Tokenizers United at QIAS 2025 Shared Task: A Retrieval-Augmented Generation Pipeline for Islamic Knowledge Assessment

Mohamed Samy[2], **Mayar Boghdady**[1], **Marwan El Adawi**[1], **Mohamed Nassar**[1], **Ensaf Hussein**[1]
[1] **School of Information Technology and Computer Science, Nile University**
[2] **Faculty of Computer and Information Sciences, Ain Shams University**
MBoghdadi@nu.edu.eg, M.Mahmoud2179@nu.edu.eg, M.Ali2265@nu.edu.eg, EnMohamed@nu.edu.eg
mohamedsamyy02@gmail.com

Abstract

This paper presents the approach and results for Sub Task 2: General Islamic Knowledge Question Answering at QIAS 2025, a shared task designed to evaluate the capabilities of Large Language Models (LLMs) in answering multiple-choice questions across diverse domains of Islamic knowledge, including theology, jurisprudence, biography, and ethics. A Retrieval-Augmented Generation (RAG) system powered by the Gemini language model was developed for this task.

In the proposed system, the *retriever module* performs semantic search over curated classical Islamic sources to identify passages relevant to each input question, while the *generator module* leverages the LLM to reason over the retrieved evidence and generate a final answer. This integration of evidence retrieval with contextual reasoning enables accurate responses across diverse knowledge areas.

On the official test set, the system achieved an accuracy of 87%, ranking 5th out of 10 participating teams in QIAS 2025 Sub Task 2. These results demonstrate the effectiveness of combining retrieval-based evidence with generative reasoning in specialized religious domains, highlighting the potential of RAG architectures for *high-stakes, knowledge-intensive question answering tasks* and confirming their robustness in the QIAS 2025 benchmark.

1 Introduction

Automated assessment of Islamic knowledge is a critical task requiring both linguistic proficiency and deep domain expertise. It faces challenges from the complexity of Arabic morphology and orthography, the breadth of Islamic sources, and the demand for trustworthy responses in educational contexts.

Within the context of the QIAS2025 Shared Task (Sub Task 2: Islamic Assesment) , existing methods based on generic LLMs, classical retrieval, or translation pipelines often fail to capture domain-specific semantics, suffer from hallucinations, and lack grounding in authoritative sources. This highlights a gap between current capabilities and the requirements of knowledge- intensive domains such as Islamic studies.

To address this, a Retrieval-Augmented Generation (RAG) framework is proposed, combining Muffakir embeddings for evidence retrieval with Gemini 2.5 Flash Lite for generative reasoning. Preprocessed texts are segmented into enriched units for efficient retrieval, ensuring grounded and accurate responses. Experiments show the system achieves 84% precision on development data and 87% on the official test set, outperforming baselines in the QIAS 2025 evaluation (Bouchekif et al., 2025a).

The main contributions are: (1) a curated Arabic knowledge base for Islamic studies, (2) integration of retrieval with a state-of-the-art LLM, and (3) empirical validation in high-stakes assessment tasks under the QIAS 2025 benchmark (Bouchekif et al., 2025a).

2 Related Work

Several studies have explored the development of Islamic Question Answering (QA) systems, following either retrieval-based methods or knowledge-based approaches enhanced with semantic processing. An early example is (Mohamed et al., 2015), which introduced *Al-Bayan*, a knowledge-based Arabic answer selection system for Islamic sciences that participated in SemEval-2015 Task 3. By combining a Quranic ontology enriched with Tafseer resources, keyword matching, and a decision tree classifier, the system achieved an accuracy of 74.53% and a macro-F1 score of 67.65%.

A broader perspective is provided in (Alnefaie

Proceedings of The Third Arabic Natural Language Processing Conference, pages 960–965

Question	Answers	Level	Label
ما هو القول القديم للشافعي في صوم أيام التشريق؟	A) لا يجوز صومها مطلقاً. B) يجوز صومها للمتمتع إذا عدم الهدي عن الأيام الثلاثة الواجبة في الحج. C) يجوز صومها لمن لم يجد الهدي فقط. D) يجوز صومها للمسافر فقط.	advanced	B
أنا سورة قصيرة، نزلتُ كاملة بسبب كلمة غضب قالها قريب للنبي -صلى الله عليه وسلم- رداً على دعوته. فما اسمي؟	A) سورة الكافرون B) سورة المسد (تبت) C) سورة النصر D) سورة الهمزة	beginner	B
أنا اختلاف ألفاظ الوحي المنزل في الحروف أو كيفيتها من تخفيف وتشديد وغير ذلك. فماذا أكون حسب تعريف الزركشي؟	A) التجويد B) القرآن C) أسباب النزول D) القراءات	advanced	D
بماذا عرفت أخت أنس بن النضر أخاها الذي استشهد يوم أحد؟	A) بوجهه B) بيده C) ببناته D) كل الأجوبة خطأ	intermediate	C
اختر الآية من سورة الضحى التي تدل على معنى ععما صرمك فتركك، وما أبغضك منذ أحبك"	A) ألم يجدك يتيماً فآوى B) ووجدك ضالاً فهدى C) ووجدك عائلاً فأغنى D) ما ودعك ربك وما قلى	intermediate	D

Table 1: Sample of Islamic knowledge assessment questions with answer options (A–D), difficulty level, and correct label.Latin labels are forced with `\textlatin{}` to avoid RTL localization.

et al., 2023), which presented a comprehensive survey of Islamic QA systems drawing on Qur'an, Hadith, and Fatwa sources. Their evaluation classified systems into traditional retrieval-based and knowledge-based categories, with deep learning models such as AraBERT, AraElectra, and mT5 showing promise but remaining highly dependent on dataset quality. The survey applied thirteen evaluation criteria, concluding that most current systems suffer from limited coverage, lack of public availability, and difficulty in handling non-factoid questions.

Recent contributions have expanded the scope of Islamic QA to general knowledge domains including theology, jurisprudence, biography, and ethics. (Qamar et al., 2024) introduced a large-context Islamic QA dataset for non-factoid questions, derived from Qur'an, Tafsir, and Hadith. Domain-specific legal reasoning has also been addressed; for example, (Al-Qurishi et al., 2022) proposed *AraLegal-BERT*, a BERT model fine-tuned on Arabic legal texts to enhance QA in Islamic jurisprudence. Other benchmarks include (Malhas, 2023), which developed *QuranQA* for span selection tasks, and (Premasiri et al., 2022), which introduced *MadinaQA* for beginner and intermediate Islamic studies. Advances in Retrieval-Augmented Generation (RAG) were demonstrated by (Alan et al., 2024), who presented *MufassirQAS*, while (Rizqullah et al., 2023) proposed *QASiNa*, targeting QA over Sirah Nabawiyah texts.

Work has also extended beyond Arabic into Persian, with (Ghafouri et al., 2023), (Etezadi and Shamsfard, 2021), and (Zeinalipour et al., 2025) developing QA systems and benchmarks for multi-hop and multiple-choice reasoning. Domain-specific applications include Islamic inheritance law, where (Bouchekif et al., 2025b) provided benchmarks and evaluations of large language models for legal reasoning.

Taken together, these studies highlight the increasing interest in combining knowledge-based and deep learning approaches to address the challenges of Islamic QA. The literature underscores the importance of multilingual support, robust reasoning across complex religious texts, and domain-specific legal knowledge representation, while also pointing to the potential of modern language models and Retrieval-Augmented Generation to advance the field.

2.1 Task Setup: QIAS 2025 Shared Task

The (Bouchekif et al., 2025a) shared task has been established as a benchmark competition to evaluate systems for Islamic Question Answering. It consists of multiple subtasks designed to assess models in handling diverse domains of Islamic knowledge. This work focuses on **(Bouchekif et al., 2025a) Subtask 2: General Islamic Knowledge QA**, which targets multiple-choice question answering across domains includ-

ing theology, jurisprudence, biography, and ethics.

Subtask 2 attracted participation from ten international teams. Evaluation was based on system accuracy in selecting the correct option among four candidates. The system presented in this paper ranked **5th out of 10** with an accuracy of 0.875, demonstrating the competitiveness of lightweight RAG pipelines against more complex architectures.

2.2 Dataset

To evaluate the proposed approach, the **QIAS 2025** (Bouchekif et al., 2025a) dataset is used. This benchmark includes multiple-choice questions on *Qur'anic studies, Hadith, Fiqh, Islamic history, and Arabic linguistics*, each annotated with difficulty level (*beginner, intermediate, advanced*) and the correct label. The dataset covers both factual recall and higher-order reasoning, enabling assessment of comprehension and semantic interpretation in Islamic knowledge. Table 1 shows sample questions with answer options, difficulty levels, and correct labels, highlighting the diversity of jurisprudential, exegetical, and historical content.

In addition to the question–answer pairs, the QIAS organizers provide a collection of classical Islamic reference works that serve as the textual backbone for knowledge-intensive tasks. These include:

- *Usul al-Fiqh and Legal Maxims*
- *Al-Itqan fi Ulum al-Qur'an* (The Perfect Guide to the Sciences of the Qur'an)
- *Al-Sirah wa al-Shama'il* (Prophetic Biography and Characteristics)
- *Tashnif al-Masamih bi-Jam' al-Jawami'* (A Comprehensive Collection of Jurisprudential Principles)
- *Manhaj al-Naqd fi Ulum al-Hadith* (Methodology of Criticism in the Sciences of Hadith)

These books were segmented into smaller chunks and indexed to form the system's knowledge base, allowing retrieval-augmented generation to ground answers in authoritative Islamic sources.

3 Methodology

The Islamic knowledge assessment system is designed as a multistage pipeline that combines Arabic text preprocessing, embedding-based retrieval, and large language model (LLM) generative reasoning. The overall workflow is depicted in Figure 1, which describes the stages from raw document ingestion to final response generation. The system architecture begins with a query question, followed by query embedding, vector search for relevant knowledge chunks, prompt construction, LLM-based reasoning, and ultimately the production of an answer.

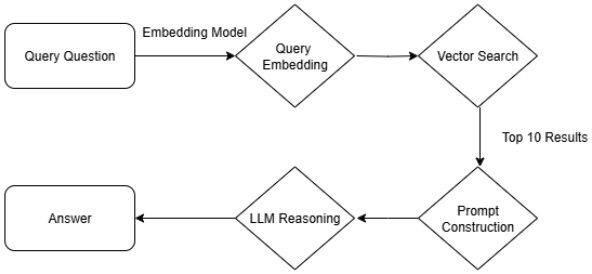

Figure 1: Proposed pipeline of the Islamic knowledge assessment system.

Document ingestion is performed by extracting text from diverse file formats. The extracted content undergoes cleaning, which involves the removal of diacritics, Tatweel, unwanted symbols, phone numbers, emails, and URLs. Text normalization for punctuation and whitespace is applied to ensure consistency. After cleaning, the text is split into overlapping chunks, which may be based on words, sentences, or paragraphs. Each chunk is annotated with metadata such as keywords, positional information, and unique identifiers.

Embedding-based retrieval constitutes the second stage of the pipeline. The preprocessed text chunks are transformed into dense vector representations using the `Muffakir Embedding` model. These embeddings are stored in a vector database, allowing efficient similarity-based retrieval when a query is introduced. The user query is also converted into an embedding, which is compared against the stored vectors to identify the most semantically relevant passages.

Prompt construction serves as the bridge between retrieval and reasoning. Once relevant chunks are retrieved, they are assembled together with the user query into a structured prompt. This ensures that the LLM receives not only the query

but also the most contextually aligned passages, enabling grounded and accurate responses.

Generative reasoning is executed by large language models that process the constructed prompt. Several models were evaluated, including `Silma`, `Qwen3 1.7B/8B`, `Aya`, and `Allam`. Among these, `Gemini 2.5 Flash Lite` demonstrated superior performance in generating coherent and contextually faithful answers. A flash reranker was also tested for post-retrieval refinement; however, direct retrieval combined with Gemini exhibited more reliable outcomes.

The final architecture integrates these components into a Retrieval-Augmented Generation (RAG) system. Beginning with the query question, the process proceeds through query embedding, vector search, prompt construction, and LLM-based reasoning, which culminates in the generation of the final answer. This structured pipeline enables the system to leverage external knowledge repositories while preserving the fluency and reasoning ability of modern LLMs. Table 1 provides illustrative examples of multiple choice questions (MCQs) produced by the system, including their correct labels and difficulty levels.

4 Results

in Subtask 2 (Accuracy: 0.93). Detailed scores and rankings are shown in Table 5 (Appendix B As part of the QIAS 2025 Shared Task (Bouchekif et al., 2025a), this system was evaluated on Subtask 2. On the development set, accuracies ranged between 44.29% and 84.29% across different configurations (Table 2). The highest score (84.29%) was obtained using `Gemini 2.5 Flash Lite` with `Muffakir_Embedding` and direct similarity search (Top-K = 10), showing that lightweight, well-aligned components can surpass more complex pipelines.

Model size did not consistently translate into better results, as the Qwen3 models showed variable performance across scales (54.43–78.00%). Embedding choice was the most decisive factor: `Muffakir_Embedding` consistently outperformed other embeddings, while `silma-embedding-matryoshka-v0.1` achieved the lowest accuracy (44.29%). Retrieval strategy also proved critical, with direct retrieval outperforming reranking approaches (Flash, BGE). Chain-of-thought prompting offered only modest improvements compared to embedding and retrieval methods.

Overall, the findings highlight that domain-specific embeddings, lightweight LLMs, and simple retrieval mechanisms are more effective than scaling models or adding complex reasoning layers. Our optimized configuration—Qdrant with cosine similarity, a chunk size of 400 characters with overlap of 100 characters, Top-10 retrieval, and 768-dimensional `Muffakir_Embeddings`—achieved 87.00% accuracy on the held-out test set, confirming strong generalization. The complete set of hyperparameters for this configuration is summarized in (Table 3).

A breakdown by difficulty level (Table 4) shows that performance was highest on beginner questions (89.14%), followed by intermediate (83.43%), while advanced questions proved more challenging (75.43%). This suggests that while the system is robust in handling straightforward queries, further optimization is needed to improve reasoning in complex or nuanced scenarios.

5 Conclusion

Within the framework of the QIAS 2025 (Bouchekif et al., 2025a) Shared Task , specifically Subtask 2, this study demonstrates that effective automated assessment in the domain of Islamic knowledge can be achieved through a carefully optimized Retrieval-Augmented Generation (RAG) pipeline. The experiments confirm that domain-specific embeddings—particularly `Muffakir_Embedding`—when paired with a lightweight yet capable LLM such as `Gemini 2.5 Flash Lite`, significantly outperform larger, general-purpose models. Contrary to common assumptions, complex reranking strategies and large-scale models did not yield superior results; in this case, direct retrieval with cosine similarity achieved the highest accuracy of 87%.

The findings underscore three key lessons for high-stakes, domain-specific QA systems: (1) high-quality, domain-tuned embeddings are critical for precision, (2) retrieval quality has a greater impact than advanced reasoning prompts or reranking layers, and (3) computational efficiency and scalability can be maintained without sacrificing accuracy.

In the official results of the shared task, the system achieved a ranking of **5th out of 10 teams** in (Bouchekif et al., 2025a) Subtask 2, highlighting

Table 2: Development set accuracies across configurations. "Reranker" indicates post-retrieval reranking. "CoT" indicates Chain-of-Thought prompting.

LLM Model	Embedding model	Reranker / CoT	Acc (%)
Qwen3 (0.6B)	Arabic-Triplet-Matryoshka-V2	-	54.4
Qwen3 (1.7B)	Arabic-Triplet-Matryoshka-V2	-	62.4
Gemini 2.5 Flash Lite	**Muffakir_Embedding**	**-**	**84.3**
Gemini 2.5 Flash Lite	Muffakir_Embedding	Flash reranker	77.9
Qwen3 (8B)	mohamed2811/Muffakir_Embedding	-	58.4
Aya (8B)	silma-embedding-matryoshka-v0.1	-	44.3
Qwen3 (8B)	Muffakir_Embedding	-	78.00
Qwen3 (8B)	silma-embedding-matryoshka-v0.1	BGE-reranker-v2-m3	69.0
Qwen3 (8B)	mohamed2811/Muffakir_Embedding	BGE-reranker-v2-m3 + CoT	75.0

Table 3: Key hyperparameters (final configuration).

Component	Value
Chunk size	400 characters
Overlap	100 characters
Top-K retrieval	10 (based on cosine similarity)
Embedding dim	768 (Muffakir_Embedding)
Vector count	∼15,000
HNSW: m	64
HNSW: ef_construct	1024
HNSW: full_scan_threshold	0
HNSW: payload_m	96
Optimizers: indexing_threshold	14,000
Optimizers: default_segment_number	40
Optimizers: max_optimization_threads	4

Table 4: Accuracy by difficulty level.

Level	Wrong	Correct	Accuracy (%)
Advanced	43	132	75.43
Beginner	38	312	89.14
Intermediate	29	146	83.43

its competitiveness in a multilingual and domain-sensitive evaluation setting. Future research directions include expanding the knowledge base to additional Islamic sciences, incorporating multilingual capabilities for cross-lingual assessment, and integrating adaptive difficulty calibration to further enhance learner evaluation.

Limitations

A primary limitation of this work lies in the restricted computational resources, which prevented extensive experimentation with larger and more advanced reasoning models that could potentially achieve higher accuracy. In addition, the evaluation of larger and more precise embedding models, as well as the use of computationally intensive reranking strategies, was not feasible under the available setup. These constraints may have capped the system's performance ceiling, suggesting that future studies with greater resources could further enhance both retrieval quality and answer generation.

Acknowledgments

The organizers of QIAS 2025: Islamic Q&A Shared Task are gratefully acknowledged for providing the datasets and guidance that enabled this work. The QIAS benchmark,(Bouchekif et al., 2025a) covering general Islamic knowledge, served as the basis for system development and evaluation. Appreciation is also extended to the scholars and annotators for curating and verifying the bilingual datasets, and to the Association for Computational Linguistics (ACL) for offering a platform to disseminate this research.

References

Muhammad Al-Qurishi, Sarah AlQaseemi, and Riad Soussi. 2022. Aralegal-bert: A pretrained language model for arabic legal text. *arXiv preprint arXiv:2210.08284*.

Ahmet Yusuf Alan, Enis Karaarslan, and Ömer Aydin. 2024. A rag-based question answering system proposal for understanding islam: Mufassirqas llm. *CoRR*, abs/2401.15378.

Sarah Alnefaie, Eric Atwell, and Mohammad Ammar Alsalka. 2023. Islamic question answering systems survey and evaluation criteria. *International Journal on Islamic Applications in Computer Science and Technology*, 11(1):9–18.

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, Ara-*

bicNLP 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 (Arabic-NLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Romina Etezadi and Mehrnoush Shamsfard. 2021. A knowledge-based approach for answering complex questions in persian. *CoRR*, abs/2107.02040.

Arash Ghafouri, Hasan Naderi, Mohammad Aghajani Asl, and Mahdi Firouzmandi. 2023. Islamicpcqa: A dataset for persian multi-hop complex question answering in islamic text resources. *CoRR*, abs/2304.11664.

Rana Malhas. 2023. Fine-tuning arabic qa models for qur'an qa task. In *Proceedings of the 2022 Workshop on Open-Source Arabic Corpora and Tools (OSACT 2022)*.

Reham Mohamed, Maha Ragab, Heba Abdelnasser, Nagwa M. El-Makky, and Marwan Torki. 2015. Albayan: A knowledge-based system for arabic answer selection. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 226–230. Association for Computational Linguistics.

Damith Premasiri, Tharindu Ranasinghe, Wajdi Zaghouani, and Ruslan Mitkov. 2022. Dtw at qur'an qa 2022: Utilizing transfer learning with transformers for question answering in a low-resource domain. In *Proceedings of the Qur'an QA 2022 Shared Task*.

Faiza Qamar, Seemab Latif, and Rabia Latif. 2024. A benchmark dataset with larger context for non-factoid question answering over islamic text. *arXiv preprint arXiv:2409.09844*.

Muhammad Razif Rizqullah, Ayu Purwarianti, and Alham Fikri Aji. 2023. Qasina: Religious domain question answering using sirah nabawiyah. In *arXiv preprint*.

Kamyar Zeinalipour, Neda Jamshidi, Fahimeh Akbari, Marco Maggini, Monica Bianchini, and Marco Gori. 2025. Persianmcq-instruct: A comprehensive resource for generating multiple-choice questions in persian. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM)*, pages 344–372. Association for Computational Linguistics.