A Recursive Prompt-Driven Reasoning Approach for Multi-Hop Arabic Fatwa Question Answering

Manal Al-Qahtani[1], Bader Alkhamees[2], Mourad Ykhlef[3]
Department of Information Systems-College of Computer and Information Sciences,
King Saud University, Riyadh 12372, Saudi Arabia[1, 2, 3]
Department of Information Systems-College of Computer Science and Information Technology,
Shaqra University, Riyadh 15526, Saudi Arabia[1]

***Abstract*—Answering Arabic fatwa inquiries represents a highly critical yet challenging task in Natural Language Processing, requiring consideration of multiple jurisprudential conditions, interpretation of interconnected concepts, and synthesis of evidence distributed across multiple religious sources. Existing Arabic question answering systems, including conventional single-step question answering (QA) models and retrieval-augmented generation (RAG) frameworks, typically perform retrieval only once and therefore struggle to capture the intermediate reasoning steps required for complex religious inquiries, often resulting in incomplete reasoning and factual hallucinations. To address these limitations, this study proposes a recursive prompt-driven reasoning approach for multi-hop Arabic fatwa question answering that integrates instruction-based supervision, task-specific language models, hybrid retrieval, and recursive reasoning to support end-to-end multi-hop question answering. The proposed method employs an instruction-based supervision strategy constructed from the MAFQA dataset to train specialized question decomposition and question answering models. To support evidence grounding, a hybrid retrieval module combines dense semantic retrieval, sparse lexical matching, and cross-encoder reranking to retrieve both few-shot demonstrations and supporting evidence at each reasoning hop, while a relevance-guided context update mechanism retains only the most informative contextual evidence throughout the recursive reasoning process. Comprehensive experiments conducted on the MAFQA dataset demonstrate the effectiveness of the proposed approach. The hybrid retrieval module consistently outperforms individual dense and sparse retrieval methods, while task-specific fine-tuning significantly improves end-to-end performance across Arabic sequence-to-sequence models and large language models under realistic inference conditions. The fine-tuned GPT-4.1 configuration achieves the best overall performance, obtaining an F1-token score of 31.88, a ROUGE-L score of 26.64, and a Relevance score of 92.39. Furthermore, the proposed approach surpasses representative state-of-the-art Arabic question answering systems, retrieval-augmented generation methods, and reasoning-based approaches, achieving a 36.5% relative improvement in F1-token over the strongest reasoning baseline. These results demonstrate that integrating recursive question decomposition, hybrid retrieval, and dynamic context updating improves performance under the evaluated multi-hop Arabic fatwa question answering setting.**

***Keywords*—Arabic question answering; Arabic fatwa question answering; Arabic natural language processing; multi-hop reasoning; retrieval-augmented generation; hybrid retrieval; large language models; question decomposition**

I. Introduction

Recent advances in natural language processing (NLP) and large language models (LLMs) have significantly enhanced the capabilities of question answering (QA) systems across diverse domains [1]. Unlike traditional information retrieval systems that return a list of potentially relevant documents, modern QA systems aim to directly provide concise and accurate answers to inquiries expressed in natural language [2], [3]. These developments have driven substantial progress in multilingual and domain-specific QA applications, including healthcare [4], education [5], law [6], and religion [7]. However, despite the success of modern QA systems, Arabic question answering remains relatively underexplored compared to English and other high-resource languages, lagging behind the increasing demand for intelligent Arabic-language information access. This disparity is primarily attributed to the limited availability of large-scale benchmark datasets, domain-specific resources, and reasoning-oriented evaluation frameworks [8], [9]. These limitations are particularly acute in specialized domains where answering questions requires domain expertise, deep contextual understanding, and complex reasoning [10]. Within Arabic QA, fatwa question answering represents a highly challenging and critical domain. Fatwa inquiries seek Islamic rulings and religious guidance on diverse topics such as worship, financial transactions, family matters, and contemporary social issues [11]. Unlike conventional factoid questions, fatwa inquiries often demand the consideration of multiple conditions, the synthesis of evidence from disparate religious sources, and the interpretation of interconnected jurisprudential concepts. Consequently, addressing these inquiries involves complex reasoning processes that extend far beyond simple information retrieval or direct answer generation [12], [13].

Despite recent progress in Arabic language models and retrieval-augmented generation (RAG) systems, existing approaches remain limited in their ability to address complex Arabic fatwa questions. Direct QA systems typically attempt to generate answers in a single step based solely on the input question. While such approaches may perform adequately for simple questions, they often struggle to capture the intermediate reasoning chains required for complex, multi-hop inquiries [14], [15]. Similarly, while standard RAG systems improve factual grounding by incorporating external evidence, conventional single-step RAG frameworks generally retrieve information only once and generate answers in a single phase. As a result, they frequently fail to effectively integrate evidence distributed

across multiple reasoning steps [16], [17]. Furthermore, recent studies demonstrate that LLMs can generate hallucinated content, unsupported claims, or incomplete reasoning chains when answering complex, domain-specific questions, raising concerns regarding reliability and factual grounding [18]. These challenges underscore the critical need for QA systems capable of performing explicit, multi-step reasoning while maintaining strong contextual grounding.

To address these limitations, this study proposes a recursive prompt-driven reasoning approach for multi-hop Arabic fatwa question answering. The proposed method recursively decomposes complex fatwa questions into interpretable subquestions and performs recursive reasoning through intermediate answer generation and dynamic context updating. To support contextual grounding, a hybrid retrieval module combines dense semantic retrieval, sparse lexical retrieval, and cross-encoder reranking to dynamically retrieve both few-shot demonstrations and supporting evidence at each reasoning step. Furthermore, an instruction-based supervision strategy is employed to train specialized models for question decomposition and answer generation, enabling them to learn explicit reasoning behaviors and intermediate inference patterns. By combining recursive question decomposition, hybrid retrieval, contextual grounding, and answer synthesis within a unified reasoning process, the proposed approach generates more accurate, contextually relevant, and faithful answers for complex multi-hop Arabic fatwa question answering tasks.

II. Related Work

Arabic question answering (QA) has attracted increasing research attention in recent years due to the growing demand for intelligent Arabic-language information access. Early Arabic QA systems primarily relied on keyword matching, information retrieval techniques, and ontology-based semantic methods. In the Islamic domain, Kamal et al. [19] proposed a fatwa QA system based on Term Frequency–Inverse Document Frequency (TF-IDF) and Latent Semantic Indexing (LSI), achieving a recall of 95% and a mean reciprocal rank (MRR) of 0.92 on a manually collected fatwa corpus. Similarly, Maraoui et al. [20] developed an Arabic QA system for Hadith and Tafsir that employed XML-based knowledge representations and formal query processing, achieving an accuracy of 92% on a manually constructed evaluation set. Although these approaches demonstrated promising results, their dependence on handcrafted resources limited their scalability and adaptability to increasingly complex QA scenarios. To overcome the limitations of purely lexical retrieval methods, several studies explored semantic approaches based on ontologies and machine learning techniques. Abdelnasser et al. [21] introduced Al-Bayan, a Quranic QA system that combined Support Vector Machines (SVMs) with ontology-based semantic similarity to retrieve relevant Quranic verses, achieving an accuracy of 85%. Likewise, Sheker et al. [22] proposed an ontology-based approach for Islamic fatwa QA using Ibn Othaimeen's fatwa collection, reporting a precision of 72% and recall of 59%. Although these approaches improved semantic matching compared with purely lexical methods, they depended heavily on manually constructed knowledge resources and exhibited limited scalability and domain coverage.

The emergence of pretrained language models (PLMs) significantly advanced Arabic QA research. Several studies employed transformer-based architectures for Quranic QA using the Quran Reading Comprehension Dataset (QRCD) [23]. Premasiri et al. [24] investigated transfer learning approaches by fine-tuning multiple Arabic and multilingual transformer models—including AraBERT, CAMeLBERT, mBERT, and AraELECTRA—on the QRCD dataset. In addition, ensemble learning techniques were explored to combine predictions from different models and improve answer selection performance. Among the evaluated architectures, AraELECTRA achieved the strongest performance, reporting a partial Reciprocal Rank (pRR) score of 49%. These findings highlighted the effectiveness of Arabic pretrained representations for Quranic QA while demonstrating the potential benefits of model ensembling. Furthermore, ElKomy and Sarhan [25] proposed an ensemble framework integrating several Arabic BERT variants, including ARBERT, MARBERT, and AraBERT, for Quranic QA. Their approach employed majority voting to identify the most probable answer candidates, followed by a post-processing stage designed to refine the selected predictions. Experimental results demonstrated that the proposed ensemble approach substantially improved performance, achieving an Exact Match (EM) score of 27% and a pRR score of 75%, thereby outperforming individual models.

Similarly, Alnajjar and Mika [26] investigated the use of multilingual BERT for Quranic QA by fine-tuning the model on the QRCD [23] dataset and further enhancing it through additional pretraining on Islamic-domain corpora. Their findings revealed that domain-adaptive pretraining significantly improved performance, with the enhanced model achieving a pRR score of 70%. The authors also incorporated a post-processing mechanism to verify the validity and length of generated answers, further contributing to performance gains. Aftab and Malik [27] explored several pretrained transformer models for Quranic QA and proposed multiple strategies to enhance model generalization. Specifically, they employed decoupled weight decay regularization to mitigate overfitting and utilized data augmentation techniques by incorporating examples from the Arabic Question Answering and Comprehension Corpus (AQQAC [28]) to expand the QRCD [23] training data. Their experimental findings indicated that fine-tuning Arabic pretrained models yielded superior performance compared with multilingual models or models trained from scratch, with AraBERT outperforming competing approaches by 30.8% and 26.8% in terms of pRR and F1-score, respectively.

Furthermore, Alsaleh et al. [29] fine-tuned and evaluated models including AraBERT, CAMeLBERT, and ArabicBERT using the QRCD [23] dataset. Their experimental results demonstrated that AraBERT achieved the strongest performance among the evaluated models, obtaining a pRR score of 44% and an F1-score of 0.42. Nevertheless, the authors characterized the overall results as only "fair," suggesting considerable room for improvement. In particular, the relatively low EM score of 0.16 highlighted the difficulty of accurately identifying exact answer spans from Classical Arabic Quranic text. More recently, large language models (LLMs) have demonstrated strong capabilities in generative QA and

reasoning-intensive tasks. Within the Islamic domain, Basem et al. [30] investigated instruction-tuned LLMs, including Gemini and DeepSeek, combined with Arabic prompting strategies and post-processing techniques for Quranic QA. Their approach achieved a pAP@10 score of 0.637, outperforming several conventional transformer-based baselines. Saadaoui et al. [31] explored the integration of chain-of-thought (CoT) prompting with GPT-4, BART, and LLaMA models, demonstrating that structured prompting substantially enhanced reasoning performance and achieved F1-scores reaching 95% under specific evaluation settings. Similarly, Eltanbouly et al. [32] evaluated several state-of-the-art LLMs, including GPT-4o, Claude, Gemini, DeepSeek, and Mistral, for Quranic passage retrieval tasks. Their findings demonstrated that appropriately designed prompting strategies enabled these models to achieve performance comparable to specialized retrieval systems, attaining a Mean Average Precision (MAP) score of 0.535 and an MRR score of 0.842. Although these findings highlight the reasoning potential of LLMs, recent studies have also emphasized their susceptibility to hallucinations, unsupported claims, and incomplete reasoning processes, particularly within knowledge-sensitive domains where factual correctness is essential.

Retrieval-Augmented Generation (RAG) has emerged as a promising paradigm for improving the factual grounding of LLM-generated responses by incorporating external evidence during answer generation. Ahmad et al. [33] introduced a hybrid RAG framework for Islamic knowledge QA that combines sparse retrieval, dense semantic retrieval, and cross-encoder reranking within a unified retrieval pipeline. The proposed approach was evaluated on two subtasks using Fanar and Mistral language models. Experimental findings indicated that hybrid retrieval substantially improved QA performance, with the Fanar model obtaining the highest results, achieving accuracies of 45% and 80% on Subtasks 1 and 2, respectively. These results demonstrate that the hybrid retrieval approach provides meaningful enhancements for Islamic knowledge QA compared to baseline LLM performance. Similarly, Alnefaie et al. [34] investigated the impact of retrieval augmentation on Quranic QA by comparing standard GPT-4 against GPT-4 enhanced with RAG using the QUQA [35] dataset. Their results demonstrated that retrieval augmentation substantially improved both answer generation and Quranic verse identification. Specifically, the RAG-enhanced system achieved an EM score of 0.266 and an F1-score of 0.772 for text-based answers. Furthermore, for Quranic verse retrieval, the proposed approach attained an F1@1 score of 0.333, outperforming standard GPT-4 without retrieval support. Mohammed et al. [18] investigated the use of retrieval augmentation and reranking techniques to improve the quality of AI-assisted Islamic fatwa generation. Their study compared three configurations: a standalone LLM, an LLM enhanced with RAG, and an LLM with RAG plus a reranker. The results indicated that incorporating reranking into the RAG pipeline improved response consistency and reduced the occurrence of hallucinated information across multiple evaluation metrics, including BERTScore, hallucination rates, completeness, and irrelevance measures. These findings emphasize the importance of evidence refinement in sensitive domains such as Islamic fatwa generation. To address hallucination and weak evidence grounding in Islamic QA, Bhatia et al. [36] introduced ISLAMICFAITHQA, a bilingual benchmark for evaluating correctness, hallucination, and abstention behaviors in generative models. In addition, the authors developed an Agentic RAG framework that enables language models to utilize structured tools to collect authoritative evidence before generating citation-supported responses. Their findings demonstrated that retrieval augmentation consistently improved answer quality, whereas Agentic RAG produced the most substantial gains over conventional RAG approaches, achieving a state-of-the-art accuracy of 57.3% using Fanar-2-27B.

Collectively, these studies demonstrate that retrieval augmentation can substantially improve factual grounding, answer accuracy, and robustness in Islamic QA. Nevertheless, most existing RAG-based approaches perform retrieval and generation either in a single stage or through predefined retrieval pipelines, providing limited support for explicit multi-hop reasoning. Consequently, complex questions that require integrating evidence across multiple reasoning steps remain challenging. Despite the increasing interest in multi-hop QA, decomposition-based reasoning strategies remain largely underexplored in Arabic QA literature. One of the few studies investigating this direction was conducted by Abdelazim et al. [37], who investigated different reasoning strategies using the ACQAD [38] dataset under both Long Context Window (LCW) and RAG settings. The authors employed Chain-of-Thought prompting to decompose complex questions into intermediate subquestions, thereby facilitating stepwise reasoning across multiple evidence passages. Their framework utilized GPT-4-turbo as the primary language model and examined the effectiveness of decomposition-based reasoning in conjunction with long-context processing and retrieval augmentation. Experimental results demonstrated that incorporating question decomposition substantially improved performance compared with direct QA approaches, increasing accuracy from 75% to 92% under the LCW setting.

III. METHODOLOGY

A. Overview of the Proposed Approach

This study proposes a recursive prompt-driven reasoning approach for multi-hop Arabic fatwa question answering (QA) that integrates instruction-based supervision, task-specific language models, and hybrid retrieval mechanisms within a unified reasoning process. Designed to address the limitations of conventional systems when handling complex fatwa inquiries, the proposed approach explicitly models multi-step reasoning, evidence integration, and contextual synthesis. As illustrated in Fig. 1, the proposed approach consists of four primary components.

The first component is an instruction-based supervision module that transforms complex fatwa reasoning processes into structured instruction–response pairs for question decomposition and answer generation. This supervision strategy explicitly models intermediate reasoning steps and provides training signals for downstream models. The second component comprises two task-specific language models: a Question Decomposition (M_{QD}) model, which decomposes complex questions into interpretable subquestions, and a Question

Answering (M_{QA}) model, which generates context-grounded sub-answers and synthesizes the final response. Both models are trained under a unified instruction-following paradigm. The third component is a hybrid retrieval module that combines dense semantic retrieval, sparse lexical retrieval, and cross-encoder reranking. This module dynamically retrieves relevant demonstrations and supporting evidence at each reasoning step, providing contextual grounding for both question decomposition and answer generation. By leveraging the complementary strengths of semantic and lexical retrieval, the retrieval module improves retrieval quality and ensures access to evidence that is both semantically relevant and lexically precise.

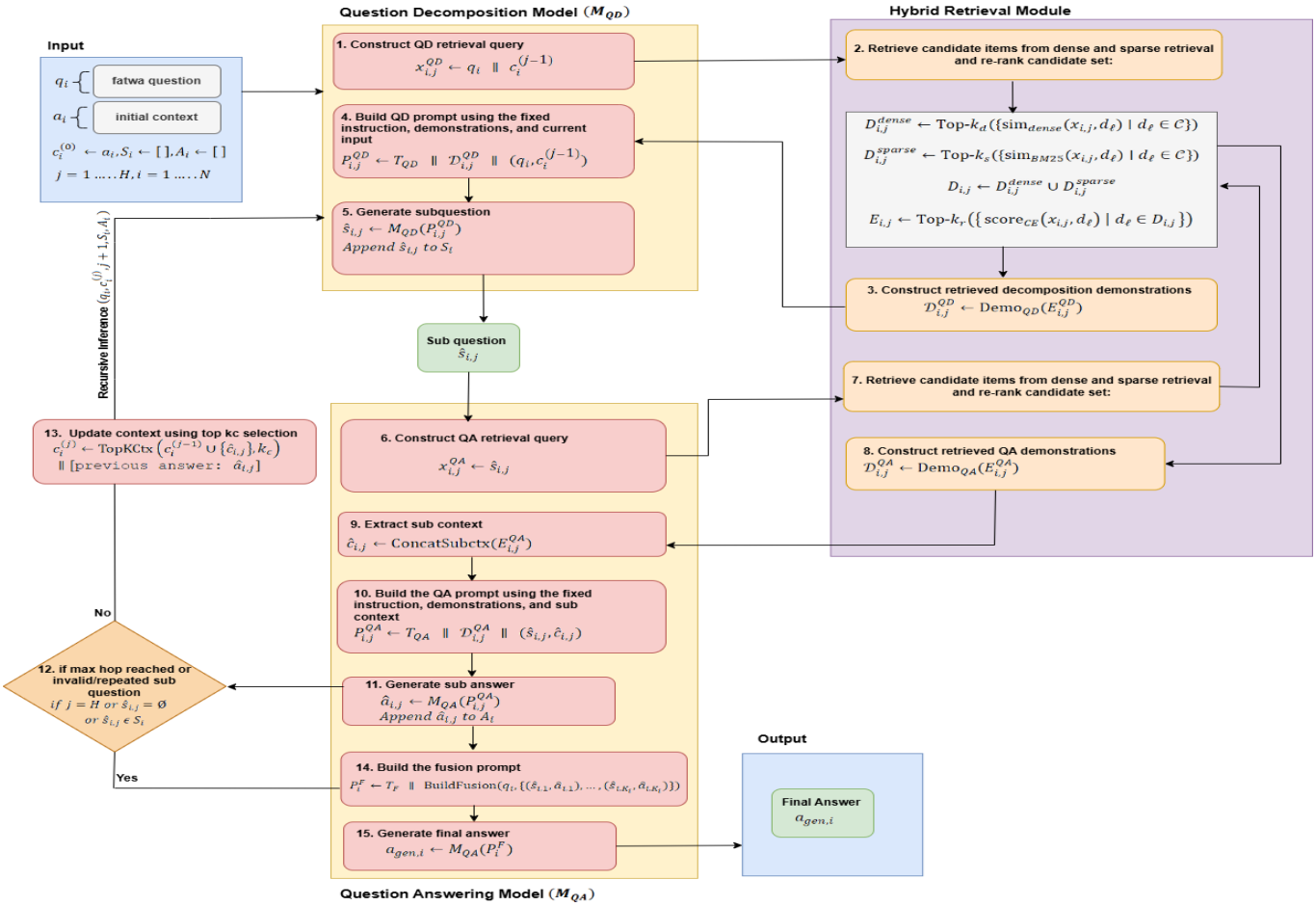

Fig. 1. The architecture of the proposed approach.

The final component is a recursive reasoning process that orchestrates the interaction among the decomposition, retrieval, and answer generation modules. Given a complex fatwa question, the proposed method recursively generates sub-questions, retrieves supporting evidence, produces intermediate answers, and updates the reasoning context before proceeding to subsequent reasoning steps. Once the recursive reasoning process terminates, the generated sub-answers are aggregated and used to construct a final answer that is coherent, contextually grounded, and faithful to the retrieved evidence. Through the coordinated interaction of these components, the proposed approach transforms complex fatwa questions into a sequence of interpretable reasoning steps, enabling structured multi-hop reasoning and dynamic evidence integration. This design enables the generation of more accurate and contextually relevant answers while improving reasoning transparency and reducing the likelihood of unsupported or hallucinated responses.

B. *Instruction-Based Supervision*

To support explicit multi-hop reasoning, the proposed approach adopts an instruction-based supervision strategy derived from the MAFQA [12] dataset—a benchmark for multi-hop Arabic fatwa question answering (QA) that contains complex questions, intermediate reasoning chains, subquestions, sub-answers, and supporting evidence passages. Unlike conventional QA datasets that provide only question–answer pairs, MAFQA [12] explicitly captures the intermediate reasoning structures required to address complex fatwa inquiries. This rich annotation enables the construction of supervision signals for both question decomposition and answer generation within a unified instruction-following paradigm.

Each dataset instance is represented as a structured reasoning chain consisting of a complex fatwa question, a set of intermediate subquestions, their corresponding sub-answers, supporting evidence passages, and a synthesized final answer.

Following data preprocessing and normalization, instruction-based supervision is constructed for two core tasks: question decomposition (QD) and question answering (QA).

Formally, for each dataset instance i, the supervision process maps a complex question into a sequence of intermediate reasoning steps:

$$q_i \rightarrow \{(s_{i,j}, c_{i,j}, a_{i,j})\}_{j=1}^{K_i} \quad (1)$$

where, $s_{i,j} \in SQ_i$ denotes the j -th sub-question, $a_{i,j} \in SA_i$ represents the corresponding sub-answer, and $c_{i,j}$ denotes the context associated with that reasoning step. The context is constructed from retrieved evidence grounded in the supporting passages associated with the original question. The variable K_i denotes the total number of reasoning steps required to answer the complex question q_i. This formulation explicitly models the intermediate reasoning structure underlying complex fatwa analysis and provides fine-grained supervision signals for both decomposition and answer generation. By exposing the model to structured reasoning chains during training, the framework encourages explicit reasoning behavior, progressive evidence integration, and improved contextual grounding throughout the inference process.

1) Question decomposition supervision: Question decomposition is formulated as a structured generation task in which a complex fatwa question is transformed into an ordered sequence of sub-questions together with their associated contexts. The objective of this task is to explicitly model the intermediate reasoning steps required to answer complex multi-hop questions. For each training instance, the input consists of natural-language instruction requesting the decomposition of a complex question into a sequence of sub-questions. To facilitate instruction-following behavior, the prompt is augmented with semantically similar decomposition demonstrations retrieved from the training set using dense semantic retrieval. These demonstrations provide examples of how complex questions can be decomposed into coherent reasoning chains and serve as in-context guidance during training.

The target output is the gold decomposition represented as an ordered set of sub-question–context pairs. By learning this mapping, the model acquires the ability to generate interpretable reasoning steps that progressively simplify complex fatwa questions into manageable sub-problems. The resulting decomposition serves as the foundation for the subsequent retrieval and answer generation stages.

2) Question answering supervision: The question answering task is formulated as a context-grounded generation problem in which the model learns to generate answers conditioned on both a question and its supporting evidence. The supervision strategy is designed to encourage the generation of faithful responses that remain grounded in the provided context while minimizing unsupported or hallucinated content. For each training instance, the input consists of a natural-language instruction, a question, and its associated context. Similar to the decomposition task, semantically related demonstration examples are incorporated into the prompt to provide in-context guidance. The target output is the corresponding gold answer obtained from the dataset annotations.

The proposed approach learns two complementary answer generation functions. The first generates intermediate sub-answers for individual sub-questions using their associated evidence passages. The second synthesizes a final answer by combining the original complex question with the generated reasoning chain consisting of sub-question–sub-answer pairs. Through this hierarchical supervision strategy, the model learns not only to generate context-grounded intermediate answers but also to aggregate multiple reasoning steps into a coherent final response. Consequently, the resulting supervision strategy directly supports the recursive multi-hop reasoning process described in Section E.

C. Task-Specific QD and QA Models

Following the construction of instruction-based supervision, the proposed approach trains two task-specific language models to learn different stages of the multi-hop reasoning process. The proposed method adopts a modular design in which question decomposition and answer generation are learned separately. This design allows each model to specialize in its respective task and facilitates more effective handling of complex fatwa questions that require multiple reasoning steps.

Both tasks are formulated under a unified instruction-following paradigm, where the model receives natural-language instruction together with task-specific inputs and generates the corresponding output. Furthermore, semantically similar demonstration examples are dynamically retrieved and incorporated into the prompt to provide in-context guidance during both training and inference. This design enables the models to learn structured reasoning patterns from relevant examples while maintaining consistency with the instruction-based supervision strategy described in Section B.

At the end of the training stage, two task-specific models are obtained:

$\boldsymbol{M}_{\mathrm{QD}}$: Question Decomposition Model

$\boldsymbol{M}_{\mathrm{QA}}$: Question Answering Model

The Question Decomposition model M_{QD} is responsible for transforming a complex fatwa question into an ordered sequence of sub-question–context pairs that explicitly represent the intermediate reasoning steps required to answer the original query. The Question Answering model M_{QA} is responsible for generating context-grounded answers for individual sub-questions and synthesizing the final answer to the original complex question. Together, M_{QD} and M_{QA} form the core learning components of the proposed approach. During inference, the decomposition model generates the intermediate reasoning structure, while the answering model produces evidence-grounded sub-answers and the final synthesized response. This modular design enables the proposed approach to support explicit multi-hop reasoning and facilitates the recursive reasoning process described in Section E.

D. Hybrid Retrieval Module

To support context-aware reasoning throughout the inference process, the proposed approach incorporates a hybrid retrieval module that combines dense semantic retrieval, sparse lexical retrieval, and cross-encoder reranking. The retrieval module is invoked recursively during both the question decomposition (QD) and question answering (QA) stages to retrieve relevant demonstrations and supporting evidence that adapt to the evolving reasoning state of the model. The retrieval process serves two complementary purposes. During the question decomposition stage, it retrieves semantically similar complex questions together with their corresponding decompositions, which are incorporated as few-shot demonstrations to guide sub-question generation. During the question answering stage, it retrieves both supporting evidence passages and question–answer demonstrations for each generated sub-question, enabling the generation of context-grounded answers.

1) Index construction: To prevent information leakage, all retrieval indices are constructed exclusively from the training split. Two task-specific indices are maintained within the proposed approach. The first index supports the question decomposition task and stores representations of complex fatwa questions together with their associated contextual information. This index is used to retrieve semantically similar examples whose gold decompositions serve as in-context demonstrations during decomposition. The second index supports the question answering task and is constructed at a finer granularity using sub-question–context pairs extracted from the training data. This design enables the retrieval of reasoning steps and evidence passages that are directly relevant to the generated sub-questions during recursive inference.

*2) Retriever architecture:*The retrieval architecture combines dense semantic retrieval and sparse lexical retrieval to leverage their complementary strengths. Previous studies have shown that dense retrievers are highly effective at capturing semantic relevance but may overlook exact lexical matches and domain-specific terminology, whereas BM25-based retrieval preserves important lexical cues and keyword overlap that may not be fully reflected in dense embedding representations [39]. Furthermore, hybrid retrieval architectures that integrate dense and sparse retrieval have been shown to consistently improve retrieval effectiveness across various information retrieval tasks, particularly for domain-specific queries and specialized terminology [40], [41]. Motivated by these findings, the proposed approach employs a hybrid retrieval strategy that combines semantic and lexical retrieval before applying cross-encoder reranking. Dense retrieval captures semantic similarity between queries and candidate items using multilingual sentence embeddings, enabling the retrieval of conceptually related examples even when lexical overlap is limited. Sparse retrieval, on the other hand, employs BM25-based lexical matching to preserve exact term overlap and domain-specific terminology. This combination is particularly important in Arabic fatwa question answering, where semantically relevant evidence may not always share the same wording as the query, while precise religious terminology often plays a critical role in determining the final ruling.

Given a query generated during the reasoning process, retrieval is performed independently using both dense and sparse retrievers. The resulting candidate sets are merged to form a unified candidate pool:

$$D_{i,j} = D_{i,j}^{dense} \cup D_{i,j}^{sparse} \tag{2}$$

where, $D_{i,j}^{dense}$ and $D_{i,j}^{sparse}$ denote the candidate sets returned by the dense and sparse retrievers, respectively, at the reasoning step j for instance i.

3) Cross-encoder re-ranking: To further improve retrieval quality, the merged candidate set is refined using a cross-encoder re-ranking model. Unlike bi-encoder retrieval methods that independently encode queries and documents, the cross-encoder jointly processes each query–candidate pair and produces a relevance score that captures fine-grained semantic interactions between them. Formally, each candidate is assigned a reranking score:

$$\text{score}_{CE}(x_{i,j}, d_\ell) = \text{CE}(x_{i,j}, d_\ell) \tag{3}$$

where, v$x_{i,j}$ denotes the retrieval query at the reasoning step j, d_ℓ represents a candidate item, and $CE(\cdot)$ denotes the cross-encoder reranking model. The candidates are subsequently ranked according to their cross-encoder scores, and the top-ranked items are retained as the final retrieval set used during inference. During question decomposition, the retrieved items provide decomposition demonstrations that guide the generation of structured sub-question–context pairs. During question answering, they supply both supporting evidence and question–answer demonstrations that enable the generation of context-grounded sub-answers and final responses. By integrating dense retrieval, sparse retrieval, and cross-encoder reranking within a unified architecture, the proposed hybrid retrieval module retrieves and reranks supporting evidence throughout the recursive reasoning process, thereby supporting both decomposition and answer generation stages.

E. Recursive Multi-Hop Reasoning Process

At inference time, the proposed approach executes a recursive, multi-hop reasoning process that integrates question decomposition, hybrid retrieval, and question answering (QA) within a unified function. Rather than treating these components as isolated pipeline stages, this process serves as the core reasoning mechanism of the proposed approach, operationalizing the instruction-based supervision strategy, leveraging the fine-tuned task-specific models, and dynamically integrating the hybrid retrieval module. The complete formal execution flow is presented in Algorithm 1.

Algorithm 1: Recursive Prompt-Driven Multi-Hop Reasoning Process

Input: Complex question q_i, initial context a_i, maximum reasoning hops H, retrieval hyperparameters (k_d, k_s, k_r), and context-selection hyperparameter k_c

Output: Generate final answer $a_{gen,i}$

Constants:

I_{QD}: question decomposition retrieval index, I_{QA}: question answering retrieval index, $C = I_{QD} \cup I_{QD}$: unified retrieval corpus, T_{QD}, T_{QA}, T_F: fixed instruction templates

Initialization: $j \longleftarrow 1,\ c_i^{(0)} \leftarrow a_i,\ S_i \leftarrow [],\ A_i \leftarrow []$

Function: RecursiveInfer $(q_i,\ c_i^{(j-1)},\ j,\ S_i,\ A_i)$

if $j > H$ **then**

$K_i = |S_i|$

$P_i^F \longleftarrow T_F \parallel BuildFusion\ (q_i, \{(\hat{s}_{i,1}, \hat{a}_{i,1}), \dots, (\hat{s}_{i,k_i}, \hat{a}_{i,k_i})\})$

$a_{gen,i} \longleftarrow M_{QA}(P_i^F)$

return $a_{gen,i}$

Question Decomposition (QD)

$x_{i,j}^{QD} \longleftarrow q_i \parallel c_i^{(j-1)}$

$D_{i,j}^{dense} \longleftarrow Top - k_d\ (sim_{dense}(x_{i,j}^{QD}, d_\ell), d_\ell \in C)$

$D_{i,j}^{sparse} \longleftarrow Top - k_s\ (sim_{BM25}(x_{i,j}^{QD}, d_\ell), d_\ell \in C)$

$D_{i,j} \longleftarrow D_{i,j}^{dense} \cup D_{i,j}^{sparse}$

$E_{i,j}^{QD} \longleftarrow Top - k_r\ (score_{CE}(x_{i,j}^{QD}, d_\ell), d_\ell \in D_{i,j})$

$D_{i,j}^{QD} \longleftarrow Demo_{QD}(E_{i,j}^{QD})$

$P_{i,j}^{QD} \longleftarrow T_{QD} \parallel D_{i,j}^{QD}(q_i,\ c_i^{(j-1)})$

$\hat{s}_{i,j} \longleftarrow M_{QD}(P_{i,j}^{QD})$

if $\hat{s}_{i,j} = \emptyset\ or\ \hat{s}_{i,j} \in S_i$ **then**

$K_i = |S_i|$

$P_i^F \longleftarrow T_F \parallel BuildFusion\ (q_i, \{(\hat{s}_{i,1}, \hat{a}_{i,1}), \dots, (\hat{s}_{i,k_i}, \hat{a}_{i,k_i})\})$

$a_{gen,i} \longleftarrow M_{QA}(P_i^F)$

return $a_{gen,i}$

$Append\ \hat{s}_{i,j}$ to S_i

Question Answering (QA)

$x_{i,j}^{QA} \longleftarrow \hat{s}_{i,j}$

$D_{i,j}^{dense} \longleftarrow Top - k_d\ (sim_{dense}(x_{i,j}^{QA}, d_\ell), d_\ell \in C)$

$D_{i,j}^{sparse} \longleftarrow Top - k_s\ (sim_{BM25}(x_{i,j}^{QA}, d_\ell), d_\ell \in C)$

$D_{i,j} \longleftarrow D_{i,j}^{dense} \cup D_{i,j}^{sparse}$

$E_{i,j}^{QA} \longleftarrow Top - k_r\ (score_{CE}(x_{i,j}^{QA}, d_\ell), d_\ell \in D_{i,j})$

$\hat{c}_{i,j} \longleftarrow ConcatSubctx\ (E_{i,j}^{QA})$

$D_{i,j}^{QA} \longleftarrow Demo_{QA}(E_{i,j}^{QA})$

$P_{i,j}^{QA} \longleftarrow T_{QA} \parallel D_{i,j}^{QA}(\hat{s}_{i,j},\ \hat{c}_{i,j})$

$\hat{a}_{i,j} \longleftarrow M_{QA}(P_{i,j}^{QA})$

if $\hat{a}_{i,j} = \emptyset$ **then**

$\hat{a}_{i,j} \longleftarrow$ fallback answer

$Append\ \hat{a}_{i,j}\ to\ A_i$

Context Update

$c_i^{(j)} \leftarrow TopKCtx\left((c_i^{(j-1)} \cup \{\hat{c}_{i,j}\}), k_c\right) \parallel$ [previous answer: $\hat{a}_{i,j}$]

return RecursiveInfer $(q_i,\ c_i^{(j)},\ j+1,\ S_i,\ A_i)$

1) Formal initialization and core logic: Given a complex input question q_i and an initial context a_i, the inference process is initialized by setting the reasoning hop index $j \leftarrow 1$, the initial running context $c_i^{(0)} \leftarrow a_i$, and initializing empty lists for sub-questions $S_i \leftarrow []$ and intermediate sub-answers $A_i \leftarrow []$. The system then invokes the recursive function, which operates over sequential reasoning hops bounded by the maximum hop hyperparameter H. If the current hop index exceeds this limit $j > H$, the recursion terminates and triggers the final answer synthesis routine. Let $K_i = |S_i|$ denote the total number of valid hops executed. The system constructs a final fusion prompt P_i^F by combining the fixed instruction template T_F with the

aggregated chain of all generated sub-question and sub-answer pairs via the $BuildFusion$ function. The final ground answer $a_{gen,i}$ is then synthesized using the fine-tuned QA model M_{QA}.

$$a_{gen,i} \longleftarrow M_{QA}(P_i^F) \quad (4)$$

If $j \leq H$, the proposed approach systematically executes three consecutive phases within the active reasoning hop:

- Question Decomposition (QD): To isolate the next logical reasoning step, a decomposition query $x_{i,j}^{QD}$ is formulated by concatenating the original question with the context accumulated up to the previous step.

$$x_{i,j}^{QD} \longleftarrow q_i \parallel c_i^{(j-1)} \quad (5)$$

- This query is routed to the hybrid retrieval module to extract lexically and semantically relevant candidates from the unified corpus C. After cross-encoder re-ranking, the selected passages are converted into in-context demonstrations to build the final decomposition prompt $P_{i,j}^{QD}$. Finally, the fine-tuned decomposition model M_{QD} processes this prompt to output the candidate sub-question.

$$\hat{s}_{i,j} \longleftarrow M_{QD}\left(P_{i,j}^{QD}\right) \quad (6)$$

- Early Stopping Criteria: To avoid redundant computation and prevent infinite reasoning loops, an early stopping condition is evaluated. If the generated sub-question $\hat{s}_{i,j}$ is empty Ø or matches any sub-question already present in the S_i list, build the final fusion prompt P_i^F, and return the final generated answer $a_{gen,i}$. If the sub-question is valid and unique, it is appended to the S_i list.

- Evidence Retrieval and Sub-Answer Generation: Once a unique sub-question $\hat{s}_{i,j}$ is committed, it serves as the foundation for the QA query.

$$x_{i,j}^{QA} \longleftarrow \hat{s}_{i,j} \quad (7)$$

Using this query, the proposed approach executes an identical hybrid retrieval and cross-encoder re-ranking pipeline over the unified corpus C to extract the top context passages. An evidence-specific sub-context $\hat{c}_{i,j}$ is formed by concatenating these top passages.

$$(\hat{c}_{i,j} \longleftarrow ConcatSubctx\ (E_{i,j}^{QA})) \quad (8)$$

These elements are assembled into the QA prompt block $P_{i,j}^{QA}$. The fine-tuned QA model then evaluates the prompt to generate a sub-answer:

$$\hat{a}_{i,j} \longleftarrow M_{QA}\left(P_{i,j}^{QA}\right) \quad (9)$$

- Fallback Mechanism: To prevent an empty intermediate output from being propagated to subsequent reasoning steps, the generated sub-answer is evaluated immediately after generation. If the QA model produces a non-empty response, the generated response is retained as the intermediate sub-answer. If the model output is empty, the fallback mechanism is activated and the predefined fallback answer "No answer found" is assigned to the current intermediate sub-answer:

$$\hat{a}_{i,j} \longleftarrow fallbackanswer \quad (10)$$

The resulting sub-answer is then appended to the intermediate sub-answer list A_i, after which the recursive procedure proceeds to the subsequent context-update step. Thus, the fallback mechanism is triggered specifically by an empty intermediate QA output and replaces the empty output with the predefined fallback value before the reasoning process continues.

2) Context update mechanism: A key component of the recursive reasoning process is the context update mechanism, which controls how information is propagated across sequential reasoning steps while preventing the context buffer from expanding unboundedly. Let the candidate passage pool at reasoning step j be defined as the union of the previous step's context and the newly extracted sub-context passages:

$$\mathcal{P}_{i,j} = c_i^{(j-1)} \cup \{\hat{c}_{i,j}\} \quad (11)$$

Instead of monotonically accumulating all historical context, the proposed approach applies a relevance-based selection function, $TopKCtx$, formally defined as:

$$c_i^{(j)} \leftarrow TopKCtx\left(\mathcal{P}_{i,j}, k_c\right) \parallel [\text{previous answer: } \hat{a}_{i,j}]$$
$$\text{subject to} \quad \left|\text{TopKCtx}\left(\mathcal{P}_{i,j}, k_c\right)\right| = \min\left(k_c, \left|\mathcal{P}_{i,j}\right|\right) \quad (12)$$

where, k_c is a context-selection hyperparameter that determines the maximum number of top-ranked context segments retained at each reasoning step. The string literal of the current sub-answer $\hat{a}_{i,j}$ is trailing-concatenated to anchor the generation trajectory. The relevance scoring function evaluates individual candidate segments $x \in \mathcal{P}_{i,j}$ using the cosine similarity between the dense semantic representations of the active QA query $x_{i,j}^{QA}$ and each candidate text chunk:

$$\text{Score}(x) = \cos\left(\Phi\left(x_{i,j}^{QA}\right), \Phi(x)\right) = \frac{\Phi\left(x_{i,j}^{QA}\right) \cdot \Phi(x)}{\left\|\Phi\left(x_{i,j}^{QA}\right)\right\| \|\Phi(x)\|} \quad (13)$$

where, $\Phi(.)$ denotes the dense sentence embedding model utilized by the retrieval module. Thus, $TopKCtx$ filters and retains the top k_c context segments from $\mathcal{P}_{i,j}$ that are most semantically aligned with the current sub-question query $x_{i,j}^{QA}$. Following this context update, the recursive function advances the system state by recursively executing the next hop $j + 1$ with the updated context $c_i^{(j)}$. This tightly coupled design guarantees that each step remains grounded in verified evidence, bounding the context to minimize hallucination while maximizing the interpretability of the multi-hop reasoning chain.

IV. Experimental Evaluation

A. Dataset Preparation and Preprocessing

All experiments were conducted using the MAFQA dataset, which contains complex Arabic fatwa questions alongside expert-annotated intermediate reasoning chains, supporting

contexts, and reference answers. The dataset was specifically designed to support both question decomposition (QD) and question answering (QA) tasks within the proposed recursive reasoning approach. Prior to training, the dataset was processed through a comprehensive preprocessing pipeline to enhance data quality and ensure consistent textual representations. Duplicate instances with identical question text were removed to prevent data leakage and redundant supervision. Arabic text was subsequently normalized by removing diacritics and elongation characters (tatweel), unifying orthographic variants (e.g., mapping ٱ, إ, and آ to ا), standardizing character forms (such as ۀ), eliminating invisible Unicode characters, and removing extraneous whitespace or formatting artifacts. These preprocessing steps reduce lexical sparsity and improve tokenization consistency across different language models. To ensure reproducibility, the processed dataset was randomly shuffled using a fixed seed and partitioned into training, validation, and test sets following an 80/10/10 split. The training partition was used for model optimization and retrieval index construction, the validation set was employed for hyperparameter selection and early stopping, and the test set was reserved exclusively for final evaluation. Constructing all retrieval indices solely from the training split prevented information leakage during retrieval-augmented inference, ensuring a rigorous evaluation protocol.

Following preprocessing, the dataset was converted into task-specific training instances corresponding to the two learning objectives described in Section III. For the QD task, each instance maps a complex fatwa question to its structured decomposition, consisting of intermediate subquestions and their associated contexts. For the QA task, each instance pairs a question (or a generated subquestion) with its corresponding supporting context and reference answer. Source sequences were truncated to a maximum length of 512 tokens, and target sequences were limited to 256 tokens to satisfy model input constraints. Padding tokens in the target sequences were masked using the value -100 to exclude them from loss computation during supervised fine-tuning.

B. Baseline Models

To rigorously evaluate the proposed approach, a diverse set of baseline models was selected, encompassing Arabic-specialized sequence-to-sequence (seq2seq) architectures, multilingual encoder–decoder models, and instruction-tuned large language models (LLMs). This selection facilitates a systematic comparison across models with varying pretraining strategies, linguistic coverage, and reasoning capabilities, thereby providing a comprehensive assessment of the proposed recursive reasoning approach. The first category comprises Arabic seq2seq models, specifically AraBART [42], AraT5-base [43], and Arabic-T5-small [44]. These models were pretrained on large-scale Arabic corpora to capture the nuanced linguistic characteristics of the language. AraBART [42] is an Arabic adaptation of the BART architecture employing a six-layer encoder and a six-layer decoder, whereas AraT5-base [43] and Arabic-T5-small [44] belong to the T5 family and were optimized for Arabic text generation through extensive pretraining on domain-specific datasets. These architectures serve as strong language-specific baselines for evaluating both question decomposition and answer generation. The second category consists of multilingual seq2seq models, represented by mT5-small [45] and mT5-base [46]. Both models belong to the multilingual T5 (mT5) family and were pretrained on the large-scale mC4 corpus covering 101 languages. Their multilingual pretraining enables cross-lingual knowledge transfer, making them suitable baselines for assessing the necessity of language-specific pretraining in complex Arabic fatwa question answering.

To evaluate the performance of contemporary instruction-following architectures, the experiments also include GPT-4.1 [47] and GPT-3.5-turbo [48]. These instruction-tuned LLMs excel at complex reasoning and natural language generation via prompt-based interaction. Given its advanced contextual understanding, GPT-4.1 [47] provides an authoritative baseline for multi-hop question decomposition and answer generation, while GPT-3.5-turbo [48] offers a lightweight yet competitive alternative for assessing instruction-tuned models within the proposed approach.

C. Training Settings and Hyperparameters

To ensure fair and reproducible experimentation, all evaluated models were trained and evaluated under standardized experimental settings whenever applicable. Sequence-to-sequence (seq2seq) models were implemented using the Hugging Face Transformers framework, whereas GPT-based models were fine-tuned through the OpenAI fine-tuning API. Separate models were trained for the Question Decomposition (QD) and Question Answering (QA) tasks using the instruction-based supervision described in Section III. For the seq2seq models, optimization was performed using the AdamW optimizer with a learning rate of 5×10^{-5} for five epochs. Model selection was based on the lowest validation loss, and early stopping was employed to mitigate overfitting. During inference, deterministic beam search was adopted with a beam size of four, together with repetition control mechanisms to improve generation quality. The GPT-based models (GPT-4.1 and GPT-3.5-turbo) were fine-tuned using conversational JSONL datasets constructed from the training split. To ensure deterministic and reproducible generation, stochastic sampling was disabled by setting the temperature to 0.0 and top-(p) to 1.0 during inference.

The proposed recursive prompt-driven reasoning approach employed identical retrieval and inference settings across all evaluated models. Hybrid retrieval combined dense semantic retrieval using the intfloat/multilingual-e5-large embedding model with sparse BM25 retrieval, followed by reranking using the BAAI/bge-reranker-v2-m3 cross-encoder. Retrieval indices were constructed exclusively from the training partition to prevent information leakage during evaluation. Recursive reasoning was limited to a maximum of three reasoning hops, while identical retrieval parameters were adopted throughout all experiments. The principal training, decoding, and retrieval hyperparameters are summarized in Table I.

The experiments were conducted in a Google Colab environment equipped with an NVIDIA T4 GPU. The GPU was used to execute the locally deployed Hugging Face models and other local components of the framework, including dense embedding generation, retrieval, reranking, and evaluation. GPT-based models were accessed through the OpenAI API and

therefore did not require local GPU deployment. Consequently, the local computational requirements of the proposed framework are primarily associated with the retrieval and supporting processing components, while GPT-based model inference was performed remotely through the API.

TABLE I. TRAINING, INFERENCE, AND RETRIEVAL HYPERPARAMETERS USED THROUGHOUT THE EXPERIMENTS

Category	Parameter	Value
Global Settings	Random seed	42
	Training epochs	5
	Maximum reasoning hops ((H))	3
	Maximum QD generation tokens	160
	Maximum QA generation tokens	128
Seq2Seq Models	Optimizer	AdamW
	Learning rate	(5\times10^{-5})
	Weight decay	0.01
	Training/evaluation batch size	4 / 4
	Maximum source/target length	512 / 256
	Evaluation strategy	Per epoch
	Early stopping	Validation loss
	Beam size	4
	No-repeat n-gram	3
	Repetition penalty	1.1
Retrieval Module	Dense retrieval (FAISS) top-(k)	20
	Sparse retrieval (BM25) top-(k)	50
	Candidates reranked	16
	Retrieved demonstrations	3
	Top-(k) context segments	5
GPT-based Models	Temperature	0.0
	Top-(p)	1.0

D. Evaluation Metrics

The proposed approach is evaluated from two complementary perspectives: end-to-end final answer generation and retrieval effectiveness. The evaluation protocol combines lexical, semantic, and retrieval metrics to comprehensively assess the quality of the generated final answers and the effectiveness of the hybrid retrieval module. Given the rich morphology and lexical variation of the Arabic language, the selected metrics measure not only surface-level textual similarity but also semantic alignment between the generated and reference answers.

1) Lexical and semantic similarity metrics: The quality of the generated final answers is evaluated using a combination of lexical overlap and semantic similarity metrics. Lexical overlap is calculated using BLEU and ROUGE-L. The Bilingual Evaluation Understudy (BLEU) score calculates precision based on shared n -grams, incorporating a brevity penalty (BP) to penalize unfairly compressed generations. Formally, it is computed as:

$$\text{BLEU} = BP \cdot \exp\left(\frac{1}{N}\sum_{n=1}^{N} w_n \ln p_n\right) \quad (14)$$

where, N is the maximum n-gram size (typically 4), w_n represents the baseline weight for each n-gram order, and p_n refers to the clipped precision for n-grams of order n. The brevity penalty (BP) is defined as:

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{\left(1-\frac{r}{c}\right)} & \text{if } c \le r \end{cases} \quad (15)$$

where, r is the token count of the reference text and c is the token count of the candidate output. The clipped n-gram precision (p_n) prevents score inflation from repetitive generated words and is expressed as:

$$p_n = \frac{\sum_{C\in\{\text{Candidates}\}}\sum_{\text{n-gram}\in C}\text{Count}_{\text{clip}}(\text{n-gram})}{\sum_{C'\in\{\text{Candidates}\}}\sum_{\text{n-gram}'\in C'}\text{Count}(\text{n-gram}')} \quad (16)$$

The ROUGE-L score focuses on recall, leveraging the Longest Common Subsequence (LCS) to capture the longest ordered series of tokens shared between the candidate and target sequences:

$$\text{ROUGE-L} = \frac{\text{LCS}(X,Y)}{m} \quad (17)$$

where, X represents the predicted text string, Y represents the gold reference string, and m is the total length of the reference text. To complement lexical evaluation, BERTScore is employed to compute a token-level semantic match by performing greedy cosine alignments over dense token representations within a multilingual embedding space. Let x and y denote the token embedding sequences of the predicted and reference text, respectively. Precision (BERT-P), recall (BERT-R), and F1-score (BERT-F) are computed as follows:

$$BERT - P = \frac{1}{|x|}\sum_{x_i \epsilon\, x} max_{y_j \in y}\, x_i^T y_j \quad (18)$$

$$BERT - R = \frac{1}{|y|}\sum_{y_i \epsilon\, y} max_{x_i \in x}\, x_i^T y_j \quad (19)$$

$$BERT - F = \frac{2\,(BERT-P.BERT-R)}{BERT-P + BERT-R} \quad (20)$$

where, $x_i^T y_j$ denotes the cosine similarity between token-level embeddings.

2) Relevance metric: This metric calculates the cosine similarity between the pooled sentence embeddings of the candidate answer (A_i') and the gold reference answer (A_i):

$$\text{Relevance} = \frac{1}{N}\sum_{i=1}^{N}\cos\big(E(A_i'), E(A_i)\big) \quad (21)$$

where, $\text{E}(\cdot)$ denotes the sentence embedding function and N represents the total number of evaluated QA instances. A higher relevance value indicates stronger semantic alignment with the reference answer.

3) Retrieval evaluation metrics: Because retrieval operations ground both the decomposition and answering tracks, the hybrid retrieval engine is evaluated independently using standard information retrieval performance indicators.

Recall@k measures the ratio of search operations where at least one correct target (supporting context passage or prompt demonstration) is positioned within the top-k returned elements. Let Q denote the collection of queries, R_i^k the set of top-k returned candidates for an active query i, and G_i the ground-truth target set:

$$\text{Recall@k} = \frac{1}{|Q|}\sum_{i=1}^{|Q|} 1\left(R_i^k \cap G_i \neq \emptyset\right) \quad (22)$$

where, $1(\cdot)$ represents an indicator function returning 1 if the intersection is non-empty, and 0 otherwise.

The positional quality of the ranking results is evaluated using the Mean Reciprocal Rank (MRR), which focuses on the explicit rank index (r_i) of the first true matching candidate returned for a query i:

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|} \frac{1}{r_i} \quad (23)$$

If a query yields no correct matches within the ranked output pool, its reciprocal rank defaults to zero. Higher MRR indices indicate that the relevant supporting evidence passages and few-shot demonstrations are positioned earlier in the ranked retrieval list, reflecting better retrieval precision.

E. Evaluation of the Hybrid Retrieval Module

The hybrid retrieval module constitutes a fundamental component of the proposed approach, as it provides the demonstrations and supporting evidence required during both the Question Decomposition (QD) and Question Answering (QA) stages. To assess its effectiveness, the retrieval module is evaluated independently before conducting the end-to-end experiments. The proposed retrieval module combines dense semantic retrieval using the intfloat/multilingual-e5-large embedding model, sparse lexical retrieval using BM25, and cross-encoder reranking using BAAI/bge-reranker-v2-m3. Retrieval is performed recursively during inference, where the retrieval query dynamically evolves according to the current reasoning step. During the QD stage, the retrieval module returns semantically related decomposition demonstrations, whereas during the QA stage it retrieves supporting evidence passages and question–answer demonstrations for the generated sub-question. To ensure a fair evaluation, all retrieval indices are constructed exclusively from the training partition. Retrieval performance is evaluated on three complementary tasks: QD demonstration retrieval, QA demonstration retrieval, and evidence context retrieval. For each task, the proposed hybrid retrieval module is compared against two individual retrieval baselines: sparse lexical retrieval using BM25 and dense semantic retrieval using the multilingual E5-large encoder. Performance is measured using Recall@5, Recall@10, and Mean Reciprocal Rank (MRR@5). Table II summarizes the retrieval performance across the three retrieval tasks.

The experimental results demonstrate that dense semantic retrieval consistently outperforms sparse lexical retrieval across all retrieval tasks. The improvements achieved by the multilingual E5-large encoder indicate that semantic representations are substantially more effective than lexical matching for retrieving decomposition demonstrations, intermediate reasoning examples, and supporting evidence in Arabic multi-hop fatwa question answering. Across all three retrieval tasks, the proposed hybrid retrieval module achieves the best overall performance. For QD demonstration retrieval, the hybrid approach attains a Recall@5 of 0.9762 and an MRR@5 of 0.9185, outperforming both BM25 and dense retrieval alone. Similar improvements are observed for QA demonstration retrieval, where the proposed method achieves 0.9598 Recall@5 and 0.9008 MRR@5, indicating that combining semantic retrieval with lexical matching and reranking enables more accurate retrieval of semantically related reasoning demonstrations. Evidence context retrieval represents the most challenging retrieval task because relevant supporting passages often exhibit limited lexical overlap with the input query while containing implicit reasoning patterns and domain-specific religious terminology. Although dense retrieval substantially improves retrieval quality over BM25, the proposed hybrid module achieves the highest retrieval accuracy, increasing MRR@5 from 0.6757 obtained by dense retrieval alone to 0.7310 after cross-encoder reranking. This improvement demonstrates that reranking effectively refines candidate ordering by capturing deeper query–document interactions, thereby providing more contextually relevant evidence for subsequent reasoning steps.

TABLE II. RETRIEVAL PERFORMANCE COMPARISON ACROSS THE EVALUATED RETRIEVAL STRATEGIES

Retrieval Method		**Metrics**		
		Recall@5	***Recall@10***	***MRR@5***
QD Demonstration Retrieval	BM25 (sparse)	0.869	0.8929	0.7954
	E5-large (dense)	0.9405	0.9762	0.8903
	Proposed Hybrid Retrieval Module	0.9762	0.9762	0.9185
QA Demonstration Retrieval	BM25 (sparse)	0.7789	0.8191	0.681
	E5-large (dense)	0.9447	0.9648	0.8745
	Proposed Hybrid Retrieval Module	0.9598	0.9698	0.9008
Evidence Context Retrieval	BM25 (sparse)	0.5327	0.608	0.4379
	E5-large (dense)	0.8241	0.8794	0.6757
	Proposed Hybrid Retrieval Module	0.8693	0.8894	0.731

Overall, these results demonstrate that the proposed hybrid retrieval module effectively combines the complementary strengths of dense semantic retrieval, sparse lexical matching, and cross-encoder reranking. The resulting improvements in retrieval accuracy and ranking quality provide stronger contextual grounding for both question decomposition and answer generation, thereby establishing a reliable foundation for the recursive multi-hop reasoning process.

F. End-to-End Evaluation of the Proposed Method

The proposed approach is evaluated under a realistic end-to-end setting to assess its effectiveness in solving complex Arabic fatwa questions through the complete recursive reasoning process. To simulate practical deployment conditions, the initial context is dynamically retrieved using the input question rather than being directly supplied from the dataset. Consequently, the proposed approach performs recursive question decomposition, evidence retrieval, intermediate answer generation, dynamic context updating, and final answer synthesis within a unified inference process. This evaluation therefore reflects the cumulative effect of error propagation across all reasoning stages and provides a realistic assessment of the approach's robustness. To evaluate the effectiveness of task-specific fine-tuning, each model is compared against its corresponding pretrained baseline using identical retrieval settings and inference configurations. Performance is assessed using the lexical and semantic evaluation metrics introduced in Section D. The evaluated models include Arabic-specific sequence-to-sequence models, multilingual sequence-to-sequence models, and instruction-tuned GPT-based models. Table III presents the end-to-end performance of all evaluated models on the final answer generation task.

The results demonstrate that task-specific fine-tuning consistently improves end-to-end reasoning performance across nearly all evaluated architectures despite the substantially more challenging retrieved-context setting. Because the initial context, intermediate evidence, and generated sub-answers are produced recursively during inference, inaccuracies introduced at early reasoning stages may propagate to subsequent steps. Consequently, the reported performance reflects the behavior of the complete reasoning pipeline rather than isolated component performance. Among the Arabic-specific sequence-to-sequence models, AraT5-base exhibits the largest improvement after fine-tuning. Its F1 score increases from 4.84 to 16.43, while ROUGE-L improves from 4.51 to 13.75 and the relevance score rises from 75.38 to 86.60. Similarly, AraBART achieves consistent improvements across nearly all evaluation metrics, particularly in semantic similarity and contextual relevance, whereas Arabic-T5-small demonstrates substantial gains in F1, ROUGE-L, BERTScore, and relevance despite minor decreases in BLEU scores. These observations suggest that fine-tuning primarily improves semantic fidelity rather than exact lexical matching. The multilingual mT5 models remain the weakest performers under the realistic end-to-end setting. Although fine-tuning produces measurable improvements, both mT5-small and mT5-base achieve substantially lower lexical and semantic scores than the Arabic-specific models. Their comparatively lower performance may reflect the multilingual nature of these models, which distributes representational capacity across multiple languages, whereas Arabic-specific models are more specialized in modeling Arabic linguistic characteristics. This difference may be particularly important for MAFQA because of its domain-specific terminology and multi-step reasoning requirements. Differences in model capacity, pretraining data distribution, and tokenization may also contribute to the observed performance gap. Furthermore, the recursive decomposition, retrieval, and answer-generation stages may introduce accumulated noise that particularly affects the smaller multilingual models. Overall, these results indicate an advantage for the evaluated Arabic-specific models over the evaluated multilingual models on the MAFQA dataset.

TABLE III. COMPARATIVE END-TO-END PERFORMANCE OF BASELINE AND FINE-TUNED MODELS ON FINAL ANSWER GENERATION TASK

Model		Metrics								
		F1-token	***BLEU-1***	***BLEU-2***	***ROUGE-2***	***ROUGE-L***	***BERT-P***	***BERT-R***	***BERT-F***	***Relevance***
AraBART	Baseline	12.71	09.14	03.01	04.27	09.66	84.13	86.71	85.65	77.23
	Fine-tuned	15.01	10.67	05.16	04.20	12.24	88.27	86.44	87.33	86.48
Arabic-T5-small	Baseline	07.98	06.39	02.69	01.62	05.94	82.01	84.90	83.42	81.99
	Fine-tuned	11.89	05.40	02.43	02.66	09.82	88.86	85.22	86.99	84.89
AraT5-base	Baseline	04.84	00.52	00.14	0	04.51	82.28	78.01	80.07	75.38
	Fine-tuned	16.43	09.95	05.03	04.73	13.75	89.51	86.32	87.87	86.60
GPT-4.1	Baseline	20.15	15.96	09.05	07.09	15.41	88.00	88.24	88.02	89.67
	Fine-tuned	31.88	26.52	18.63	16.51	26.64	90.14	90.35	90.22	92.39
GPT-3.5 -turbo	Baseline	24.86	20.51	12.39	09.90	19.64	89.36	89.25	89.29	91.61
	Fine-tuned	25.52	20.04	12.49	10.90	20.95	89.86	88.93	89.37	90.34
mT5-base	Baseline	01.42	00.50	00.21	00.24	01.35	79.61	81.78	80.63	78.35
	Fine-tuned	02.40	01.38	00.60	00.44	02.21	76.32	81.54	78.76	78.52
mT5-small	Baseline	02.06	00.85	00.35	00.32	01.94	80.77	82.06	81.40	79.03
	Fine-tuned	03.01	01.51	00.61	00.54	02.77	78.90	82.32	80.53	79.66

The best overall performance is achieved by the GPT-based models, particularly GPT-4.1. After fine-tuning, GPT-4.1 attains the highest scores across almost all evaluation metrics, achieving an F1 score of 31.88, BLEU-1 of 26.52, ROUGE-L of 26.64, BERT-F of 90.22, and a relevance score of 92.39. These results indicate that large instruction-tuned language models benefit substantially from task-specific supervision while maintaining strong reasoning capabilities under recursive multi-hop inference. Although GPT-3.5-turbo also demonstrates competitive performance, its improvements after fine-tuning are

comparatively modest, suggesting that its pretrained model already provides relatively strong zero-shot reasoning capabilities.

Although the best F1-token (31.88) and ROUGE-L (26.64) scores remain modest in absolute terms, these values should be interpreted in light of both the difficulty of the task and the characteristics of the evaluation metrics. MAFQA consists of complex Arabic fatwa questions that require multi-step reasoning, integration of evidence distributed across supporting passages, and synthesis of a final response. Under the end-to-end setting, the system must also retrieve supporting evidence and generate intermediate sub-questions and answers, allowing errors at earlier reasoning stages to propagate to the final response. Moreover, F1-token and ROUGE-L primarily measure lexical overlap and may assign relatively low scores when generated answers express the correct information using wording that differs from the reference. This is particularly relevant to Arabic because of its morphological and lexical variability and the possibility of multiple valid answer formulations. The substantially higher BERT-F (90.22) and Relevance (92.39) scores achieved by the fine-tuned GPT-4.1 model support this interpretation, indicating stronger semantic correspondence than is reflected by lexical overlap alone.

Overall, the end-to-end evaluation confirms the effectiveness of the proposed recursive reasoning approach under realistic deployment conditions. Despite the inherent challenges posed by error propagation across decomposition, retrieval, and answer generation stages, task-specific fine-tuning consistently improves lexical accuracy, semantic similarity, and contextual relevance. Furthermore, the superior performance of GPT-4.1 demonstrates the advantages of combining instruction-based supervision, hybrid retrieval, and recursive multi-hop reasoning within a unified reasoning process for complex Arabic fatwa question answering.

G. Ablation Study

To quantify the contribution of the major components of the proposed method, an ablation study was conducted using the best-performing configuration identified in the previous section, namely the fine-tuned GPT-4.1 model. All ablation experiments were performed under the same end-to-end retrieved-context setting, using identical dataset splits, retrieval settings, inference constraints, and evaluation metrics to ensure a fair comparison. Three ablation configurations were considered. The first removes the recursive decomposition and multi-hop reasoning mechanism, allowing the model to generate the final answer directly from the original question without producing intermediate sub-questions or reasoning steps. The second replaces the proposed hybrid retrieval module with a dense semantic retriever, thereby removing sparse lexical retrieval and cross-encoder reranking. The third evaluates the pretrained GPT-4.1 model without task-specific fine-tuning, while preserving the remaining components of the proposed method. The complete proposed method, integrating recursive decomposition, hybrid retrieval, and task-specific fine-tuning, serves as the reference configuration. Table IV summarizes the quantitative results of the ablation study.

The results demonstrate that all three components contribute positively to the overall performance of the proposed method. The largest performance degradation occurs when recursive decomposition and multi-hop reasoning are removed. In this configuration, the F1 score decreases from 31.88 to 16.21, while BLEU-2 and ROUGE-L decline from 18.63 to 6.42 and from 26.64 to 13.27, respectively. Similarly, the relevance score decreases from 92.39 to 83.28. These substantial reductions indicate that explicitly decomposing complex fatwa questions into intermediate reasoning steps is the primary factor enabling accurate multi-hop reasoning. Removing the hybrid retrieval module also results in noticeable performance degradation. Replacing the hybrid retriever with dense retrieval alone reduces the F1 score to 24.36, while BLEU-2 and ROUGE-L decrease to 11.27 and 19.86, respectively. Although the reduction in BERTScore is relatively small, the decline in lexical and relevance metrics indicates that combining dense retrieval, sparse lexical matching, and cross-encoder reranking provides more accurate demonstrations and supporting evidence than dense retrieval alone. The without fine-tuning configuration similarly exhibits lower performance than the complete proposed method. The pretrained GPT-4.1 model achieves an F1 score of 20.15, which is more than eleven points lower than the fine-tuned model. Consistent reductions are also observed across BLEU, ROUGE-L, BERTScore, and Relevance, demonstrating that task-specific fine-tuning substantially improves the model's ability to perform recursive reasoning and generate contextually grounded answers for complex Arabic fatwa questions. Overall, the ablation study confirms that each component of the proposed method contributes to its effectiveness. Among them, recursive decomposition and multi-hop reasoning have the greatest impact on end-to-end performance, while hybrid retrieval and task-specific fine-tuning further enhance contextual grounding, semantic consistency, and answer quality.

TABLE IV. ABLATION STUDY RESULTS OF THE PROPOSED METHOD UNDER THE END-TO-END RETRIEVED-CONTEXT SETTING

Setting	Metrics								
	F1-token	***BLEU-1***	***BLEU-2***	***ROUGE-2***	***ROUGE-L***	***BERT-P***	***BERT-R***	***BERT-F***	***Relevance***
Without decomposition and recursive reasoning	16.21	12.37	06.42	05.74	13.27	87.65	88.13	87.61	83.28
Without hybrid retrieval	24.36	20.11	11.27	09.44	19.86	88.41	89.63	88.52	91.23
Without finetuning	20.15	15.96	09.05	07.09	15.41	88.00	88.24	88.02	89.67
Proposed method	31.88	26.52	18.63	16.51	26.64	90.14	90.35	90.22	92.39

H. Comparison with State-of-the-Art Methods

To further assess the effectiveness of the proposed method, a comparison with a diverse set of state-of-the-art (SOTA) question answering systems was conducted, representing different architectural paradigms, including multilingual instruction-tuned large language models (LLMs), Arabic-focused LLMs, retrieval-augmented generation (RAG) systems, and reasoning-based approaches. All methods were evaluated using the same MAFQA test split under identical end-to-end retrieved-context conditions to ensure a fair comparison. The instruction-tuned LLMs were evaluated in a zero-shot setting without task-specific fine-tuning, while retrieval-based approaches dynamically retrieved supporting evidence during inference. Performance was assessed using the lexical, semantic, and relevance metrics described in Section D.

The compared systems include five multilingual instruction-tuned LLMs (GPT-5 [49], Gemini-2.5 [50], DeepSeek-LLM-7B [51], Qwen-7B [52], and Mistral-7B [53]), two Arabic instruction-tuned LLMs (SILMA-9B-Instruct-v1.0 [54] and AceGPT-v2-8B [55]), a standard retrieval-augmented generation (RAG) system [56], and the reasoning-based LCW-QD [37] approach proposed by Abdelazim *et al.* [37], which combines question decomposition with long-context reasoning for Arabic multi-hop question answering. Together, these systems represent the principal directions in contemporary Arabic question answering and provide strong baselines for evaluating the effectiveness of the proposed recursive reasoning approach. Table V presents the quantitative comparison between the proposed method and the evaluated state-of-the-art Arabic QA systems.

The proposed method consistently achieves the highest performance across all lexical, semantic, and relevance metrics. Specifically, it obtains an F1-token score of 31.88, BLEU-1 of 26.52, BLEU-2 of 18.63, ROUGE-2 of 16.51, and ROUGE-L of 26.64. It also achieves the highest semantic similarity and contextual relevance, with BERT-F and Relevance scores of 90.22 and 92.39, respectively. These results demonstrate the effectiveness of integrating recursive question decomposition, hybrid retrieval, dynamic context updating, and task-specific fine-tuning within a unified multi-hop reasoning process. Among the multilingual instruction-tuned LLMs, Qwen-7B achieves the strongest overall performance, obtaining an F1-token score of 19.94, ROUGE-L of 16.53, and a Relevance score of 89.58. GPT-5, AceGPT-v2-8B, and Mistral-7B also demonstrate competitive performance, indicating that modern instruction-tuned LLMs possess strong zero-shot reasoning capabilities. Nevertheless, their performance remains substantially below that of the proposed method, suggesting that large-scale pretraining and instruction tuning alone are insufficient for effectively addressing complex multi-hop Arabic fatwa questions.

Among the Arabic-specific models, AceGPT-v2-8B achieves the strongest overall performance, whereas SILMA-9B-Instruct-v1.0 exhibits relatively weak lexical performance despite achieving the highest BLEU-1 score among the baseline systems. This observation suggests that high unigram overlap does not necessarily correspond to better semantic quality or more complete answers. The retrieval-augmented baseline (RAG) improves contextual grounding compared with conventional sequence-to-sequence models, achieving a Relevance score of 88.10. However, its lexical performance remains considerably lower than that of the proposed method, with an F1-token score of 12.95 and ROUGE-L of 9.09. These findings indicate that retrieval augmentation alone is insufficient for effectively solving complex multi-hop fatwa questions without explicit reasoning mechanisms to organize and integrate the retrieved evidence.

TABLE V. COMPARISON OF THE PROPOSED METHOD WITH STATE-OF-THE-ART ARABIC QUESTION ANSWERING SYSTEMS ON THE MAFQA DATASET

Model Category		Metrics						
		F1-token	***BLEU-1***	***BLEU-2***	***ROUGE-2***	***ROUGE-L***	***BERT-F***	***Relevance***
Multilingual Instruction-Tuned LLMs	GPT-5	18.37	14.46	03.67	04.51	12.16	87.76	90.28
	Deepseek-llm-7b	12.79	11.87	02.89	03.08	10.57	86.37	82.12
	Gemini-2.5	08.24	17.21	03.96	02.63	07.11	84.21	80.69
	Qwen-7B	19.94	16.85	09.77	07.69	16.53	87.53	89.58
	Mistral-7B	17.68	14.58	07.85	06.10	14.39	87.24	88.72
Arabic Instruction-Tuned LLMs	SILMA-9B-Instruct-v1.0	06.05	27.46	08.50	02.07	05.35	81.51	80.14
	AceGPT-v2-8B	18.01	16.45	06.28	06.18	14.28	87.82	89.55
Reasoning-based approach	LCW-QD	12.95	08.81	04.15	03.52	09.09	85.22	88.10
Retrieval-Augmented Systems (RAG)		23.36	20.59	09.53	10.23	19.18	88.22	90.26
Proposed method		31.88	26.52	18.63	16.51	26.64	90.22	92.39

Among all baseline systems, the reasoning-based LCW-QD approach achieves the strongest performance, obtaining an F1-token score of 23.36, ROUGE-L of 19.18, BERT-F of 88.22, and a Relevance score of 90.26. Compared with this strongest baseline, the proposed method improves the F1-token score from 23.36 to 31.88, corresponding to a relative improvement of approximately 36.5%. Likewise, ROUGE-L increases from 19.18 to 26.64 (approximately 38.9%), while BERT-F and Relevance improve from 88.22 to 90.22 and from 90.26 to 92.39, respectively. These improvements demonstrate that

recursive reasoning, dynamic context updating, and hybrid retrieval provide substantial benefits beyond those achieved through question decomposition and long-context reasoning alone.

Overall, the results demonstrate that while instruction-tuned language models, retrieval-augmented systems, and decomposition-based reasoning approaches provide competitive performance, the proposed method consistently achieves the strongest results across lexical, semantic, and relevance-based metrics. The findings indicate that integrating recursive reasoning with hybrid retrieval and dynamic contextual grounding enables more accurate, contextually relevant, and semantically faithful answer generation for complex multi-hop Arabic fatwa question answering.

V. Conclusion and Future Directions

This study addressed the challenge of answering complex Arabic fatwa questions, where accurate responses require integrating evidence across multiple sources. The experimental results demonstrate that explicitly modeling intermediate reasoning substantially improves answer quality compared with conventional single-step question answering approaches. By recursively decomposing complex questions, retrieving evidence at each reasoning step, and progressively refining the reasoning context, the proposed recursive prompt-driven reasoning approach improves answer quality and semantic relevance under the evaluated multi-hop Arabic fatwa question answering setting.

The experimental evaluation provides several important findings. First, combining dense semantic retrieval, sparse lexical retrieval, and cross-encoder reranking consistently produces higher-quality supporting evidence than individual retrieval strategies, highlighting the effectiveness of complementary retrieval mechanisms for knowledge-intensive Arabic question answering. Second, task-specific fine-tuning substantially improves end-to-end performance across diverse model architectures, particularly Arabic-specialized models and large language models operating under realistic inference conditions. Third, the ablation study demonstrates that the combined recursive decomposition and multi-hop reasoning process was the primary contributor to the observed performance gains, while hybrid retrieval and task-specific fine-tuning also contribute to the overall performance. Finally, comparisons with representative state-of-the-art methods confirm that the proposed method achieves superior overall performance on the MAFQA dataset across lexical, semantic, and relevance-based evaluation metrics. Beyond these performance improvements, this study demonstrates the effectiveness of integrating explicit multi-hop reasoning with retrieval-augmented language models for domain-specific Arabic question answering. The proposed approach provides an interpretable reasoning process in which intermediate reasoning steps, retrieved evidence, and generated responses remain tightly connected throughout inference, thereby improving both answer transparency and factual grounding.

Future work will focus on several directions. First, the generalizability of the proposed approach will be investigated through evaluation on additional Arabic multi-hop question answering datasets and other knowledge-intensive domains as suitable benchmarks become available. Second, the recursive control mechanism will be further examined through the evaluation of semantic stopping criteria for identifying redundant or unproductive reasoning steps beyond the current termination conditions. Third, alternative context-selection strategies, including history-aware, diversity-aware, evidence-coverage-based, and unpruned approaches, will be compared with the current context update mechanism to assess retained-evidence rates and their impact on final QA performance. Finally, future research will investigate more advanced retrieval and evidence-selection techniques, including agent-based retrieval approaches, and evaluate emerging reasoning-capable language models.

References

[1] P. Shailendra, R. C. Ghosh, R. Kumar, and N. Sharma, "Survey of large language models for answering questions across various fields," in Proc. 10th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS), vol. 1, Mar. 2024, pp. 520–527.

[2] R. S. Roy and A. Anand, *Question answering for the curated web: Tasks and methods in QA over knowledge bases and text collections*. Cham, Switzerland: Springer Nature, 2022.

[3] N. Sabry, A. M. Sauber, and P. El-Kafrawy, "Surveying question answering systems: A comparative study," Egypt. J. Lang. Eng., vol. 9, no. 1, pp. 22–31, Apr. 2022, doi: 10.21608/EJLE.2022.118590.1029.

[4] A. Zechariah, B. Krishna, H. M. Thomas, J. Mammen, D. M. Sharma, P. Krishnamurthy, et al., "Patient-centric question answering: Overview of the shared task at the second workshop on NLP and AI for multilingual and healthcare communication," in Proc. 2nd Workshop NLP AI Multilingual Healthcare Commun. (NLP-AI4Health), 2025, pp. 55–68.

[5] L. E. Chen, S. Y. Cheng, and J. S. Heh, "Chatbot: A question answering system for student," in Proc. Int. Conf. Adv. Learn. Technol. (ICALT), Jul. 2021, pp. 345–346.

[6] W. Huang, J. Jiang, Q. Qu, and M. Yang, "AILA: A question answering system in the legal domain," in Proc. 29th Int. Joint Conf. Artif. Intell. (IJCAI), Jan. 2020, pp. 5258–5260.

[7] M. Basem, I. Oshallah, A. Hamdi, and A. Mohamed, "Few-shot prompting for extractive Quranic QA with instruction-tuned LLMs," in Proc. Intell. Methods, Syst., Appl. (IMSA), Jul. 2025, pp. 24–29.

[8] A. Alrayzah, F. Alsolami, and M. Saleh, "Challenges and opportunities for Arabic question-answering systems: Current techniques and future directions," PeerJ Comput. Sci., vol. 9, Art. no. e1633, 2023.

[9] M. M. Biltawi, S. Tedmori, and A. Awajan, "Arabic question answering systems: Gap analysis," IEEE Access, vol. 9, pp. 63876–63904, 2021.

[10] M. Abu-Daoud, L. Kharouf, O. E. Hajj, D. E. Samad, M. Al-Omari, J. Mallat, et al., "MedAraBench: Large-scale Arabic medical question answering dataset and benchmark," arXiv:2602.01714, 2026.

[11] F. Alnizar, "Fatwa and the making and renewal of Islamic law. By Omer Awass. New York: Cambridge University Press, 2023. Pp. 284. ISBN: 9781009260923," J. Law Relig., vol. 40, no. 1, pp. 108–110, 2025.

[12] M. A. Al-Qahtani, B. F. Alkhamees, and M. Ykhlef, "MAFQA: A dataset for benchmarking multi-hop Arabic fatwa question answering," Data, vol. 11, no. 3, Art. no. 64, 2026.

[13] N. X. Phuc and T. D. Van, "PuxAI at QIAS 2025: Multi-agent retrieval-augmented generation for Islamic inheritance and knowledge reasoning," in Proc. Third Arabic Nat. Lang. Process. Conf.: Shared Tasks, Nov. 2025, pp. 905–913.

[14] Z. Jiang, J. Araki, H. Ding, and G. Neubig, "Understanding and improving zero-shot multi-hop reasoning in generative question answering," in Proc. 29th Int. Conf. Comput. Linguistics (COLING), Oct. 2022, pp. 1765–1775.

[15] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, "Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions," in Proc. 61st Annu. Meeting Assoc. Comput. Linguistics (ACL), Jul. 2023, pp. 10014–10037.

[16] L. Wang, H. Chen, N. Yang, X. Huang, Z. Dou, and F. Wei, "Chain-of-retrieval augmented generation," Adv. Neural Inf. Process. Syst., vol. 38, pp. 59888–59915, 2026.

[17] X. Xiao, H. Y. Huang, R. Liu, and J. Xie, "MASS-RAG: Multi-agent synthesis retrieval-augmented generation," in Findings Assoc. Comput. Linguistics: ACL 2026, Jul. 2026, pp. 9865–9883.

[18] M. Y. Mohammed, S. A. Ali, S. K. Ali, A. A. Majeed, and E. H. Mohamed, "Aftina: Enhancing stability and preventing hallucination in AI-based Islamic fatwa generation using LLMs and RAG," Neural Comput. Appl., vol. 37, no. 25, pp. 20957–20982, 2025.

[19] A. I. Kamal, M. Abd Azim, and M. Mahmoud, "Enhancing Arabic question answering system," in Proc. Int. Conf. Comput. Intell. Commun. Netw. (CICN), Nov. 2014, pp. 641–645.

[20] H. Maraoui, K. Haddar, and L. Romary, "Arabic factoid question-answering system for Islamic sciences using normalized corpora," Procedia Comput. Sci., vol. 192, pp. 69–79, 2021.

[21] H. Abdelnasser, M. Ragab, R. Mohamed, A. Mohamed, B. Farouk, N. M. El-Makky, et al., "Al-Bayan: An Arabic question answering system for the Holy Quran," in Proc. EMNLP Workshop Arabic Nat. Lang. Process. (ANLP), Oct. 2014, pp. 57–64.

[22] M. Sheker, S. Saad, R. Abood, and M. Shakir, "Domain-specific ontology-based approach for Arabic question answering," J. Theor. Appl. Inf. Technol., vol. 83, no. 1, p. 43, 2016.

[23] R. Malhas, W. Mansour, and T. Elsayed, "Qur'an QA 2022: Overview of the first shared task on question answering over the Holy Qur'an," in Proc. 5th Workshop Open-Source Arabic Corpora Process. Tools Shared Tasks Qur'an QA Fine-Grained Hate Speech Detect., Jun. 2022, pp. 79–87.

[24] D. Premasiri, T. Ranasinghe, W. Zaghouani, and R. Mitkov, "DTW at Qur'an QA 2022: Utilizing transfer learning with transformers for question answering in a low-resource domain," in Proc. 5th Workshop Open-Source Arabic Corpora Process. Tools Shared Tasks Qur'an QA Fine-Grained Hate Speech Detect., Jun. 2022, pp. 88–95.

[25] M. ElKomy and A. M. Sarhan, "TCE at Qur'an QA 2022: Arabic language question answering over the Holy Qur'an using a post-processed ensemble of BERT-based models," in Proc. 5th Workshop Open-Source Arabic Corpora Process. Tools Shared Tasks Qur'an QA Fine-Grained Hate Speech Detect., Jun. 2022, pp. 154–161.

[26] K. Alnajjar and M. Hämäläinen, "Harnessing multilingual resources for question answering in Arabic," arXiv:2205.08024, 2022.

[27] E. Aftab and M. K. Malik, "eRock at Qur'an QA 2022: Contemporary deep neural networks for Qur'an-based reading comprehension question answering," in Proc. 5th Workshop Open-Source Arabic Corpora Process. Tools Shared Tasks Qur'an QA Fine-Grained Hate Speech Detect., Jun. 2022, pp. 96–103.

[28] M. M. A. Alqahtani, "Quranic Arabic semantic search model based on ontology of concepts," Ph.D. dissertation, School Comput., Univ. Leeds, Leeds, U.K., 2019.

[29] A. Alsaleh, S. Althabiti, I. Alshammari, S. Alnefaie, S. Alowaidi, A. Alsaqer, et al., "LK2022 at Qur'an QA 2022: Simple transformers model for finding answers to questions from the Qur'an," in Proc. 5th Workshop Open-Source Arabic Corpora Process. Tools Shared Tasks Qur'an QA Fine-Grained Hate Speech Detect., Jun. 2022, pp. 120–125.

[30] M. Basem, I. Oshallah, A. Hamdi, and A. Mohamed, "Few-shot prompting for extractive Quranic QA with instruction-tuned LLMs," in Proc. Int. Conf. Intell. Methods, Syst., Appl. (IMSA), Jul. 2025, pp. 24–29.

[31] Z. Saadaoui, G. Tlig, and F. Jarray, "LLMs-based approach for Quranic question answering," in Proc. 20th Int. Conf. Web Inf. Syst. Technol. (WEBIST), Nov. 2024, pp. 112–118.

[32] S. Eltanbouly, S. Albatarni, S. Hassanein, and T. Elsayed, "Can LLMs directly retrieve passages for answering questions from the Qur'an?," in Proc. Third Arabic Nat. Lang. Process. Conf., Nov. 2025, pp. 203–210.

[33] M. A. Ahmad, M. Ballout, R. A. Ahmad, and E. Bruni, "Transformer Tafsir at QIAS 2025 shared task: Hybrid retrieval-augmented generation for Islamic knowledge question answering," in Proc. Third Arabic Nat. Lang. Process. Conf.: Shared Tasks, Nov. 2025, pp. 899–904.

[34] S. Alnefaie, E. Atwell, and M. A. Alsalka, "Using the retrieval-augmented generation technique to improve the performance of GPT-4 in answering Quran questions," in Proc. 6th Int. Conf. Nat. Lang. Process. (ICNLP), Mar. 2024, pp. 377–381.

[35] S. Alnefaie, E. Atwell, and M. A. Alsalka, "HAQA and QUQA: Constructing two Arabic question-answering corpora for the Quran and Hadith," in Proc. 14th Int. Conf. Recent Adv. Nat. Lang. Process. (RANLP), Sep. 2023, pp. 90–97.

[36] G. Bhatia, H. Mubarak, M. Jarrar, G. Mikros, F. A. Zaraket, M. Alhirthani, et al., "From RAG to agentic RAG for faithful Islamic question answering," in Findings Assoc. Comput. Linguistics: ACL 2026, Jul. 2026, pp. 26469–26488.

[37] H. Abdelazim, T. Begemy, A. Galal, H. Sedki, and A. Mohamed, "Multi-hop Arabic LLM reasoning in complex QA," Procedia Comput. Sci., vol. 244, pp. 66–75, 2024.

[38] A. H. Sidhoum, M. H. Mataoui, F. Sebbak, and K. Smaïli, "ACQAD: A dataset for Arabic complex question answering," in Proc. Int. Conf. Cyber Secur., Artif. Intell. Theor. Comput. Sci., Dec. 2022.

[39] S. Wang, S. Zhuang, and G. Zuccon, "BERT-based dense retrievers require interpolation with BM25 for effective passage retrieval," in Proc. ACM SIGIR Int. Conf. Theory Inf. Retrieval (ICTIR), Jul. 2021, pp. 317–324.

[40] K. C. Chakravarthy, "Semantic–lexical fusion: Improving retrieval accuracy for AI-driven knowledge systems," Int. J. Comput. Exp. Sci. Eng., vol. 11, no. 4, pp. 8555–8564, 2025.

[41] P. J. Sager, A. Kamaraj, B. F. Grewe, and T. Stadelmann, "Deep retrieval at CheckThat! 2025: Identifying scientific papers from implicit social media mentions via hybrid retrieval and re-ranking," arXiv:2505.23250, 2025.

[42] Hugging Face, "AraBART," [Online]. Available: https://huggingface.co/moussaKam/AraBART. [Accessed: Aug. 19, 2025].

[43] Hugging Face, "AraT5-base," [Online]. Available: https://huggingface.co/UBC-NLP/AraT5-base. [Accessed: Aug. 19, 2025].

[44] Hugging Face, "Arabic-T5-small," [Online]. Available: https://huggingface.co/flax-community/arabic-t5-small. [Accessed: Aug. 19, 2025].

[45] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, et al., "mT5: A massively multilingual pre-trained text-to-text transformer," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol. (NAACL-HLT), Jun. 2021, pp. 483–498.

[46] Hugging Face, "mT5-base," [Online]. Available: https://huggingface.co/google/mt5-base. [Accessed: Jul. 4, 2026].

[47] OpenAI, "GPT-4.1 model," OpenAI API Documentation, [Online]. Available: https://developers.openai.com/api/docs/models/gpt-4.1. [Accessed: Jul. 4, 2026].

[48] OpenAI, "GPT-3.5 Turbo," OpenAI API Documentation, [Online]. Available: https://developers.openai.com/api/docs/models/gpt-3.5-turbo. [Accessed: Jul. 4, 2026].

[49] OpenAI, "Introducing GPT-5," Aug. 7, 2025. [Online]. Available: https://openai.com/index/introducing-gpt-5/. [Accessed: Jul. 4, 2026].

[50] Google DeepMind, "Gemini 2.5: Our most intelligent AI model," Mar. 25, 2025. [Online]. Available: https://blog.google/innovation-and-ai/models-and-research/google-deepmind/gemini-model-thinking-updates-march-2025/. [Accessed: Jul. 4, 2026].

[51] X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, et al., "DeepSeek LLM: Scaling open-source language models with longtermism," arXiv:2401.02954, 2024.

[52] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, et al., "Qwen technical report," arXiv:2309.16609, 2023.

[53] Mistral AI, "Mistral-7B-Instruct-v0.2," Hugging Face Model Card, 2023. [Online]. Available: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2. [Accessed: Jul. 4, 2026].

[54] Silma AI, "SILMA-9B-Instruct-v1.0," Hugging Face Model Card, 2024. [Online]. Available: https://huggingface.co/silma-ai/SILMA-9B-Instruct-v1.0. [Accessed: Jul. 4, 2026].

[55] J. Liang, Z. Cai, J. Zhu, H. Huang, K. Zong, B. An, et al., "Alignment at pre-training! Towards native alignment for Arabic LLMs," Adv. Neural Inf. Process. Syst., vol. 37, pp. 13872–13896, 2024.

[56] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," Adv. Neural Inf. Process. Syst., vol. 33, pp. 9459–9474, 2020.