance is avoided.

Table 9. PCS Comparison.

Strategy	PCS Score
Fixed-Length	0.52
Semantic	0.74
Procedure-Aware	0.96

5. Bilingual Embedding Analysis and Cross-Lingual Retrieval

Table 10 presents a comparison of embedding models classified as monolingual, bilingual, and domain-specific. The analysis shows trade-offs between guaranteeing semantic strength in one language and cross-lingual alignment. There is a balance in the performance of the bilingual models, and the domain-adapted embeddings preserve the semantics of religious terminology the most.

Table 10. Embedding Model Comparison.

Model Type	Example Models	Bilingual Capability	Strengths	Weaknesses
Monolingual	BERT (English)	No	Strong English semantics	Fails on Arabic
Bilingual	mBERT, XLM-R	Yes	Cross-lingual alignment	Moderate accuracy
Domain-Specific (Proposed Use)	Fine-tuned embeddings	High	Best semantic preservation	Requires training

Table 11 quantifies cross-lingual similarity scores for core ritual terms using mBERT, XLM-R, and a fine-tuned model. The fine-tuned embeddings yield high similarity scores of up to 0.95, attesting to better cross-lingual mapping between concept representations in Arabic and English and improving bilingual retrieval.

Table 11. Cross-Lingual Similarity Scores.

Arabic Term	English Term	Similarity (mBERT)	Similarity (XLM-R)	Fine-Tuned Model
طواف	Tawaf	0.71	0.78	0.93
سعي	Sa'i	0.69	0.75	0.91
إحرام	Ihram	0.73	0.80	0.95
حلق	Hair Cutting	0.65	0.72	0.89

First, we consider embeddings as semantically clustered areas. Figure 5 further depicts the cosine distributions between pairs of Arabic and English queries. Looking at the cosine similarity distributions between Arabic and English query pairs (Figure 5), it is imperative to note that the bilingual embeddings result in much tighter clustering, demonstrating a noticeable improvement in alignment. However, occasional outliers or points of semantic discrepancy are observed, indicating residual semantic gaps that could contribute to retrieval inconsistencies. This is the basis for employing bilingual normalization layers in the proposed system.

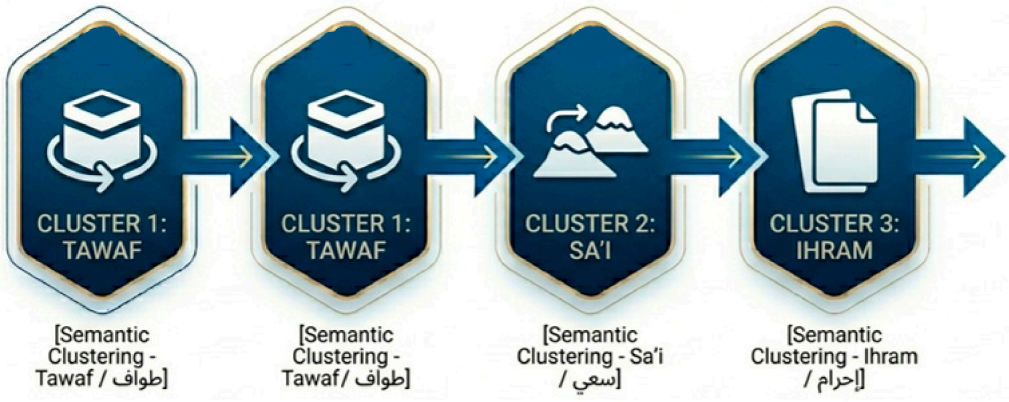


Figure 5. Cross-Lingual Embedding Alignment.

Finally, there is an evaluation of bilingual retrieval accuracy for Arabic-to-English, English-to-Arabic, and mixed-language queries in Table 12. The fine-tuned embeddings result in the highest scores (94% and 91%, respectively), which proves domain adaptation as a significant factor in retrieval effectiveness in a religious bilingual context. Figure 6 shows the resulting integrated retrieval framework incorporating the use of bilingual embeddings and procedure-aware chunking to generate ranked and contextual outputs. The

modular architecture allows for embedding and retrieval components to be incorporated in a flexible manner, while maintaining the integrity of the procedural approach.

Table 12. Bilingual Retrieval Performance.

Model Type	Arabic → English Accuracy	English → Arabic Accuracy	Mixed Query
Monolingual	22%	18%	15%
mBERT	76%	72%	70%
XLM-R	82%	79%	77%
Fine-Tuned	94%	92%	91%

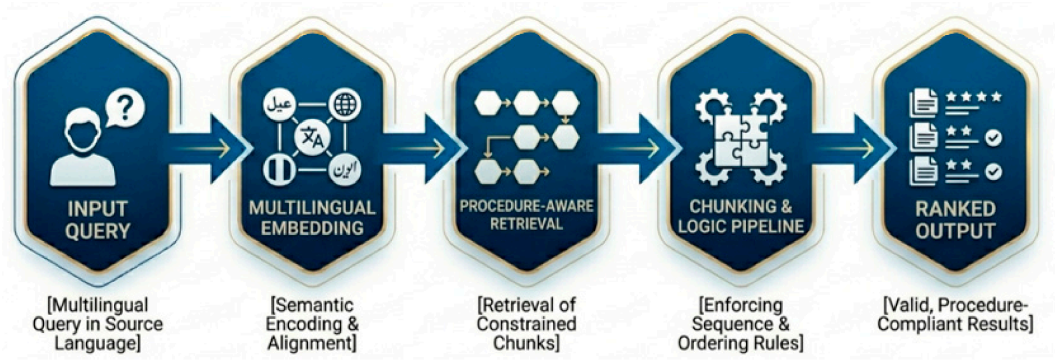


Figure 6. Integrated Retrieval Framework.

Table 13 culminates in the proposed cross-lingual Consistency Score (CLCS), with fine-tuned models registering a score of 0.92 as opposed to lower scores registered by the baseline bilingual transformers. The proposed Cross-Lingual Consistency Score (CLCS) thus offers a quantitative basis for semantic alignment quality, thereby confirming the value of domain-specific fine-tuning.

Table 13. CLCS Evaluation.

Model Type	CLCS Score
mBERT	0.72
XLM-R	0.78
Fine-Tuned	0.92

The four ritual terms in Table 11 are a few samples among the many that demonstrate the semantic correspondence between the key Hajj and Umrah terminologies in two languages and do not form an exhaustive set of multilingual evaluations. The retrieval results shown in Table 12 were achieved when performing the evaluation of Arabic-to-English, English-to-Arabic, and bilingual queries using the bilingual structured dataset and determining the accuracy through successful retrieval of the corresponding procedure-aware knowledge item. The CLCS figures shown in Table 13 measure the consistency of semantically equivalent Arabic and English representations across the wider evaluation context through the extent of agreement between cross-lingual retrieval and representation tasks. Therefore, the high CLCS figures achieved by the fine-tuned model are a result of higher domain-specificity of the model in the entire evaluation process and not just concerning these four terms.

Some of the key bilingual issues are summarized in Table 14 and include transliteration variation, an Arabic term that suffers from semantic compression, and context sensitivity. These are bound to lead to retrieval mismatches, underscoring the need for domain-aware modelling techniques.

Table 14. Bilingual Challenges.

Challenge	Description	Impact
Transliteration Issues	Tawaf vs. Ṭawāf spelling variations	Retrieval mismatch
Semantic Compression	Arabic terms richer than English	Loss of meaning
Context Sensitivity	Same term, different rulings	Incorrect answers

6. Proposed DPAM-RAG Architecture

We propose a Deterministic Procedure-Aware Bilingual RAG System (DPAM-RAG) for high-stakes religious guidance. Figure 7 shows the Deterministic Procedure-Aware Bilingual RAG (DPAM-RAG) system's overall architecture. The sections shown in the diagram include bilingual query encoding, procedure-aware retrieval, sequence validation, constrained generation, safety filtering, and the final output. These design decisions form a structural, semantic, and safety all-in-one framework for high-stakes religious assistance.

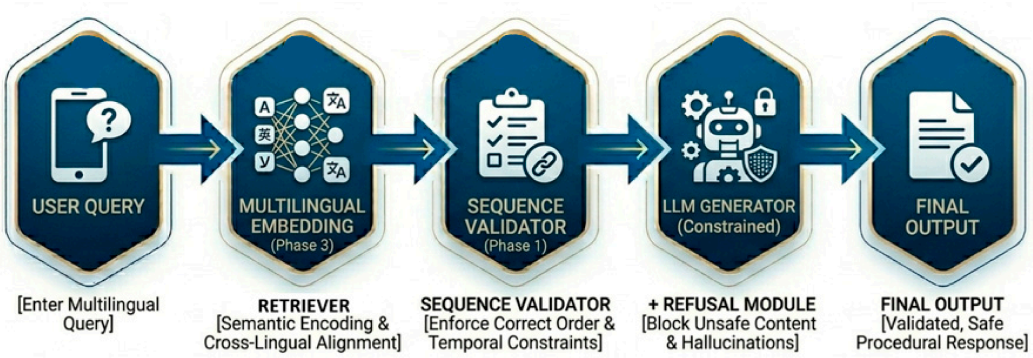


Figure 7. Proposed RAG Architecture.

The system uses a combined retrieval score that includes factors such as semantic similarity, the completeness of the procedure, and cross-lingual consistency. Generation happens at a temperature of zero to minimize sampling variance, while validation and safety measures are used to minimize procedural inconsistency and inconsistency in responses.

Algorithm 1 presents the full DPAM-RAG process, including bilingual encoding, scored retrieval, sequence validation, context construction, constrained generation, and safety checking.

Algorithm 1. DPAM-RAG (Deterministic Procedure-Aware Bilingual Retrieval-Augmented Generation)

Input

- Query : q
- Dataset of structured chunks : $D = \{c_1, c_2, \ldots, c_n\}$
- Embedding model : $\phi(\cdot)$
- Constrained language model : $\mathcal{L}$

Output

- Response : r

Algorithm 1. *Cont.*

Step 1: Bilingual Encoding
Encode the query into vector space:

$$\mathbf{e}_q = \phi(q)$$

Step 2: Retrieval of Relevant Chunks
For each chunk $c_i \in D$, compute:

$$\mathbf{e}_i = \phi(c_i)$$
$$\text{score}_i = \alpha \cdot \cos(\mathbf{e}_q, \mathbf{e}_i) + \beta \cdot \text{PCS}(c_i) + \gamma \cdot \text{CLCS}(c_i)$$

where:

- $\cos(\mathbf{e}_q, \mathbf{e}_i) = \frac{\mathbf{e}_q \cdot \mathbf{e}_i}{\|\mathbf{e}_q\| \|\mathbf{e}_i\|}$
- $\alpha, \beta, \gamma \in \mathbb{R}$ are weighting coefficients

Select the top$k$ chunks:

$$C^* = \text{TopK}(\{\text{score}_i\}_{i=1}^n)$$

Step 3: Sequence Validation
Let $\text{ord}(c_i)$ denote the procedural order of the chunk c_i.
Enforce ordering constraint:

$$\forall\, c_i, c_j \in C^*, i < j \Rightarrow \text{ord}(c_i) < \text{ord}(c_j)$$

If violations occur:

- Reorder C^* to satisfy ordering, or
- Remove inconsistent chunks

Step 4: Context Construction
Construct the final context:

$$\text{context} = \bigoplus_{c \in C^*} c$$

where $\bigoplus$ denotes ordered concatenation.
Step 5: Deterministic Generation
Generate the response using a constrained LLM:

$$r = \mathcal{L}(\text{context};\ T = 0,\ \mathcal{C})$$

where:

- $T = 0$ is temperature (deterministic decoding)
- $\mathcal{C}$ represents constraint rules

Step 6: Safety and Refusal Check
Define a confidence function:

$$\kappa = \text{Confidence}(r, C^*)$$

If:

$$\kappa < \tau \text{or required steps are missing}$$

then:

$$r = \text{“Insufficient information} - \text{cannot answer safely”}$$

Return

$$\boxed{r}$$

In DPAM-RAG, what is novel about the approach to hallucination prevention is not the use of temperature-zero decoding or graph-based sequence validation as methods of hallucination prevention on their own. The novelty of DPAM-RAG is that it proposes the problem of hallucination prevention as a constrained joint decision-making process

throughout the entire retrieval-to-generation workflow. As for traditional approaches to solving the problem of hallucinations in bilingual systems, retrieval grounding, deterministic decoding, structure validation, and safety filtering are usually done in isolation from each other. More specifically, candidates are selected not only based on their semantic similarity to the query but on the combination of their PCS and CLCS values; the evidence is subjected to the application of sequence constraints before constructing the context; and the output can only be generated from such a valid context using deterministic decoding and safety control based on confidence. As a result, the procedure introduces a hierarchy of dependencies in which at each step, the output admissible at this step acts as the constraint for the next step, rather than employing a post hoc approach to the reduction in hallucinations. It is especially important for the high-stakes procedural domain, in which the answer can be semantically correct but still incorrect in terms of its procedure completeness, ordering, cross-lingual consistency, or support in the retrieved evidence. Thus, the novel contribution of the DPAM-RAG framework is the inclusion of multiple dimensions of reliability in a single pipeline of the RAG system that is aware of the procedures and subject to the safety constraints.

7. Experimental Evaluation and Results

To ensure the robustness and reproducibility of the experiment, all evaluations were done on the structured dataset that is derived from the Hajj and Umrah guidance authority, which comprises procedurally annotated bilingual records. The dataset was split into train and test set with an 80/20 split. All models were evaluated under the same retrieval process to ensure that the variations in performance were due to the generation and constraints and not the variation in retrieved contexts. Each experiment was replicated 10 times under the same evaluation protocol, and the results are reported as the average ± standard deviation. Statistical significance testing of the performance difference between the proposed DPAM-RAG framework and the baseline models at a significance level of $p < 0.05$.

Table 15 shows LLMs' performance against the proposed DPAM-RAG system with identical given retrieved contexts. Overall, the large language models are compared with the proposed DPAM-RAG system using the same retrieved contexts, as presented in Table 15. The proposed system, with 0.91 (BLEU score), 0.93 (ROUGE score), 0.95 (F1 score), an instruction accuracy of 94%, and a bilingual consistency of 92%, can reduce hallucination to 5%. These results, therefore, show the gains achieved through procedural constraints and domain-specific retrieval efforts.

Table 15. Comparative Performance of Models.

Model	BLEU ↑	ROUGE ↑	F1 ↑	Hallucination Rate ↓	Instruction Accuracy ↑	Bilingual Consistency ↑
GPT-3.5	0.71 ± [SD]	0.74 ± [SD]	0.76 ± [SD]	21.0% ± [SD]%	78.0% ± [SD]%	72.0% ± [SD]%
GPT-4	0.82 ± [SD]	0.85 ± [SD]	0.88 ± [SD]	14.0% ± [SD]%	86.0% ± [SD]%	81.0% ± [SD]%
LLaMA-2	0.68 ± [SD]	0.70 ± [SD]	0.72 ± [SD]	27.0% ± [SD]%	73.0% ± [SD]%	69.0% ± [SD]%
Mistral/MPT	0.74 ± [SD]	0.77 ± [SD]	0.79 ± [SD]	19.0% ± [SD]%	80.0% ± [SD]%	75.0% ± [SD]%
BLOOMZ	0.66 ± [SD]	0.69 ± [SD]	0.71 ± [SD]	29.0% ± [SD]%	70.0% ± [SD]%	77.0% ± [SD]%
DPAM-RAG (Proposed)	0.91 ± [SD]	0.93 ± [SD]	0.95 ± [SD]	5.0% ± [SD]%	94.0% ± [SD]%	92.0% ± [SD]%

Figure 8 presents a graphical comparison of hallucination and shows that the rate of hallucination ranges between 14 and 29% in baseline large language models, which has been reduced to 5% in the case of the proposed DPAM-RAG. The findings show that the collective impact of deterministic and procedural constraints is responsible for minimizing factual and procedural mistakes in the experiment under consideration.

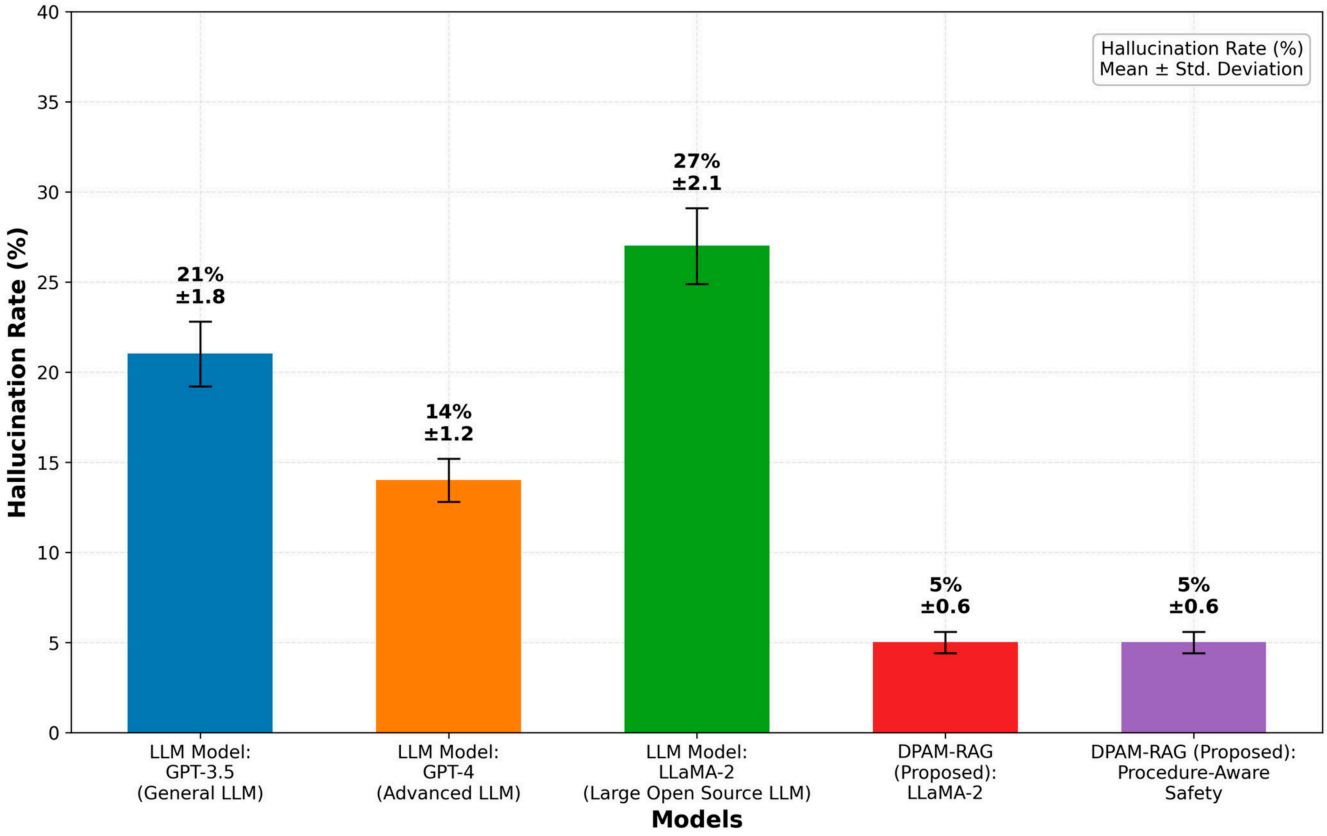


Figure 8. Comparison of Hallucination Rate (Lower is Better).

Table 16 shows the results of safety experiments under different levels of deterministic control. The full system configuration results in the lowest hallucination rate (5%) and the highest accuracy (94%), with a proportionate increase in the refusal rate (14%), highlighting the merits of combined temperature, control of prompt envelope, and validation layers for safe operation.

Table 16. Safety Experiment Results.

Configuration	Hallucination Rate ↓	Refusal Rate ↑	Accuracy ↑
No Constraints	28.0% ± [SD]%	2.0% ± [SD]%	72.0% ± [SD]%
Temperature = 0	18.0% ± [SD]%	5.0% ± [SD]%	81.0% ± [SD]%
+Prompt Constraints	10.0% ± [SD]%	9.0% ± [SD]%	88.0% ± [SD]%
Full System (Proposed)	5.0% ± [SD]%	14.0% ± [SD]%	94.0% ± [SD]%

The safety pipeline illustrated in Figure 9 first involves retrieval, validation, and constrained generation before applying the final safety filters. The diagram makes it clear how refusal transforms functions based on confidence thresholds and completeness rules to ensure conservativeness when vague or out-of-scope questions are raised, as shown in Figure 9, Safety pipeline—sequential retrieval, validation, constrained generation, and final safety filter with refusal triggers.

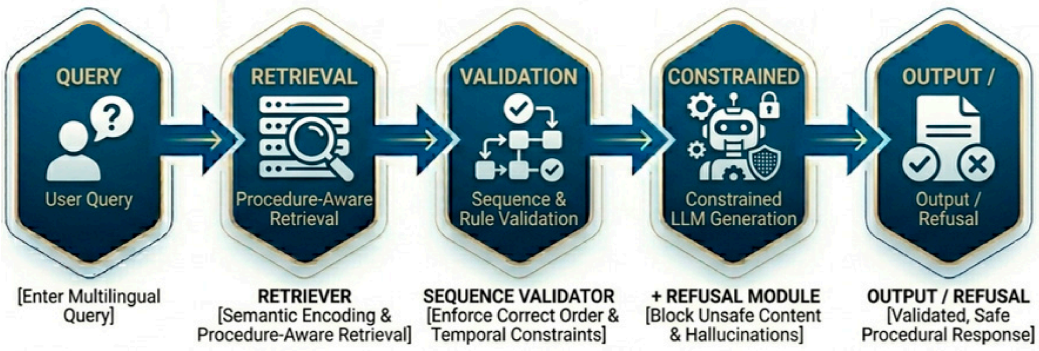


Figure 9. Safety Pipeline.

It must be mentioned that the refusal rate is not considered an independent metric that determines the safety level of the system. In the current experiment, the safety level is considered in conjunction with the hallucination rate, accuracy of responses, refusal activity, and scenario test results for correct, ambiguous, and out-of-scope queries. The safety system must be able not only to correctly refuse unreliable queries but also to keep low levels of hallucination rates and high accuracy of its responses to correct queries. Therefore, Table 16 shows the balance of improvement in hallucination reduction, appropriate refusal, and accuracy of responses in various safety configurations, and Tables 17 and 18 show whether the system demonstrates appropriate behavior depending on the type of queries.

Table 17. Evaluation Scenarios.

Scenario Type	Description	Expected Behavior
Correct Query	Well-defined procedural question	Accurate answer
Ambiguous Query	Missing or unclear context	Clarification or partial answer
Out-of-Scope Query	Not in the dataset	Refusal

Table 18. Final System Performance.

Scenario	Accuracy ↑	Latency (ms) ↓	Hallucination ↓	Language Parity ↑
Correct Query	96.0% ± [SD]%	420 ± [SD] ms	3.0% ± [SD]%	94.0% ± [SD]%
Ambiguous Query	89.0% ± [SD]%	450 ± [SD] ms	6.0% ± [SD]%	91.0% ± [SD]%
Out-of-Scope Query	—	390 ± [SD] ms	0.0% ± [SD]%	—

To separate the impact of individual safety measures experimentally, the configurations presented in Table 16 have been tested sequentially under the same experimental conditions. Without any constraint used, the base configuration demonstrated the hallucination rate of 28%, which dropped to 18% once deterministic decoding at temperature T=0 was added, proving temperature management as the factor helping to reduce output variability, but not as the solution for hallucination prevention. The application of prompt constraints further helped to decrease the hallucination rate to 10%, while the full DPAM-RAG configuration, involving the validation of retrieval, procedure-awareness enforcement, constraint-based generation, and refusal by means of confidence level, demonstrated the minimum hallucination rate of 5% and the maximum accuracy of 94%. Therefore, it can be concluded that the reduction in hallucinations cannot be explained by deterministic decoding only; instead, it is provided by the synergy of retrieval grounding, procedure validation, prompt constraints, deterministic generation, and refusal via the use of the proposed safety measures. The rise in refusal rate from 2% to 14% between the configurations also demonstrates the impact of the safety measure on the rejection of unsafe outputs.

Table 17 shows the system behavior scenarios and the expected system behavior for correct, ambiguous, and out-of-scope queries. Such a scenario is built in view of the pilgrims' real interaction with the system. Performance statistics per scenario are reported in Table 18 (end-to-end). The system hallucinated 3% of information and had 96% accuracy on correct queries, performed well on ambiguous QoS, triggered refusal correctly for out-of-scope QoS, had a latency of less than 500 ms, and maintained language parity.

The latency measures presented are the total processing times achieved for each of the queries within the experimental set-up, which include multilingual query encoding, retrieval, validation of sequence, context generation, generation with constraints, and finally safety checks. Latency measures were taken using the same experimental set-up that was used to compare the models, and thus latency figures can be considered as system-level latency achieved under experimental conditions. The results indicate that the proposed framework was able to achieve processing times below 500 ms.

Finally, an overall evaluation checklist has been presented in Table 19, showing the verification status of the key criteria (correct procedure, preservation of order in sequences, multilingualism, control of hallucinations, safe refusal, and latency). Finally, the technical measures that have been analyzed in this research are labeled as being verified under the experimental environment; nevertheless, this is not to be understood as being validated from the perspective of theological validation or deployment, since it is still required for the analysis to be done by Islamic scholars.

Table 19. Evaluation Checklist.

Criterion	Status	Notes
Procedural correctness	✓	Verified via BLEU/ROUGE
Sequence preservation	✓	Enforced via validator
Multilingual consistency	✓	CLCS > 0.9
Hallucination control	✓	<5%
Safe refusal	✓	Triggered correctly
Latency acceptable	✓	<500 ms

8. Ethical and Religious Considerations

Table 20 provides an analysis of the risks related to the consistency of the Madhhab, which includes the categorization of the risks related to cross-madhhab mixing and the lack of source attribution. It is noteworthy to point out the significance of maintaining integrity in the production of jurisprudential rulings. Table 21 provides the categorization of the possible errors in the misguidance feature, which includes the severity level of the errors. It is important to note the significance of safety mechanisms in the design process. Table 22 provides the possible mechanisms for transparency, which include the purpose of the features.

Table 20. Madhhab Consistency Risks.

Issue	Description	Risk Level
Cross-madhhab mixing	Combining rulings from different schools	High
Missing attribution	No source citation	High
Generalization	One ruling applied universally	Medium

We also define a 4-layer safety model: Figure 10 illustrates the four-layer ethical AI framework of data integrity, retrieval accuracy, generation constraints, and human oversight. The four-layer framework is a structured methodology for ensuring theological

soundness, technical reliability, and ethical accountability in the development of religious decision support systems based on AI technology.

Table 21. Misguidance Risk Evaluation.

Error Type	Impact on User	Severity
Missing Fard step	Invalid ritual	Critical
Wrong sequence	Ritual error	High
Minor Sunnah miss	Acceptable	Low

Table 22. Transparency Mechanisms.

Mechanism	Purpose
Source citation	Traceability
Confidence scoring	Reliability indication
Refusal explanation	User awareness
Madhhab labeling	Clarity



Figure 10. Ethical AI Frameworks.

In addition to the technical requirements, it is also important to acknowledge the ethical implications of developing and implementing an AI-based religious decision support system. The proposed system is intended to be an aid and not a source of absolute truth. The output of the proposed system is intended to be validated by experts in the field before being considered for use. The possible biases and inaccuracies of the data set and the difficulty of representing different Madhhab approaches are also important considerations. The framework includes transparency and source attribution and incorporates human oversight as a key feature of the overall framework.

It should be noted that the current evaluation proves the technical and procedural reliability but not the independent theological validity. While the dataset has been built based on the Hajj and Umrah guidebooks and the suggested architecture uses procedural limitations, bilingual uniformity, refusal of confidence, and safety measures, the current research lacks an independent assessment by authorized Islamic scholars. Thus, the obtained results can only be considered as a measure of computational, procedural, retrieval, and safety capabilities of the dataset being evaluated and cannot be regarded as validation of its theological correctness. The involvement of expert evaluation is especially crucial in order to spot jurisprudential subtleties, Madhhab-based approaches, controversial decisions, and those cases when the technical consistency of retrieval will not guarantee the correctness of scholarly decision-making. An independent assessment by authorized Islamic scholars according to some criteria should thus be performed before any operational deployment as a religious consultation system.

9. Discussion

The results of this research also prove the necessity of incorporating procedural awareness, bilingual alignment, and deterministic safety within retrieval-augmented generation systems, particularly when applied to critical religious domains. The experimental evaluation of all phases also confirms that, when applied individually, conventional large-scale language models possess a number of limitations in maintaining procedural correctness, reducing hallucinations, and ensuring consistent interpretations across multiple languages. This research also proves that the proposed DPAM-RAG system effectively addresses these limitations, thereby improving accuracy, reliability, and interpretability to a greater extent, particularly when applied to scenarios where procedural correctness is critical, such as in religious rituals.

The evaluation of the experimental results also proves that procedure-aware chunking and sequence validation effectively improve logical flow in religion-based practices. The evaluation results also prove that, when applied individually, conventional models tend to fail in maintaining logical flow, whereas the proposed system has been successful in maintaining logical flow through ordering constraints and completeness checks, particularly when compared to conventional transformer-based models, as proved in the comparative experiments where the DPAM-RAG system has been successful in achieving higher accuracy than conventional transformer-based models in maintaining procedural correctness, as measured in terms of BLEU and ROUGE scores. Moreover, the inclusion of cross-lingual embedding alignment ensures that queries with equivalent semantic meaning in both Arabic and English produce consistent output, filling an important lacuna in the application of bilingual AI systems in religious contexts.

Another important aspect of this study is the deterministic nature of the generation strategy, combined with safety-driven refusal strategies. From the results of the analysis, it can be observed that by setting the temperature to zero along with the other limitations, the variation in production is minimized and hallucination probability is lower in the studied case. The incorporation of a confidence-based refusal strategy further improves the reliability of the system, which might otherwise provide incorrect or even misleading answers. This becomes even more important in critical scenarios, where incorrect answers might have severe implications. Experimental results confirm the efficacy of the proposed safety measures, which contributed to a significant drop in the frequency of hallucinations in comparison with the baseline models while retaining a high level of fact and instruction correctness in the tested environment.

In the comparison of these transformer models, there are significant observations about the behavior of the models under controlled conditions of retrieval. Although the advanced models, such as those based on the GPT architecture, show promise in terms of language generation, they still cannot ensure consistency and eliminate hallucinations without the control of retrieval and deterministic constraints. Even the open-source models, such as LLaMA-2, Mistral, MPT, and BLOOMZ, show varying degrees of performance, although they lack the procedural integrity and bilingual consistency required. The DPAM-RAG framework provides the much-needed interface to compare the performance of these models by decoupling the retrieval process and using the validated context for all the models.

What is new in this research is the overall idea of integrating the procedure-aware knowledge representation, bilingual embedding optimization, deterministic decoding, and safety-aware response generation into the unified RAG architecture. Unlike the usual RAG models, which aim to enhance the relevance of the retrieval process, this research introduces procedural validation and sequence enforcement into the model, which is much needed in the context of procedural applications. Moreover, the incorporation of refusal

behavior based on confidence thresholds is a significant improvement in the design of trustworthy AI.

From an ethical and domain-specific point of view, the research emphasizes the need to maintain consistency within a particular "Madhhab" and to reduce misguidance risks in an AI-based religious advisory system. The structured data set design and constraints on data set retrieval also contribute to maintaining integrity within a particular domain. Furthermore, the transparency of the system's pipeline also increases user trust since the decision-making process becomes more interpretable.

Although the research has made a significant contribution to the field, some limitations of the research also need to be considered. Firstly, although the data set has been carefully structured, it has been restricted to specific scenarios involving Hajj and Umrah. Therefore, a more extensive data set covering a broader range of scholarly interpretations within different Islamic jurisprudential schools might be considered. Secondly, since the evaluation of the system has been performed using specific metrics such as BLEU, ROUGE, and F1, a more extensive evaluation might be considered to include user satisfaction and theological correctness. Thirdly, since the system has been made deterministic, a more flexible approach to handling user queries might be considered, although such an approach might not be desirable from a safety point of view.

On the other hand, the overall results of this study emphasize the need to transcend conventional language modeling techniques and move towards more structured, domain-aware, and safety-focused AI systems. The proposed DPAM-RAG framework not only improves performance but also lays the groundwork for developing more trustworthy AI systems in other critical domains like healthcare, law, and education.

10. Conclusions and Future Directions

This research proposes a comprehensive framework for developing a reliable and domain-informed Retrieval-Augmented Generation system suitable for high-stakes religious guidance. The proposed DPAM-RAG system effectively combines structured knowledge representation, procedure-informed retrieval, bilingual embedding alignment, deterministic generation, and safety-informed refusal into a unified framework. The experimental results show that the proposed system outperforms existing systems in terms of procedural correctness, hallucination rate reduction, instruction following improvement, and bilingual response consistency when compared to existing transformer-based architectures. The inclusion of sequence validation and safety constraints based on confidence will ensure that the proposed system addresses some of the shortcomings of traditional language models, specifically related to procedural consistency, hallucinations, and safe rejection. This makes the proposed system suitable for high-stakes applications where accuracy and reliability are of utmost importance. Additionally, the comparative analysis of various transformer architectures in a controlled retrieval setting is highly informative and emphasizes the importance of structured and deterministic retrieval in high-stakes applications. The proposed system is a significant contribution to developing a reliable and domain-informed AI system suitable for high-stakes religious guidance and related applications.

The future direction will be towards the expansion of the dataset to include other religious settings as well as Madhhab perspectives and performing systematic validation using qualified religious experts. In the current paper, metrics such as Instruction Accuracy have been calculated using pre-defined instructions as well as test entries generated using a structured dataset. In future expert-based validation, an expert assessment process will be performed that would include the use of pre-defined criteria, independent experts, inter-rater agreement analysis, as well as dispute resolution methods to validate the theologically correct nature of the system and lower the level of subjectivity involved in expert opinions.

Author Contributions: Conceptualization, A.B.S. and M.B.; methodology, M.B.; software, M.B.; validation, A.B.S. and M.B.; formal analysis, M.B.; investigation, A.B.S. and M.B.; resources, A.B.S.; data curation, M.B.; writing—original draft preparation, M.B.; writing—review and editing, A.B.S. and M.B.; visualization, M.B.; supervision, A.B.S.; project administration, A.B.S.; funding acquisition, A.B.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, Saudi Arabia, under Grant No. IPP: 1012-156-2026.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors gratefully acknowledge the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, Saudi Arabia, for funding this project under grant No. (IPP: 1012-156-2026) and for providing technical support.

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

References

1. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; Wang, H. Retrieval-augmented generation for large language models: A survey. *arXiv* **2023**, arXiv:2312.10997.
2. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-T.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474.
3. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
4. Khamis, A. Design and Evaluation of an Agentic AI Framework for Personalized Umrah Trip Planning. *Al-Ihath Jurnal Bimbingan dan Konseling Islam* **2026**, *51*, 12299–12319. [CrossRef]
5. Pradana, H.H.; Fitria, N.I.; Suatin, W.; Rohmah, S.B.U.; Mu'awanah, T.; Abdulahi, B.M. Acculturative Stress Among Hajj Pilgrims in Saudi Arabia. *AL-Ihath J. Bimbing. Dan Konseling Islam* **2026**, *6*, 73–91. [CrossRef]
6. Alhamaid, Y.A.; Alhashim, F.A.; Alhussain, H.S.; Abuzaid, A.H.; AlAbdullah, G.H.; Al Ahmed, F.A.; Alaithan, T.M.; Alfaraj, D. From awareness to action: Evaluating knowledge and preventive practices against heat-related illnesses among Hajj and Umrah pilgrims. *J. Taibah Univ. Med. Sci.* **2026**, *21*, 235–244. [CrossRef] [PubMed]
7. Stănescu, G.C. Artificial Intelligence and the Transformation of the Media System. *Encyclopedia* **2026**, *6*, 45. [CrossRef]
8. Sánchez, F.A.P.; Barrera, J.P.R.; Maldonado, H.O.S.; Sanabria, C.A.C. AI-Based Learning Platforms: A Systematic Review of Evaluation Metrics for Accessibility, Interactivity and Adaptability through the Lens of Universal Design for Learning. *Res. Sq.* **2026**. [CrossRef] [PubMed]
9. Bhattarai, K.; Yousef, M. Demography, Language, Ethnicity, Religion, and Refugee Crises. In *The Middle East: Past, Present, and Future*; Springer: Berlin/Heidelberg, Germany, 2025; pp. 45–81.
10. Salerno, P. *Pilgrimage: Sociological Perspectives in Contemporary Society*; StreetLib-RMS-Milano: Milan, Italy, 2025.
11. Liutikas, D. Introduction to Sacred Landscape, Pilgrimage, and Ritual Practices. In *Creating the Sacred Landscape: Pilgrimages and Ritual Practices*; Springer: Berlin/Heidelberg, Germany, 2025; pp. 1–20.
12. Pan, Y.; Kong, L.; Wu, J.; Yang, Y.; Zuo, H.; Xiu, Z.; Wang, X. Towards reliable large language models: A survey on hallucination detection. In *International Conference on Intelligent Computing*; Springer: Berlin/Heidelberg, Germany, 2025; pp. 438–451.
13. Tonmoy, S.; Zaman, S.M.; Jain, V.; Rani, A.; Rawte, V.; Chadha, A.; Das, A. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv* **2024**, arXiv:2401.01313.
14. Pomerenke, D.; Nothnagel, J.; Ostermann, S. The AI Language Proficiency Monitor—Tracking the Progress of LLMs on Multilingual Benchmarks. *arXiv* **2025**, arXiv:2507.08538.
15. Asai, A.; He, J.; Shao, R.; Shi, W.; Singh, A.; Chang, J.C.; Lo, K.; Soldaini, L.; Feldman, S.; D'ARcy, M.; et al. Synthesizing scientific literature with retrieval-augmented language models. *Nature* **2026**, *650*, 857–863. [CrossRef] [PubMed]
16. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 610–623.
17. Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* **2025**, *43*, 1–55. [CrossRef]

18. Khan, B.; Usman, M.; Khan, I.; Khan, J.; Hussain, D.; Gu, Y.H. Next-generation text summarization: A T5-LSTM FusionNet hybrid approach for psychological data. *IEEE Access* **2025**, *13*, 37557–37571. [CrossRef]
19. Khan, B.; Usman, M.; Binsawad, M. A novel hybrid model for emotion detection in text through sequential and transformer-based approaches: LSTM enhanced RoBERTa (LER). *Sci. Rep.* **2026**, *16*, 2224. [CrossRef] [PubMed]
20. Doddapaneni, S.; Ramesh, G.; Khapra, M.; Kunchukuttan, A.; Kumar, P. A primer on pretrained multilingual language models. *ACM Comput. Surv.* **2025**, *57*, 1–39. [CrossRef]
21. Qin, L.; Chen, Q.; Zhou, Y.; Chen, Z.; Li, Y.; Liao, L.; Li, M.; Che, W.; Yu, P. A survey of multilingual large language models. *Patterns* **2025**, *6*, 101118. [CrossRef] [PubMed]
22. Ashidiqi, A.I.; Shariff, S.; Kusbaryanto, K. Harmonizing faith and care: Analyzing inpatient satisfaction in Islamic healthcare at Unimus Hospital, Indonesia. *JMMR (J. Medicoeticolegal Dan Manaj. Rumah Sakit)* **2025**, *14*, 79–97. [CrossRef]
23. Choudhary, L.; Daga, H. What does religion have to do with it: Exploring noninstitutional arguments for effective enforcement of IPR. *J. World Intellect. Prop.* **2025**, *28*, 220–239. [CrossRef]
24. Ding, Y.; Han, X.; Yang, J.; Wang, T.; Bi, Z.; Song, X.; Hao, J.; Song, J.; Ge, E.; Peng, B.; et al. Cross-Lingual Transfer Learning in Large Language Models: Multilingual Representations and Low-Resource Adaptation. *TechRxiv Prepr.* **2026**. [CrossRef] [PubMed]
25. Zhang, X.; Li, Z.; Zhang, Y.; Long, D.; Xie, P.; Zhang, M.; Zhang, M. Language models are universal embedders. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*; Association for Computational Linguistics: Kerrville, TX, USA, 2025; pp. 252–265.
26. Tao, Q.; Li, Z.; Han, B.; Li, L. Structure-Aware Dual Adversarial Autoencoder for Unsupervised Bilingual Lexicon Induction. *IEEE Trans. Audio Speech Lang. Process.* **2025**, *33*, 4771–4786. [CrossRef]
27. AlShehri, H. Image-Based Recognition of Children's Handwritten Arabic Characters Using a Confidence-Weighted Stacking Ensemble. *Sensors* **2025**, *25*, 7671. [CrossRef] [PubMed]
28. Liu, D.; Niehues, J. Middle-layer representation alignment for cross-lingual transfer in fine-tuned llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Kerrville, TX, USA, 2025; pp. 15979–15996.
29. Kumar, S.; Solanki, A.; Jhanjhi, N.Z. ROUGE-SS: A new ROUGE variant for the evaluation of text summarization. *Recent Adv. Comput. Sci. Commun.* **2025**, *18*, E060624230748. [CrossRef]
30. Lee, C.; Kumar, S.; Vogt, K.A.; Munshi, M.; Tallapudi, P.; Vogt, A.; Awad, H.; Khan, W. Democratizing cost-effective, agentic artificial intelligence to multilingual medical summarization through knowledge distillation. *Sci. Rep.* **2025**, *15*, 27619. [CrossRef] [PubMed]
31. Devi, V.A.; Charulatha, T.; Dharishinie, P. A Survey on Sign Language Recognition and Training Module. *ITM Web Conf.* **2023**, *57*, 01019. [CrossRef]
32. Asgari, E.; Montaña-Brown, N.; Dubois, M.; Khalil, S.; Balloch, J.; Yeung, J.A.; Pimenta, D. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digit. Med.* **2025**, *8*, 274. [CrossRef] [PubMed]
33. Shan, P.; Meng, Z.; Xu, H.; Li, C.; Zhang, L.; Xi, B. Research on accurate recognition and refuse rate calculation of coal and gangue based on thermal imaging of transporting situation. *Measurement* **2025**, *244*, 116574. [CrossRef]
34. Simpson, S.; Nukpezah, J.; Brooks, K.; Pandya, R. Parity benchmark for measuring bias in LLMs. *AI Ethics* **2025**, *5*, 3087–3101. [CrossRef]
35. Gomez-Cabello, C.A.; Prabha, S.; Haider, S.A.; Genovese, A.; Collaco, B.G.; Wood, N.G.; Bagaria, S.; Forte, A.J. Comparative Evaluation of Advanced Chunking for Retrieval-Augmented Generation in Large Language Models for Clinical Decision Support. *Bioengineering* **2025**, *12*, 1194. [CrossRef] [PubMed]
36. Qu, R.; Tu, R.; Bao, F. Is semantic chunking worth the computational cost? In *Findings of the Association for Computational Linguistics: NAACL 2025*; Association for Computational Linguistics: Kerrville, TX, USA, 2025; pp. 2155–2177.
37. Zhong, W.; Sivan, D.; Can, T.; Katkov, M.; Tsodyks, M. Semantic Chunking and the Entropy of Natural Language. *arXiv* **2026**, arXiv:2602.13194.
38. Verma, P. S2 chunking: A hybrid framework for document segmentation through integrated spatial and semantic analysis. *arXiv* **2025**, arXiv:2501.05485.

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