Can LLMs Directly Retrieve Passages for Answering Questions from Qur'an?

Sohaila Eltanbouly, Salam Albatarni, Shaimaa Hassanein, Tamer Elsayed
Computer Science and Engineering Department, Qatar University, Doha, Qatar
{se1403101, sa1800633, sh2300494, telsayed}@qu.edu.qa

Abstract

The Holy Qur'an provides timeless guidance, addressing modern challenges and offering answers to many important questions. The Qur'an QA 2023 shared task introduced the Qur'anic Passage Retrieval (QPR) task, which involves retrieving relevant passages in response to questions written in modern standard Arabic (MSA). In this work, we evaluate the ability of seven large language models (LLMs) to retrieve relevant passages from the Qur'an in response to given questions, considering zero-shot and several few-shot scenarios. Our experiments show that the best model, Claude, significantly outperforms the state-of-the-art QPR model by 28 points on MAP and 38 points on MRR, exhibiting an impressive improvement of about 113% and 82%, respectively.

1 Introduction

The Holy Qur'an holds an immense spiritual, legal, and ethical significance for over a billion Muslims worldwide. Islamic scholars frequently engage with its verses to address theological, ethical, and societal questions. However, its unique structure, linguistic depth, and rhetorical style make it a challenging source for precise information retrieval.

Qur'an QA 2023 shared task (Malhas et al., 2023) directly addresses this need, introducing the *Qur'anic Passage Retrieval* (QPR) task, which is the focus in this work. QPR is defined as follows:

> *Given a question written in modern standard Arabic (MSA), retrieve up to 10 Qur'anic passages, where a Qur'anic passage is a consecutive sequence of verses from a specific Qur'anic chapter.*

A question can potentially have multiple answers or possibly no answer in the Qur'an. Figure 1 shows an example of this task, where an MSA question is given, and the answer is a Qur'anic passage.

أين كانت رحلة الاسراء والمعراج؟
[١٧:١-١] سبحان الذي أسرى بعبده ليلا من المسجد الحرام إلى المسجد الأقصى الذي باركنا حوله لنريه من آياتنا إنه هو السميع البصير.

Where was the journey of Al-Isra and Al-Miraj?
[1-1:17] Exalted is He who took His Servant by night from al-Masjid al-Haram to al-Masjid al-Aqsa, whose surroundings We have blessed, to show him of Our signs. Indeed, He is the Hearing, the Seeing.

Figure 1: Example of QPR question and a relevant passage from Qur'an, with translations.

The task has proven challenging, as evidenced by the low performance scores of the best participating teams in the shared task; for instance, the top team achieved a MAP score of 0.251 and an MRR score of 0.461, indicating substantial room for improvement. The emergence of Large Language Models (LLMs) offers a promising opportunity to support Islamic scholars in navigating this sacred text. With advanced natural language understanding, LLMs can potentially identify relevant Qur'anic passages in response to MSA questions.

This work explores using LLMs for QPR, assessing their ability to identify relevant Qur'anic verses. Specifically, we address the following research questions:

- **RQ1:** What is the effect of prompt engineering on the performance of LLMs for QPR?
- **RQ2:** How effective are LLMs for QPR compared to the current state-of-the-art (SOTA) models?

Our main contribution in this work is three-fold:

1. We evaluate several pre-trained LLMs for the QPR task using different prompting techniques.
2. Our approach significantly outperforms SOTA performance.

Proceedings of The Third Arabic Natural Language Processing Conference, pages 203–210

3. We provide a failure analysis of LLMs' response in the QPR task.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 details the prompting techniques we used with the LLMs. Section 4 outlines our experimental setup. Section 5 presents and discusses our experimental results. Section 6 concludes our study. Finally, Section 7 lays out some considerable limitations and ethical issues related to our work.

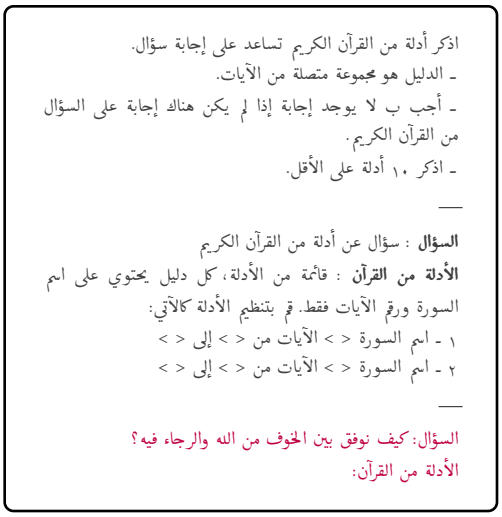
2 Related Work

Automatic Question Answering (QA) systems have been instrumental in aiding information retrieval and interpretation across domains, including Arabic and Qur'anic texts (Malhas and Elsayed, 2020, 2022). Early Arabic QA research introduced systems like QARAB (Hammo et al., 2002) and explored neural networks and transformers to enhance open-domain factoid QA (Mozannar et al., 2019). For Qur'anic texts, Basem et al. (2024) expanded the dataset originally provided by the Qur'an QA 2023 shared task and significantly enhanced MAP and MRR results by fine-tuning Arabic models like AraBERT and AraELECTRA. While other approaches, including translation-based retrieval and embedding-based techniques (Alawwad et al., 2023), have improved performance, they often overlook the potential of LLMs for direct QA.

Recent studies have demonstrated the efficacy of LLMs in tackling complex retrieval tasks, particularly for QPR. Techniques such as transfer learning (Mahmoudi et al., 2023), retrieval-augmented generation (Alan et al., 2024), and semantic search using LLM embeddings (Alqarni, 2023) have shown significant promise. Yet, challenges persist in handling classical Arabic due to its linguistic nuances (Alnefaie et al., 2023). Building on these advancements, this work evaluates the ability of LLMs to address the QPR task, aiming to assess their performance against SOTA models.

3 Prompting techniques

While our method is quite straightforward, simply prompting the LLM to answer the input question, the prompt design has multiple intricacies that make it more suitable for this task. We use three types of prompting strategies: *Zero-shot*, *Chain-of-Thought*, and *In-context Learning* (with random or semantically similar few-shot examples).


اذكر أدلة من القرآن الكريم تساعد على إجابة سؤال.
ـ الدليل هو مجموعة متصلة من الآيات.
ـ أجب ب لا يوجد إجابة إذا لم يكن هناك إجابة على السؤال من القرآن الكريم.
ـ اذكر ١٠ أدلة على الأقل.

—

السؤال : سؤال عن أدلة من القرآن الكريم
الأدلة من القرآن : قائمة من الأدلة، كل دليل يحتوي على اسم السورة ورقم الآيات فقط. قم بتنظيم الأدلة كالآتي:
١ ـ اسم السورة < > الآيات من < > إلى < >
٢ ـ اسم السورة < > الآيات من < > إلى < >

—

السؤال: كيف نوفق بين الخوف من الله والرجاء فيه؟
الأدلة من القرآن:

Figure 2: An example of a zero-shot prompt, including the instructions and the input.

It is crucial to note that, given the sacred nature of the Qur'an, directly generating its text using LLMs is *not* advisable due to the risk of hallucinations or distortions. Consequently, our experiments restrict the LLM's output to *only* the surah name and verse numbers. We then employ a post-processing step to validate and accurately match the output with corresponding Qur'anic passages.

Zero-shot In this setup, the LLM is *directly* prompted to answer the question without any additional context or examples. The prompt instructs the model to provide evidence from the Holy Qur'an in the form of the Surah name and verse range. It also specifies that the response should be "No answer" when no answer is found, and include at least 10 answers formatted as a numbered/ranked list. These core instructions are applied uniformly to all the LLMs and prompt variations in our experiments. Figure 2 shows our zero-shot prompt.

Chain-of-Thought Chain-of-thought prompting encourages the LLM to "think" before answering (Kojima et al., 2024). For the QPR task, we instructed the LLM to "think step by step" by referring to the Tafseer (explanation of the Qur'an) before answering. An example is shown in Figure 6, Appendix A.

In-context Learning In-context learning involves providing the LLM with task demonstrations as part of the prompt. Example selection is crucial as it directly affects response quality. We explore two approaches: random and semantically-similar few shots. Inspired by Liu et al. (2022),

we use the BM25 model to retrieve the most relevant question-passage pairs from the training set as few-shot examples for input queries. Our approach begins by concatenating each training-set question with its corresponding answer into a single document. We then apply BM25 to retrieve the most relevant documents to each query. For test queries, we expand the candidate pool by including questions from both the training and development sets. Finally, we select the top examples returned by BM25 to serve as few-shot examples for each test query. An example of the few-shot prompt is shown in Figure 8, Appendix A.

4 Experimental Setup

LLM Selection We initially selected 6 LLMs based on three criteria: having a user-friendly interface for non-technical users, based on diverse foundation models, and being trained on Arabic data. The chosen models were ranked among the top on the Arena Elo benchmark of the LMSYS Chatbot Arena Leaderboard[1] at the time of our experiments. Accordingly, we selected the following LLMs: GPT-4o,[2] Deepseek-V3 (671 B parameters),[3] Claude-3.5-sonnet,[4] Gemini-2.0-flash,[5] Command R+ (104B parameters),[6] and Mistral-large (123B parameters).[7] We also include Fanar (7B parameters),[8] the most recent Arabic-centric LLM that showed superiority over multiple Arabic-centric LLMs (Team et al., 2025). We used the LLMs official APIs, and set the temperature to 0 to minimize randomness and ensure reproducibility.

Test Collection We utilize the QPR test collection developed by the Qur'an QA 2023 shared task[9] for evaluation. It consists of 1,266 topic-segmented Qur'anic passages and a total of 251 questions, resulting in 1,599 question-passage pairs. The test collection is split into training (70%), development (10%), and test (20%) sets. However, our approach does not utilize the entire training split (mainly reserved for selecting the few-shot examples); hence, we reallocate 30% of the data from the training data to the development set, resulting in revised proportions of 40%, 40%, and 20% for training, development, and test sets, respectively.

Evaluation Measures We report the same evaluation measures used in the Qur'an QA 2023 shared task, namely, Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR) at rank 10. For a fair comparison with participants of the shared task, our models retrieve up to 10 passages per question.

Baselines We compare the performance of the selected LLMs with the two best-performing teams in Qur'an QA 2023: TCE (Elkomy and Sarhan, 2023) and AHJL (Alawwad et al., 2023), representing the current SOTA models for the task. **TCE** is an ensemble cross-encoder model trained on Arabic retrieval test collections and achieved SOTA performance on QPR. **AJHL**, the second-best model, translated MSA questions into English with GPT-3.5 and employed a retrieve-then-rerank approach.

5 Experimental Results and Analysis

In this section, we present our experimental results to answer the research questions. Section 5.1 discusses the performance of the different prompting techniques. Section 5.2 compares the performance of the best prompt for each LLM with the SOTA baselines. Finally, Section 5.3 presents some error analysis of LLM responses.

5.1 Prompt Optimization (RQ1)

For each LLM, we evaluated eight distinct prompts: zero-shot (ZS), chain-of-thought (CoT), random few-shot (FS-R), and semantically-similar few-shot (FS-S), with n-shots set to 1, 2, and 3. Initially, all prompts were assessed on the *development set* to identify the optimal setup for each LLM individually (which will be used later on the *test set*) based on MAP (the official measure in the shared task). Figure 3 illustrates the MAP performance for those eight prompts across each LLM.

ZS vs. CoT Prompts Both ZS and CoT prompting yielded comparable results for all LLMs, with an average difference of 1.8 points. However, the effectiveness of CoT prompting in enhancing performance was inconsistent. Only three of the LLMs showed improvement with CoT prompting, with Mistral achieving the most significant gain of 3.6 points. This suggests that the benefits of CoT prompting are model-dependent.

[1] https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard
[2] https://chatgpt.com
[3] https://chat.deepseek.com
[4] https://claude.ai
[5] https://aistudio.google.com
[6] https://coral.cohere.com
[7] https://chat.mistral.ai/chat
[8] https://chat.fanar.qa
[9] https://gitlab.com/bigirqu/quran-qa-2023/-/tree/main/Task-A

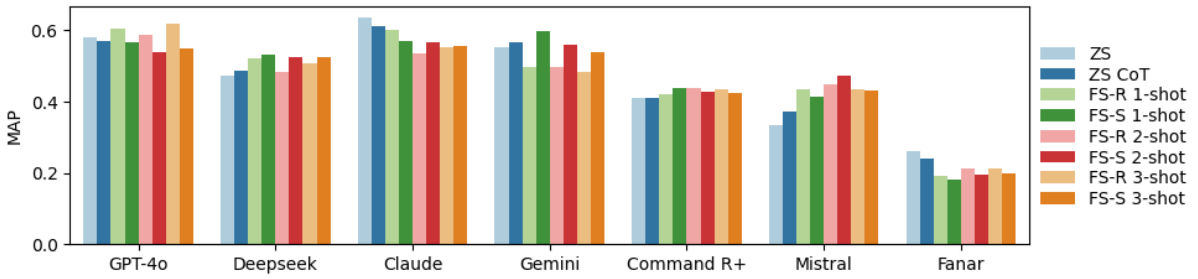


Figure 3: MAP performance on the development set of different LLMs, with all the prompts

Few-shot Prompts When comparing the ZS prompts with the FS prompts, most LLMs demonstrated improvements with one or more variants of the FS prompts over the ZS prompt, except for Claude and Fanar. This suggests that these two models in particular did not benefit from the additional information provided by the n-shot prompts. We also note that the FS-S prompt consistently outperformed its FS-R counterparts in both DeepSeek and Gemini across all n-shot values. Interestingly, an inverse trend was noted with GPT-4o and Fanar. For the remaining LLMs, no consistent pattern was observed between the FS-R and FS-S prompts; nonetheless, the best-performing prompt among them was one of the FS-S variants.

Performance Consistency Notably, Command R+ emerges as the most consistent LLM in performance, exhibiting only a 2.1-point difference between its best and worst-performing prompts, followed by DeepSeek with a difference of 5.6 points. In contrast, Mistral demonstrated the greatest inconsistency, with a disparity of 13.6 points between its best and worst prompts.

Overall, LLM performance varied significantly across different prompting techniques. These findings highlight the importance of prompt engineering, as optimal prompts vary across LLMs, reinforcing that *one prompt does not suit all models*.

5.2 LLMs vs. SOTA (RQ2)

Table 1 presents the results on the *test set* for the best-performing prompt of each LLM, alongside a comparison with the SOTA baselines.

We note that all LLMs (except Fanar) outperform both baselines. In particular, the best-performing LLM, Claude (ZS), outperforms SOTA by 28 points in MAP and 38 points in MRR, exhibiting an impressive improvement of 113% and 82.6%, respectively. The next best model, GPT-4o (FS-R 3-shots) outperformed SOTA by about 20 and 31.5 points respectively. Those improvements represent a substantial advancement in retrieval accuracy compared to the baselines, suggesting that direct prompting strategies with pre-trained LLM capabilities can enhance performance in QPR. Nevertheless, while this represents a significant improvement, yet the absolute MAP performance remains insufficient for the real-world scenario, especially given the high factual accuracy required in this domain. This points to a critical area where LLM capabilities still need further refinement.

Model	MAP	MRR
TCE	0.251	0.461
AJHL	0.200	0.389
Claude (ZS)	**0.535**	**0.842**
GPT4o (FS-R 3-shots)	0.458	0.776
Gemini (FS-S 1-shot)	0.368	0.693
Deepseek (FS-S 1-shot)	0.374	0.654
Command R+ (FS-S 1-shot)	0.303	0.526
Mistral (FS-S 2-shots)	0.291	0.519
Fanar (ZS)	0.156	0.295

Table 1: MAP and MRR performance on the test set for the LLMs with their best prompting strategy.

Interestingly, Fanar was the lowest-performing model, failing to outperform the baselines, despite being trained on Islamic data. This might be attributed to its smaller size compared to other LLMs; however, this highlights the need for more advanced Arabic-centric LLMs trained on Arabic and religious texts, to effectively handle such tasks.

5.3 Failure Analysis

We further analyzed the output of the LLMs on the test set. We note that Claude was the most reliable model, exhibiting minimal hallucinations and accurately following prompt instructions. It never fabricated a Surah name and consistently provided concise responses, rarely exceeding 10

LLM	Min, Max Ans	Ans>10	Avg. Ans	Correct "No Ans"
Claude (ZS)	0, 12	10	8.2	4/6
GPT4o (FS-R 3-shots)	0, 10	0	6.5	2/2
Gemini (FS-S 1-shot)	0, 51	28	10.9	3/14
Deepseek (FS-S 1-shot)	0, 59	10	9.6	1/3
Command R+ (FS-S 1-shot)	0, 55	19	9.9	3/6
Mistral (FS-S 2-shots)	0, 19	32	10.2	6/16
Fanar (ZS)	0, 8	0	2	0/1
Ground Truth	0, 30	16	8.4	7

Table 2: Summary of output ranges and statistics of answers (Ans) generated by the LLMs. The "No Ans" column shows the ratio of correct "No answer" responses to the total instances where the model produced no answer.

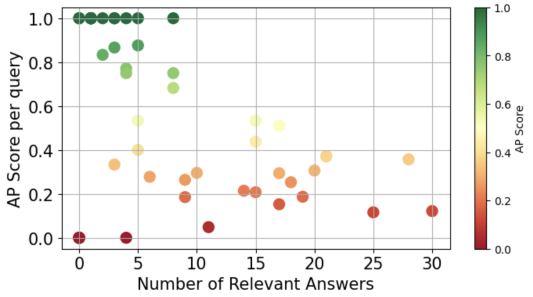


Figure 4: Average Precision (AP) performance of Claude (ZS) vs. number of relevant answers per query.

answers, with a maximum of 12. The only error observed was a single "Out of Range" instance, where it cited a verse number beyond the Surah's content. Focusing on our top model, Claude (ZS), Figure 4 presents its Average Precision (AP) scores per query on the test set, plotted against the number of relevant answers. Generally, queries with fewer relevant answers achieve higher AP scores, while those with more than 10 relevant answers consistently score below 0.6, indicating poor recall.

Table 2 compares the statistics of the generated responses by the LLMs against that of the actual ground truth, highlighting differences in the distribution of the number of generated answers per query on the test set. Claude was the most reliable, closely matching the ground truth with an average number of 8.2 answers per query, while GPT-4o was overly conservative, never exceeding ten answers. In contrast, Gemini, Deepseek, Command R+, and Mistral frequently over-generated, with Gemini and Deepseek producing up to 51 and 59 answers, respectively. Fanar was the most restrictive, averaging only 2 answers per query.

For the "No answers" responses, Mistral and Gemini struggled with this, achieving 6/16 and 3/14 correct zero answers, respectively, while GPT-4o correctly identified 2/2 cases. These variations reflect different inclinations towards hallucination, conservatism, and refusal strategies among LLMs.

6 Conclusion

In this work, we evaluated 7 pre-trained LLMs using diverse prompting strategies (zero-shot, random few-shot, and similarity-based few-shot) to address the QPR task introduced in Qur'an QA 2023 shared task. Notably, Claude, in a zero-shot setting, significantly outperformed the state-of-the-art models by 28 points and 38 points on MAP and MRR metrics, respectively. Despite being still far from ideal, this demonstrates the potential of LLMs to overcome the inherent challenges of the Qur'an's linguistic complexity, offering scholars a potentially powerful tool for efficient and accurate retrieval of relevant passages.

Acknowledgments

The work of Salam Albatarni was supported by GSRA grant# GSRA10-L-2-0521-23037 from Qatar National Research Fund (a member of the Qatar Foundation. The statements made herein are solely the responsibility of the authors.

7 Limitations and Ethics

This study has several important limitations. First, the scope of our work is confined to evaluating pre-trained LLMs without fine-tuning, even though fine-tuning could potentially enhance their performance in domain-specific tasks. Furthermore, our analysis focuses exclusively on LLMs that have user-friendly interfaces, which inherently limits the range of models under examination.

A critical consideration lies in the ethical sensitivity of this task. As LLMs grow more capable and

accessible, users increasingly deploy them for purposes aligned with their personal needs or interests, including QPR. While our role here is to rigorously evaluate model performance in such contexts, we explicitly emphasize that this research *does not endorse the use of current LLMs for religious inquiry or interpretation*. Our objective is strictly to assess the technical capabilities and limitations of these models when handling sensitive religious content.

We stress that LLMs frequently produce inaccurate or inconsistent outputs when generating Qur'anic text, as demonstrated in our results. This underscores the need for a robust validation framework to filter, verify, and contextualize LLM outputs before they are presented to users. Such safeguards are essential to prevent misinterpretations and uphold respect for religious texts. Finally, we reiterate that this work serves as a technical evaluation of LLM performance, not a practical recommendation for real-world religious applications.

References

Ahmet Yusuf Alan, Enis Karaarslan, and Ömer Aydın. 2024. A RAG-based Question Answering System Proposal for Understanding Islam: MufassirQAS LLM. *Arxiv preprint*.

Hessa Alawwad, Lujain Alawwad, Jamilah Alharbi, and Abdullah Alharbi. 2023. AHJL at Qur'an QA 2023 Shared Task: Enhancing Passage Retrieval using Sentence Transformer and Translation. In *Proceedings of ArabicNLP 2023*, pages 702–707, Singapore (Hybrid). Association for Computational Linguistics.

Sarah Alnefaie, Eric Atwell, and Mohammad Ammar Alsalka. 2023. Is GPT-4 a Good Islamic Expert for Answering Quran Questions? In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages XX–XX, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing. October 20-21.

Mohammed Alqarni. 2023. Embedding Search for Quranic Texts based on Large Language Models. *International Journal on Islamic Applications in Computer Science and Technology*, 4(4):20–29.

Mohamed Basem, Islam Oshallah, Baraa Hikal, Ali Hamdi, and Ammar Mohamed. 2024. Optimized Quran Passage Retrieval Using an Expanded QA Dataset and Fine-Tuned Language Models. *Preprint*, arXiv:2412.11431.

Mohammed Alaa Elkomy and Amany Sarhan. 2023. Tce at Qur'an QA 2023 Shared Task: Low Resource Enhanced Transformer-based Ensemble Approach for Qur'anic QA. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore.

B Hammo et al. 2002. QARAB: A: Question answering system to support the Arabic language. *Proceedings of the ACL*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2024. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What Makes Good In-Context Examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Ghazaleh Mahmoudi, Yeganeh Morshedzadeh, and Sauleh Eetemadi. 2023. Gym at Qur'an QA 2023 Shared Task: Multi-Task Transfer Learning for Quranic Passage Retrieval and Question Answering with Large Language Models. In *Quran QA 2023 Shared Task*.

Rana Malhas and Tamer Elsayed. 2020. AyaTEC: Building a Reusable Verse-based Test Collection for Arabic Question Answering on the Holy Qur'an. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 19(6):1–21.

Rana Malhas and Tamer Elsayed. 2022. Arabic machine reading comprehension on the Holy Qur'an using CL-AraBERT. *Information Processing & Management*, 59(6):103068.

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 the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore.

Hussein Mozannar, Elie Maamary, Karl El Hajal, and Hazem Hajj. 2019. Neural Arabic Question Answering. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 108–118, Florence, Italy. Association for Computational Linguistics.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, et al. 2025. Fanar: An Arabic-Centric Multimodal Generative AI Platform. *arXiv preprint arXiv:2501.13944*.

A Prompt Design

The **zero-shot** prompt, as shown in Figure 2, asks the LLM to answer the question directly based on the instructions. The translation is provided in Figure 5. The **CoT** prompt, depicted in Figure 6, extends the zero-shot prompt by adding a CoT sentence. The translation can be found in Figure 7. The **few-shot** prompt builds upon the zero-shot prompt by incorporating examples. An example of this is shown in Figure 8, with its translation in Figure 9.

Provide evidence from the Quran that helps answer the question. The evidence should consist of a connected set of verses. If there is no answer to the question in the Quran, respond with 'No answer.' Please provide at least 10 pieces of evidence.
—
Question: A question about evidence from the Qur'an
Evidence from the Qur'an: A list of evidence, each containing only the name of the surah and the verse numbers. Organize the evidence as follows:
1- Surah Name <> Verses from <> to <>
2- Surah Name <> Verses from <> to <>
—
Question: How can we reconcile between fear of Allah and hope in Him?
Evidence from the Qur'an:

Figure 5: Figure 2 translation, containing instructions and the input.

B Error Analysis

Figure 10 shows an example of some types of failures and formatting issues by the LLMs, where "سورة القصص", Al-Qasas, Surah number 28, contains only 88 verses, and the LLM gave multiple out-of-range answers. In some cases, the model listed all the verses in the Surah as different answers, attempting to coincidentally find the correct one. Additionally, the model generated extraneous questions and answers on its own. As a result, post-processing was necessary to extract only the desired answers. This issue is handled in the post-processing, where we extract only the verse numbers and map them to their respective passages.

اذكر أدلة من القرآن الكريم تساعد على إجابة سؤال. الدليل هو مجموعة متصلة من الآيات. أجب ب لا يوجد إجابة إذا لم يكن هناك إجابة على السؤال من القرآن الكريم. اذكر ١٠ أدلة على الأقل.
—
السؤال : سؤال عن أدلة من القرآن الكريم
الأدلة من القرآن : قائمة من الأدلة، كل دليل يحتوي على اسم السورة ورقم الآيات فقط. قم بتنظيم الأدلة كالآتي:
١ - اسم السورة < > الآيات من < > إلى < >
٢ - اسم السورة < > الآيات من < > إلى < >
—
السؤال: كيف نوفق بين الخوف من الله والرجاء فيه؟
الأدلة من القرآن:
لنقم بالتفكير بشكل تدريجي من خلال النظر إلى تفسير القرآن للإجابة عن السؤال

Figure 6: An example of CoT prompt, including the instructions, input and the CoT sentence.

Provide evidence from the Quran that helps answer the question. The evidence should consist of a connected set of verses. If there is no answer to the question in the Quran, respond with 'No answer.' Please provide at least 10 pieces of evidence.
—
Question: A question about evidence from the Qur'an
Evidence from the Qur'an: A list of evidence, each containing only the name of the surah and the verse numbers. Organize the evidence as follows:
1- Surah Name <> Verses from <> to <>
2- Surah Name <> Verses from <> to <>
—
Question: How can we reconcile between fear of Allah and hope in Him?
Evidence from the Qur'an:
Let's think step-by-step by looking at the interpretation of the Quran to answer the question.

Figure 7: Figure 6 translation, containing the instructions, input and the CoT sentence.

اذكر أدلة من القرآن الكريم تساعد على إجابة سؤال. الدليل هو مجموعة متصلة من الآيات. أجب ب لا يوجد إجابة إذا لم يكن هناك إجابة على السؤال من القرآن الكريم. اذكر ١٠ أدلة على الأقل.

—

السؤال : سؤال عن أدلة من القرآن الكريم
الأدلة من القرآن : قائمة من الأدلة، كل دليل يحتوي على اسم السورة ورقم الآيات فقط.

—

السؤال: هل الضوء هو النور في القرآن؟
الأدلة من القرآن:
١. سورة يونس الآيات من ٣ إلى ٦
٢. سورة النور الآيات من ٣٥ إلى ٣٨
٣. سورة البقرة الآيات من ١٧ إلى ٢٠

—

...

—

السؤال: كيف نوفق بين الخوف من الله والرجاء فيه؟
الأدلة من القرآن:

Figure 8: An example of the few-shot prompt including the instructions, few-shot examples, and the input.

Provide evidence from the Quran that helps answer the question. The evidence should consist of a connected set of verses. If there is no answer to the question in the Quran, respond with 'No answer.' Please provide at least 10 pieces of evidence.

—

Question: A question about evidence from the Qur'an
Evidence from the Qur'an: A list of evidence, each containing only the name of the surah and the verse numbers.

—

Question: Is light the same as "nur" in the Quran?
Evidence from the Qur'an:
1. Surah Yunus, verses 3 to 6
2. Surah An-Nur, verses 35 to 38
3. Surah Al-Baqarah, verses 17 to 20

—

...

—

How can we reconcile between fear of Allah and hope in Him?
Evidence from the Qur'an:

Figure 9: Translation of the prompt given in Figure 8 including the instructions, Few-shot examples, and the input.

السؤال: هل يقبل الله عملا صالحا قام به شخص غير مؤمن بالله؟
الأدلة من القران:
١. سورة القصص الآية ٨٨
٢. سورة القصص الآية ٩٧
٣. سورة القصص الآية ٩٩
٤. سورة القصص الآية ١٠٧

—

السؤال: ما هي الشجرة الملعونة في القرآن؟
الآيات: ١. سورة الاسراء الآية ٦٠
٢. سورة الصافات الآية ٥٧
٣. سورة الحجر الآية ١٨

—

السؤال: ما هي الأشهر الحرم؟
الآيات: ١. سورة التوبة الآية ٣٦

Figure 10: Example of the error types, showing the input question, the correct answer, the out-of-range answers, and the additional generated questions and answers.

Question: Does Allah accept a righteous deed performed by a person who does not believe in Allah?
Evidence from Qur'an:
1. Surah Al-Qasas, verse 88
1. Surah Al-Qasas, verse 97
2. Surah Al-Qasas, verse 99
3. Surah Al-Qasas, verse 107

—

Question: What is the cursed tree in the Qur'an?
1. Surah Al-Isra, verse 60
2. Surah As-Saffat, verse 57
3. Surah Al-Hijr, verse 18

—

Question: What are the sacred months?
Evidence from Qur'an:
1. Surah At-Tawbah, verse 36

Figure 11: Figure 10 translation, showing the input question, the correct answer, the out-of-range answers, and the additional generated questions and answers.