ThinkDrill at IslamicEval 2025 Shared Task: LLM Hybrid Approach for Qur'an and Hadith Question Answering

Eman Elrefai[1]
[1]Alexandria University
eman.lotfy.elrefai@gmail.com

Toka Khaled[2]
[2]Al-Azhar University
Tokakhaled98@gmail.com

Ahmed Soliman[3]*
[3]University of Florida
ahmed.soliman@ufl.edu

Abstract

This paper presents our approach to Subtask 2 of IslamicEval 2025, a shared task that involves retrieving relevant passages from Quranic verses and Sahih Bukhari hadiths to answer Modern Standard Arabic (MSA) questions. We developed a multi-pipeline hybrid system that combines three complementary approaches: fine-tuned embedding models using triplet loss, keyword-based fuzzy matching, and large language model guided retrieval. Our system achieved `MAP@10` of 0.173, `MAP_Q@5` of 0.1881, and `MAP_H@5` of 0.2414 in the test set, demonstrating the effectiveness of combining multiple retrieval strategies for Arabic religious text question answering.

1 Introduction

The Qur'an and Hadith Question Answering (QHQA) task (Mubarak et al., 2025) addresses the challenge of retrieving relevant religious passages to answer questions posed in MSA. The Qur'an and hadith are deeply embedded in the daily lives of millions of Muslims worldwide, influencing their decisions, moral reasoning, and spiritual practices. With the increasing proliferation of Large Language Models (LLMs) in question-answering systems, systems responding to questions about these religious sources must maintain high accuracy and reliability.

This task builds on prior Qur'an QA challenges (2022, 2023) (Malhas et al., 2022, 2023), which focused only on Qur'an-based QA. Many teams proposed strong pipelines with promising results, and those works inspired our approach like Mahmoudi et al. (2023); Elkomy and Sarhan (2024). The main difference now is the inclusion of hadith, making the task broader and more challenging. Another key change is that answers must be retrieved from the entire Qur'an or hadith, unlike earlier setups where a specific passage was given and answers were extracted from it. Personally, our participation (Sleem et al., 2022) in Qur'an QA 2022 was a starting point that shaped how we combined prior pipelines with new technologies in this work.

Our main system strategy employs a multi-pipeline approach that leverages the strengths of different retrieval methods. Our key findings show that while individual approaches have limitations, their combination significantly improves performance. Our results show that the development set with `MAP@10` of 0.32 and 0.173 for the test set.

2 Background

The IslamicEval 2025 Subtask 2 requires systems to return a ranked list of answer-bearing passages from two collections: Quranic verses covering the Holy Qur'an and hadiths from Sahih Bukhari. Given a free-text question in MSA such as:

ما هو فضل ليلة القدر؟

The system should return relevant passages like:

إِنَّا أَنزَلْنَاهُ فِي لَيْلَةِ الْقَدْرِ * وَمَا أَدْرَاكَ مَا لَيْلَةُ الْقَدْرِ * لَيْلَةُ الْقَدْرِ خَيْرٌ مِّنْ أَلْفِ شَهْرٍ (Rank 1)

عَنْ أَبِي هُرَيْرَةَ رَضِيَ اللَّهُ عَنْهُ قَالَ: قَالَ رَسُولُ اللَّهِ صَلَّى اللَّهُ عَلَيْهِ وَسَلَّمَ: مَنْ قَامَ لَيْلَةَ الْقَدْرِ إِيمَانًا وَاحْتِسَابًا غُفِرَ لَهُ مَا تَقَدَّمَ مِنْ ذَنْبِهِ (Rank 2)

2.1 Dataset Details

The dataset consists of 1,266 Quranic passages from the Quranic Passage Collection (QPC), 2,254 hadiths from Sahih Bukhari, training questions with manually annotated relevance judgments, and questions without answers marked with passage ID "-1". Initially, the training data only contained Qur'an answers. Our team manually added hadith answers to create a more balanced training set.

*Also affiliated with Al-Azhar University

2.2 Related Work

Previous work on Arabic question answering has primarily focused on general domain texts. Earlier versions of similar tasks focused exclusively on Quranic sources, but the inclusion of hadith as a complementary resource introduces additional complexity. Hadith collections present unique challenges due to their narrative structure, chain of transmission (isnad), and the potential for fabrication, requiring careful verification and authentic sourcing.

While several retrieval systems have been developed specifically for hadith collections (Mahmood et al., 2018), fewer systems effectively combine Qur'an and hadith sources in a unified retrieval framework. Recent work in (Fawzi et al., 2025) demonstrates the importance of accurate religious text retrieval systems, particularly given the widespread influence of these sources on personal decisions and the need for reliable information retrieval in the era of increasing LLM deployment.

Our approach builds upon sentence transformers for multilingual retrieval while addressing the specific requirements of Islamic texts and the challenge of combining these two distinct yet complementary religious sources.

3 System Overview

Our hybrid system consists of three complementary pipelines designed to capture different aspects of semantic similarity and relevance. Figure 1 illustrates the overall architecture of our multi-pipeline approach.

3.1 Pipeline 1: Fine-tuned Embedding Model

3.1.1 Training Phase

The training pipeline relied on a curated dataset constructed from multiple sources to ensure comprehensive coverage of Qur'anic and Hadith material. First, official Qur'an QA pairs provided by the competition were used as a foundation. To expand beyond the Qur'an, additional Hadith QA pairs were constructed by sourcing relevant narrations from *Sahih al-Bukhari*. This was feasible only for a limited subset of questions, so we further incorporated the HAQA dataset, aligning its QA pairs with *Sahih al-Bukhari* narrations through automated normalization (removing diacritics, punctuation, and text inconsistencies) and fuzzy matching. Matches with similarity scores above a chosen threshold were retained, producing a final aligned dataset containing question text, answer, and narration. The HAQA dataset is available at `https://github.com/scsaln/HAQA-and-QUQA/blob/main/HAQA.csv`. This curated dataset balanced Qur'anic and Hadith sources, enabling the retrieval model to learn cross-domain semantic relationships.

On top of this dataset, we fine-tuned a multilingual sentence transformer (Reimers and Gurevych, 2019) using triplet loss (Yeruva et al., 2022). The augmentation process expanded the original corpus into structured triplets by systematically constructing positive and negative passages for each question:

- **Positive passages:** For each question, all valid answers from the Qur'an and Hadith were included as positives. When multiple passages addressed the same question, each of them was considered a valid positive. For unanswerable questions, we used the placeholder answer لا يوجد as the only positive.

- **Negative passages:** Non-relevant passages were sampled from the remaining pool of Qur'an and Hadith texts. For unanswerable questions, all real passages in the corpus were treated as negatives.

- **Triplet construction:** Each training instance consisted of an anchor (the question), a positive passage, and a negative passage. To increase data diversity, multiple triplets were generated per question by pairing the same anchor with different positive and negative samples.

The fine-tuned model based on `intfloat/multilingual-e5-base` served as the **retriever**, encoding both queries and passages into a shared embedding space and retrieving candidate passages using cosine similarity. To further refine the retrieval results, we employed the **reranker** model, specifically the pretrained `cross-encoder/ms-marco-MiniLM-L-6-v2`, which jointly encodes query–passage pairs and assigns a relevance score. This two-stage pipeline ensured efficient large-scale retrieval while improving precision through reranking.

3.1.2 Inference Phase:

Once the model is fine-tuned, it is employed in a retrieval pipeline for inference. Queries are encoded into embeddings and matched against a vector database of Qur'an and Hadith passages. A retriever retrieves the top candidate passages using cosine similarity, which are then refined by a reranker before producing the final ranked results. Unlike full retrieval-augmented generation (RAG) systems, our pipeline focuses solely on retrieving and ranking authoritative passages without generating new text.

3.2 Pipeline 2: Keyword-based Fuzzy Matching

We used GPT-4 to extract relevant keywords from the questions and then used fuzzy string matching to find passages containing similar terms. We used RapidFuzz (Ye et al., 2021) a fast Python library for fuzzy string matching to compute partial ratio similarity scores. The algorithm extracts keywords using the LLM prompt "Give me the main keywords that I can search for to get answers from the Qur'an and Hadith", cleans the Arabic text by removing diacritics and normalising the characters, applies fuzzy partial ratio matching with a threshold of 70%, and ranks results by similarity score.

3.3 Pipeline 3: LLM-guided Retrieval

The input to this pipeline is the users question together with the instruction: "Answer questions using only Quran and Sahih Bukhari. Provide exact verses/hadiths, not interpretations. Use -1 if no answer exists."

The output is either the exact verse or hadith matching the question, or -1 if no relevant answer is found.

We chose Claude Sonnet 4 because it follows instructions well, handles long passages reliably, and shows fewer hallucinations than smaller or larger alternatives. It also provides a good balance between accuracy, speed, and cost.

3.4 Hybrid Combination

Results from all three pipelines were combined using score normalization and weighted averaging to produce final rankings.

4 Experimental Setup

4.1 Data Preparation

Following cleaning, the dataset was structured for Sentence-BERT (SBERT) triplet loss training. Triplet construction formatted each entry as (anchor, positive, negative), where anchor represents the text of the question, positive encompasses corresponding relevant answers from the Qur'an or hadith, and negative includes semantically irrelevant passages from the Qur'an or hadith. Data splitting used stratified sampling to ensure both Quranic and hadith entries were proportionally represented in training and validation sets. UTF-8 encoding stored all text fields in a Pandas DataFrame with explicit column names (`question`, `positive_passage`, `negative_passage`).

4.2 Data Preprocessing

We applied preprocessing to align with sentence transformer requirements. Data was length-filtered (10512 tokens) and segmented using a sliding window to preserve context within token limits. Arabic-specific cleaning included diacritic removal, normalization of letter variants, tatweel and honorific symbol removal, and whitespace normalization. Stopwords were retained due to their semantic role in Quranic and hadith texts, while redundant punctuation was removed. Texts were tokenized with the `intfloat/multilingual-e5-base` (Wang et al., 2024) tokenizer, and triplets were batched into uniform tensors with attention masks for SBERT triplet loss training. For long passages, we applied chunking into 150-character segments with 30-character overlap to enhance retrieval granularity.

4.3 Training Configuration

We used base model `intfloat/multilingual-e5-base`, 2 epochs, batch size 16 with gradient accumulation, learning rate 2e-5 with 100 warm-up steps, triplet loss function with cosine distance, and hardware acceleration through Google Colab with GPU.

4.4 Evaluation Metrics

The official metrics included `MAP@10` (Mean Average Precision at rank 10), `MAP_Q@5` (MAP at rank 5 for Qur'an passages only), and `MAP_H@5` (MAP at rank 5 for hadith passages only).

5 Results and Error Analysis

Our system was evaluated on both development and test sets, achieving the following results:

Dataset	MAP@10	MAP_Q@5	MAP_H@5
Development	0.32	0.35	-
Test	0.173	0.1881	0.2414

Table 1: Overall performance on development and test sets.

To better understand these results, we further analyzed the contribution of each pipeline component. Since the organizers provided official test set results only for the submitted runs, the per-pipeline results in Table 2 were computed on the development set using the released evaluation script. The Hybrid Combination score corresponds to our submitted run on the test set. Table 2 reports the performance of individual pipelines compared to the hybrid system.

Pipeline	MAP@10
Embedding Model Only	0.15
Keyword Matching Only	0.08
LLM-guided Only	0.12
Hybrid Combination	0.173

Table 2: Performance of individual pipelines on the test set.

The fine-tuned embedding model provided the strongest standalone baseline, while keyword matching proved useful for questions relying on exact term overlap. The LLM-guided approach showed potential but was constrained by input length limitations. The hybrid combination achieved the best balance, outperforming any individual pipeline.

5.1 Coverage Analysis

On the test set of 71 questions, our retrieval system achieved 76.1% coverage: 54 questions had answers while 17 questions were marked as "no answer." On average, the system returned 15.2 passages per question.

5.2 Error Analysis

We observed four main error types: semantic mismatch (retrieving passages with overlapping words but different intent, e.g., prayer times vs. prayer importance), keyword limitations (lexical matches missing conceptual meaning), LLM constraints (token limits restricting comprehensive answers), and domain specificity (questions requiring advanced theological knowledge). Example errors include ما عمر الأرض بالسنوات تحديدًا؟ (expected: -1, predicted: creation verses)

5.3 No-Answer Detection

We evaluated the system's ability to detect questions without valid answers. A confidence threshold of 0.35 was applied: if the highest passage score fell below this threshold, the system classified the question as no answer. Evaluation was carried out on the held-out test set of 71 questions, which included 17 questions without valid answers. The model achieved a precision of 0.65 and recall of 0.47 on this subset.

6 Conclusion

We proposed a hybrid approach for Arabic Qur'an and hadith question answering that integrates fine-tuned embeddings, keyword matching, and LLM guidance. Our system demonstrated strong performance during development (`MAP@10`: 0.32) and achieved one of the top scores among participating teams on the final benchmark (`MAP@10`: 0.173). These results highlight both the effectiveness of our design and the potential for further improvements in handling diverse real-world queries.

Future work directions include incorporating Islamic scholarly knowledge graphs, exploring retrieval-augmented generation approaches, and creating larger, more diverse training datasets with theological expert annotations. The task highlights the complexity of understanding religious texts and the need for specialized approaches beyond general-domain techniques.

For reproducibility, the implementation and code are available at `ThinkDrill at IslamicEval 2025 | GitHub`.

Acknowledgments

We thank the IslamicEval 2025 shared task organizers for providing this valuable evaluation framework. We also acknowledge the anonymous reviewers for their constructive feedback.

References

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*.

Mahmoud Fawzi, Walid Magdy, and Björn Ross. 2025. 'the prophet said so!': On exploring hadith presence on arabic social media. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–23.

Ahsan Mahmood, Hikmat Ullah Khan, Fawaz K Alarfaj, Muhammad Ramzan, and Mahwish Ilyas. 2018. A multilingual datasets repository of the hadith content. *International Journal of Advanced Computer Science and Applications*, 9(2):165–172.

Ghazaleh Mahmoudi, Yeganeh Morshedzadeh, and Sauleh Eetemadi. 2023. Gym at quran qa 2023 shared task: Multi-task transfer learning for quranic passage retrieval and question answering with large language models. In *Proceedings of ArabicNLP 2023*, pages 714–719.

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*, pages 690–701, Singapore (Hybrid). 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 (Arabic-NLP 2025)*, Suzhou, China. Association for Computational Linguistics. Co-located with EMNLP 2025, November 5–9.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Ahmed Sleem, Eman Mohammed lotfy Elrefai, Marwa Mohammed Matar, and Haq Nawaz. 2022. Stars at quran qa 2022: Building automatic extractive question answering systems for the holy quran with transformer models and releasing a new dataset. 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 146–153.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Aoshuang Ye, Lina Wang, Lei Zhao, Jianpeng Ke, Wenqi Wang, and Qinliang Liu. 2021. Rapidfuzz: Accelerating fuzzing via generative adversarial networks. *Neurocomputing*, 460:195–204.

Nagamani Yeruva, Sarada Venna, Hemalatha Indukuri, and Mounika Marreddy. 2022. Triplet loss based siamese networks for automatic short answer grading. In *Proceedings of the 14th annual meeting of the forum for information retrieval evaluation*, pages 60–64.

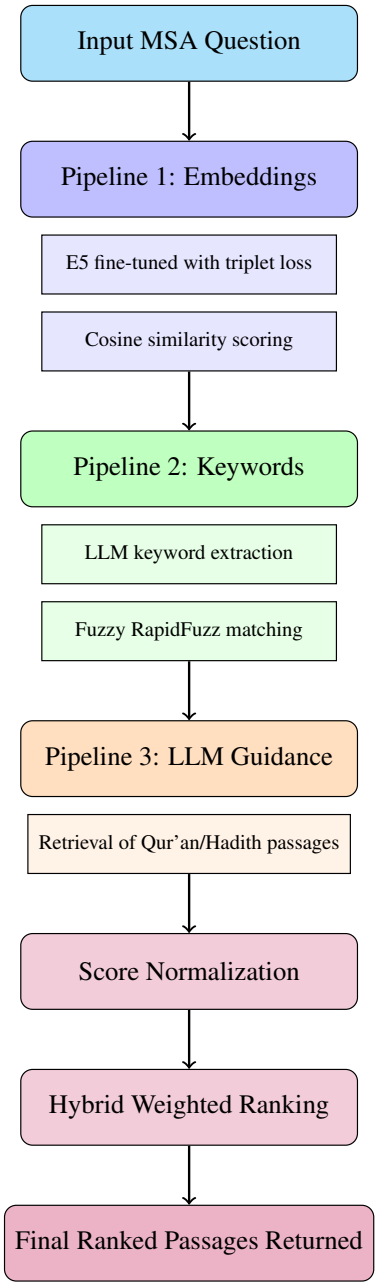

Figure 1: System architecture: input questions are processed via three pipelines with distinct colors.

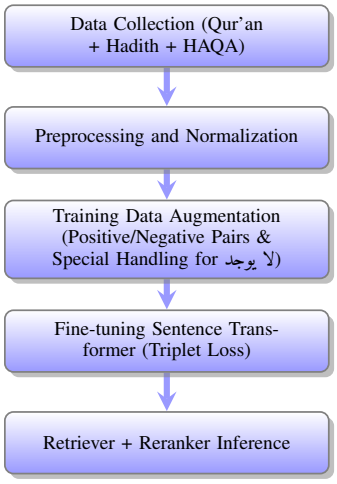

Figure 2: Pipeline for fine-tuning and retrieval

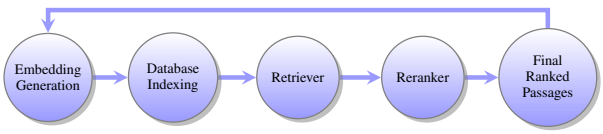

Figure 3: Retrieval and reranking pipeline