SEA-Team at QIAS 2025: Enhancing LLMs for Question Answering in Islamic Texts

Sanaa Alowaidi
University of Leeds
King Abdulaziz University
saalowaidi@kau.edu.sa

Eric Atwell
University of Leeds
e.s.atwell@leeds.ac.uk

Mohammed Ammar Alsalka
University of Leeds
em.a.alsalka@leeds.ac.uk

Abstract

This paper presents our participation in the QIAS 2025 shared tasks, namely Islamic Inheritance Reasoning and Islamic Knowledge Assessment sub-tasks. We propose an Islamic Retrieval-Augmented Generation (RAG) system that integrates multiple knowledge sources and semantic retrieval methods. Our evaluation compares multilingual general-purpose models and Arabic-centric models, using the accuracy metric. Results show that multilingual models consistently outperform Arabic-language models. The Mistral-large achieved the highest accuracy in Task 1 (72%) using basic RAG with our augmented knowledge resource, while GPT-4o with RAG and K2R retrieval achieved the best score in Task 2 (87.71%). These findings highlight the effectiveness of RAG in enhancing LLM performance for complex Islamic reasoning and knowledge assessment tasks.

1 Introduction

Large language models (LLMs) demonstrate strong capabilities in understanding, interpreting, and generating text that is close to human language. Several powerful multilingual general-purpose models have emerged, such as GPT-4o and Mistral-large. There are several Arabic-centric models developed recently, including Falcon (Almazrouei et al., 2023), ALLaM (Bari et al., 2024), Mistral SABA, and Fanar (Abbas et al., 2025), which are trained on specialized Arabic and Islamic knowledge resources. Arabic and religious texts present significant challenges for LLMs due to their linguistic complexity and the sensitive nature of Islamic teachings. Recently, Retrieval-Augmented Generation (RAG) has emerged as one of the most effective NLP techniques for question-answering. It enhances LLMs ability by retrieving relevant information from external knowledge sources and then using it to generate more accurate responses (Lewis et al., 2020; Oche et al., 2025). RAG is significant for domain-specific applications where accuracy and reliability are critical (Han et al., 2024).

Prior studies have applied RAG to various Islamic domains, including Quranic teachings, Turkish Islamic knowledge, and historical Islamic medical texts (Alnefaie et al., 2024; Alan et al., 2025; Sayeed et al., 2025). Moreover, promising research has focused recently on enhancing the retrieval stage of the RAG pipeline through query expansion and reformulation strategies in English (Yang et al., 2025; Wang et al., 2024).

To the best of our knowledge, RAG techniques have not yet been evaluated for Islamic inheritance reasoning or Islamic knowledge assessment. We address this gap by contributing to Task 1: Islamic Inheritance Reasoning and Task 2: Islamic Knowledge Assessment, as introduced in the QIAS 2025 shared task (Bouchekif et al., 2025a). Task 1 evaluates an LLMs ability to answer questions requiring precise reasoning and calculations based on Islamic jurisprudence. Task 2 assesses the accuracy of LLMs in answering general Islamic questions across multiple disciplines. Both tasks are challenging not only because they require advanced reasoning, but also because they include questions of varying difficulty levels that reflect the depth and complexity of Islamic knowledge. In this paper, we investigate strategies to enhance LLMs for Islamic QA, addressing the following research questions: how does the combination of few-shot prompting and RAG techniques affect the LLMs? How does the type of LLM, general multilingual or Arabic-centric, affect the accuracy of RAG? Does the size of the knowledge resource affect the performance of the RAG system? What is the effect of applying semantic retrievals through query expansions and reformations on LLMs?

Proceedings of The Third Arabic Natural Language Processing Conference, pages 940–946

The paper is organized as follows: In Section 2, related works are reviewed. Section 3 details the datasets used. Section 4 describes the proposed system structure, while Section 5 describes the implementation setup. In Section 6, the results are presented and discussed. Finally, the paper concludes with a summary and suggestions for future work.

2 Related Works

Several studies have examined the application of LLMs to Islamic knowledge. Alnefaie et al. (2023) used GPT for question answering on a Quran dataset. Bouchekif et al. (2025b) evaluated several multilingual LLMs and Arabic LLMs with zero-shot prompting on an inheritance dataset. These works highlighted key limitations of LLMs, including hallucination and misinterpretation. More recent research has explored using RAG techniques to improve LLM performance. Alnefaie et al. (2024) applied RAG to the GPT-4 model in the Quranic Question Answer dataset. Alan et al. (2025) introduced the MufassirQA system, which enhances the ChatGPT-3.5 Turbo model with RAG by using religious knowledge resources in the Turkish language. Furthermore, Sayeed et al. (2025) investigated the use of RAG with LLaMA-3, Mistral-7B, and Qwen-2 to answer medical questions based on an old Islamic medical text. These studies found that RAG consistently outperforms baseline LLMs and emphasized that its performance is highly dependent on the quality of the retrieval and knowledge resources. However, the performance of RAG in Islamic domain-specific knowledge remains largely underexplored. Furthermore, promising research has recently focused on improving the retrieval stage of the RAG pipeline through query expansion and reformulation strategies in English (Yang et al., 2025; Wang et al., 2024; Li et al., 2024). In this study, we extensively explore the RAG in both Arabic and multilingual LLMs. In addition, study the effect of different retrieval strategies that incorporate query expansion and reformulation methods.

3 Datasets

In this paper, we use the four officially published datasets [1] corresponding to the two subtasks of the QIAS 2025 shared task.

3.1 Task 1: Islamic Inheritance Reasoning (Ilm al-Mawrth)

The Islamic Inheritance dataset comprises 22,000 multiple-choice questions (MCQs). Each question includes six answer choices with only one correct label (Bouchekif et al., 2025b). The questions are classified into two levels of difficulty: beginner and advanced. The dataset was divided into 20,000 for the training set, 1,000 for the validation set, and 1,000 for the test set. In addition, the fatwa dataset is used as a supplementary knowledge resource. It consists of 3165 fatwas from Islamic websites covering general legal, ethical, and social topics.

3.2 Task 2: General Islamic Knowledge

The first dataset consists of 1700 question pairs in MCQ format covering Hadith criticism, Quranic sciences, legal theory, and prophetic biography. Each question has four answer choices, with one correct answer. The data distribution is 700 question pairs for the validation set and 1,000 for the test set. The questions are categorized into three complexity levels: beginner, intermediate, and advanced. Moreover, a supplementary Islamic corpus was used as an external knowledge resource for the RAG system. It comprises unsupervised data of relevant Islamic texts. The corpus includes approximately 50 Islamic books in MS Word format, all of which are directly related to the evaluation dataset topics.

4 System Overview

The proposed system adopts the RAG architecture (Lewis et al., 2020; Wang et al., 2024; Oche et al., 2025) and consists of three main phases [2]: Knowledge Resource Preparation, Retrieval, and Answer Generation, as illustrated in Appendix A. The Knowledge Resource Preparation phase is conducted offline, where documents are preprocessed and converted into vector representations. This phase includes four modules: loading, chunking, embedding, and indexing. First, the input documents are loaded and preprocessed to produce normalized, cleaned text. Next, the chunking module divides the documents into smaller units. This step is essential for improving retrieval effectiveness, enabling embedding storage, and addressing the

[1] https://gitlab.com/islamgpt1/qias_shared_task_2025

[2] in The code is available in our repository:https://github.com/S-Alowaidi/SEA-RAG_Enhancing-LLMs

context-length limitations of LLMs. In our experiments, we used token-aware recursive chunking to segment documents into semantically coherent units, ensuring token compatibility with the embedding model's tokenization. Following the recommendations in (Wang et al., 2024), we set the splitter to 500 tokens per chunk with a 50-token overlap. The embedding module then transforms each chunk into a high-dimensional dense vector using OpenAI Embeddings, enabling efficient semantic similarity searches. Finally, the indexing module stores these embeddings in a FAISS (Facebook AI Similarity Search) vector database. The Retrieval and Answer Generation phases are executed in real time. At query time, the question is embedded, and the retrieval module searches for the most relevant chunks in the vector index. We propose three semantic retrieval methods: basic similarity search and two semantically enhanced strategies. For the enhanced methods, keywords are extracted offline using GPT-4o. Candidate keywords are filtered to remove noise by eliminating stop words, short terms, and duplicates, as well as semantically irrelevant items using cosine similarity (threshold = 0.3) against the original question. Basic Similarity Search: retrieves chunks in a single pass using the original query. Keyword-Augmented Two-Stage Retrieval (K2R): performs parallel retrieval using the original query and semantically filtered keywords, then merges and deduplicates the retrieved chunks. Multi-Query Reformulation with Keywords (MQR-K): reformulates each keyword with the query into a complete sub-question in Arabic, retrieves semantically similar chunks in parallel, and merges the results for diversification. In the Answer Generation phase, the retrieved context is combined with the question and its answer choices in a structured prompt, which is then sent to the LLMs. The output undergoes a post-generation validation step to ensure compliance with the single-letter MCQ answer format.

5 Experimental Setup

All experiments were run on Google Colab and used the LangChain framework. Embeddings are generated using text-embedding-3-large (OpenAI) and stored in a FAISS flat index with cosine similarity. The retrieval module is configured to return the top-$k = 4$ chunks per query. To address the third research question, in Task 1, we evaluate two datasets. The first is the fatwa dataset as a baseline knowledge resource for RAG. In addition, we proposed an expanded dataset that combines the fatwa dataset with MCQ training data by including question and correct-answer pairs as additional contextual knowledge. In the generation phase, we evaluate several LLMs in three-shot prompting, including GPT-4o [3], Mistral-large-latest [4], Fanar Islamic-RAG [5], Allam-7B [6], and Mistral-SABA-24B [7].

Model	3-Shot	Fatwa	Expand-K	K2R
Fanar	53.3	57.8	62.8	59.5
Mistral-S	43.8	50.1	55.6	53.2
Mistral-L	61.5	66.0	72.0	70.0
ALLaM	47.8	50.6	53.2	43.5
GPT-4o	57.5	58.9	61.6	63.4

Table 1: Task 1 results comparing Few-Shot Prompting, Basic RAG (Fatwa only), Expanded Knowledge (Fatwa + Train), and RAG with K2R.

Model	Few-Shot	Basic	K2R	MQR
Fanar	29.43	57.86	54.14	55.71
Mistral-S	66.07	75.86	76.29	72.43
Mistral-L	78.29	83.86	77.57	76.00
ALLaM	71.29	77.29	79.43	77.43
GPT-4o	83.71	83.86	87.71	85.57

Table 2: Task 2 results comparing Few-Shot prompting, Basic RAG, K2R retrieval RAG, and MQR-K retrieval RAG.

6 Results and Discussion

Task 1: Table 1 shows the results on the development set. The Mistral-large model achieved the highest accuracy 72.0% when using RAG with the expanded fatwa dataset. Comparing the fatwa-only dataset to the augmented version, the Arabic-centric models Mistral-SABA-24B and Fanar Islamic-RAG benefited the most, with gains of 6.3 and 4.5 points, respectively. ALLaM-7B showed the least improvement. On the other hand, Mistral-large had the second-highest gain 6.0 points, while GPT-4o's improvement was relatively small at 2.7 points. This highlights how the reasoning ability of multilingual models can be greatly enhanced by adding domain-specific knowledge. **Test set:** For the test set, we selected

[3] via OpenAI API:https://platform.openai.com/
[4] via Mistral API:https://mistral.ai/
[5] via Fanar API:https://fanar.qa/
[6] via: https://huggingface.co/transformers
[7] via Groq API

Model	T1Ed	T1K2R	T2R	T2K2R
Mistral-L	61.1	63.0	87.7	89.1
GPT-4o	59.9	57.1	87.8	89.0

Table 3: Results on test set for Task1: T1Ed refers to Expanded Knowledge,T1K2R refers to K2R retrieval RAG. Task2:T2R refers to Basic RAG, T2K2R refers to K2R retrieval RAG

the best-performing approaches based on the results from the development set. As presented in Table 3, Mistral-large achieved the highest accuracy using the K2R method, reaching 63%. Unlike the development set, its performance improved in the test set. In contrast, GPT's performance with the K2R method declined slightly in the test data.

Task2:

Table 2 shows the accuracy results for different retrieval methods in answering general Islamic questions. GPT-4o achieved the highest accuracy of 87.71% using the K2R retrieval method, outperforming its baseline RAG. This is expected since GPT-4o has been trained on a large amount of data, including Islamic knowledge. Mistral-large achieved the second-highest accuracy in the baseline RAG 83.86%. However, its performance dropped slightly with the K2R and MQR-K retrieval methods (77.57% and 76.00%, respectively. The performance of Arabic-centric models varied across retrieval methods. ALLaM-7B and Mistral-SABA performed best with K2R, while Fanar achieved its best results, 55.71%, with MQR-K.

Test set: For the test set, we chose two strategies based on their performance during the development phase. The table 3 indicates that Mistral-large and GPT-4o achieved very similar results, both reaching approximately 89% with the K2R method. Therefore, the expanded query could be a promising approach to enhancing RAG.

Overall Analysis Based on the results, it is clear that RAG performance is heavily dependent on the quality of the retrieved contexts. Enhancing retrieval with the K2K approach outperformed basic RAG retrieval for all models. However, the performance continued to fluctuate compared to the knowledge-enrichment approach and depended on the nature of the task and the model used. For example, in task 1, the nature of inheritance law texts often shares similar keywords (e.g., wife, paternal mother, heirs). Hence, refining the query by broadening keywords could produce a wider context that distracts the LLMs. In the case of multiple-query reformulation (MQR), we observed a general drop in performance for most models, except Fanar, which showed a slight improvement. This may be due to the static reformulation method used, which can cause loss of semantic meaning and introduce noise. The results show that, in general, Arabic-centric models benefit from higher recall when broadening the context by expanding queries with keywords. In contrast, stronger models perform better with fewer but more relevant contexts.

7 Conclusion

This work presents our contributions to the QIAS 2025 shared task, focusing on Task 1: Islamic Inheritance Reasoning and Task 2: Islamic Knowledge Assessment. We propose an Islamic RAG system that leverages multiple knowledge sources and retrieval methods, utilizing more than five different LLMs. Our experimental results show that multilingual general-purpose models outperform Arabic-language models in both tasks. For Task 1, Mistral-large achieved the best performance (72%), while for Task 2, GPT-4o delivered the strongest results in general Islamic knowledge reaching (87.71%). Among the Arabic models, Fanar performed best in Task 1 by 62.8%, and ALLaM-7B led in Task 2 by 79.43%. We also observed that expanding the knowledge sources in Task 1 improved the performance of all models, with the most notable gains for Arabic models such as Fanar and ALLaM-7B.

Regarding the use of RAG with a semantic retrieval strategy, results indicate that semantic retrieval RAG generally outperformed three-shot prompting across all models and both tasks. However, its advantage over basic RAG varied according to the nature of the task data and the model.

Future research should explore alternative query expansions and reformulation approaches, such as using LLMs to generate more semantically relevant queries dynamically. In addition, investigating other RAG enhancement techniques, including re-ranking and document summarization, may yield further improvements. Finally, we emphasize the importance of developing high-quality Islamic knowledge sources to improve model relearning effectively.

Acknowledgments

The authors would like to thank the competition organizers. We also thank KAU for its support.

References

Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur A. Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed K. Elmagarmid, Mohamed Y. Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, Majd Hawasly, and 22 others. 2025. Fanar: An arabic-centric multimodal generative ai platform. *arXiv preprint arXiv:2501.13944*.

Ahmet Yusuf Alan, Enis Karaarslan, and Ömer Aydın. 2025. Improving llm reliability with rag in religious question-answering: Mufassirqas. *Turkish Journal of Engineering*, 9(3):544–559.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *Preprint*, arXiv:2311.16867.

Sarah Alnefaie, Eric Atwell, and Mohammad Ammar Alsalka. 2023. Is GPT-4 a good islamic expert for answering Quran questions? In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 124–133, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Sarah Alnefaie, Eric Atwell, and Mohammed Ammar Alsalka. 2024. Using the retrieval-augmented generation technique to improve the performance of gpt-4 in answering quran questions. In *2024 6th International Conference on Natural Language Processing (ICNLP)*, pages 377–381.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan AlRashed, Faisal A. Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, and 6 others. 2024. Allam: Large language models for arabic and english. *Preprint*, arXiv:2407.15390.

Abdessalam Bouchekif, Samer Rashwani, Mohammed Ghaly, Mutaz Al-Khatib, Emad Mohamed, Wajdi Zaghouani, Heba Sbahi, Shahd Gaben, and Aiman Erbad. 2025a. Qias 2025: Overview of the shared task on islamic inheritance reasoning and knowledge assessment. In *Proceedings of The Third Arabic Natural Language Processing Conference, ArabicNLP 2025, Suzhou, China, November 5–9, 2025*. Association for Computational Linguistics.

Abdessalam Bouchekif, Samer Rashwani, Heba Sbahi, Shahd Gaben, Mutaz Al-Khatib, and Mohammed Ghaly. 2025b. Assessing large language models on islamic legal reasoning: Evidence from inheritance law evaluation. In *Proceedings of The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Binglan Han, Teo Susnjak, and Anuradha Mathrani. 2024. Automating systematic literature reviews with retrieval-augmented generation: a comprehensive overview. *Applied Sciences*, 14(19):9103.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Zhicong Li, Jiahao Wang, Zhishu Jiang, Hangyu Mao, Zhongxia Chen, Jiazhen Du, Yuanxing Zhang, Fuzheng Zhang, Di Zhang, and Yong Liu. 2024. Dmqr-rag: Diverse multi-query rewriting for rag. *arXiv preprint arXiv:2411.13154*.

Agada Joseph Oche, Ademola Glory Folashade, Tirthankar Ghosal, and Arpan Biswas. 2025. A systematic review of key retrieval-augmented generation (rag) systems: Progress, gaps, and future directions. *arXiv preprint arXiv:2507.18910*.

Mohammad Amaan Sayeed, Mohammed Talha Alam, Raza Imam, Shahab Saquib Sohail, and Amir Hussain. 2025. From rag to agentic: Validating islamic-medicine responses with llm agents. *arXiv preprint arXiv:2506.15911*.

Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, Ruicheng Yin, Changze Lv, Xiaoqing Zheng, and Xuanjing Huang. 2024. Searching for best practices in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17716–17736, Miami, Florida, USA. Association for Computational Linguistics.

Qimin Yang, Huan Zuo, Runqi Su, Hanyinghong Su, Tangyi Zeng, Huimei Zhou, Rongsheng Wang, Jiexin Chen, Yijun Lin, Zhiyi Chen, and Tao Tan. 2025. Dual retrieving and ranking medical large language model with retrieval augmented generation. *Scientific Reports*, 15(1):18062.

A System Architecture

A comprehensive description of the proposed RAG system is illustrated in Figure 1.

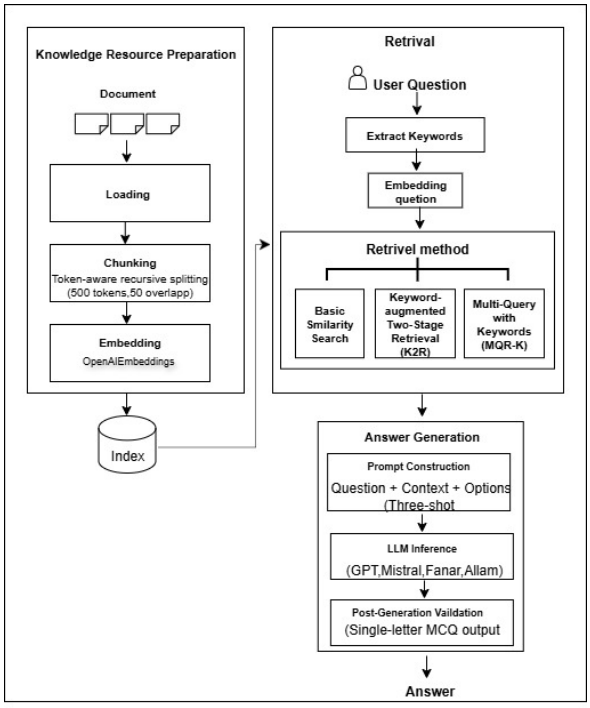

Figure 1: The proposed RAG system architecture

```
Input: question q, options O = {A: o1, B: o2, C: o3, D: o4, (E: o5), (F: o6)},
 vectorstore V, LLM model M
Output: prediction answer y ∈ {A:F}
Step 1: Keyword extraction and filtering:
  K_raw ← extract_keywords(q)
  K ← filter_keywords(K_raw, q)
Step 2: Query construction:
  Q ← [q] + K  // multi-query list
Step 3: Parallel Retrieval:
  D_all ← U retrieve_topk(V, qi) for qi ∈ Q
  D ← deduplicate(D_all)
Step 4: Context building:
  C ← concatenate(contents(d) for d ∈ D)
Step 5: Prompt construction:
  P ← compose(few_shots, task_instructions, C, q, O)
Step 6: Generation and validation:
  y_raw ← LLM(M, P)
  y ← normalize_to_single_choice(y_raw, allowed=keys(O))
return y
```

Figure 2: RAG based on the Keyword-Augmented Two-Stage Retrieval (K2R) approach

B Keyword-Augmented Two-Stage Retrieval (K2R)

The K2R approach retrieves documents using a multi-query parallel FAISS search. Figure 2 describes the general steps for RAG based on the K2R method.

C Multi-Query Reformulation with Keywords (MQR-K)

This approach is based on reformulating the question using a fixed Arabic template to generate one query per keyword. Figure 3 explains the general steps for RAG based on the MQR-K method.

Model	Beg.	Int.	Adv.
Mistral-L | T2R	90.57	85.33	76.67
GPT-4o | T2R	90.29	90.00	74.00
Mistral-L | T2K2R	92.29	85.33	78.00
GPT-4o | T2K2R	92.57	87.33	74.00
Mistral-L | T1Ed	75.20	–	47.00
GPT-4o | T1Ed	72.60	–	47.20
Mistral-L | T1K2R	78.80	–	47.20
GPT-4o | T1K2R	77.40	–	36.80

Table 4: Accuracy (%) across difficulty levels Beginner (Beg.), Intermediate (Int.), Advanced (Adv.). A dash (–) indicates an unavailable level. For Task 1: T1Ed refers to Expanded Knowledge, T1K2R refers to K2R retrieval RAG. Task 2:T2R refers to Basic RAG, T2K2R refers to K2R retrieval RAG

D Prompt

Figure 4 demonstrates the prompt used in the experiments. The few-shot examples refer to three examples specifically for the target task taken from the training data. The context refers to the documents retrieved using one of the retrieval methods, baseline semantic similarity, K2R, or MQR.

E In-depth Analysis

Table 4 presents the performance of various models on Task 1 and Task 2 across three difficulty levels: beginner, intermediate, and advanced. The results indicate that all models generally achieved high accuracy on beginner-level questions for both tasks.

In Task 1, the Mistral-large model answered approximately 75.20% of beginner questions, while the GPT-4o model answered about 72.60% when applying the RAG with the expanding knowledge approach. However, for advanced questions, the accuracy of most methods in answering these questions reaches only about 47%, indicating weaker performance on questions that require complex inheritance reasoning compared with simpler ones.

```
Input: question q,  options O = {A: o1, B: o2, C: o3, D: o4, (E: o5), (F: o6)},
vectorstore V,  LLM model M
Output: prediction answer y ∈ {A:F}
Step 1: Keyword extraction and filtering:
  K_raw ← extract_keywords(q)
  K ← filter_keywords(K_raw, q)
Step 2: Query construction:
  for each kw ∈ K:
    q_i ← "Given the question: {q} what information
        is available about:{kw}?" // substitute q and kw into q_i
    add q_i to Q
Step 3: Parallel Retrieval:
  D_all ← U retrieve_topk(V, qi) for qi ∈ Q
  D ← deduplicate(D_all)
Step 4: Context building:
  C ← concatenate(contents(d) for d ∈ D)
Step 5: Prompt construction:
  P ← compose(few_shots, task_instructions, C, q, O)
Step 6: Generation and validation:
  y_raw ← LLM(M, P)
  y ← normalize_to_single_choice(y_raw, allowed=keys(O))
return y
```

Figure 3: RAG based on the Multi-Query Reformulation with Keywords (MQR-K) approach

Additionally, the K2R retrieval RAG approach made substantial improvements on beginner-level questions. In contrast, for advanced questions, while the Mistral-large model maintained its accuracy in the K2R approach, the performance of the GPT-4o model decreased when queries were expanded with keywords.

For task 2, which focused on general Islamic knowledge, most approaches demonstrated exceptional performance, achieving accuracy rates of 92.57%, 90%, and 78% at the beginner, intermediate, and advanced levels, respectively. It is clear from the results that the K2R retrieval method achieved a notable improvement at the beginner and advanced levels across models. Moreover, the results show that while Mistral-large and GPT-4o both performed similarly overall, the Mistral-large model often slightly outperformed the GPT-4o model on advanced questions.

```
prompt = f"""
{few_shot_examples}
Context:
{context}
You are a specialist in Islamic sciences.
Your task is to answer multiple-choice
questions by selecting the correct option.
Question:
{question}
{options_text}
Please respond using only one English letter from:
{valid_letters}
Do not write any explanation or additional text.
""".strip()
```

Figure 4: The prompt used for the RAG system