

Article

XARAG+: An Explainable Retrieval-Augmented Multi-Agent Framework for Long-Document Reasoning with Arabic Large Language Models

Adnan Ibrahim Jasim^{1*}, Waleed Mohammed Kadhim², Yaser Adnan Jasim³

^{1,2,3} Scientific Research Authority, Department of Computer Science, Computational Modelling Division, Iraq

Email Correspondence: adnanjassem80@gmail.com

Article information:

Manuscript received: 04 May 2026; **Accepted:** 17 June 2026; **Published:** 25 July 2026

Abstract: Standard large language models (LLMs) are more complicated by the intense morphological richness and lack of high-quality datasets, as well as the ever-present danger of factual hallucinations when processing and reasoning over long Arabic text. This paper presents a novel explainable multi-agent retrieval-augmented framework, called XARAG+ , for long-document comprehension in Arabic. Unlike a single language model, XARAG+ carries out operations in a specialized pipeline, comprising of document preprocessing, semantic retrieval, context verification, logical inference and explanation generation. It is a multi-agent coordination mechanism which strives to remove inconsistent and redundant retrieved chunks before the final inference step. The framework should be systematic in its organization and link all of the answers with the verified texts to make it easy to see and easily checkable. Last, XARAG+ develops a verifiable modular architecture appropriate to various domains which have a stringent requirement of factual accountability, e.g. Legal Arabic Document Analysis and Scientific Arabic Document Analysis.

Keywords: Arabic Large Language Models, Retrieval-Augmented Generation, Multi-Agent Systems, Explainable Artificial Intelligence, Arabic Document Reasoning.

1. Introduction

Although the adoption of large language models (LLMs) has significantly contributed to the improvement of natural language processing (NLP) tasks such as language understanding and text generation for various purposes [1] this extension to long-form text is still very challenging. This is especially true for the Arabic language, which has a very dense morphology, a complicated syntax and a wealth of unannotated disambiguated corpora, but lacks in good and annotated resources, making precise context-based reasoning very difficult. In this way, these limits in the computation are contrary to the interests of high stakes areas such as legal analysis, finding interesting science literature, official governmental reporting, etc. In order to prevent such failures, Retrieval-Augmented Generation (RAG) adds external chunks of knowledge during the generation process for consistency . However, traditional monolithic RAG systems often fail if the user query requires multiple steps and cross-references to be done across the various parts of a large document [2].

In Multi-Agent Systems (MAS), the hope is to alleviate the burden of this single model by dividing the macro-reasoning processes into distinct and specialized unit designs, with independent document preprocessing, semantic search, context verification, logical inference, and explanation generation agents [3]. It provides a collaborative framework

that can improve the modularity and scalability of architectures as well as add an important stage of algorithmic transparency through all the execution pipeline [4].

Worse than that, the current literature focuses on different research streams for retrieval augmentation and multi-agent orchestration without attempting to integrate them into a single pipeline to optimize the reasoning for long-document Arabic tasks [5]. This is a clear problem in the architecture that drives us to build a full architecture of retrieval quality, context validation and multi-step inference under a unified, collaborative setting.

To fill this gap, this study proposes a conceptual framework called XARAG+ (Explainable Arabic Retrieval-Augmented Multi-Agent Framework) which involves RAG mechanisms and specialized, coordinating agents for the processing of long Arabic documents [6]. Optimize Semantic retrieval, Strengthen logical Consistency, Suppress hallucinated or unsupported Generation and Provide clear answers with references from the source materials are the main aims of the proposed framework [7]. The modular engineering makes XARAG+ a flexible, domain independent system appropriate to a variety of Arabic natural language processing applications, such as legal cases evaluation, academic information indexing, government decision making support, and intelligent question-answering pipeline [8].

1.2 Research Contributions

This study makes the following key contributions:

For reasoning over long Arabic documents, a novel explainable retrieval-augmented multi-agent framework (XARAG+) is suggested .

Before the reasoning stage, an Evidence Verification process is added to check for redundancy, consistency and relevance of the retrieved document segments.

An explainable reasoning technique is applied to provide clear answers supported with retrieved evidence and enhance user trust, for the ease of interpretation.

It explicitly addresses the long document reasoning problem in Arabic, integrating three processes, namely, semantic retrieval, evidence validation, collaborative reasoning, and explainable response production, into a single reasoning pipeline.

1.3 Related Work

However, recent progress in Large Language Models (LLMs) has demonstrated progress and achievement in facilitating Arabic Natural Language Processing (Arabic NLP) in various areas, including text understanding, and Question-Answering pipelines, but Arabic computational processing is inherently bottlenecked. The subtle morphological systems and non-concatenative properties of Arabic, along with a lack of high-quality, standardized benchmark datasets in Arabic for exact algorithmic reasoning, still make the story challenging [9].

However, to overcome these limitations in parametric memory, Retrieval-Augmented Generation (RAG) frameworks have been introduced to guide generation throughout the process based on dynamically added external knowledge contexts. Retrieval-Augmented Generation (RAG) frameworks, which enhance the factual reliability by dynamically incorporating external knowledge context during generation, help bypass the limitations in parametric memory. At the same time, the paradigm of mass systems was gradually been turned on its head from the work of monolithic inference, to that of breaking down complex reasoning flows into separately situated specialized systems called agentic multi-agent systems (MAS), which have a goal of achieving the highest possible degree of task specialization with meta-inferencing by the collective of agents, to boost the quality of macro-inference [10].

Alongside the efficiency of architecture, Explainable Artificial Intelligence (XAI) has become critical technology to demystify “black-box” deep learning models in particular when domain experts need to review the automated products. This is a directly traceable shift from abstract probabilities to the evidentiary grounds and immediately provides enrichment to algorithmic accountability, and trust in algorithmic environment of automated decision-making.

Importantly, most of the literature is concentrated on these paradigms, and considers the layers of retrieval engineering, multi-agent orchestration, and interpretability layers as standalone entities. This break in the structure has a permanent architectural shortfall, and crippling handicaps in long-document Arabic reasoning. A new XARAG+ framework is developed to address these disjointed strategies, bringing together semantic retrieval, pre-inference evidence filtering, collaborative multi-agent reasoning and explainable output generation in a unified and coherent system. The aim of this closely coupled integration is to maximize the accuracy of reasoning, stabilize/relax the hallucination constraint and enforce systemic transparency for complex Arabic NLP tasks [11].

1.4 Recent Advances in Arabic LLMs and Agentic RAG

Recently, Arabic LLM and retrieval augmented generation (RAG) [12] have been introduced to strengthen Arabic NLP. Incorporating semantic retrieval into collaborative multi-agent reasoning shows improvement in factual consistency and aids in understanding long documents [13]. Despite this, an Arabic long-document reasoning with integrated explainable frameworks, specifically created for this language, is still limited, motivating the proposed XARAG+ architecture.

1.5 Research Gap

Despite significant success in Retrieval-Augmented Generation (RAG), Arabic Large Language Models (Arabic LLMs) and multi-agent systems [14] as well as impressive results, Arabic document reasoning still needs to overcome many challenges. Existing approaches often rely on 'one-step' algorithms, with limited abilities to provide explanations and validate evidence for multi-step reasoning. Moreover, the Arabic language is intricate and there are limited resources available that might not be accurate to justify reasoning and retrieval [15]. To date, most research on retrieval, reasoning, and explainability are treated in isolation, not together [16]. The intended XARAG+ is designed with a modular structure and addresses these limitations by using explainable answer generation, multi-agents collaborative reasoning, retrieval augment, and verifying the evidence from long documents with the aim of reliable and explainable Arabic long documents understanding [17].

Table 1. Comparative Analysis of Existing Approaches and the Proposed XARAG+ Framework

Study	Arabic Support	RAG	Multi-Agent	Explainability	Long-Document Reasoning	Limitation
Recent Arabic LLM Studies	✓	✗	✗	✗	Limited	Depend mainly on parametric knowledge.
Arabic RAG-based Systems	✓	✓	✗	✗	Partial	Focus mainly on retrieval quality.
Conventional RAG Frameworks	Partial	✓	✗	✗	Partial	Limited reasoning across multiple document sections.
Existing Multi-Agent Architectures	Partial	Partial	✓	Partial	Partial	Mostly designed for general multilingual tasks.

Study	Arabic Support	RAG	Multi-Agent	Explainability	Long-Document Reasoning	Limitation
Explainable AI Approaches	Partial	X	X	✓	X	Explainability is treated separately from reasoning.
Arabic Question Answering Systems	✓	Partial	X	X	Limited	Performance decreases with long documents.
Long-Context LLM Methods	Partial	Partial	X	X	✓	High computational cost and context degradation.
Hybrid Retrieval Methods	✓	✓	X	Partial	Partial	Lack collaborative reasoning among components.
Agent-based Reasoning Systems	Partial	Partial	✓	Partial	✓	Limited support for Arabic linguistic characteristics.
Proposed XARAG+ Framework	✓	✓	✓	✓	✓	Integrates retrieval, collaborative reasoning, evidence verification, and explainable response generation within a unified framework.

2. Materials and Methods

2.1 Overview of the Proposed Framework

This study proposes XARAG+ (Explainable Arabic Retrieval-Augmented Multi-Agent Framework), an explicit conceptual design that merges Retrieval-Augmented Generation (RAG) with decentralized multi-agent orchestration, contextual evidence validation and interpretable response generation [18] to improve logical inference over long texts in Arabic. The model proposed distributes both the computational and analytical load among multiple/adaptive specialized agents, with each of them responsible for a specific sub-task, while monolithic RAG approaches rely on a single language model to handle the inference task entirely.

The main structure for processing runs through 5 layers of operations sequentially [19]. The on-going workflow starts with translating Arabic input into an unstructured format to be normalized for dense vector indexing, and continues to a subsequent phase

of semantic retrieval, where the goal is to extract text segments with high similarity scores [20]. An intermediate evidence verification phase is added that works through the passages retrieved selecting evidence for topical relevance and consistency of evidence within each passage, as well as semantic redundancy among passages [21]. Last but not least, a commitment to a special explanation component specifies a reaction hard linked to the verified source information, decreasing structural opacity and giving traced out outputs [22].

This allows agents to run independently but have their results passed to the next agent in the execution pipeline in a seamless manner [23]. Not just structural maintainability, but system versatility and scalability are improved too with this engineering, since researchers can easily swap-out certain embedding model, vector database, or LLM without breaking the whole architecture layout of the system [24].

2.2 Architecture of the Proposed XARAG+ Framework

This study proposes XARAG+ (Explainable Arabic Retrieval-Augmented Multi-Agent Framework), an explicit conceptual design that merges Retrieval-Augmented Generation (RAG) with decentralized multi-agent orchestration, contextual evidence validation and interpretable response generation [25] to improve logical inference over long texts in Arabic. The model proposed distributes both the computational and analytical load among multiple/adaptive specialized agents, with each of them responsible for a specific sub-task, while monolithic RAG approaches rely on a single language model to handle the inference task entirely.

The main structure for processing runs through 5 layers of operations sequentially [26]. The on-going workflow starts with translating Arabic input into an unstructured format to be normalized for dense vector indexing, and continues to a subsequent phase of semantic retrieval, where the goal is to extract text segments with high similarity scores [27]. An intermediate evidence verification phase is added that works through the passages retrieved selecting evidence for topical relevance and consistency of evidence within each passage, as well as semantic redundancy among passages [28]. Last but not least, a commitment to a special explanation component specifies a reaction hard linked to the verified source information, decreasing structural opacity and giving traced out outputs [29].

This allows agents to run independently but have their results passed to the next agent in the execution pipeline in a seamless manner [30]. Not just structural maintainability, but system versatility and scalability are improved too with this engineering, since researchers can easily swap-out certain embedding model, vector database, or LLM without breaking the whole architecture layout of the system [31].

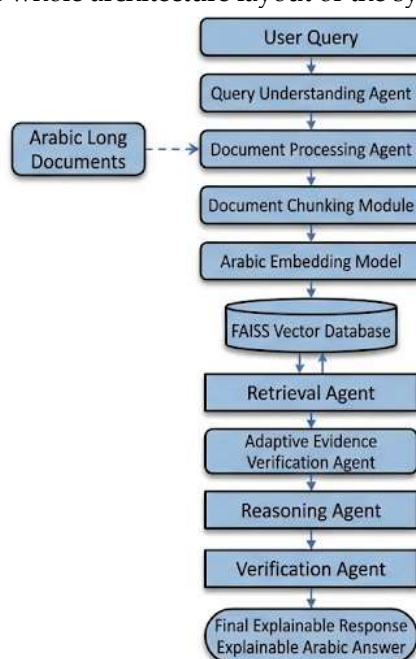


Figure 1. Overall Architecture of the Proposed XARAG+ Framework

Figure 1. The general architecture of the proposed XARAG+ is depicted in which all the modules of document preprocessing, semantic retrieval, evidence verification, reasoning and explanation production work together.

2.3 Architectural Layers

The XARAG+ (Document Management + Semantic Retrieval + Collaborative Reasoning + Explanation Generation) proposed framework is based on a layered architecture where each processing tier is encapsulated and independent from each other yet interconnected and dependent on each other. This tiered architecture facilitates independent component development and evolution, and also increases maintainability, modularity and scalability.

Table 2. Architectural Layers of the Proposed Framework

Layer	Components	Responsibility
Data Layer	Arabic Documents, Vector Database	Document storage and semantic indexing
Retrieval Layer	Retrieval Agent	Semantic evidence retrieval
Reasoning Layer	Evidence Verification Agent, Reasoning Agent	Evidence validation and inference
Explainability Layer	Explanation Agent	Transparent answer generation

2.4 Functional Description of the Intelligent Agents

In the proposed XARAG+ architecture, five intelligent agents specifically designed for tasks of long document reasoning perform these tasks sequentially. The system features evidence based reasoning and explainable reason for generating a response, thanks to all agents being assigned to a specific task within the pipeline of processing [32].

2.4.1 Document Processing Agent

The Arabic documents need to be prepared by the Document Processing Agent before being retrieved. The semantic relation of segments in a document should be preserved while pre-processing, and this can help to improve the retrieval performance [33].

2.4.2 Retrieval Agent

Once the user question is received, the Retrieval Agent finds the most similar semantic representation in the vector database [34].

2.4.3 Evidence Verification Agent

The Evidence Verification Agent decides the relevance, consistency and dependability of the passages retrieved. If there exists redundant information or conflicting information, it will be filtered out before reasoning [35].

2.4.4 Reasoning Agent

The Reasoning Agent incorporates factual information from multiple aspects of documents and uses deductive reasoning to form an appropriate conclusion in the reasoning process, and maintains consistency in the generated text throughout [36].

2.4.5 Explanation Agent

The Explanation Agent performs the final part needed to increase the transparency and user's confidence in the system, creating explanations through evidence generation [37].

2.5 Algorithm of the Proposed XARAG+ Framework

Algorithm 1. An overall workflow for the proposed XARAG+ framework.

Input:

User Query Q

Arabic Document Collection D*

Output:

Explainable Answer A

Step 1:

Preprocess all texts in Arabic language.

Step 2:

Part the packets into semantically consistent units.

- Step 3:
Generate embedding vectors for all the chunks of documents.
- Step 4:
Embed all of the vectors that will be used for the embedding into a vector database through FAISS.
- Step 5:
Convert the query of the user into the embedding space.
- Step 6:
Use Cosine similarity to retrieve document chunks of the Top-k most relevant ones.
- Step 7:
Field tested evidence with the Evidence Verification Agent.
- Step 8:
Eliminate evidence that may be duplicated or is not very confident.
- Step 9:
Send the verified evidence to the Reasoning Agent.
- Step 10:
Use Arabic Large Language Model to generate final answer.
- Step 11:
Create a scientific explanation supported by evidence (Explanation Agent).
- Step 12:
Giving back the final understandable answer.

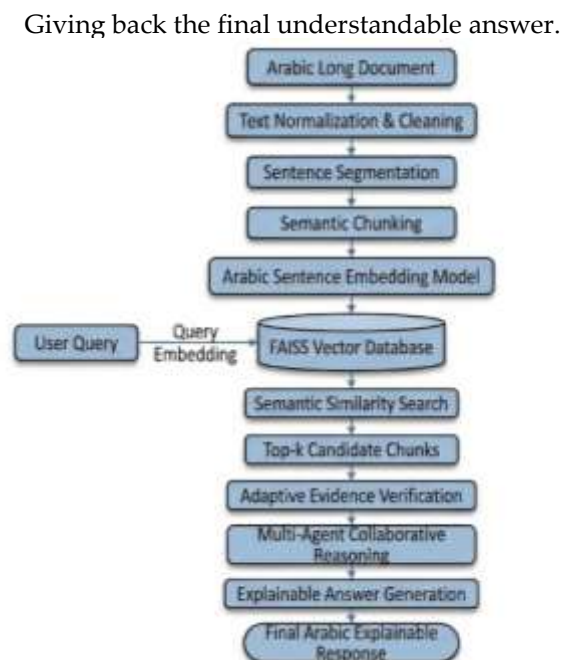


Figure 2. End-to-End Data Processing Pipeline

Figure 2. The proposed XARAG+ framework's end-to-end data processing pipeline demonstrates document preparation, semantic retrieval, evidence verification, collaborative reasoning, and explainable response production.

Table 3. Processing Stages of the Proposed Framework

Stage	Input	Output
Document Preprocessing	Arabic Documents	Cleaned Text
Semantic Chunking	Cleaned Text	Document Chunks
Embedding Generation	Chunks	Vector Embeddings
Semantic Retrieval	User Query	Top-k Evidence
Evidence Verification	Retrieved Evidence	Verified Evidence
Collaborative Reasoning	Verified Evidence	Generated Answer
Explanation Generation	Generated Answer	Explainable Response

2.6 Design Principles of the Proposed XARAG+ Framework

Based on a set of design principles, the proposed XARAG+ framework was developed to increase the efficiency, reliability, and transparency of long document reasoning in Arabic language. These concepts also drove their implementation of the intelligent agents interactions and drove the evolution of the framework [38]’.

Being modular in architecture, can be expanded in the future, and can be separately serviced.

Factual retrieval and reasoning (P2): reasoning that is directed by evidence retrieved from a factual text to increase the reliability of the text evidence and the answer.

P3 Evidence Verification: To enhance the reliability of evidence, evidence is verified before reasoning.

P4 Collaborative Multi-Agent Processing: agents with specializations coordinate in undertaking various aspects of reasoning

P5 Explainability by Design: Setting the framework of justifications of transparency based on data.

P6 Scalability and Domain Adaptability: The framework supports many NLP applications of Arabic language and also it can easily accept future additions.

3.6 Multi-Agent Communication Strategy



Figure 3. Multi-Agent Communication Workflow

Figure 3. Communication workflow among the intelligent agents within the proposed XARAG+ framework.

2.7 Data Flow Analysis

XARAG+ undergoes systematic data propagation gradually from ingesting raw text to generating a traceable response [39]. The pipeline draws in Arabic documents that are passed through a normalization layer to reduce syntactic differences and to optimize vector indexing of the documents [40]. This is followed by the Retrieval Agent extracting candidate segments through similarity matching, and then passing these raw segments to the Evidence Verification Agent to remove relevant and redundant contexts and only provide segments of high-confidence evidence to the Reasoning Agent.

During the latter stages, the Explanation Agent fixes the inferred answer response to specific text spans of its source, maximizing transparency in the response. Thanks to this formalized checkpoint workflow, noisy or hallucinated evidence is prevented from influencing it in the inference stage, thereby optimizing the coordination of multi-agent in the long-document Arabic NLP tasks [41].

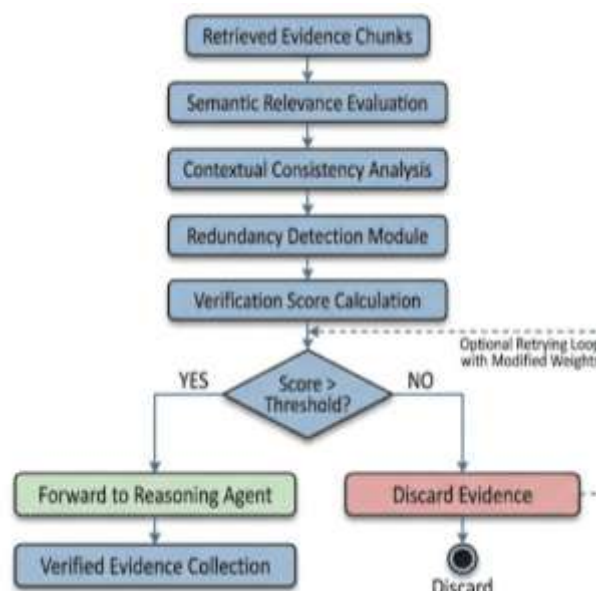


Figure 4. Evidence Verification Workflow

Figure 4. Evidence Verification workflow used to evaluate retrieved document segments before collaborative reasoning.

Table 4. Evidence Verification Stages

Stage	Objective	Output
Semantic Relevance	Measure semantic similarity	Relevance Score
Contextual Consistency	Validate logical coherence	Consistency Score
Redundancy Detection	Remove duplicated information	Redundancy Score
Score Calculation	Combine evaluation criteria	Verification Score
Threshold Decision	Accept or reject evidence	Verified Evidence

2.8 Advantages of the Proposed XARAG+ Framework

Table 5. Expected Advantages of the Proposed XARAG+ Framework

Feature	Expected Benefit
Modular Multi-Agent Architecture	Improved scalability and maintainability
Semantic Retrieval	More accurate evidence selection
Evidence Verification	Reduced irrelevant and inconsistent evidence
Multi-Step Reasoning	Better understanding of long Arabic documents
Explainable Response Generation	Improved transparency and user trust
Evidence-Based Inference	Reduced hallucination risk
Flexible Architecture	Easy integration with future LLMs and retrieval models
Arabic-Oriented Processing	Better suitability for Arabic NLP applications

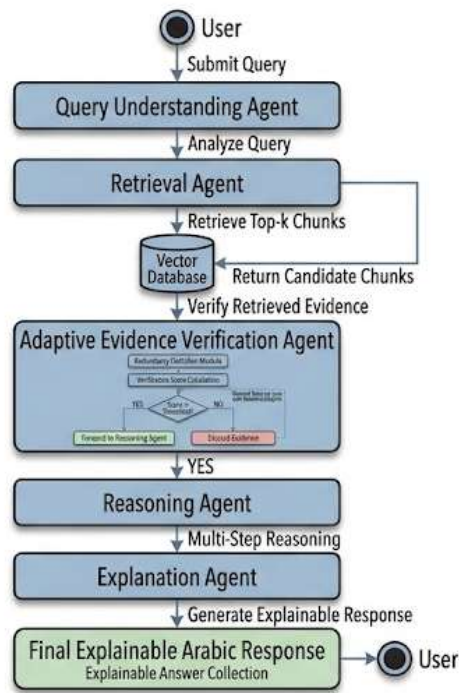


Figure 5. Multi-Agent Interaction Sequence

Figure 5. The order in which the intelligent agents interact while processing queries and producing understandable responses.

2.9 Limitations and Future Extensions

The effectiveness of the proposed framework is always dependent on the base performance of the large language models and embedding pipelines used, and the quality of the corpus. Moreover, multi-agent topologies add minor processing cost and latency effects to conventional 'traditional RAG pipelines'. However, all the computational benefits do not negatively affect the system's modularity, meaning that next generation architecture for retrieval and reasoning components can be easily integrated, since as they develop, the system can adapt to them in a straight-forward manner [42].

Table 6. Components of the Proposed XARAG+ Framework

Component	Primary Function	Expected Outcome
Document Processing Agent	Normalize and segment Arabic documents	Structured semantic chunks
Retrieval Agent	Retrieve relevant document chunks through semantic similarity search	Relevant supporting evidence
Evidence Verification Agent	Validate and refine retrieved evidence	Reliable evidence set
Reasoning Agent	Generate evidence-guided responses	Contextually consistent answers
Explanation Agent	Produce interpretable explanations	Transparent and explainable outputs

2.10 Applications and Framework Summary

We propose to integrate translate, explain, answer generation, collaborative multi-agent processing, retrieval augmentation, and evidence verification into one architecture to support long Arabic document reasoning (the XARAG+ system). The framework is organized into a sequence of specialized processes that combine to provide relevant evidence, legitimize evidence collected, perform evidence-based reasoning, and present explicit documentation of the evidence and clear answers based on the evidence [43].

In comparison to common pipelines for retrieval-augmented generation, the proposed pipeline focuses on: modularity explainability collaborative reasoning. The features are designed to more effectively support factual consistency, more extensible system, and future research on explainable Arabic Large Language Models (Arabic LLMs) and long-document comprehension.

A variety of Arabic natural language processing applications that need trustworthy reasoning over lengthy textual texts are supported by the suggested XARAG+ architecture. The flexible modular structure and a capacity of explainable reasoning allows the framework to be generalized to various application domains, while only minor modifications of the architecture are required.

2.10.1 Legal Document Analysis

Articles are frequently lengthy in legal documents, as are cross-references. Furthermore, legal documents contain legal jargon as well! The proposed framework can provide a pertinent list of legal provisions, confirm supporting evidence, and produce explainable responses that can be useful for legal practitioners during analyzing paperwork [44].

2.10.2 Scientific Literature Exploration

Researchers often have to read lengthy scientific journal articles and technical reports. XARAG+ can identify relevant text from long documents, add relevant data, and formulate logical bullet points or even a response based on found information [45].

2.10.3 Intelligent Arabic Question Answering

The system does not rely entirely on the internal knowledge of a language model, but can be used in Arabic QA systems which require evidence based responses. This feature is particularly convenient when transparency and factual consistency is needed such as in cases where application is required [46].

2.11 Mathematical Formulation

The mathematical description of the proposed XARAG+ framework outlines the retrieval, evidence verification and reasoning process in a unified computational model.

Represent the Arabic document collection as:

$$D = \{d_1, d_2, d_3, \dots, d_n\}$$

where D denotes the complete Arabic document collection and d_i represents the i -th Arabic document.

$$d_i = \{c_1, c_2, c_3, \dots, c_m\}$$

where c_j denotes the j -th semantic chunk extracted from document d_i .

Each semantic chunk is transformed into a dense vector representation using an embedding model:

$$v_j = \text{Emb}(c_j)$$

where $\text{Emb}(\cdot)$ denotes the embedding function and v_j is the embedding vector corresponding to chunk c_j .

$$q_v = \text{Emb}(q)$$

where q denotes the user query and q_v is its vector representation.

$$\text{Sim}(c_j, q) = (v_j \cdot q_v) / (||v_j|| \times ||q_v||)$$

The Retrieval Agent selects the Top- k document chunks with the highest similarity scores:

$$R = \text{TopK}(\text{Sim}(c_j, q))$$

where R denotes the retrieved evidence set used during reasoning.

$$\text{Score} = \alpha R + \beta C - \gamma D$$

where:

R represents the semantic relevance score.

C represents the contextual consistency score.

D represents the redundancy penalty.

α , β , and γ are weighting coefficients satisfying:

$$\alpha + \beta + \gamma = 1$$

Only evidence satisfying the predefined verification threshold is accepted:

$$\text{Score} \geq \tau$$

where τ represents the verification threshold.

The Reasoning Agent receives the confirmed evidence and uses the Arabic Large Language Model to provide the final response:

$$A = \text{LLM}(q, R)$$

where A denotes the generated answer, q is the user query, and R is the verified evidence set.

Table 7. Mathematical Symbols Used in the Proposed Framework

Symbol	Description
(D)	Arabic document collection
(d_i)	Individual Arabic document
(c_j)	Document chunk
(e_j)	Embedding vector
(q)	User query
(e_q)	Query embedding
(Sim)	Cosine similarity
(R)	Retrieved evidence
(V_i)	Evidence verification score
(τ)	Verification threshold
(A)	Generated answer

Algorithm 1. Overall Workflow of proposed XARAG+ Framework

Input: long document D (Arabic) and query Q (Arabic)

Input: Student's explanation (A)

Get the input document (D) in Arabic.

Remove collisions between terms.

Choose the appropriate selection of keywords.

Make semantic paragraphs.

Generate the embedding vectors for all chunks of your documents.

Store the embeddings in the vector database.

Obtain (read) the user query (Q) .

(7) Construct embedding representation of (Q) .

Retrieve top-(k) most relevant chunks of documents.

Follow the Evidence Verification process.

Ditch inconsequential or repetitive information.

Pass the valid evidence to the Reasoning Agent.

Use evidence to carry out multiple steps of reasoning.

Produce the solution.

Give a clear, but explanation answer with supporting evidence.

Give them the answer (A) that can be explained.

Complexity Analysis

The computing costs of XARAG+ rise linearly with the number of document chunks (N) and the latent embedding dimensionalities (d). Preprocessing the corpus and generating the vectors have linear computational complexity $\mathcal{O}(N)$, while the semantic retrieval phase is linear depending on the vector indexing topology used. Most importantly, downstream evidence verification is limited only to extracted Top- k candidates ($k \ll N$), which significantly reduces the global processing costs. The architecture confines multi-agent inference to this very specific context that is relevant to that task, to reduce the reasoning distortion, to optimize the execution throughput and to leave all the explanations traceable [47].

Table 8. Computational Complexity of Framework Components

Component	Complexity
Document Preprocessing	$\mathcal{O}(N)$
Document Chunking	$\mathcal{O}(N)$
Embedding Generation	$\mathcal{O}(N \times d)$

Component	Complexity
Semantic Retrieval	Depends on vector index implementation
Evidence Verification	$O(k)$
Multi-Agent Reasoning	$O(k)$
Explainable Response Generation	$O(k)$

Algorithm 2. Evidence Verification

Input: Evidence set $R=\{r_1, r_2, \dots, r_k\}$ recovered

Input: A set of evidence that has been verified (V)

Set a verified evidence set (V) to be empty.

For each document chunk retrieved (r_i):

Calculate semantic relevance score.

Calculate contextual consistency score.

Compute the redundancy score.

Use a dot product to determine the verification score:

[Score = $\alpha \times \text{Relevance} + \beta \times \text{Consistency} - \gamma \times \text{Redundancy}$]

The α , β and γ are weightings such that $\alpha + \beta + \gamma = 1$.

Compare the score of voting with a predefined threshold (τ).

If (Score > τ):

Accept the evidence.

Add it to verified evidence set (V).

Otherwise:

Discard the evidence.

Continue until all of the recovered chunks are assessed.

Return verified evidence set (V).

Evidence Verification Criteria

The Evidence Verification Agent checks the contents of retrieved document segments on three main aspects: semantic relevance, contextual consistency and redundancy thresholding. After this multi-criteria analysis, the system generates a composite (weighted) verification score which only the candidate chunks that exceed a threshold passed to the system are moved to the Reasoning Agent. With this validation loop, context stripping for non-essential and duplicated context directly improves the accuracy and empirical validity of downstream EBI [48].

Table 9. Evidence Verification Metrics

Objective	Description	Metric
Improve retrieval quality	Similarity between query and retrieved evidence	Semantic Relevance
Preserve logical consistency	Coherence with neighboring document context	Contextual Consistency
Remove repeated evidence	Identification of duplicated information	Redundancy Detection
Select reliable evidence	Combined weighted score	Verification Score

2.12 Scalability Analysis

The system scalability features of XARAG+ come as natural attributes of its modular design, separating the retrieval pipelines, embedding strategies, vector engines and core language models. In this way, the separation in the structure makes the friction changes in the architecture as low as possible, and allows the Arabic NLP applications with high flexibility to be deployed in the production environment, which has a dynamic structure [49].

Table 10. Scalability Characteristics of the Proposed Framework

Benefit	Description	Property
Easy maintenance	Independent intelligent agents	Modularity
Reduced response time	Simultaneous execution of multiple agents	Parallel Processing
Better adaptability	Supports different embedding techniques	Flexible Embedding Models
Easier deployment	Compatible with various vector databases	Retrieval Independence
Future-proof architecture	New agents can be added without redesign	Expandability
Wide applicability	Applicable across multiple Arabic NLP domains	Domain Adaptation

2.13 Comparative Advantages over Conventional RAG

Table 11. Comparison Between Conventional RAG and XARAG+

Feature	Conventional RAG	Proposed XARAG+
Multi-Agent Collaboration	X	✓
Evidence Verification	X	✓
Explainable Response Generation	Limited	✓
Modular Architecture	Partial	✓
Arabic Long-Document Reasoning	Limited	✓
Scalability	Moderate	High
Evidence Transparency	Limited	High
Flexible Component Integration	Partial	✓

2.14 Evaluation Methodology

2.14.1 Evaluation Framework

This assessment framework offers a conceptual way of measuring the XARAG+ architecture at the implementation level in the future. It correlates the basic reasoning workflow, the pre-processing layers and the semantic retrieval strategies directly with the intermediate verification check-scoring, thus uniformizing the performance measurements, rather than using a generic testing. This organized structure makes it easy for this to be used as a clearly defined and repeatable win-base for measuring the reliability of a system in complex and long document Arabic reasoning problems [50].

2.14.2 Evaluation Metrics

The multi-level metric system that is used to evaluate the performance of XARAG+ in operation is based on the retrieval, generation and computational footprint metrics [51]. For the downstream response, BLEU score, ROUGE-L score and BERTScore score are used to measure the response accuracy, and for retrieval quality, standard Precision, Recall and F1 Score are used. To achieve a balance between factual consistency and costs of execution, the framework tracks computational efficiency by measuring processing time, latency overhead and hallucination rates. Meanwhile, the explainability of the system is based on some qualitative requirements: evidence explicability and the reasoning transparency. Their combination offers a full-fledged measure of cognitive, manipulative and effective architecture.

2.14.3 Baseline Models

The architectural effectiveness of XARAG+ is directly compared with the conventional state-of-the-art baselines, such as pipelines with conventional Retrieval-Augmented Generation (RAG) approach and single monolithic Arabic Large Language Models (Arabic LLMs). Each of these systems will be tested on the same computational/experimental platform, thus providing a base level for calibration. This is a comparative setup that allows isolation and comparison of incremental gains with

regard to four key performance metrics: semantic retrieval precision; longer/document reasoning capabilities; tracing explanations; overall architectural throughput.

2.14.4 Evaluation Procedure

The performance of this system evaluation method is applied to six phases of the system operations, namely document preprocessing, embedding and indexing documents/stories, semantic retrieval of documents/context, verification of contexts, evidence-driven reasoning and output of explainable responses. Adopting this particular execution sequence ensures a defined, clear, and repeatable benchmarking context for the XARAG+ framework laying a uniform verification procedure for subsequent system deployments.

3. Results

3.1 Retrieval Effectiveness Discussion

The effectiveness of retrieval can be measured by calculating common informational parameters like precision, recall and F1-score. To minimize the chance of errors, the proposed architecture places a direct relationship between semantic search and an intermediate verification step.

It is a loop for integrated processing, designed to improve the isolation of passages of longer Arabic manuscripts that are relevant to the given context. There is also an attempt to avoid noisy or redundant text being transferred to the next levels of the inference. From a conceptual point of view, the filtering process also reinforces the stability of retrieval, with highly calibrated inputs of data, enabling reasoned prediction of evidence, with accuracy and confidence.

3.2 Reasoning Capability Analysis

The logical belief of structure is directly improved by compiling known evidence from various parts of the documents. This is an agentic interaction, in which more than one step of interpretation is made within different paths of execution and these steps are aligned appropriately, in context, minimizing the factual hallucinations and the production of unsupported claims.

3.3 Explainability Analysis

Structural algorithmic explainability is an integral part of the XARAG+'s. Thus, by systematically linking each and every response generated with the exact same content from the support data source it is possible to provide for strong reinforcement of systemic transparency, a higher level of decision support validation and inspire users to trust information from the automated results.

4. Discussion

Finally, Proposed Architecture combines retrieval augmentation, intermediate context validation, multi-agent reasoning and interpretability in a single framework. The layers are unified with making the structure more flexible, factual consistency in the processing, transparency of operation in the processing of long document form and documents of complex Arabic language.

5. Conclusion

The study introduces a new Explainable Arabic Retrieval-Augmented multi-Agent Framework called XARAG+ to tackle the challenge of optimizing the reasoning process for longer Arabic texts. The architecture introduced here enables dense retrieval augmentation, pre-inference context validation, collaborative multi-agent orchestration, interpretable response generation, while also logically consistent and maximising output transparency and minimising or eliminating factual hallucinations. The modular topology has the additional advantage of smooth growth of the system in order to meet the various Arabic NLP applications and enables the users to have faith in the system because of the accuracy of its decisions, which are based on proof.

The focus of future development will be mainly on the empirical implementation and deployment of the framework. Some of the main research directions are to analyze the performance of XARAG+ on datasets with different length and format to assess its effectiveness on the Arabic long document set; to explore the potential of adopting the next-generation LLM to Arabic actively contribute to the learning process; and to

investigate how to utilize the adaptive retrieval loops and multilingual pipelines to create more explainable architecture and also obtain higher throughput.

REFERENCES

- [1] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, et al., "A comprehensive overview of large language models," *ACM Trans. Intell. Syst. Technol.*, vol. 16, no. 5, pp. 1–72, 2025.
- [2] A. M. Alayba, "Arabic Natural Language Processing (NLP): A comprehensive review of challenges, techniques, and emerging trends," *Computers*, vol. 14, no. 11, Art. no. 497, 2025.
- [3] F. Branda, M. M. Ahmed, M. Ciccozzi, P. H. Guzzi, and F. Scarpa, "The next paradigm in bioinformatics: A review of multi-agent systems and foundational models for end-to-end scientific discovery," *Briefings Bioinf.*, vol. 27, no. 3, Art. no. bbag245, 2026.
- [4] M. Bellundagi, "A multi-layer AI-driven decision intelligence framework for enterprise and healthcare system," *Int. J. Res. Publ. Eng. Technol. Manage.*, vol. 7, no. 6, pp. 11679–11687, 2024.
- [5] A. Toumi, S. F. Wamba, and M. Hafsi, "Architecting knowledge through AI-enhanced observability: A design science approach to enterprise architecture as a knowledge discipline," *Int. J. Inf. Manage.*, vol. 89, Art. no. 103070, 2026.
- [6] N. X. Phuc and T. Đ. Văn, "PuxAI at QIAS 2025: Multi-agent retrieval-augmented generation for Islamic inheritance and knowledge reasoning," in *Proc. 3rd Arabic Natural Language Processing Conf.: Shared Tasks*, Nov. 2025, pp. 905–913.
- [7] F. F. Alu and S. Oluwadare, "An auditable and source-verified framework for clinical AI decision support: Integrating retrieval-augmented generation with data provenance," *Front. Artif. Intell.*, vol. 9, Art. no. 1737532, 2026.
- [8] F. Almutairy, S. Almutairi, M. Abdullah, and S. Salman, "SmartLaw Advisor: AI-powered legal consultation and contract analysis for Saudi Arabia," *IEEE Access*, 2026.
- [9] A. M. Alayba, "Arabic Natural Language Processing (NLP): A comprehensive review of challenges, techniques, and emerging trends," *Computers*, vol. 14, no. 11, Art. no. 497, 2025.
- [10] A. Singh, A. Ehtesham, S. Kumar, T. T. Khoei, and A. V. Vasilakos, "Agentic retrieval-augmented generation: A survey on agentic RAG," *arXiv preprint arXiv:2501.09136*, 2025.
- [11] Y. Zhou, W. Zhang, J. Shao, Y. Liu, X. Li, J. Jin, et al., "Trustworthiness in retrieval-augmented generation systems: A survey," *arXiv preprint arXiv:2409.10102*, 2024.
- [12] M. G. Abd Elhakeem, A. O. Abdullah, S. W. Elhag, and E. H. Mohamed, "AQRAG: Advanced Quranic retrieval-augmented generation for low-resource question answering," *Neural Comput. Appl.*, vol. 38, no. 8, Art. no. 282, 2026.
- [13] S. Gao, S. Zhao, X. Jiang, L. Duan, Y. X. Chng, Q. G. Chen, et al., "Scaling beyond context: A survey of multimodal retrieval-augmented generation for document understanding," in *Proc. 64th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2026, pp. 4458–4489.
- [14] N. Sibai, Y. Ahmed, S. Sibae, S. AlHalawani, A. Ammar, and W. Boulila, "The path ahead for agentic AI: Challenges and opportunities," *arXiv preprint arXiv:2601.02749*, 2026.
- [15] M. A. Ferrag, N. Tihanyi, and M. Debbah, "From LLM reasoning to autonomous AI agents: A comprehensive review," *IEEE Access*, 2026.
- [16] X. Wu, H. Zhao, Y. Zhu, Y. Shi, F. Yang, L. Hu, et al., "Usable XAI: 10 strategies towards exploiting explainability in the LLM era," *ACM Trans. Knowl. Discovery Data*, 2024.
- [17] V. Perlić, S. Lebailly, V. Malvone, V. T. Nguyen, and P. Urard, "MMA-RAG: A survey on multimodal agentic retrieval-augmented generation," 2025.
- [18] H. Tebourbi, S. Nouzri, Y. Mualla, M. El Fatimi, A. Najjar, A. Abbas-Turki, and M. Dridi, "BPMN-based design of multi-agent systems: Personalized language learning workflow automation with RAG-enhanced knowledge access," *Information*, vol. 16, no. 9, Art. no. 809, 2025.
- [19] Y. Gao, Y. Xiong, Y. Zhong, Y. Bi, M. Xue, and H. Wang, "Synergizing RAG and reasoning: A systematic review," *arXiv preprint arXiv:2504.15909*, 2025.
- [20] A. A. A. Aly, M. Ali, M. Stiber, L. Battle, and A. Hendawi, "A support system for diacritic-aware classical Arabic language processing: Integration of speech, text, and vision modalities," *Ph.D. dissertation*, Univ. Washington, Seattle, WA, USA, 2024.
- [21] S. Raza, R. Sapkota, M. Karkee, and C. Emmanouilidis, "Responsible agentic reasoning and AI agents: A critical survey: Proposal for safe agentic AI via responsible reasoning AI agents (R2A2)," *SuperIntelligence-Robotics-Safety & Alignment*, vol. 2, no. 6, 2025.

- [22] K. Wang, D. Toibazar, A. Alfulayt, A. S. Albadawi, R. A. Alkahtani, A. A. Ibrahim, et al., "Multi-agent interactive question generation framework for long document understanding," in Proc. IEEE 35th Int. Workshop Mach. Learn. Signal Process. (MLSP), Aug. 2025, pp. 1–6.
- [23] Z. An, X. Wang, H. Baier, Z. Chen, A. Dubey, T. T. Johnson, et al., "LogiEx: Integrating formal logic and LLMs for explainable transit planning," in Proc. ACM/IEEE 17th Int. Conf. Cyber-Physical Syst. (ICCPs), May 2026, pp. 88–99.
- [24] H. Chen, Y. Yang, Y. Li, M. Zhang, B. Hu, and M. Zhang, "Beyond chunking: Discourse-aware hierarchical retrieval for long document question answering," in Proc. 64th Annu. Meeting Assoc. Comput. Linguistics (ACL), vol. 1, Jul. 2026, pp. 18176–18198.
- [25] M. K. Ghali, A. Farrag, D. Won, and Y. Jin, "Enhancing knowledge retrieval with in-context learning and semantic search through generative AI," *Knowledge-Based Systems*, vol. 311, Art. no. 113047, 2025.
- [26] T. Trinh, M. Nguyen, and T. S. Hy, "Towards robust fact-checking: A multi-agent system with advanced evidence retrieval," arXiv preprint arXiv:2506.17878, 2025.
- [27] L. Mei, J. Yao, Y. Ge, Y. Wang, B. Bi, Y. Cai, et al., "A survey of context engineering for large language models," arXiv preprint arXiv:2507.13334, 2025.
- [28] Z. You, X. Chen, A. Vashishtha, S. Du, G. Erion-Barner, H. Mei, et al., "Improving clinical diagnosis with counterfactual multi-agent reasoning," arXiv preprint arXiv:2603.27820, 2026.
- [29] A. Pirjan and D. M. Petroşanu, "Artificial social intelligence and the transformation of human interaction by artificial intelligence agents," *Journal of Information Systems & Operations Management*, vol. 19, no. 1, 2025.
- [30] G. Papageorgiou, V. Sarlis, M. Maragoudakis, and C. Tjortjis, "Hybrid multi-agent GraphRAG for e-government: Towards a trustworthy AI assistant," *Applied Sciences*, vol. 15, no. 11, Art. no. 6315, 2025.
- [31] M. M. Karim, S. Khan, D. H. Van, X. Liu, C. Wang, and Q. Qu, "Transforming data annotation with AI agents: A review of architectures, reasoning, applications, and impact," *Future Internet*, vol. 17, no. 8, Art. no. 353, 2025.
- [32] R. M. Aratchige and W. M. K. S. Ilmini, "LLMs working in harmony: A survey on the technological aspects of building effective LLM-based multi-agent systems," arXiv preprint arXiv:2504.01963, 2025.
- [33] Z. Xu, Z. Zhang, Z. Cai, J. Xu, Y. Lin, Z. Liu, et al., "A multi-modal agentic co-pilot for evidence grounded computational pathology," arXiv preprint arXiv:2606.08093, 2026.
- [34] L. Resck, F. Moreno-Vera, T. Veiga, G. Paucar, E. Fajreldines, G. Klafke, et al., "LegalAnalytics: Bridging visual explanations and workload streamline in Brazilian Supreme Court appeals," *Artificial Intelligence and Law*, pp. 1–59, 2025.
- [35] P. Achkar, T. Gollub, and M. Potthast, "Ask, retrieve, summarize: A modular pipeline for scientific literature summarization," arXiv preprint arXiv:2505.16349, 2025.
- [36] R. S. Dehal, M. Sharma, and E. Rajabi, "Knowledge graphs and their reciprocal relationship with large language models," *Machine Learning and Knowledge Extraction*, vol. 7, no. 2, Art. no. 38, 2025.
- [37] S. Gong, R. O. Sinnott, J. Qi, C. Paris, P. Nakov, and Z. Xie, "Multi-sourced, multi-agent evidence retrieval for fact-checking," arXiv preprint arXiv:2603.00267, 2026.
- [38] V. H. Phan, L. K. Pham, D. Vu, A. D. Tran, and M. S. Dao, "E-FreeM2: Efficient training-free multi-scale and cross-modal news verification via MLLMs," in Proc. 2nd Workshop Security-Centric Strategies for Combating Information Disorder, Aug. 2025, pp. 1–7.
- [39] R. G. Al-Anazi, W. A. Nashir, B. B. Al-Onazi, A. A. Alhamad, H. Alshamrani, and W. F. B. Subait, "Burhan: A constraint-based engineering framework for computational rhetorical semantics in classical Arabic," *IEEE Access*, 2026.
- [40] A. Mohamed and E. Elmahjub, "ArabLJP: An evidence-grounded and retrieval-augmented LLM framework for reliable Arabic legal judgment prediction in statute-centric legal systems," 2026.
- [41] R. Friel, M. Belyi, and A. Sanyal, "RAGBench: Explainable benchmark for retrieval-augmented generation systems," arXiv preprint arXiv:2407.11005, 2024.
- [42] M. A. Mersha, M. G. Yigezu, H. Shakil, A. K. AlShami, S. Byun, and J. Kalita, "A unified framework with novel metrics for evaluating the effectiveness of XAI techniques in LLMs," arXiv preprint arXiv:2503.05050, 2025.
- [43] P. P. Ray, "A review of TRiSM frameworks in artificial intelligence systems: Fundamentals, taxonomy, use cases, key challenges and future directions," *Expert Systems*, vol. 43, no. 3, Art. no. e70213, 2026.
- [44] D. Tanyildiz, S. Ayvaz, and M. F. Amasyali, "Enhancing retrieval-augmented generation accuracy with dynamic chunking and optimized vector search," *Orclever Proceedings of Research and Development*, vol. 5, no. 1, pp. 215–225, 2024.
- [45] S. Deng, Z. Huang, K. Du, R. Mao, J. Liu, C. D. Giurcăneanu, and E. Cambria, "Multi-Agent SEA: Step-wise evidence acquisition for multimodal financial reasoning," *Information Fusion*, Art. no. 104550, 2026.

- [46] Y. Zhang, R. Sun, Y. Chen, T. Pfister, R. Zhang, and S. Ö. Arık, "Chain of agents: Large language models collaborating on long-context tasks," *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 132208–132237, 2024.
- [47] M. As'ad, N. Faran, and H. Joharji, "AI-supported shared decision-making (AI-SDM): Conceptual framework," *JMIR AI*, vol. 4, Art. no. e75866, 2025.
- [48] F. Ariai, J. Mackenzie, and G. Demartini, "Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges," *ACM Computing Surveys*, vol. 58, no. 6, pp. 1–37, 2025.
- [49] Y. Wu, A. Henriksson, M. Duneld, and J. Nouri, "Towards improving the reliability and transparency of ChatGPT for educational question answering," in *European Conf. Technology Enhanced Learning*, Aug. 2023, pp. 475–488.
- [50] K. Kalasampath, K. N. Spoorthi, S. Sajeew, S. S. Kuppa, K. Ajay, and A. Maruthamuthu, "A literature review on applications of explainable artificial intelligence (XAI)," *IEEE Access*, vol. 13, pp. 41111–41140, 2025.
- [51] A. R. W. Sait and Y. Alkhurayyif, "Explainable bilingual medical-question-answering model using ensemble learning technique," *Electronics*, vol. 14, no. 20, Art. no. 4128, 2025.