Benchmarking Generative AI on Quranic Knowledge

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper evaluates the performance of large language models (LLMs) and embedding-based retrieval systems in answering Quranic questions, a task demanding both semantic understanding and theological grounding. The Quran's complex rhetorical structure, contextual depth, and inter-verse coherence pose challenges for general-purpose models. To address this, we introduce a human-reviewed benchmark of 881 multiple-choice questions derived from 200 Quranic verses, stratified by five cognitive reasoning levels (using Bloom's Taxonomy) and four familiarity tiers based on verse perplexity. We assess model performance on two tasks: (1) multiple-choice QA (semantic comprehension), and (2) verse identification (reference grounding). Results show that instruction-tuned LLMs such as Fanar-1-9B achieve 41% accuracy on MCQs and 15.6% top-1 verse identification accuracy, with a marked decline from low-complexity ("Remember") to high-complexity ("Evaluate") questions. Conversely, a dense retriever achieves 45.1% top-5 accuracy and an MRR of 0.341, with particularly strong performance on familiar and low-level questions (e.g., 73% on "Remember", 57% on low-perplexity verses).

1 Introduction

The Quran serves as the foundational religious text for over 1.9 billion Muslims worldwide [12], offering spiritual, moral, and legal guidance. With its complex linguistic structure, comprising metaphor, allegory, and nuanced rhetorical forms [3], Quranic interpretation demands deep contextual and theological understanding. These qualities pose significant challenges to automated systems attempting semantic retrieval or question answering (QA).

Traditional keyword-based retrieval systems often return irrelevant or superficial results when applied to Quranic content [15]. Early QA systems, such as Al-Bayan [1], employed rule-based methods combined with shallow learning techniques but lacked robustness for semantically complex queries. With the rise of neural models, more sophisticated approaches have emerged, leveraging resources like the Quranic Arabic Corpus and datasets like AyaTEC [5] and QRCD [14].

Recent advances in large language models (LLMs) and Retrieval-Augmented Generation (RAG) have renewed interest in Quranic QA [2]. However, systems like QuranGPT[1] raise concerns regarding theological accuracy and lack of source attribution. Existing datasets such as AyaTEC [5] offer limited evaluative depth, they lack cognitive-level labeling, enforce single-verse answers even when multiple are valid, and fail to account for model familiarity with specific topics. These limitations hinder robust assessment of reasoning ability, generalization, and contextual alignment.

This paper proposes a two-part benchmark and evaluation framework that systematically assesses (i) semantic understanding via multiple-choice questions and (ii) contextual verse identification. Ques-

[1]https://www.qurangpt.com/

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

tions are categorized along two dimensions: cognitive reasoning (using Bloom's Taxonomy) and linguistic difficulty (using verse perplexity).

Contributions. Our key contributions are: **(i) Quranic QA Benchmark:** A curated dataset of 881 manually reviewed questions mapped to 200 Quranic verses stratified by familiarity of the verse and the cognitive demand. **(ii) Dual Evaluation Tasks:** A comprehensive framework assessing both semantic comprehension (MCQ task) and contextual reference identification. **(iii) Model Analysis:** Empirical evaluation of four LLMs and a dense retriever model across tasks, showing the superior performance of embedding-based retrieval systems in finding the context.

2 Related Work

Quranic QA has seen significant evolution, from rule-based approaches to ontology-based systems and neural architectures. Al-Bayan [1] pioneered Quranic QA using support vector machines and handcrafted rules. Later systems incorporated semantic ontologies and syntactic parsers, improving contextual relevance [15].

Recent approaches leverage LLMs and RAG pipelines. MufassirQAS [2] integrates RAG with theological grounding using domain-specific knowledge. However, tools like QuranGPT lack transparency in verse sourcing, raising reliability concerns. Despite these advances, no prior benchmark combines cognitive taxonomy with linguistic complexity to evaluate Quranic QA systems systematically.

3 Benchmark Construction

3.1 Corpus and Preprocessing

We use the full Quranic text from Tanzil[2], including Arabic, English translation, and tafsir from *Al-Mukhtasar* [13]. The corpus is organized verse-wise and preprocessed into single-verse and three-verse contextual chunks.

3.2 Perplexity-Based Verse Stratification

To quantify linguistic complexity, we calculate verse-level perplexity using a reference LLM. Verses are divided into four bins: Low (most familiar verses), Medium, High, and Very High. We sample 200 verses (50 per bin) for benchmark construction.

3.3 MCQ Generation via Bloom's Taxonomy

Using Bloom's Taxonomy [10], we generate five questions per verse, targeting cognitive levels: Remember, Understand, Apply, Analyze, and Evaluate. Each MCQ consists of one correct verse and three distractors. After manual curation, we retain 881 high-quality questions.

3.4 Similar-Verses Dataset

Each benchmark verse is mapped to a set of semantically related verses using the Ayat Similarity tool [11]. These pairs support soft-matching for reference detection evaluation, which computes vector similarity based on n-gram features of root, lemma, and surface forms.

A sample is shown in Figure 1 illustrates the structure and cognitive complexity of questions in our benchmark. Each question is grounded in a specific Quranic verse and aligned with one of Bloom's cognitive levels. In this case, the item targets the "Analyze" level and requires semantic discrimination between thematically similar verses. The respondent must break down conceptual elements (e.g., disbelief, signs, interpretation) across multiple options to determine which verse best exhibits the misreading of observable signs due to internal psychological bias.

[2] `https://tanzil.net/download/`

Sample Question

Question ID: 19

Query: What ayah serves as the strongest evidence of how disbelief leads to misinterpretation of divine signs?

Multiple Choice Options:

(A) [52:44] – وَإِن يَرَوْا كِسْفًا مِّنَ السَّمَاءِ سَاقِطًا يَقُولُوا سَحَابٌ مَّرْكُومٌ

(B) [36:14] – إِذْ أَرْسَلْنَا إِلَيْهِمُ اثْنَيْنِ فَكَذَّبُوهُمَا فَعَزَّزْنَا بِثَالِثٍ

(C) [34:27] – قُلْ أَرُونِيَ الَّذِينَ أَلْحَقْتُم بِهِ شُرَكَاءَ كَلَّا

(D) [40:33] – يَوْمَ تُوَلُّونَ مُدْبِرِينَ مَا لَكُم مِّنَ اللَّهِ مِنْ عَاصِمٍ

Correct Answer: A (Option 1)

Bloom's Taxonomy Level: Analyze

Verse Reference: 52:44

Perplexity Classification: High (8.65)

This question exemplifies the "Analyze" level of Bloom's taxonomy by requiring students to distinguish between verses that are thematically close but semantically distinct. The respondent must analyze each verse's structure and intent to determine which most accurately reflects the misreading of divine signs due to disbelief.

Similar Verses:

- **Surah Ash-Shu'arā' [26:187]:**
  فَأَسْقِطْ عَلَيْنَا كِسَفًا مِّنَ السَّمَاءِ إِن كُنتَ مِنَ الصَّادِقِينَ
- **Surah Saba' [34:9]:**
  أَفَلَمْ يَرَوْا إِلَىٰ مَا بَيْنَ أَيْدِيهِمْ وَمَا خَلْفَهُم مِّنَ السَّمَاءِ وَالْأَرْضِ ۚ إِن نَّشَأْ نَخْسِفْ بِهِمُ الْأَرْضَ
  أَوْ نُسْقِطْ عَلَيْهِمْ كِسَفًا مِّنَ السَّمَاءِ ۚ إِنَّ فِي ذَٰلِكَ لَآيَةً لِّكُلِّ عَبْدٍ مُّنِيبٍ

Figure 1: example multiple-choice question from the benchmark. the correct answer is grounded in a specific verse, while distractors are selected from semantically related but contextually distinct verses.

4 Evaluation Tasks

4.1 Task 1: Multiple-Choice QA

Given a question, models must choose the correct verse among four options. This tests semantic comprehension and reasoning. All are evaluated in zero-shot mode using `lm-evaluation-harness` [6].

Table 1: (a) MCQ QA accuracy (%) by Bloom's level and verse perplexity. (b) Verse Identification accuracy (%) by Bloom's level and perplexity. Highest values per row are bolded.

Task	Model	Remember	Understand	Apply	Analyze	Evaluate	Avg.	Low PPL	Med. PPL	High PPL	V.High PPL
Task (a) MCQ											
DeepSeek		43.0	36.6	37.6	37.7	**36.7**	38.3	7.0	21.6	**46.9**	**76.5**
Gemma		39.8	36.1	29.7	34.7	32.2	34.5	24.3	21.6	36.5	60.0
Qwen		40.3	35.0	31.5	32.9	34.5	34.8	18.5	24.2	40.3	61.5
Fanar		**45.7**	**44.8**	**35.8**	**42.9**	35.6	**41.0**	**64.6**	**30.0**	38.9	27.5
Task (b) Verse ID											
DeepSeek		14.0	11.0	12.1	11.8	9.0	11.6	20.6	8.4	2.4	14.0
Gemma		7.5	7.1	7.3	8.2	6.8	7.4	14.0	3.5	0.0	11.5
Qwen		15.6	12.0	12.1	12.2	18.6	14.1	25.5	9.7	3.3	16.0
Fanar		40.3	37.7	**37.0**	36.5	**36.7**	37.6	53.1	27.8	37.4	30.5
Dense Retriever		**73.1**	**53.6**	30.9	**37.1**	27.7	**44.5**	**56.8**	**45.8**	**39.3**	**36.0**

4.2 Task 2: Verse Identification

Given a question, models return the most relevant verse(s) without options. Answers are evaluated against the Similar-Verses set using MRR and top-k accuracy.

5 Model and Retriever Setup

5.1 Evaluated Language Models for Task 1,2

We evaluate four instruction-tuned LLMs: (i) DeepSeek-R1-Distill-3B [7], (ii) Gemma-3-4B-IT [9], (iii) Qwen2.5-1.5B-Instruct [4], and (iv) Fanar-1-9B-Instruct [8] on both Tasks.

5.2 Embedding-Based Retriever for Task 2

These verse-level and contextual chunks were used to construct a searchable knowledge base, supporting both dense and sparse retrieval. For dense retrieval, chunks were encoded using the transformer-based model `BAAI/bge-m3` and indexed with FAISS to enable efficient cosine similarity search. In parallel, a sparse retrieval index was built using BM25 for keyword-based matching. Given a query, top-N results were retrieved from both indices. The final ranking was computed using Reciprocal Rank Fusion (RRF), which promotes candidates appearing near the top of either list. Retrieved chunks were then mapped back to their original verse references. We evaluate top-5 accuracy using the Similar-Verses gold set.

6 Results

We evaluated system performance on two tasks: (a) multiple-choice QA (semantic comprehension) and (b) verse identification (reference grounding). Results are reported across Bloom's cognitive levels and verse perplexity tiers to examine model behavior under varying reasoning depth and textual familiarity.

6.1 Task (a): Multiple-Choice QA

Table 1 (top half) summarizes MCQ accuracy across Bloom's levels and perplexity bands. Fanar-1-9B leads overall, achieving the highest accuracy in "Remember" (45.7%) and "Understand" (44.8%) categories. It also dominates on low-perplexity verses (64.6%), indicating strong performance when the language and semantics are familiar. However, Fanar's performance drops sharply as perplexity increases, falling to 27.5% on very high-perplexity questions.

Conversely, DeepSeek-R1-Distill-3B exhibits the opposite trend. Despite weaker performance on simpler content, it performs best on high (46.9%) and very high-perplexity (76.5%) verses. This suggests that DeepSeek's learned representations may generalize better to unfamiliar or structurally complex inputs. Qwen and Gemma show moderate performance across all axes but fail to outperform Fanar or DeepSeek in any major category.

Table 2: Mean Reciprocal Rank (MRR) and total number of correct verse identifications for each model on the verse identification task. MRR values are also broken down by Bloom's cognitive levels to assess reasoning depth. The dense retriever achieves the highest MRR and coverage across all categories, highlighting its advantage in grounding responses in relevant Quranic references.

Model	MRR	# Correct	Remember	Understand	Apply	Analyze	Evaluate
DeepSeek	0.083	102	0.089	0.076	0.086	0.087	0.075
Qwen	0.091	123	0.111	0.079	0.081	0.074	0.108
Gemma	0.039	65	0.043	0.038	0.036	0.044	0.034
Fanar	0.156	332	0.163	0.152	0.155	0.158	0.151
Dense Retriever	**0.341**	**397**	**0.563**	**0.397**	**0.252**	**0.272**	**0.192**

6.2 Task (b): Verse Identification (Top-1 Accuracy)

Table 1 (bottom half) presents top-1 verse identification accuracy across the same evaluation axes. The dense retriever outperforms all LLMs, achieving 73.1% on "Remember" and maintaining the lead across all Bloom levels and perplexity bands. Its top-1 accuracy drops to 27.7% for "Evaluate" and 36.0% on very high-perplexity items, yet remains consistently superior. More detailed results are provided in Section A.

Among the LLMs, Fanar-1-9B again performs best overall, with 40.3% on "Remember" and 36.7% on "Evaluate". It also maintains relatively balanced performance across perplexity tiers. Qwen shows stronger retrieval alignment than DeepSeek in most bands, while Gemma performs weakest overall, struggling particularly with complex or less familiar questions.

6.3 Task (b): Verse Identification (MRR)

Table 2 presents Mean Reciprocal Rank (MRR) for verse identification. The dense retriever again leads with an MRR of 0.341 and the highest correct retrieval count (397/881). Notably, it achieves an MRR of 0.563 on "Remember" and 0.397 on "Understand", indicating effective semantic retrieval even without generative reasoning.

Fanar ranks second (MRR = 0.156), more than doubling the next best LLM (Qwen at 0.091). Fanar consistently scores highest across all Bloom levels among LLMs. DeepSeek and Qwen trail closely behind, with Qwen showing slightly better MRR on "Evaluate". Gemma continues to underperform, with the lowest MRR (0.039) and weakest cognitive generalization.

Overall, Fanar demonstrates superior performance across both tasks, particularly in low-to-medium cognitive complexity and familiar language contexts. However, its decline on high-perplexity questions reveals limitations in generalization. DeepSeek's unusual strength on high-perplexity items suggests robustness to linguistic unfamiliarity, though it lags in accuracy on simpler tasks.

The dense retriever remains the most reliable system for factual and context-grounded verse identification, outperforming all LLMs in both top-1 accuracy and MRR. While LLMs like Fanar show promising results with instruction tuning, their limitations in grounding and semantic alignment underline the importance of retrieval augmentation–especially for theologically precise domains such as Quranic QA.

7 Conclusion and Future Work

This study introduced a new evaluation benchmark for Quranic question answering, designed to probe semantic understanding and contextual fidelity across both cognitive reasoning levels and verse familiarity. Through a two-task framework-multiple-choice QA and verse identification– we benchmarked four LLMs and a dense retriever, revealing key insights into the limitations and strengths of current models.

Overall, instruction-tuned LLMs demonstrated limited reasoning depth, with accuracy declining significantly as cognitive complexity increased. For example, Fanar-1-9B, the strongest LLM, achieved 45.7% on "Remember" questions but only 35.6% on "Evaluate", a pattern closely aligned with human learning curves. The same trend held across verse perplexity: performance dropped consistently on less familiar verses, suggesting that familiarity (as measured by model perplexity) is a meaningful proxy for difficulty.

In contrast, the dense retriever outperformed all LLMs on the verse identification task, achieving 73.1% accuracy on low-level queries and 0.341 MRR overall. Furthermore, we found that successful verse identification strongly correlates with MCQ correctness, reinforcing the importance of contextual retrieval in faith-sensitive QA.

Future work will expand this benchmark to include open-ended generative answers, multi-hop reasoning, and paraphrased questions to evaluate semantic robustness. Fine-tuning dense retrievers on tafsir-rich corpora and exploring hybrid RAG architectures remain promising directions. Ultimately, this line of research aims to develop Quranic QA systems that are not only linguistically competent but also theologically sound, context-aware, and cognitively aligned with how humans reason over sacred texts.

References

[1] Heba Abdelnasser, Maha Ragab, Reham Mohamed, Alaa Mohamed, Bassant Farouk, Nagwa M. El-Makky, and Marwan Torki. Al-bayan: An arabic question answering system for the holy quran. In *ANLP@EMNLP*, 2014.

[2] Ahmet Yusuf Alan, Enis Karaarslan, and Omer Aydin. A rag-based question answering system proposal for understanding islam: Mufassirqas llm. 2024.

[3] Hessa Abdulrahman Alawwad, Lujain Alawwad, Jamilah Alharbi, and Abdullah Alharbi. Ahjl at quran qa 2023 shared task: Enhancing passage retrieval using sentence transformer and translation. In *ARABICNLP*, 2023.

[4] Alibaba. Qwen2.5-1.5b-instruct. `https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct`, 2024. Accessed: 2025-07-06.

[5] Muhammad Huzaifa Bashir, Aqil M. Azmi, Haq Nawaz, Wajdi Zaghouani, Mona T. Diab, Ala I. Al-Fuqaha, and Junaid Qadir. Arabic natural language processing for quranic research: a systematic review. *Artificial Intelligence Review*, 56:6801–6854, 2021.

[6] Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Jason Phang, and ... lmevaluationharness: An open source library for independent, reproducible, and extensible evaluation of language models. Preprint on arXiv and open-source repository, 2024. `https://github.com/EleutherAI/lm-evaluation-harness`.

[7] DeepSeek. Deepseek-r1-distill-llama-3b. `https://huggingface.co/suayptalha/DeepSeek-R1-Distill-Llama-3B`, 2024. Accessed: 2025-07-06.

[8] A Fanar-Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, et al. Fanar: An arabic-centric multimodal generative ai platform. *arXiv preprint arXiv:2501.13944*, 2025.

[9] Google. Gemma-3-4b-it. `https://huggingface.co/google/gemma-3-4b-it`, 2024. Accessed: 2025-07-06.

[10] David R Krathwohl. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218, 2002.

[11] Language ML. Ayat similarity tool for quranic verse comparison. `https://ayat.language.ml/`, 2024. Accessed: 2025-07-06.

[12] Pew Research Center. Islam was the worlds fastest-growing religion from 2010 to 2020, 2025.

[13] Qul. Al-mukhtasar tafsir - qul tarteel project. `https://www.qul.org.au/tafsir`, 2024. Accessed: 2025-07-06.

[14] Muhammad Razif Rizqullah, Ayu Purwarianti, and Alham Fikri Aji. Qasina: Religious domain question answering using sirah nabawiyah. *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, pages 1–6, 2023.

[15] Fandy Setyo Utomo, N Suryana, and Mohd Sanuzi Azmi. Question answering systems on holy quran: A review of existing frameworks, approaches, algorithms and research issues. *Journal of Physics: Conference Series*, 1501, 2020.

A Verse Identification Results

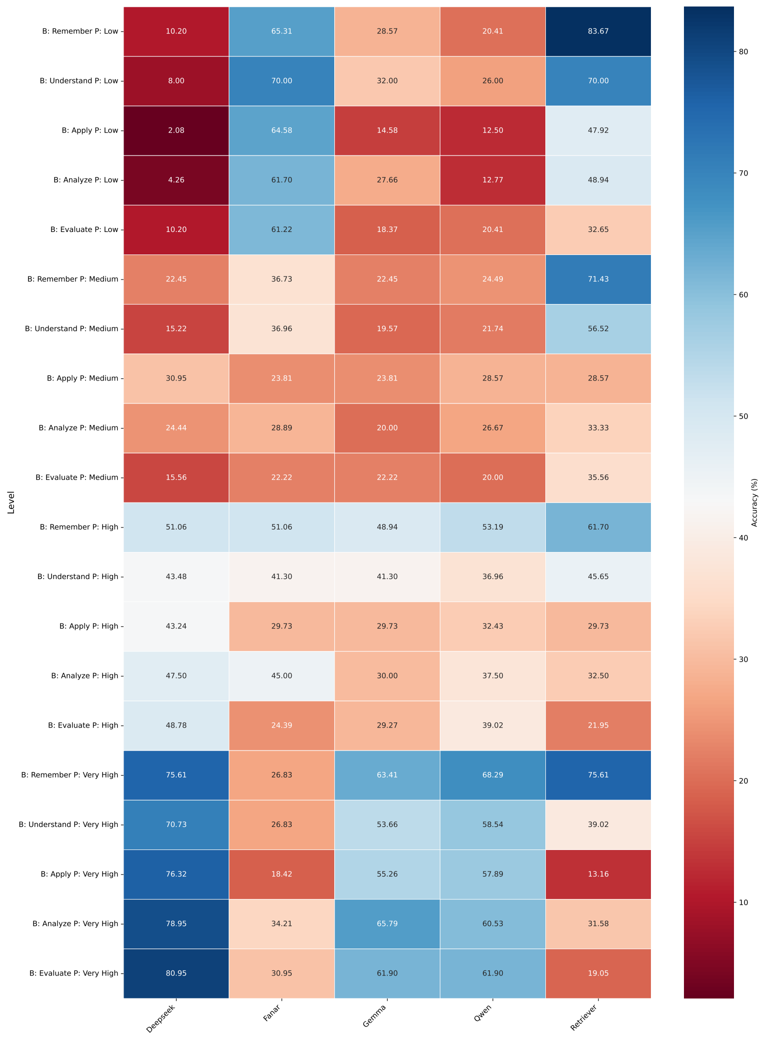

Figure 2: Heatmap showing model performance across Bloom's taxonomy levels (B: Remember, Understand, Apply, Analyze, Evaluate) and perplexity strata (P: Low, Medium, High, Very High). Each cell represents verse-level accuracy (for a given model. The five models evaluated are *Fanar*, *Deepseek*, *Gemma*, *Qwen*, and *Retriever*.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification:

   In abstract and introduction, we summarize the problem statement and the results of the experiments conducted. Details on the results of the benchmark can be found in (section 6). Evaluation tasks are described in (section 4).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

   Justification:

   Although not explicitly mentioned, several limitations are implied by the way the study is designed. Since the benchmark only uses 881 questions from 200 verses, it clearly doesnt cover the Quran comprehensively. The evaluation is also limited to two tasks MCQ answering and verse identification implying that more realistic settings like open-ended or multi-hop QA remain untested.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification:

   The work is primarily empirical and benchmark-driven, focusing on constructing a Quranic QA dataset, stratifying questions by Blooms Taxonomy and perplexity, and evaluating LLMs and retrievers on multiple-choice QA and verse identification tasks

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification:

   All details needed to reproduce the results are explained in the paper. The construction of benchmark, including datasets, are mentioned in (section 3), while details about the retriever are mentioned in (section 5.2).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification:

(`https://github.com/rawan-rm2208263/quranic_rag_repo`)

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification:

Our paper mentions information about the models used and the size of each, as well as the retriever setup in (section 5)

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification:

   The results are presented as accuracy scores, top-k retrieval performance, and mean reciprocal rank (MRR) across Blooms levels and perplexity bands

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification:

   Our paper describes the models evaluated (DeepSeek, Gemma, Qwen, Fanar) and the retriever setup (BAAI/bge-m3 with FAISS, BM25), but it does not provide details on the compute resources used.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

The research did not cause directly or indirectly harm to any party.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification:

We considered the holiness of the Quranic text, so we included only four Bloom's Taxonomy levels, excluding the Generate level (section 3.3). Our research is focused on pure Quranic text, which a huge population is concerned with, as mentioned in (section 1).

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification:

There is no risk associated with this paper. It introduces a comprehensive QA dataset and benchmarks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification:

    Our paper explicitly cites all external resources: Quranic text from Tanzil (CC-BY 3.0), tafsir from Qul Tarteel Project, and the Ayat Similarity tool. The evaluated models (DeepSeek, Gemma, Qwen, Fanar) are linked to their Hugging Face repositories, which provide license information, and the lm-evaluation-harness toolkit is credited under its MIT License. No proprietary datasets or restricted tools are used, and all assets are employed in line with their stated terms of use.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification:

    New assets are documented in the Contributions section.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification:

No human subjects were included in the experiments.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification:

    No human subjects were included in the experiments.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification:

    The study benchmarks four instruction-tuned LLMs (DeepSeek-R1-Distill-3B, Gemma-3-4B-IT, Qwen2.5-1.5B-Instruct, and Fanar-1-9B-Instruct) on two tasks: multiple-choice QA and verse identification. Their roles, evaluation setup, and comparative performance are documented in detail, with results analyzed across Blooms Taxonomy levels and verse perplexity bands. Since LLM evaluation is central to the contribution, their usage is explicitly declared and described.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.