# Evaluating and Enhancing Large Language Models for Islamic Question Answering

Md. Naimul Haque, Mostahid Hasan Fahim, Md. Minhajul Haque, Md. Fuad Al Amin,
Dr. M. Shahidur Rahman, Dr. Mohammad Reza Selim, Md. Ataullha Saim

Department of Computer Science and Engineering,
Shahjalal University of Science and Technology (SUST), Sylhet, Bangladesh
Email: naimul@example.com, fahim@example.com, selim@example.com

***Abstract*—This paper evaluates the performance of large language models (LLMs) in the context of Islamic question answering, a specialized domain that requires both factual accuracy and cultural sensitivity. We investigate two models, Mistral-7B and LLaMA-3.2-3B, assessing their ability to answer multiple-choice and short-answer questions derived from a curated dataset of 13,000 Islamic questions. Unlike general-purpose evaluations, our work emphasizes the unique challenges posed by religious content, including the necessity to align with authentic sources and avoid interpretive errors. The study applies prompt engineering, targeted fine-tuning and RAG to enhance output quality. Evaluation was conducted using manual and LLM-as-a-judge approach. Results show significant improvements with prompt engineering: LLaMA-3.2-3B improved from 38.78% to 64.19%, while Mistral-7B improved from 41.84% to 70.05%. Fine-tuning further raised Mistral-7B performance to 68.11%. Still, the accuracy is not acceptable in such a sensitive field. Hence, we implemented a RAG pipeline to provide context from authentic religious sources. Our research found Retrieval Augmented Generation(RAG) difficult due to several challenges. As a result RAG slightly improved accuracy, which is far behind our expectation. These findings highlight the situation of domain-specific adaptation of LLMs. It also shows the potential for improvement and further research on such sensitive contexts where trustworthy responses are critical.**

***Index Terms*—Large Language Models, Islamic Question Answering, Prompt Engineering, Fine-Tuning, Domain-Specific Evaluation, Religious NLP, Retrieval Augmented Generation(RAG), LLM-as-Judge**

## I. Introduction

Large Language Models (LLMs) such as Gemini, GPT-4, and other transformer-based systems have revolutionized information access and problem solving across numerous domains. These models demonstrate impressive fluency and reasoning capabilities when applied to general knowledge, yet their reliability diminishes in highly specialized domains. In particular, religious knowledge poses unique challenges due to its reliance on sacred texts, cultural traditions, and centuries of scholarly interpretation.

Islamic studies are of great global significance, with over 2 billion Muslims worldwide seeking guidance from the Qur'an, Hadith, and the scholarly tradition. Misinformation in this domain is not only academically problematic but can also lead to misunderstandings and the misrepresentation of religious values. For example, when asked "Who is called the living martyr of Islam?" a leading LLM incorrectly responded with Hazrat Ammar ibn Yasir (RA) instead of Talha Ibn Ubaydillah (RA). Such errors highlight the need for specialized evaluations.

This study evaluates two popular open-source LLMs, Mistral-7B and LLaMA-3.2-3B, to determine their performance on Islamic question answering. We also observed improvement on some proposed interventions such as prompt engineering, fine-tuning and Retrieval Augmented Generation(RAG).This work adds to both the AI and Islamic studies fields by showing the pros and cons, chances, and things to think about when using advanced NLP systems in religious context.

## II. Related Work

Research on question answering in religious domains has gained momentum in the last decade. Abdelnasser et al. introduced Al-Bayan, a Quran-specific QA system that achieved promising accuracy using retrieval-based techniques. Later initiatives, such as Qur'an QA 2022, created benchmarks for Arabic question answering, attracting a wide range of approaches and highlighting the importance of standardized datasets.

Beyond the Qur'an, Hadith-focused QA systems have also emerged. Abdi et al. proposed semantic similarity-based retrieval methods for Hadith, addressing the linguistic and contextual mismatch between user queries and textual evidence. Similarly, Neamah et al. employed hybrid approaches combining machine learning with traditional similarity measures, demonstrating significant gains in retrieval quality.

The introduction of LLMs has shifted the paradigm toward generation-based answers, but challenges remain. Rizqulah et al. stressed that models like ChatGPT often fail in Islamic QA due to interpretive errors, proposing domain-specific datasets such as QASiNa. AbuBakar et al. identified persistent Anti-Muslim bias in modern SOTA models in their responses. Meanwhile, FatwaSet provides a corpus for broader religious NLP tasks.

However, evaluations of large-scale, general-purpose LLMs on Islamic datasets remain sparse. Our work fills this gap by systematically benchmarking LLMs on a large curated dataset, applying prompt engineering, fine-tuning and RAG.

## III. Methodology

### A. Dataset Construction

TABLE I: Dataset Statistics

| Category | Number of Questions |
|---|---|
| Multiple Choice (MCQs) | 2,000 |
| Short Answer (Q/A) | 11,000 |
| **Total** | **13,000** |

Our dataset of 13,000 QAs was drawn from authentic Islamic sources and educational repositories. Sources encompassed translations of the Qur'an, verified materials on the Five Pillars of Islam, Sindhi Tutorials, and Islamic Studies MCQs for FPSC/PPSC exams. This dataset provides a balance between factual, historical, and ritual-based questions. The structure of two example data is shown in Table II.

TABLE II: Structure of Data

| **Question** | **Solution** | **Topic** |
|---|---|---|
| Who created the heavens and the earth? | Allah | Aqidah |
| Who was the first prophet? | Prophet Adam(AS) | Prophet |
| To which Prophet the Taurath was revealed by Allah? | Ebrahim(AS)<br>**Musa(AS)**<br>Esa(AS)<br>Muhammad(S) | Prophet |

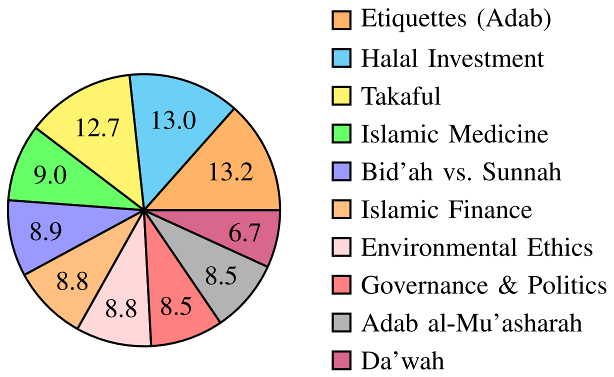

Fig. 1: Distribution of Top 10 Categories in %

### B. Model Selection

We chose Mistral-7B and LLaMA-3.2-3B because of their balance of availability, performance, and reproducibility. Although proprietary models such as GPT-4 outperform them in general contexts, open-source models are better suited for academic reproducibility and fine-tuning experiments.

### C. Prompt Engineering

Prompt engineering involved designing context-aware prompts with explicit constraints. Prompts were designed to include context, explicit constraints, and thematic guidance. For topic guidance we have organized all the topics of our dataset with proper guidance. Extra context from authentic sources are also augmented in the prompt while implementing the RAG pipeline. Prompt engineering proved computationally efficient while delivering significant performance gains. An example is shown below.

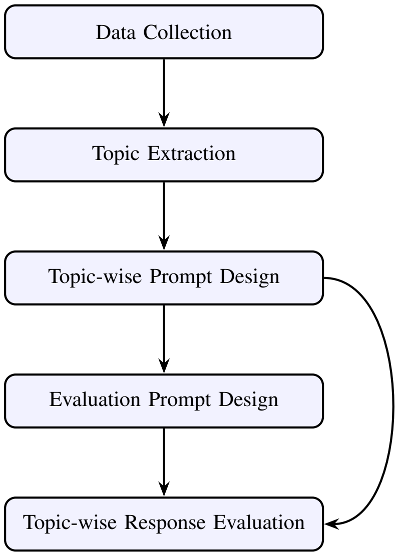

Fig. 2: Prompt engineering process.

Prompt Template

```
You are an expert Islamic scholar.
This question is about {topic}.
Guidance: {topic_guidance}
Question: {question}
Based on the provided sources above
(if any), return a short answer in
strictly not more than 5 words.
```

### D. Fine-Tuning

Fine-tuning was performed using a subset of questions. Training configurations included a batch size of 1, gradient accumulation, fp16 precision, and learning rate $5e{-}5$. We observed convergence after 20 epochs. This adaptation improved alignment but required significant computational resources.

### E. RAG Pipeline Implementation

| Authentic and fundamental sources |
|---|
| Maariful Quran |
| Tafsir al Jalalayn |
| Sahih al Bukhari |
| Sahih al Muslim |
| Al Akidah at Tahawwiyah |
| The Sealed Nacter(Biography of Muhammad[pbuh]) |
| biography of four kaliphs |

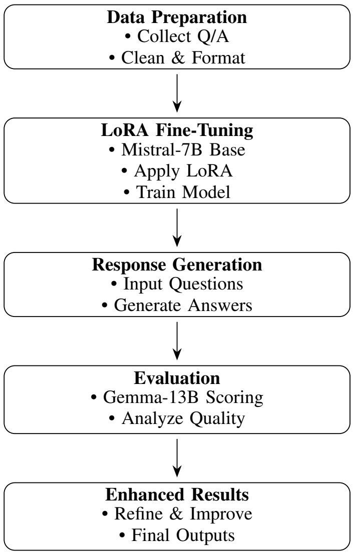

Fig. 3: Mistral-7B fine-tuning workflow

*1) Resource collection:* To provide proper grounding, we collected ten pdf books as authentic sources of islamic knowledge.

*2) Data Preparation and storing:* After loading all the pdfs, we splitted them into semantically meaningful chunks and stored into vector database. We used "all-MiniLM-L6-v2" model for embedding. For efficient retrival we cleaned data and added meaningful metadatas. In metadata, we added keywords and entities for better similarity search.

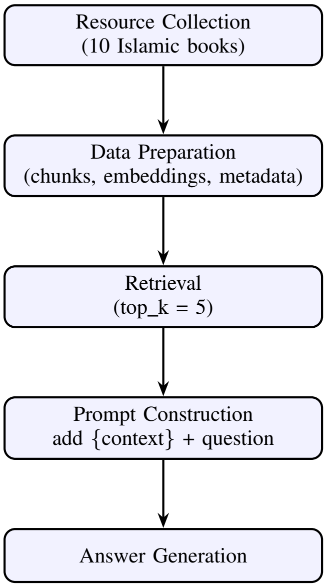

Fig. 4: RAG Pipeline

*3) Retrieval Augmented Generation:* Retrieval system embeds the query and performs similarity search on the db and retrieve related chunks. Then the model generates the answer using the context from retrieval. We used top_k = 5 in the retriever to retrieve 5 chunks with maximum similarity score.

### *F. Evaluation Framework*

In this study, we employed **Manual Evaluation (ME)** as the primary metric. Since LLM-generated answers often do not match the exact text of the reference answers, but may still convey the same meaning, human evaluators assessed responses based on semantic equivalence. A verdict of **1** was assigned for correct answers and **0** for incorrect ones. This binary evaluation ensured that meaning-preserving variations were not unfairly penalized.

TABLE III: Manual Evaluation

| Q.Id | Question | Answer | LLM Response | Verdict |
|---|---|---|---|---|
| 1 | Prophet Muhammad (PBUH) belonged to ____ family. | A. **Hashmi**<br>B. Quraishi<br>C. Makki<br>D. Madni | D. Madni | 0 |
| 2 | Who created the heavens and the earth? | Allah | Quran states "Allah" | 1 |

In addition, we used an **LLM-as-a-judge** approach for a more nuanced analysis. Specifically, the Gemma model was employed to automatically score answers on a five-point scale ranging from **1 (completely incorrect)** to **5 (fully correct and complete)**. Table IV summarizes this scoring system. This dual evaluation strategy allowed us to combine the rigor and sensitivity of human judgment with the scalability of automated LLM-based evaluation

TABLE IV: Evaluation Scale

| Score | Description |
|---|---|
| 1 | Completely incorrect or irrelevant |
| 2 | Mostly incorrect, somewhat relevant |
| 3 | Partially correct with errors |
| 4 | Mostly correct with minor issues |
| 5 | Fully correct and complete |

## IV. Results and Discussion

| Model | Baseline | Prompt Eng. | Fine-Tuned | RAG + PE |
|---|---|---|---|---|
| LLaMA-3.2-3B | 38.78% | 64.19% | – | – |
| Mistral-7B | 41.84% | 70.05% | 68.11% | 72.07% |

Baseline results indicate that both models exhibited moderate performance; however, their accuracy improved substantially with the application of prompt engineering. In particular, Mistral-7B surpassed the 70% threshold after enhancement, underscoring the effectiveness of carefully designed prompts. Furthermore, the incorporation of the RAG pipeline yielded an additional 3–8% gain in accuracy across different evaluations. Given Mistral-7B's superior performance compared to LLaMA-3.2-3B, subsequent experiments focused on fine-tuning and the integration of RAG with prompt engineering (RAG+PE).

Llama-3.2-3b and Mistral-7b. The models were assessed before and after applying prompt engineering and other techniques to improve factual accuracy and consistency.

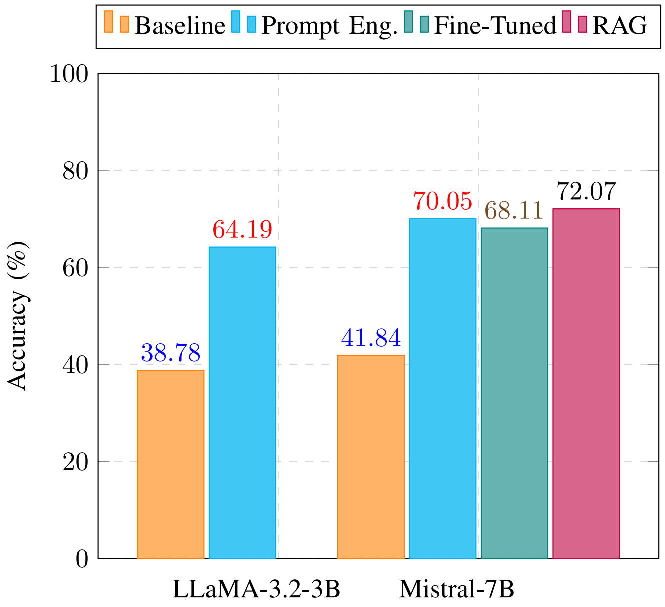

Fig. 5: Accuracy comparison across interventions.

### A. Score-Wise Evaluation Before and After

Table V shows the percentage distribution of verdict scores (1 to 5) for all models before and after prompt engineering, fine-tuning and RAG.

| Model | Score 1 | Score 2 | Score 3 | Score 4 | Score 5 |
|---|---|---|---|---|---|
| LLaMA-3.2-3B (Before) | 12.24% | 2.04% | 24.49% | 28.57% | 10.20% |
| Mistral-7B (Before) | 16.33% | 8.16% | 20.41% | 30.61% | 10.20% |
| LLaMA-3.2-3B (After) | 9.66% | 16.07% | 14.78% | 26.85% | 32.65% |
| Mistral-7B (After) | 5.56% | 10.86% | 15.69% | 15.85% | **46.34%** |
| Mistral-7B (Fine-Tuned) | 4.12% | 10.69% | 15.29% | 27.84% | 37.94% |
| Mistral-7B (RAG + PE) | 6.85% | 12.46% | 15.21% | 16.54% | **48.94%** |

TABLE V: Score Distribution (%) Before and After

### B. Observations on Score 5 (Fully Correct Answers)

After prompt engineering, fine-tuning and RAG, all models showed a significant increase in the proportion of Score 5 responses:

- **LLaMA-3.2-3B**: Increased from 10.20% to 32.65%.
- **Mistral-7B (Prompt-Engineered)**: Increased from 10.20% to 46.34%.
- **Mistral-7B (Fine-Tuned)**: Achieved 37.94%.
- **Mistral-7B (RAG+PE)**: Increased to 48.94%(Maximum).

This demonstrates that prompt engineering, finetuning, and RAG can substantially improve answer correctness and completeness on a specific domain.

### C. Observation

Prompt engineering yielded substantial improvements by encouraging concise and factual answers. For example, LLaMA-3.2-3B initially produced verbose explanations, but with prompts it achieved more accurate, direct answers. Fine-tuning and RAG improved consistency, though the marginal benefit over prompt engineering was lower.

Fine-tuning also improved alignment with religious terminology but was resource-intensive and less flexible than prompt engineering. In practice, combining both approaches could deliver the best results.

Grounding with authentic sources greatly enriches a response. However, our experiment revealed several critical challenges in implementing a proper RAG pipeline. Among them, the lack of a domain-specific embedding system and inefficient similarity search hits are the most significant issues to address. As a result, we observed many retrievals with no matches and most others with very low similarity scores. For example, "What is the purpose of fasting?" or "Who was Dhul-Qarnain (18:83)?" produced zero retrieved documents, while "Tell me about Prophet Ibrahim" had a similarity score of approximately 0.315.

### D. Error Analysis

| Question | Reference | Model Output | Verdict |
|---|---|---|---|
| Who created the heavens? | Allah | Allah created everything in detail. | 4 |
| Where was Adam placed? | Jannah | Paradise | 4 |
| **Who refused to bow to Adam?** | **Iblis** | **Angel** | **1** |

Common errors included misinterpretation of nuanced questions, verbosity beyond required short answers, and occasional hallucination in historical questions. Synonym mismatches sometimes resulted in penalties (e.g., "Jannah" vs. "Paradise" being scored lower despite equivalence). Table above shows some examples. We also visualize error proportions in Fig. 6.

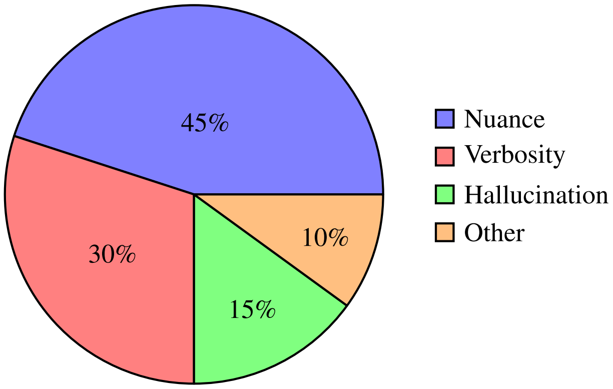

Fig. 6: Approximate distribution of error types (%).

### E. Ablation: Prompt Components

We ablate prompt components (format constraint, topic hint, evidence reminder). Results (Table VI) indicate format constraints drive the largest single gain.

TABLE VI: Prompt Component Ablation on Mistral-7B (% Accuracy)

| Variant | MCQ | Short Ans. |
|---|---|---|
| No prompt control | 48.1 | 39.6 |
| + Format constraint | 60.3 | 55.2 |
| + Topic hint | 63.0 | 57.9 |
| + Evidence reminder | 64.2 | 58.6 |
| Full (all) | **66.8** | **60.4** |

### F. Limitations

Some errors persisted, particularly in questions involving nuanced interpretations. The automatic evaluator occasionally penalized synonyms or paraphrases. Human-in-the-loop evaluation may address this issue. The RAG pipeline has a significant scope for improvement. Using more efficient chunking and better domain-specific embedding can improve accuracy greatly. Additionally, the dataset was limited to English translations, and performance across Arabic or other languages remains unexplored.

### G. Ethical and Religious Considerations

Even LLMs can efficiently generate responses, it is important to recognize the limitations of language models in issuing religious rulings. LLMs should not attempt to generate new fatwas, as they require the authority of qualified scholars. Rather, the system should obtain relevant resources from reliable sources and offer citations to books, academic publications, or validated internet sources. This method respects the authority and sensitivity of Islamic jurisprudence while guaranteeing that users receive accurate guidance.

## V. Conclusion and Future Work

This study demonstrates that LLMs, though not originally trained on religious texts, can be adapted to perform effectively in Islamic QA tasks. Prompt engineering significantly boosted accuracy without requiring costly retraining, while fine-tuning offered additional gains. Additionally, RAG has great opportunity to improve accuracy, but it needs to be pipelined carefully in a sophisticated manner. Mistral-7B, in particular, showed strong improvements. However, limitations remain, particularly in synonym handling and ignoring sensitivity.

Future directions include evaluating larger models, expanding datasets to include Arabic and other languages, human-in-the-loop evaluation for nuanced interpretations, exploring RLHF for theological accuracy, and deploying prototypes into educational apps and chatbots.

## Acknowledgment

We thank our supervisors, Dr. Mohammad Reza Selim and Dr. M. Shahidur Rahman and also the Department of CSE, SUST, for their continuous support. We also thank fellow students and colleagues for assistance in dataset compilation and validation.

## References

[1] H. Abdelnasser et al., "Al-Bayan: An Arabic Question Answering System for the Holy Quran," in *Proceedings of EMNLP*, 2014.

[2] A. Malhas et al., "Qur'an QA 2022: Shared Task on Question Answering over the Holy Quran," in *Proceedings of WANLP*, 2022.

[3] A. Abdi et al., "A Question Answering System for Hadith Texts," *Information Processing & Management*, 2019.

[4] M. Alqahtani and E. Atwell, "AQQAC: Arabic Quranic Question and Answer Corpus," *LREC*, 2017.

[5] S. Mohammed et al., "FatwaSet: A Dataset for Arabic Fatwas," in *Proceedings of LREC*, 2021.

[6] A. Rizqulah et al., "QASiNa: Islamic Question Answering Dataset on Sirah Nabawiyah," in *Proceedings of ICCSCI*, 2023.

[7] A. Abid, M. Farooqi, and J. Zou, "Persistent Anti-Muslim Bias in Large Language Models," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*, Virtual Event, USA, 2021, pp. 298–306.

[8] M. Laupichler, J. Rother, I. Grunwald Kadow, S. Ahmadi, and T. Raupach, "Large language models in medical education: Comparing ChatGPT- to human-generated exam questions," *Academic Medicine*, Dec. 2023.

[9] Y. Wu, A. Henriksson, M. Duneld, and J. Nouri, "Towards improving the reliability and transparency of ChatGPT for educational question answering," in *Responsive and Sustainable Educational Futures*, O. Viberg, I. Jivet, P. Muñoz-Merino, M. Perifanou, and T. Papathoma, Eds. Cham: Springer Nature Switzerland, 2023, pp. 475–488.

[10] R. Goyal, P. Kumar, and V. Singh, "Automated question and answer generation from texts using text-to-text transformers," *Arabian J. Sci. Eng.*, May 2023.

[11] K. S. Phogat, C. Harsha, S. Dasaratha, S. Ramakrishna, and S. A. Puranam, "Zero-shot question answering over financial documents using large language models," 2023. [Online]. Available: https://arxiv.org/abs/2311.14722

[12] D. P. Panagoulias, M. Virvou, and G. A. Tsihrintzis, "Evaluating LLM-generated multimodal diagnosis from medical images and symptom analysis," 2024. [Online]. Available: https://arxiv.org/abs/2402.01730

[13] W. Zhu, H. Liu, Q. Dong, J. Xu, S. Huang, L. Kong, J. Chen, and L. Li, "Multilingual machine translation with large language models: Empirical results and analysis," 2024. [Online]. Available: https://arxiv.org/abs/2304.04675

[14] K. Lakkaraju, S. K. R. Vuruma, V. Pallagani, B. Muppasani, and B. Srivastava, "Can LLMs be good financial advisors?: An initial study in personal decision making for optimized outcomes," 2023. [Online]. Available: https://arxiv.org/abs/2307.07422

[15] T. Globalist, "Muslims: A global perspective," Jun. 2024. [Online]. Available: https://www.theglobalist.com/muslims-islam-religion-arabic-population/

[16] J. D'Souza, H. B. Giglou, and Q. Münch, "Yescieval: Robust LLM-as-a-judge for scientific question answering," 2025. [Online]. Available: https://arxiv.org/abs/2505.14279

[17] X. Ho, J. Huang, F. Boudin, and A. Aizawa, "LLM-as-a-judge: Reassessing the performance of LLMs in extractive QA," Apr. 2025.

[18] M. AI, "LLaMA 3.2 3B instruct," 2024, version Release Date: Sep. 25, 2024. [Online]. Available: https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

[19] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7B," 2023. [Online]. Available: https://arxiv.org/abs/2310.06825