# Burhan at IslamicEval: Fact-Augmented and LLM-Driven Retrieval for Islamic QA

**Mohammad Basheer Kotit[1] , Watheq Mansour[2], Abdulhamid Touma [3], Ahmad Qadeib Alban[4]**
[1] Qatar University, Qatar
[2] The University of Queensland, Australia
[3] Syrian Virtual University, Syria
[4] Université Paris, Dauphine-PSL, France
mk1806194@qu.edu.qa, w.mansour@uq.edu.au, abdulhamid_103004@svuonline.org, Ahmad.qadeib-alban@dauphine.psl.eu

## Abstract

This paper presents our approach to the Qur'an and Hadith QA task in the IslamicEval 2025 Shared Task. Reliable retrieval requires both accuracy and context-aware answers from Qur'anic and Hadith text. To address this challenge, We combine semantic search with LLM-based re-ranking. To enhance alignment, we augment the corpus with LLM-extracted Islamic facts and paraphrased queries. An LLM-based binary classifier further verifies whether retrieved passages answer the questions. Results show improved accuracy and better alignment with user intent.

## 1 Introduction

The Holy Qur'an, a sacred and timeless text revealed over 1400 years ago in Classical Arabic, continues to attract the attention of millions of Muslims and non-Muslims for its profound teachings, legislation, and extensive body of knowledge. Therefore, developing effective systems for the Holy Qur'an, particularly for Passage Retrieval (i.e., the task of identifying and ranking candidate passages that potentially contain answers to a given question), have become a matter of paramount importance (Malhas et al., 2022) and presents unique and significant challenges (Malhas et al., 2022; Zekiye and Amroush, 2023). The challenges stem from linguistic complexities, context scarcity, and the reliability and specificity required in the Holy Qur'an. Recently, the Qur'an QA 2023 shared task dataset has further highlighted the complexity of this task, and the results revealed substantial space for further improvements (Basem et al., 2024). In continuation of this effort, Qur'an and Hadith QA 2025 is offered as subtask in IslamicEval shared task (Mubarak et al., 2025). The main difference of this year task is the addition of Hadith (Sahih Bukhari collection, in particular) to the retrieval collections, making the task more challenging.

In this paper, we present our participation in Qur'an and Hadith QA 2025 subtask and describe our proposed retrieval pipeline. The main characteristics of our system are: **(1) Augment the Qur'an and Hadith** collections with information extracted by large language models (LLMs). **(2) Employ semantic search** to form the initial retrieval list, followed by LLM-based re-ranker to prioritize the most relevant candidates. **(3) Paraphrase user queries** using LLMs to enhance semantic clarity and improve retrieval outcomes. **(4) Employ a LLM-based binary classifier** to detect questions with no answers.

### 1.1 Related Work

The task of question answering (QA) for the Holy Qur'an was introduced as a shared task in (Malhas et al., 2022). The following year, the first task of Qur'anic passage retrieval was offered as a shared task (Malhas et al., 2023). The task of Qur'anic passage retrieval has garnered significant scholarly interest due to the distinct linguistic and contextual challenges posed by the Qur'an (Basem et al., 2024). Effective systems must retrieve relevant verses to answer both factoid and non-factoid questions and bridge the linguistic gap between Modern Standard Arabic (MSA) and Classical Arabic. A further challenge lies in detecting zero-answer scenarios, where questions that have no answers within the Qur'anic passages require robust mechanisms for rejecting all the non-relevant candidates (Malhas et al., 2023).

Several teams participated in the task and employed various technique such as augmentation (Elkomy and Sarhan, 2024; Basem et al., 2024), translation (Alawwad et al., 2023). As the augmentation showed noticeable improvements, we decide to continue exploring in this direction and propose new augmentation techniques.

## 2 Background

In this section, we present the required background information about the shared task.

### 2.1 Task Definition

Our work was merely on IslamicEval Subtask 2: Qur'an and Hadith QA 2025, which is a retrieval task and a continuation of Qur'an QA 2022 and Qur'an QA 2023 shared tasks. The task is defined as follows: Given a free-text question posed in MSA, a collection of Qur'anic passages (that cover the Holy Qur'an), and a collection of Hadith from Sahih Bukhari, a system is required to retrieve a ranked list up to 20 answer-bearing Qur'anic passages or Hadith (i.e., Islamic sources that potentially enclose the answer(s) to the given question) from the two collections. The question can be a factoid or a non-factoid question. To make the task more challenging, the organizers add on purpose some questions that have no answers in the Holy Qur'an, Sahih Al-Bukhari, or both. For such cases, the ideal system should return no answers.

### 2.2 Dataset Details

The dataset proposed for the task comprises two collections: the Qur'anic Passage Collection (QPC) and the Sahih Al-Bukhari Collection (SBC) [1]. The QPC segments the 114 chapters of the Qur'an into 1,266 topical passages, while the Sahih Al-Bukhari Collection includes 2,254 Hadith. To enable training, the organizers provide 250 questions of the AyaTEC dataset along with their relevance judgments over the Qur'anic Passage collection only. The questions are divided into training (84%) and development (16%) datasets.

## 3 System Overview

In this section, we illustrate the proposed retrieval pipeline to address the task at hand.

### 3.1 Data Augmentation and Information Extraction

Our augmentation strategy involves two approaches. In the first approach, following Elkomy and Sarhan (2024), we utilize two Tafsir sources (Al-Tafsir Al-Muyassar and Tafsir Al-Jalalayn) to augment the QPC passages with relevant interpretations. We believe this step is helpful in expanding the context as the text in QPC is generally short.

In the second approach, we extract factual information from QPC and SBC passages and then append them to the original text. The intuition behind this is that many MSA questions differ linguistically from the original Qur'anic or Hadith wording, requiring a deep semantic understanding of the content. We address this need by enriching QPC and SBC through extracting explicit semantic representations using LLMs. By generating explicit facts, we bridge the linguistic gap, thereby improving semantic search recall. To achieve this goal, we develop a domain-adapted prompt (Figures 1–2) to extract key entities and relations —characters, places, events, Islamic concepts, and legal rulings— from each Qur'anic passage or Hadith. In response to our prompts, the LLM rewrites the implicit references into unambiguous descriptions, enabling the retrieval model to better align them with the query's intent. In other words, the proposed method captures both explicit and implicit meanings that standard embedding models may not explicitly state in the surface text. Finally, we pair the extracted information (IE) with its corresponding text. We refer to this pairing by QPC + IE and SBC + IE for the Qur'an and the Hadith, respectively.

### 3.2 Semantic Retrieval with LLM Re-ranking

We generate embeddings for all enriched texts in QPC and SBC using a SOTA embedding model ("text-embedding-3-large" [2]), as explained in appendix A. Query embeddings are also generated using the same model to ensure proper semantic matching.

Following this, we apply semantic search independently for each source (QPC and SBC). In the initial retrieval phase, cosine similarity is computed between the query embedding and the (QPC or SBC) embeddings. For each collection, the top 20 passages are selected, and then the two lists are merged to form the initial retrieval set (referred to Dense).

To enhance the retrieval quality, we introduce a reranking stage by pairing the candidate passages from the initial retrieval set with the user query and pass them to an LLM. The LLM, in turn, reorders the retrieved passages according to their estimated relevance to the query.

[1] https://gitlab.com/bigirqu/quran-hadith-qa-2025

[2] https://platform.openai.com

### 3.3 Query Rewriting

Towards enhancing the retrieval performance, we utilized two methods to change the query text. In the first method, we augment the query with synonym words, and in the second one, we rephrase the query to a new form. We generate a new query file for each method and per a different LLM. Following this, we feed the generated query file to the ranking pipeline described in the previous steps.

**Synonym Expansion**. Since our retrieval pipeline is mainly focused on semantic matching, we believe that adding some synonym words to the query might increase the query-passage matching. Therefor, we employ LLMs to generate one synonym word for a list of query words. The list of query words is formed after removing the Arabic stop words, such as "What", and "Why", etc. Then, each generated word is positioned after its corresponding synonym between two parentheses. Here is an example of an expanded query:

من هو النبي (الرسول) المعروف (المشهور) بالصبر (بالتحمل) ؟

**Query Rephrasing**. As LLMs are powerful writers, we decided to use their potential in paraphrasing the input query. Simply, we prompt an LLM to rewrite the given question in a better way.

### 3.4 LLM-Based No-Answer Detection

Following the reranking stage, we further refine the reranking set of passages to determine whether it contains an answer to the query. In particular, we prompt an LLM to make a binary judgment on a question-passage pair, assessing whether a given passage addresses the question explicitly or implicitly. If none of the passages in the reranked set are judged relevant, the system returns a standardized no-answer response; otherwise, the reranked list is preserved in its order.

| Dataset | MAP@10 | MAP@5 |
|---|---|---|
| QPC | 0.2761 | 0.2553 |
| QPC + jalalayn | 0.2798 | 0.2572 |
| QPC + muyassar | 0.2926 | **0.2708** |
| QPC + IE | **0.2944** | 0.2689 |
| QPC + muyassar + IE | 0.2878 | 0.2662 |

Table 1: Effect of augmenting QPC on semantic search on both the train and dev sets.

## 4 Experimental Setup

In the data augmentation phase, we used OpenAI's GPT-4o to extract factual statements from each passage in both QPC and SBC. To identify the most effective embedding model, we evaluated several Arabic embedding models, as detailed in appendix A. The "text-embedding-3-large" model demonstrated the highest overall performance and was therefore used in all subsequent experiments. Document embeddings were stored in ChromaDB [3], a persistent vector store, with cosine similarity as the distance metric.

**Retrieval** was conducted independently for the QPC and SBC datasets. For each collection, we retrieved the top 20 passages based on cosine similarity between the query and the passage embeddings, resulting in a combined list of 40 candidate passages. These were then reranked using OpenAI GPT-4o. For query paraphrasing, we test three variants of OpenAI GPT-4, namely: GPT-4o, GPT-4.1-mini, and GPT-4.1, selecting the latter for subsequent experiments due to its superior performance. Additionally, GPT-4o was employed as a binary classifier to determine whether each candidate passage was relevant to a given query.

**Evaluation** We report our evaluation result on a set *combined from the training and development* sets as we believe this gives more reliable and robust results compared to dev set only. While, we report MAP@5 and MAP@10 on the combined set, we report MAP@10, MAP_Q@5, and MAP_H@5 on the test set (as provided by the organizers).

| Model | MAP@10 | MAP@5 |
|---|---|---|
| - | 0.2944 | 0.2689 |
| **Query + Synonyms** | | |
| GPT-4.1-mini | 0.2691 | 0.2453 |
| GPT-4.1 | 0.2754 | 0.2540 |
| GPT-4o | 0.2781 | 0.2560 |
| **Paraphrased Query** | | |
| GPT-4.1-mini | 0.3007 | 0.2727 |
| GPT-4.1 | **0.3065** | **0.2815** |
| GPT-4o | 0.3026 | 0.28 |

Table 2: Results of different query expansion techniques using QPC+IE on both the train and dev sets.

[3] https://www.trychroma.com/

| Method | MAP@10 | MAP@5 |
|---|---|---|
| PQ + Dense | 0.3065 | 0.2815 |
| PQ + Dense + $RR_{\text{CE}}$ (Elkomy and Sarhan, 2024) | 0.3156 | 0.2905 |
| PQ + Dense + $RR_{\text{GPT-4}}$ | 0.4079 | 0.3898 |
| PQ + Dense + $RR_{\text{GPT-4}}$ + NAD | 0.4682 | 0.4511 |
| Dense + $RR_{\text{GPT-4}}$ + NAD | **0.4811** | **0.4660** |

Table 3: Effect of introducing reranker (RR) and no-answer detector (NAD) on performance using the paraphrased query (PQ) and augmented corpus (QPC + IE) on both the train and dev sets. GPT-4 and CE refers to GPT-4 and cross-encoder-based rerankers, respectively.

| Method | Collection | MAP@10 | MAP_Q@5 | MAP_H@5 |
|---|---|---|---|---|
| PQ + Dense + $RR_{\text{GPT-4}}$ + NAD | QPC + IE & SBC+ IE | **0.3351** | **0.3389** | **0.3876** |
| Dense + $RR_{\text{GPT-4}}$ + NAD | QPC + IE & SBC+ IE | 0.3021 | 0.3091 | 0.3461 |
| Dense + $RR_{\text{GPT-4}}$ + NAD | QPC + IE & SBC | 0.2916 | 0.3130 | 0.2936 |

Table 4: Performance of retrieval strategies for related QPCs and HAs given a query on the test set.

## 5 Results and Analysis

In this section, we present the research questions along with the experiments that answer them.

**RQ1: How does augmenting QPCs affect semantic retrieval?**

In Table 1, we present the evaluation results of the two proposed augmentation approaches (with Tafsir and with facts extracted by LLMs (IE)). It is evident that augmenting QPC provides complementary semantic signals. Notably, combining QPC with either Muyassar or IE leads to observable performance gains, confirming the benefit of pairing verse-level content with simplified or pedagogically aligned annotations. However, adding Muyassar and IE together leads to a decline in the performance. We attribute this to the semantic noise when too many interpretative strategies are combined, potentially reducing coherence in the learned embedding space. To this end, we adopt SBC + IE in subsequent experiments.

**RQ2: How effective are LLMs in reformulating the user queries?**

In Table 2, we examine the effect of introducing the paraphrasing and adding synonyms to queries using three variants of GPT-4. The results reveal a clear distinction between the effectiveness of synonym-based and paraphrased query reformulations in Quranic semantic search. Synonym-based reformulations consistently underperformed the baseline, indicating that direct lexical substitution introduces noise and query drift. In contrast, paraphrased queries yield consistent improvements across all evaluated models. These gains highlight the strength of paraphrasing in capturing deeper semantic equivalences and aligning user queries more effectively with relevant passages. We select GPT-4.1 for paraphrasing queries in the following experiments due to its superior performance.

**RQ3: How good is the LLM-based reranker? What is the best combinations of our proposed retrieval pipeline?**

Building on the best results attained from augmentation and paraphrasing queries (PQ), we examine the effect of incorporating the reranker. In Table 3, we demonstrate the effectiveness of using a finetuned cross-encoder(CE)-based (Elkomy and Sarhan, 2024) and GPT-4-based rerankers. While CE-based reranker leads to moderate improvements, a substantial gain is achieved by the LLM-based re-ranker (0.4079 vs. 0.3156 at MAP@10).

In the same table, we report the significant gains brought by integrating the NAD (No-Answer Detection) component, which filters out candidates that do not answer the query.

Building on these findings, we submit the top-two performing pipelines (last two lines of Table 3) on the test set, while adding another run without augmenting SBC to test its effect. The results on the test set are shown in Table 4. The best results are obtained when PQ is combined with the Dense + RR + NAD pipeline on the augmented collection (QPC + IE & SBC + IE), indicating the effectiveness of PQ component.

## 6 Conclusion

In this paper, we described our method for addressing Qur'an and Hadith QA 2025 shared task. We

found out that the augmenting QPC and SBC with information extracted by LLM lead to remarkable gains. After experimenting with multiple lexical- and semantic-based retrieval and reranking methods, we showed that dense search with LLM-based reranker is the best configuration. Our novel attempt to change the query surface text showed clear improvements. Finally, utilizing LLM to judge the binary relevance of a query-passage pair proved to be a promising solution in detecting questions with no answer.

## Limitations

To the best of our knowledge, resources providing tafsir or detailed explanations of Hadith from Sahih al-Bukhari are not readily available.

In addition, our preliminary experiments were conducted exclusively on the QPC dataset, as it is the only resource with ground truth annotations available. Consequently, we did not develop or evaluate our best-performing retrieval system for retrieving passages from QPC or Hadith collections. Accordingly, the findings reported at this stage are limited in scope, which reduces confidence in identifying the optimal strategy.

## Acknowledgments

We thank Dr. Joel Mackenzie for the resource support provided to Watheq Mansour.

## References

Hessa Alawwad, Lujain Alawwad, Jamilah Alharbi, and Abdullah Alharbi. 2023. Ahjl at qur'an qa 2023 shared task: Enhancing passage retrieval using sentence transformer and translation. In *Proceedings of ArabicNLP 2023*, pages 702–707.

Mohamed Basem, Islam Oshallah, Baraa Hikal, Ali Hamdi, and Ammar Mohamed. 2024. Optimized quran passage retrieval using an expanded qa dataset and fine-tuned language models. In *The International Conference of Advanced Computing and Informatics*, pages 244–254. Springer.

Mohammed Alaa Elkomy and Amany Sarhan. 2024. Tce at qur'an qa 2023 shared task: Low resource enhanced transformer-based ensemble approach for qur'anic qa. *arXiv preprint arXiv:2401.13060*.

Rana Malhas, Watheq Mansour, and Tamer Elsayed. 2022. Qur'an QA 2022: Overview of the first shared task on question answering over the holy qur'an. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 79–87, Marseille, France. European Language Resources Association.

Rana Malhas, Watheq Mansour, and Tamer Elsayed. 2023. Qur'an QA 2023 shared task: Overview of passage retrieval and reading comprehension tasks over the holy qur'an. In *Proceedings of ArabicNLP 2023, Singapore (Hybrid), December 7, 2023*, pages 690–701. Association for Computational Linguistics.

Hamdy Mubarak, Rana Malhas, Watheq Mansour, Abubakr Mohamed, Mahmoud Fawzi, Majd Hawasly, Tamer Elsayed, Kareem Darwish, and Walid Magdy. 2025. IslamicEval 2025: The First Shared Task of Capturing LLMs Hallucination in Islamic Content. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics. Co-located with EMNLP 2025, November 5–9.

Abdulrezzak Zekiye and Fadi Amroush. 2023. Aljawaab at qur'an qa 2023 shared task: Exploring embeddings and gpt models for passage retrieval and reading comprehension. In *Proceedings of ArabicNLP 2023*, pages 743–747.

## A Evaluating LLM Embeddings on QPC

As shown in Table 5, the "Muffakir embeddings" [4]—trained on culturally and religiously aligned Arabic corpora—demonstrated strong performance, outperforming general-purpose multilingual models such as "gte-multilingual-base" [5] and "multilingual-e5-large" [6] models. This suggests that domain-adapted embeddings are more effective at capturing Qur'anic semantics. Although Muffakir is smaller in scale than OpenAI's model, its competitive results underscore the advantages of domain relevance. Meanwhile, the superior performance of text-embedding-3-large is likely due to a combination of advanced model architecture, large-scale multilingual training, and task-specific optimization for retrieval.

| Model | MAP@10 | MAP@5 |
|---|---|---|
| gte-multilingual-base | 0.1542 | 0.1429 |
| multilingual-e5-large | 0.1814 | 0.1678 |
| Muffakir Embedding | 0.1994 | 0.1867 |
| text-embedding-3-large | **0.2761** | **0.2553** |

Table 5: Performance of different retrieval methods using semantic search approaches on both the train and dev sets.

## B Prompt Engineering for Factual Information Extraction from Islamic Texts

Figures 1 and 2 illustrate examples of prompt designs aimed at extracting factual information from Quranic verses and Hadith texts, respectively.

استخرج من النص القرآني التالي جميع الكيانات والمعاني الرئيسية المرتبطة به،
- مثل الشخصيات، الأماكن، الأحداث،
-المفاهيم الإسلامية، صفات الله،
-والأحكام الشرعية (إن وجدت).

**النص:**
"{QPC_text"

Figure 1: Example of a prompt designed to extract factual information from Quran verses.

استخرج من الحديث التالي المعاني والحقائق الرئيسية، مثل:
-الشخصيات المذكورة (مثل الراوي، الصحابة، أو النبي)
-الأحداث أو الأفعال التي وقعت
-الأوامر أو النواهي (إن وُجدت)
-الأحكام الشرعية أو التوجيهات الأخلاقية
-المفاهيم الإسلامية (مثل الإيمان، الإحسان، الجار...)

**النص:**
"{hadith_text}"

Figure 2: Example of a prompt designed to extract factual information from Hadith.

These prompts are specifically crafted to guide LLMs in identifying key Islamic concepts and legal rulings embedded within each Quranic passage or Hadith.

[4] https://huggingface.co/mohamed2811/Muffakir_Embedding
[5] https://huggingface.co/Alibaba-NLP/gte-multilingual-base
[6] https://huggingface.co/intfloat/multilingual-e5-large