# OntologyRAG-Q: Resource Development and Benchmarking for Retrieval-Augmented Question Answering in Qur'anic Tafsir

**Sadam Al-Azani[1,2], Maad Alowifeer[1,3], Alhanoof Alhunief [1], Ahmed Abdelali[4]**

[1]SDAIA-KFUPM Joint Research Center for Artificial Intelligence, KFUPM, Dhahran, Saudi Arabia

[2]Department of Information & Computer Science, KFUPM, Dhahran, Saudi Arabia

[3]Department of Electrical Engineering, KFUPM, Dhahran, Saudi Arabia

[4]Humain, Riyadh, Saudi Arabia

**Correspondence:** sadam.azani@kfupm.edu.sa

## Abstract

This paper introduces essential resources for Qur'anic studies: an annotated Tafsir ontology, a dataset of approximately 4,200 question-answer pairs, and a collection of 15 structured Tafsir books available in two formats. We present a comprehensive framework for handling sensitive Qur'anic Tafsir data that spans the entire pipeline from dataset construction through evaluation and error analysis. Our work establishes new benchmarks for retrieval and question-answering tasks on Qur'anic content, comparing performance across state-of-the-art embedding models and large language models (LLMs). We introduce OntologyRAG-Q, a novel retrieval-augmented generation approach featuring our custom Ayat-Ontology chunking method that segments Tafsir content at the verse level using ontology-driven structure. Benchmarking reveals strong performance across various LLMs, with GPT-4 achieving the highest results, followed closely by ALLaM. Expert evaluations show our system achieves 69.52% accuracy and 74.36% correctness overall, though multi-hop and context-dependent questions remain challenging. Our analysis demonstrates that answer position within documents significantly impacts retrieval performance, and among the evaluation metrics tested, BERT-recall and BERT-F1 correlate most strongly with expert assessments. The resources developed in this study are publicly available at https://github.com/sazani/OntologyRAG-Q.git.

## 1 Introduction

LLMs' importance across all domains is rapidly growing, revolutionizing how information is processed, generated, and applied in diverse fields—from healthcare (Zhang et al., 2025) to education (Wen et al., 2024), law (Siino et al., 2025), and beyond (Mahmud et al., 2025).

Tafsir (exegesis), the interpretation of the Qur'an, is a critical domain, especially within religious and linguistic contexts, where accuracy, nuance, and contextual sensitivity are essential (Abdelnasser et al., 2014). There is a growing use of LLMs, such as ChatGPT(OpenAI, 2023), to assist in the interpretation, analysis, and answering of questions related to Qur'anic content. However, despite their potential, these models still exhibit notable limitations in handling such content effectively (Phan et al., 2024). Scholars are primarily concerned about issues such as the reliability of generated content, potential biases, lack of source transparency, and the risk of misinterpretation. There is a clear and urgent need to develop reliable, transparent, and ethically aligned systems that can utilize the transformative potential of LLMs. This begins with the creation of robust, domain-specific resources.

One of the key challenges associated with LLMs is hallucination (Ji et al., 2023). The issue of hallucination becomes even more critical when dealing with domain-specific questions, particularly in sensitive domains such as Qur'anic content. Retrieval-Augmented Generation (RAG) has been proposed as a solution to mitigate hallucination by grounding model outputs in external knowledge sources (Tonmoy et al., 2024; Li et al., 2025). However, it is still limited when dealing multi-hop question answering (Saleh et al., 2024) and scenarios where the retrieved information may be partial or imprecise (Feng and He, 2025).

Several valuable Qur'anic ontologies have been developed, including (Dukes et al., 2010; Sherif and Ngonga Ngomo, 2015; Khan et al., 2013). However, these ontologies have primarily focused on morphological and syntactic analysis. Therefore, there is an urgent need to extend and adapt these ontologies to support and utilize the capabilities of LLMs.

This work makes several key contributions to the field of AI-driven Qur'anic Tafsir. First, we present a set of resources, including an annotated Tafsir ontology, a Tafsir-focused QA dataset, a Tafaseer



*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15552–15570

(تفاسير)[1] dataset, and a framework, offering essential tools for scholars and researchers. We also propose a RAG approach featuring a novel chunking method, termed ***Ayat-Ontology***, which applies ontology-based structuring at the verse level to enhance the accuracy and contextual depth of interpretation and analysis. The effectiveness of this approach is demonstrated through benchmark evaluations on retrieval and question-answering tasks using state-of-the-art embeddings and LLMs. Results show that the proposed method performs robustly, with GPT-4 and ALLaM (Bari et al.) achieving the highest performance. In addition, this study provides a comprehensive empirical analysis, evaluating the types of errors, retrieval efficiency, question categories, and the influence of passage positioning within the document on model responses.

## 2 Literature Review

### 2.1 Qur'anic Ontologies

Dukes et al. (2010) presented the guidelines for developing and annotating the Qur'anic Arabic Dependency Treebank (QADT), which includes two main levels of analysis: morphological annotation and syntactic representation. Their approach details the syntactic representation choice, providing comprehensive documentation for tags, inflection features, and dependency graph edge labels. This process involved multiple stages, including automatic morphological tagging, manual verification, and online collaborative annotation. Dukes and Habash (2010) manually verified and computationally analyzed the Qur'anic Arabic Corpus, contributing to a deeper understanding of the morphological aspects of Qur'anic Arabic. Zaghouani et al. (2012) provided a semantic annotation for 50 verbs previously annotated in (Dukes and Buckwalter, 2010). Sherif and Ngonga Ngomo (2015) developed the Semantic Qur'an Ontology, a multilingual Resource Description Framework (RDF) representation of Qur'an translations across 43 languages. This ontology incorporates morpho-syntactic data for in-depth linguistic analysis and comparison of translations. Khan et al. (2013) created a domain ontology focused on living creatures, including animals and birds mentioned in the Qur'an. Similarly, Al-Yahya et al. (2010) designed a computational model representing Arabic lexicons, specifically applied to the "Time nouns" in the Qur'an. Another theme-based ontology was developed by (Ta'a et al., 2014), categorizing Qur'anic content into distinct themes, with a special focus on the Syammil Al-Quran Miracle. Sharaf and Atwell (2012) developed a corpus annotating personal pronouns and their antecedents in the Quran, creating an ontological list of over 24,500 pronouns.

### 2.2 RAG LLM

In the context of Arabic RAG, El-Beltagy and Abdallah (2024) conducted a study to evaluate the effectiveness of RAG for Arabic text. The study utilized two datasets: the Arabic EduText Secondary School dataset, compiled by the authors, and the ARCD (Arabic Reading Comprehension Dataset). For retrieval tasks, the authors assessed various semantic embedding models and evaluated different LLMs for their performance. Similarly, the study in (Al-Rasheed et al., 2025) focused on Arabic lexical information retrieval, including tasks like translation, editorialization, root extraction, and part-of-speech tagging.

In parallel, advanced RAG techniques have been developed. For example, Wang et al. (2024) introduced a multiple partition paradigm for RAG, termed M-RAG, where each database partition functions as a basic unit for RAG execution. Building on this concept, we propose a novel framework that integrates LLMs with Multi-Agent Reinforcement Learning to optimize various language generation tasks, achieving significant improvements. Additionally, Niu et al. (2023) introduced RAGTruth, a corpus designed to analyze word-level hallucinations across multiple domains and tasks within standard RAG frameworks for LLM applications. RAGTruth contains nearly 18,000 naturally generated responses from various LLMs using RAG. Their findings highlight that it is possible to fine-tune smaller LLMs to achieve competitive performance in hallucination detection when compared to prompt-based methods using state-of-the-art LLMs like GPT-4.

Jeong et al. (2024) proposed an adaptive question-answering framework capable of dynamically selecting the most appropriate strategy for retrieval-augmented LLMs, ranging from simple to sophisticated methods based on the complexity of the input query. A lightweight language model was introduced that serves as a classifier to predict query complexity. The approach was evaluated on an open-domain QA dataset covering three levels of query difficulty, demonstrating improved efficiency

[1] *Tafaseer* is the plural of *Tafsir*, referring to multiple Tafsir books.

and accuracy compared to relevant baselines.

Saleh et al. (2024) proposed a zero-shot subgraph RAG approach that leverages the structured nature of knowledge graphs to effectively answer multihop questions using LLMs. Their method constructs a Cypher query based on the question to retrieve a set of semantically relevant subgraphs, which are then provided as contextual input to the LLM. The approach was evaluated on a benchmark question-answering dataset in the movie domain, demonstrating a substantial improvement in the accuracy of 2-hop and 3-hop questions when using LLAMA 8B Instruct and GPT-4 Turbo, compared to both models operating with and without standard RAG techniques.

While previous approaches largely relied on external knowledge, some efforts aim to develop knowledge-augmentation frameworks without such resources. For example, (Liao et al., 2024) proposed Awakening-Augmented-Generation (AAG), which mimics human thinking and recall to fill knowledge gaps, activating relevant information within LLMs without external sources. AAG includes two key components: Explicit awakening, which fine-tunes a context generator to produce a synthetic document as symbolic context, and Implicit awakening, which uses a hypernetwork to create adapters based on the query and synthetic document, providing parameter-based context for the LLM.

Previous work on Qur'anic ontologies has primarily addressed morphological and semantic aspects, while RAG applications in Arabic remain limited. This study introduces a comprehensive resource suite for Qur'anic Tafsir, featuring a structured ontology and a custom verse-level chunking method, OntologyRAG-Q. It also presents the first benchmark for Qur'anic retrieval and QA using state-of-the-art LLMs and embeddings. Supported by expert evaluation, the results show improved performance and address prior limitations. Compared to other advanced approaches, this work is distinct in its methodology, domain focus, and overall contributions.

## 3 Tafsir ontology and datasets

### 3.1 Tafsir ontology

Figure 1 illustrates the developed ontology for Tafsir. The ontology provides a structured representation of Tafsir-related knowledge by modeling relationships among Surahs (chapters), Ayat (verses), Tafsir texts (interpretations), Hedayat (guidance), Maqasid (key objectives), and questions and answers derived from the Tafsir.

The ontology is grounded in two core datasets: the Tafaseer dataset and the QA dataset, described in Sections 3.2 and 3.3, respectively. Each Surah/chapter is identified by its unique ID and name and includes Maqasid (key objectives) extracted from Tafisr al-Muyassar, such as the reason for its naming and its thematic purpose. At the verse level, each ayah is defined by its position (start and end numbers), its text, and the corresponding Tafsir, and Hedayat (guidance) are extracted.

From each Tafsir excerpt, a variety of questions and answers have been generated. Each question is enriched with multiple attributes and labels, supporting a multi-label classification framework where each label can span multiple classes. This ontological structure not only enables advanced semantic queries but also serves as a backbone for intelligent Tafsir-based QA systems.

### 3.2 Tafaseer dataset

Tafaseer dataset comprises a curated collection of 15 Tafsir books, systematically gathered from Al-Baḥith Al-Qur'ani[2], a well-established and reputable platform for Qur'anic research. A summary of the collected Tafsir volumes is provided in Table 6, Appendix A.

Each Tafsir source has been organized and stored in CSV format, where each file corresponds to one book and contains structured attributes, as detailed in Table 7, Appendix A. It is important to note that many Tafsir books do not interpret each verse in isolation. Instead, they often provide an exegesis that spans multiple continuous verses. In such cases, the same exegesis is duplicated for each relevant verse, with the "Related" attribute indicating the initial verse of the excerpt. This version, referred to as *Tafaseer Dataset I*, includes 6,236 samples per file, aligning with the total number of verses in the Holy Qur'an. Additionally, a second version of Tafaseer dataset has been developed, *Tafaseer Dataset II*, in which each sample corresponds to a specific excerpt, either a single verse or a sequence of related verses, as defined by the Tafsir book. Figure 2 presents the distribution of samples across the different sources in *Tafaseer Dataset II*.

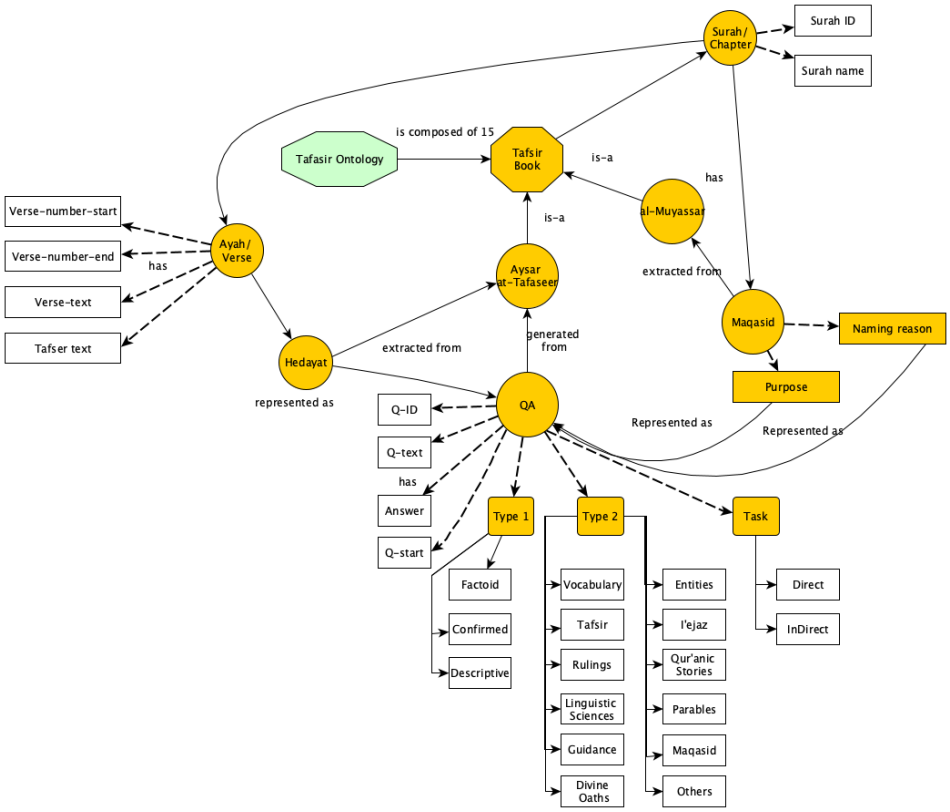


Figure 1: Developed Tafsir ontology

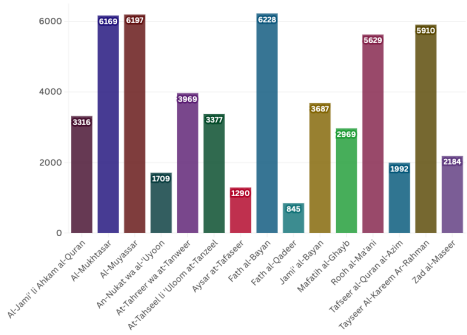


Figure 2: Statistics of the Tafaseer Dataset II

### 3.3 QA dataset

We developed a QA dataset sourced from two Tafsir books: Aysar at-Tafaseer and Tafsir Al-Muyassar. Aysar at-Tafaseer comprises 1,290 excerpts, covering all the verses of the Qur'an, while Al-Muyassar consists of 6197 excerpts. Figure 3 illustrates the overall process of the dataset development. Drawing from Aysar at-Tafaseer and considering the substantial length of the text, we selected a diverse and representative sample of Surahs (chapters) from the Qur'an. This includes Surahs from the beginning, Surah 1: Al-Fatiha (The Opening) and Surah 2: Al-Baqarah (The Cow), from the middle, Surahs 21: Al-Anbiya (The Prophets) to 30: Ar-Rum (The Romans); and from the end, Surahs 70: Al-Ma'arij (The Ascending Stairways) to 114: An-Nas (Mankind). This approach allows for a thorough evaluation of the model's ability to retrieve and interpret information from different sections of both the Qur'an and Tafsir books.

Each Surah was segmented at the verse level (excerpt), and each segment was then processed by GPT-3.5 to generate seven unique questions, using the following prompt

> Create seven questions and their corresponding answers from the given text:\n\n{text}. Answers and questions must be from the given text and should be in Arabic

We generated around 2790 QA pairs. These pairs were thoroughly reviewed, during which several categories were identified for revision:

- **Delete**: Applied when a question is found to be meaningless, the answer is incorrect, or there is ambiguity in the pair. As a result, approximately 152 questions were deleted.

[2] https://tafsir.app/

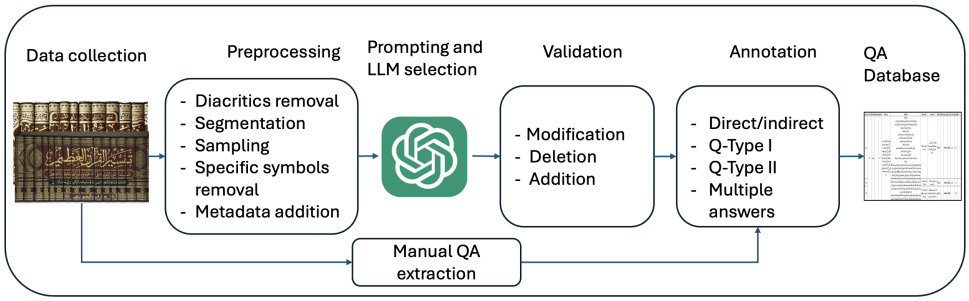


Figure 3: Process of QA-dataset development

- **Question-Modified**: Used to improve the readability, accuracy, and clarity of a question, ensuring it conveys its intended meaning more effectively. As a result, around 606 questions were modified.

- **Answer-Modified**: Employed to refine the answer, enhancing its readability and ensuring it aligns more clearly with the question and its context. About 15 answers were modified.

- **Question and Answer-Modified**: Both the question and answer are revised to improve clarity, accuracy, and meaning. As a result, 25 question-answer pairs were modified.

- **New**: Introduced when important knowledge or insights were overlooked, requiring a new question-answer pair to fill the gap. As a result, 45 new pairs were added during the first cycle.

After these revisions, we ended up with around 2,680 refined QA pairs. In addition, a total of 1,290 question–answer pairs were extracted from Aysar at-Tafaseer, focusing on verses related to Hedayat (guidance) across the entire tafsir book. Each question follows the template: **"What guidance is provided by [the verse or a set of continuous verses]?"**

Similarly, from Tafsir Al-Muyassar, we manually created 114 questions exploring the reasoning behind the naming of each Surah (chapter), framed as: **"What is the reason for naming the Surah of [Surah Name]?"** An additional 114 questions were compiled to reflect the stated purpose or overarching message of each Surah, phrased as: **"What is the purpose of the Surah of [Surah Name]?"** Table 1 provides a detailed statistical overview of the QA dataset review and refinement process, highlighting the distribution of generated, modified, deleted, and manually extracted items.

| Stage | Count |
|---|---|
| Initial Generated | 2788 |
| Removed During Revision | -152 |
| New Manually Added | +45 |
| Manual: Hedayat (Gauidance) | +1290 |
| Manual: Surah Name Reasoning | +114 |
| Manual: Surah Purpose | +114 |
| **Final Total** | **4199** |

Table 1: Statistical summary of QA dataset construction stages

### 3.4 Question classification

Classifying questions by expected answer type is essential for guiding RAG systems toward the most relevant resources and narrowing the search space. Each question was annotated with the following details:

- Direct or Indirect Question: To identify whether the question directly asks for information or implies it.

- Repeated or Multiple Answers: For questions that may have several possible answers based on context or due to repeated alliteration.

- Type I question, the questions have been classified into three categories: Factual, Descriptive, or Confirmation.

- Type II question: To define the Tafsir topic and the intent or category of the expected answer. In collaboration with domain scholars, we defined 14 question types corresponding to expected retrieved answer categories specific to Tafsir, as detailed in Appendix A, Table 9. Then, the first author, together with two part-time undergraduate research assistants at KFUPM, carried out the initial annotation of this question type. These anno-

tations were subsequently reviewed and validated by the project's internal consultant in the Islamic scholarship domain at KFUPM. The inter-annotator agreement for this stage was 72%. This procedure was excluded for manually created pairs, as they were classified directly owing to their clearly defined and unambiguous types.

## 4 Benchmark Experiments

Figure 4 illustrates the proposed approach, which consists of several key modules: chunking, embedding, retrieval, and generation.

**Ontology-based chunking.** Chunking refers to the cognitive process of breaking down large bodies of text into smaller, more manageable segments or "chunks." This technique plays a pivotal role in optimizing both the retrieval and generation components of a RAG system. Various types of chunking exist, including token-based and semantic chunking. We propose an innovative ontology-based chunking method, specifically applied at the Ayat/Verses level. In this method, each chunk represents an individual Ayah/verse or a set of verses, enriched with ontology-related information, defined above. This additional layer of contextual understanding enhances the accuracy and relevance of the system's responses, allowing for more nuanced and precise retrieval and generation processes. Adding such information explicitly into the chunk will enhance the retrieval of questions related to verses numbers, the name of Surah, and other key attributes, further improving the system's overall performance.

| Embedding Name | Model | Vector Dimension |
|---|---|---|
| Open AI | text-embedding-ada-002 (TEA2) | 1536 |
| E5 | Multilingual-E5-small (E5-small) | 384 |

Table 2: Embeddings details

**Embeddings.** Embeddings are dense vector representations that encode the semantic meaning of text, serving as a critical component in both the retrieval and generation stages of a RAG system (Gao et al., 2023). The evaluated embeddings details are presented in Table 2.

**Generative models.** For text generation, we considered ALLaM and GPT-4. Table 3 provides detailed specifications of the models used. The following prompt was employed, explicitly instructing the LLMs to generate answers based solely on the retrieved context.

> You are an expert in interpreting the Quran, specifically designed to answer users' questions. Provide answers solely based on the context provided below. Do not draw upon any external or prior knowledge or information. If the answer is not found within the given context, respond with 'I don't know.' Ensure that the Ayah (verses) are quoted verbatim as they appear in the Quran. All answers should be provided in Arabic.
> Context: {relevant_docs}
> Question: {query}
> Answer:

**Evaluation framework.** Our evaluation framework is mainly based on lexical-based metrics, including BLEU (Papineni et al., 2002) and CHRF (Popović, 2015), and semantic-based metrics, using BERTScore (Zhang et al., 2019). In the benchmarking evaluation, the focus was placed on direct questions, as RAG-based applications are inherently designed for direct question answering. Since the objective is to develop multipurpose resources, indirect questions remain useful and can be applied in related downstream tasks. The initial direct set (Direct Set I), comprising 1,178 QA pairs, was used for evaluation. Although this set was later expanded to 2,350 QA pairs (Direct Set II), the evaluation was restricted to Direct Set I due to the substantial time required for domain-expert assessment. In addition, qualitative evaluation was conducted by a domain expert in Islamic scholarship. The scholar was provided with an interface composed of "Question", "Context from which the question and answer extracted", "The Answer", "The Generated Answer from the system", the evaluation as a scale from 1-5 such that 1 is the answer is totally incorrect while 5 the answer is totally correct and accurate. Based on the domain scholar evaluation analysis, two key assessment measures were identified: **accuracy** and **correctness**. **Accuracy** refers to responses that are both correct and detailed. On the other hand, **correctness** denotes answers that are factually valid but may be less precise or contain limited explanatory detail, yet are still considered acceptable.

## 5 Results and discussion

Table 4 presents results across all chunking methods, embedding models, and LLMs. The upper section reports the ***baseline*** (semantic chunking), while the lower section shows our ***proposed method***. Search type is either Similarity or MMR (Maxi-

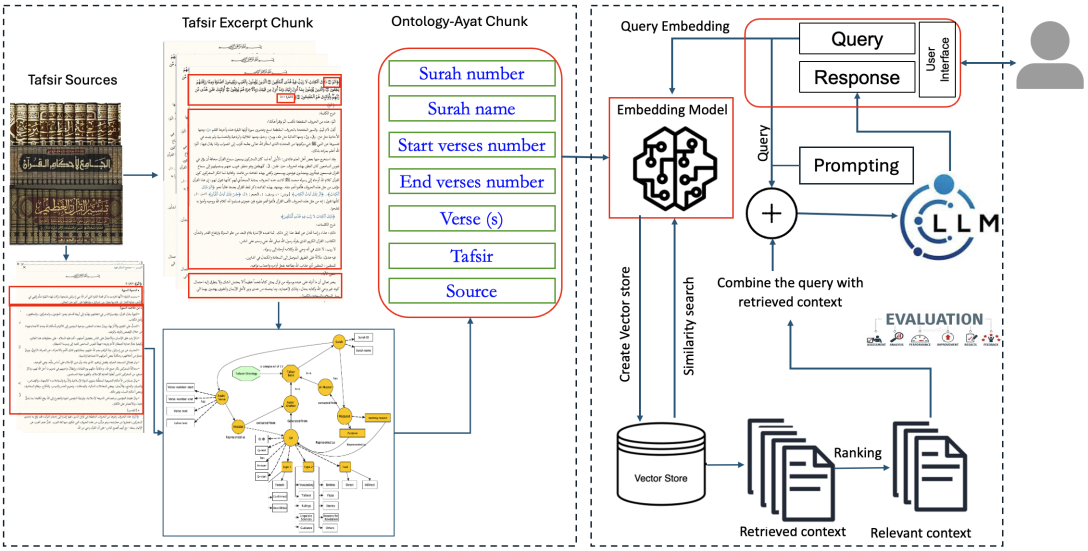


Figure 4: High-level structure of the Ayat-Ontology-based RAG system

| Model | Description | Size | Focus | Language | Organization |
|---|---|---|---|---|---|
| ALLaM | SDAIA's private Arabic language model based on LLaMA 2. | 7B | Arabic-focused NLP tasks | Arabic | SDAIA |
| GPT-4O-mini | Likely a smaller, optimized version of GPT-4. | ~8B | General-purpose conversational AI | Multilingual, focus on high efficiency | OpenAI |

Table 3: Evaluated LLMs for the generative task in the proposed approach

mal Marginal Relevance). $k$ is the number of top documents retrieved, and fetch_k is the number of documents initially retrieved for the MMR algorithm (default = 20). lambda_mult controls diversity in MMR (1 = minimum, 0 = maximum; default = 0.5). The temperature was set to zero in most cases. Among the embedding models, multilingual-e5-small achieved the best overall performance, and GPT-4O-mini achieved the best performance. Accordingly, we adopted the multilingual-e5-small embedding model for subsequent experiments.

The proposed approach demonstrates a notable advantage in handling questions that cannot be directly answered through traditional semantic chunking, such as those requiring information about the number of verses (Ayat), names of Surah (chapters), or verse references related to specific queries. This improvement is attributed to the integration of relevant contextual and structural information within each chunk.

**Qualitative analysis.** We enforce the proposed approach to explicitly state *'I don't know'* when the system lacks sufficient context to provide a valid response. This ensures relevance by preventing off-topic answers, maintains correctness by avoiding incorrect or speculative information, and upholds faithfulness by transparently acknowledging the system's limitations. In our case, this approach is enforced through prompt-based conditioning. As a result, we identified 139 cases where the system's response implicitly or explicitly indicated a lack of knowledge, effectively conveying the meaning of *"I don't know"*. We analyzed the types of questions that led to these outcomes. We found that most of these questions were more complex, often requiring specific answers derived from a given text, such as: "What is the meaning of [WORD] in a [TEXT]?" where [TEXT] may refer to a verse or an excerpt from a Tafsir. Examples of such questions are provided in Figure 7, Appendix B.

**Domain scholar evaluation and analysis.** The best-achieving responses, based on the evaluated configurations, were subsequently submitted to the domain scholar for evaluation. We identified the following key observations:

- Out of 1178, 720 cases were rated as score 5, which means that the answer is correct and accurate. In two samples, we observed that the scholar's comments indicated that the responses provided by the proposed system were superior to the label answers. Upon analyzing these cases, we found that the questions have multiple possible answers scattered across different sections of the source. Those cases are presented in Appendix B Figure 8.

- A total of 62 questions were rated as score 4 by the scholar evaluator. Most of the questions

| LLM | Chunk | Parameters | Embedings | BLEU | CHRF | BERT score | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | Precision | Recall | F1 |
| ALLaM | Semantic | k:3, MMR | TEA2 | 1.10 | 14.69 | 61.56 | 72.03 | 66.31 |
| ALLaM | Semantic | k:6, MMR, 'lambda_mult': 0.25 | TEA2 | 1.16 | 15.32 | 62.15 | 72.09 | 66.68 |
| ALLaM | Semantic | k:6, MMR, fetch_k: 50 | TEA2 | 1.06 | 15.26 | 62.43 | 72.30 | 66.93 |
| ALLaM | Semantic | k:3, MMR | E5-small | 1.21 | 15.42 | 61.30 | 71.90 | 66.10 |
| ALLaM | Semantic | k:3, Similarity | E5-small | 1.36 | 16.05 | 62.10 | 72.90 | 67.00 |
| GPT-4O-mini | Semantic | k:3, MMR | E5-small | 1.32 | 18.09 | 63.75 | 73.80 | 68.34 |
| GPT-4O-mini | Semantic | k:3, MMR | TEA2 | 1.21 | 17.69 | 63.66 | 73.21 | 68.03 |
| GPT-4O-mini | Semantic | k:6, MMR | E5-small | 3.06 | 25.50 | 68.82 | 76.01 | 72.18 |
| GPT-4O-mini | Semantic | k:6, MMR | TEA2 | 1.27 | 18.22 | 63.96 | 73.69 | 68.41 |
| ALLaM | Ayat ontology | k:6, Similarity | E5-small | 3.36 | 25.06 | 65.88 | 75.43 | 70.25 |
| GPT-4O-mini | Ayat ontology | k:6, Similarity, temperature=0.0 | E5-small | 16.93 | **41.12** | **76.15** | **79.59** | **77.71** |
| GPT-4O-mini | Ayat ontology | k:6, Similarity, temperature=0.7 | E5-small | **18.07** | 39.99 | 75.64 | 78.28 | 76.79 |

Table 4: Experimental results.

| | Accuracy | Correctness |
|---|---|---|
| Including "I don't know" | 69.52 | 74.36 |
| Excluding "I don't know" | 78.88 | 84.38 |

Table 5: Domain scholar results under different response inclusion criteria

were rated without any comments, while some received feedback due to the system providing valid answers from different excerpts in the same Tafsir source. This was especially true for questions where the answer was derived from multiple verses or different parts of the source. Other feedback indicated that the system successfully retrieved the correct corresponding verse, but without the full answer. Examples of these cases are provided in Appendix B and Figure 9.

- A total of 59 answers received a score of 3. Upon reviewing these cases, we found that none of the responses are incorrect. All are correct; however, the lower scores were mainly due to: 1) the system either presenting only the relevant verses without sufficient interpretive detail (tafsir), 2) providing a correct or alternative answer derived from another position in the same source, 3) general answers, represented by a small number of cases (which can be improved by either rephrasing the question or interactive questions), or 4) for few cases, the source and the system answers have the same meaning. Selected examples are presented in Appendix B, Figure 10.

- A total of 35 questions were rated by the scholar with a score of 2. Upon further analysis, it was found that while the system successfully retrieved related verses and provided generally correct answers, the results did not align with the label answers. Additionally, there was a case where the system offered a description of the answer rather than the direct answer, but the response was still considered correct. Selected examples of such cases are presented in Appendix B, Figure 11.

Table 5 presents the scholar evaluation accuracy and correctness with including and excluding *"I don't know"* answers.

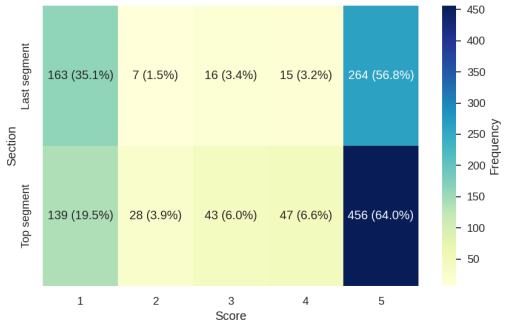


Figure 5: Frequency and percentage of retrieval scores by section.

**Position-based performance analysis.** We evaluated the impact of answer positioning within the document by dividing the main source into two sections: the initial part and the final part. The correlation between answer position and the domain scholar evaluation is shown in Figure 5. The results indicate that answer accuracy is significantly influenced by position within the document, with answers appearing in the initial section being more reliable than those in the final section.

**Evaluating metrics correlation.** One of the most critical challenges in QA tasks is the selection of appropriate evaluation metrics (Phatthiyaphaibun et al., 2024), especially for the open-ended questions. To address this, we conducted an analysis to identify which automated evaluation measures (lexical-based metrics and semantic similarity measures) most closely align with domain scholar evaluation *(accuracy and correctness)*. Our findings indicate that BERT-score, especially BERT-Recall and BERT-F1, are strongly correlated with the domain scholar evaluation. The details are depicted in Figure 6.

## 6 Conclusion

We have developed comprehensive resources for QA, model training, and generation tasks within the context of Qur'an Tafsir and provided a benchmark for QA and RAG. These resources are designed to support the advancement of AI models that can better understand, retrieve, and generate insights from given text sources, particularly the Qur'anic Tafsir. This study also introduced an anthology-based RAG approach, ***OntologyRAG-Q***, which demonstrates superior accuracy compared to the baseline methods. Through a comprehensive empirical analysis, Ayat-Ontology based chunking improves the results. The study concludes that retrieval algorithms are significantly influenced by the position of the answer within the document, particularly in large, complex texts with overlapping language (Contextual Overlap), such as Tafsir. Additionally, the study analyses different performance measures and finds that BERT-recall, followed closely by BERT-F1, shows the highest correlation with scholar evaluations.

## 7 Limitations

A limitation of this study is that it assumes all queries to have the same level of complexity. To improve accuracy and efficiency, future work should incorporate to deal with multi-step queries, such as adaptive approaches and strategies that scale from the simplest to the most sophisticated, depending on the complexity of each query. While multiple Tafsir sources were included in our dataset, we only performed the analysis using one source, as the ground truth was extracted from this particular Tafsir. This study primarily focuses on resource development and benchmark construction; consequently, the evaluation was limited to a selected subset of LLMs and embedding models. Future work can extend this analysis to a broader range of models.

## 8 Ethics Statement

From the initialization of the project, the Qur'anic text has been approached with the highest level of ethical care, reverence, and sensitivity. Ethical considerations were integrated into every stage of the research process, beginning with the careful selection of reliable and authentic sources for data collection. Throughout the project, qualified scholars were consulted and actively involved in the validation and evaluation phases to ensure religious accuracy and contextual appropriateness. The overarching aim is to develop a system that delivers accurate and contextually appropriate responses without modifying or misrepresenting the original Qur'anic verses. For the developed resources we created in this study, we explicitly specify that their intended use is limited to academic and research contexts, ensuring compatibility with the original access conditions and maintaining compliance with research-only restrictions on derivative data.

## 9 Research Impact

This research makes a significant contribution to the field of AI-driven Qur'anic studies by introducing a comprehensive framework for handling sensitive Qur'anic Tafsir data, covering the entire pipeline, from dataset construction to evaluation and error analysis. The proposed framework enables the development and assessment of systems that must operate with theological sensitivity and linguistic nuance.

The work results in the creation of new, high-quality resources for Qur'anic Tafsir, including a large-scale QA dataset, a structured Tafsir ontology, and a benchmark for retrieval and question-answering tasks. These resources can support a wide range of applications, including advanced RAG systems, machine reading comprehension, intent detection, QA modeling, and the fine-tuning or complete training of LLMs in religious or low-resource domains. At the same time, the study acknowledges field-specific limitations: the scarcity of digitized and annotated Tafsir materials, the interpretive diversity across classical commentaries, and the challenge of ensuring cultural and theological sensitivity when adapting general-purpose NLP models. Highlighting these constraints helps clarify

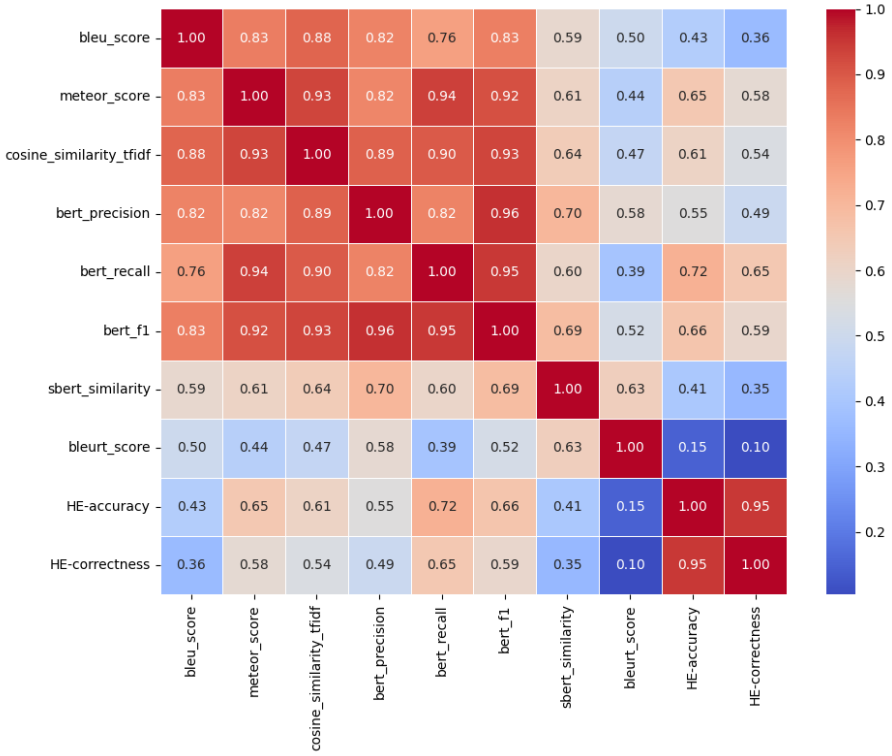


Figure 6: Correlation between evaluation measures and domain scholar evaluation (HE-accuracy and HE-correctness)

both the value and the boundaries of the presented approach.

In addition, while this research is situated in the Qur'anic domain, the methodology has broader relevance and can be extended to other structured, interpretive texts, such as classical literature and legal commentary, which share similar characteristics of layered interpretation and context-dependent reasoning. Thus, the impact of this work goes beyond Qur'anic NLP, providing a foundation for computational approaches to interpretive traditions in multiple fields.

The research also contributes to evaluation methodology by analyzing the effectiveness of different metrics and by highlighting the limitations of current approaches when applied to religious and interpretive texts. Collectively, this work lays the foundation for more reliable, interpretable, and culturally aligned NLP systems in sensitive knowledge domains.

## Acknowledgment

The authors would like to acknowledge the support provided by Saudi Data & AI Authority (SDAIA) and King Fahd University of Petroleum & Minerals (KFUPM) under SDAIA-KFUPM Joint Research Center for Artificial Intelligence (JRC-AI) grant No. JRC-UCG-09. The authors would like to express their sincere gratitude to the project consultants in the domain of Islamic scholarship, Dr. AbdulRahman Howsawi and Prof. AbdulRahman Al-Shehri, for their expert advice and scholarly input throughout the development of this work.

## References

Heba Abdelnasser, Maha Ragab, Reham Mohamed, Alaa Mohamed, Bassant Farouk, Nagwa M El-Makky, and Marwan Torki. 2014. Al-bayan: an Arabic question answering system for the holy Quran. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 57–64.

Raghad Al-Rasheed, Abdullah Al Muaddi, Hawra Aljasim, Rawan Al-Matham, Muneera Alhoshan, Asma Al Wazrah, and Abdulrahman AlOsaimy. 2025. Evaluating rag pipelines for Arabic lexical information retrieval: A comparative study of embedding and generation models. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, pages 155–164.

Maha Al-Yahya, Hend Al-Khalifa, Alia Bahanshal, Iman Al-Odah, and Nawal Al-Helwah. 2010. An ontological model for representing semantic lexicons: an application on time nouns in the holy Quran. *Arabian Journal for Science and Engineering*, 35(2):21.

M Saiful Bari, Yazeed Alnumay, Norah A Alzahrani,

Nouf M Alotaibi, Hisham Abdullah Alyahya, Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z Alsubaie, Hassan A Alahmed, Ghadah Alabduljabbar, et al. Allam: Large language models for arabic and english. In *The Thirteenth International Conference on Learning Representations*.

Kais Dukes, Eric Atwell, and Abdul-Baquee M Sharaf. 2010. Syntactic annotation guidelines for the quranic Arabic dependency treebank. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Kais Dukes and Tim Buckwalter. 2010. A dependency treebank of the Quran using traditional Arabic grammar. In *2010 the 7th International Conference on Informatics and Systems (INFOS)*, pages 1–7. IEEE.

Kais Dukes and Nizar Habash. 2010. Morphological annotation of Quranic Arabic. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2530–2536, Valletta, Malta. European Language Resources Association (ELRA).

Samhaa R El-Beltagy and Mohamed A Abdallah. 2024. Exploring retrieval augmented generation in Arabic. *Procedia Computer Science*, 244:296–307.

Tengfei Feng and Liang He. 2025. Rgr-kbqa: Generating logical forms for question answering using knowledge-graph-enhanced large language model. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3057–3070.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. *arXiv preprint arXiv:2403.14403*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Hikmat Ullah Khan, Syed Muhammad Saqlain, Muhammad Shoaib, and Muhammad Sher. 2013. Ontology based semantic search in holy Quran. *International Journal of Future Computer and Communication*, 2(6):570.

Weijie Li, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. 2025. Topology-of-question-decomposition: Enhancing large language models with information retrieval for knowledge-intensive tasks. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2814–2833.

Huanxuan Liao, Shizhu He, Yao Xu, Yuanzhe Zhang, Kang Liu, Shengping Liu, and Jun Zhao. 2024. Awakening augmented generation: Learning to awaken internal knowledge of large language models for question answering. *arXiv preprint arXiv:2403.15268*.

Doaa Mahmud, Hadeel Hajmohamed, Shamma Almentheri, Shamma Alqaydi, Lameya Aldhaheri, Ruhul Amin Khalil, and Nasir Saeed. 2025. Integrating llms with its: Recent advances, potentials, challenges, and future directions. *IEEE Transactions on Intelligent Transportation Systems*.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2023. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *arXiv preprint arXiv:2401.00396*.

OpenAI. 2023. Chatgpt. Accessed: 2025-02-15.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Hung Phan, Anurag Acharya, Sarthak Chaturvedi, Shivam Sharma, Mike Parker, Dan Nally, Ali Jannesari, Karl Pazdernik, Mahantesh Halappanavar, Sai Munikoti, et al. 2024. Rag vs. long context: Examining frontier large language models for environmental review document comprehension. *arXiv preprint arXiv*, 2407.

Wannaphong Phatthiyaphaibun, Surapon Nonesung, Peerat Limkonchotiwat, Can Udomcharoenchaikit, Jitkapat Sawatphol, Ekapol Chuangsuwanich, and Sarana Nutanong. 2024. Chie: Generative mrc evaluation for in-context qa with correctness, helpfulness, irrelevancy, and extraneousness aspects. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 154–164.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Ahmmad OM Saleh, Gökhan Tür, and Yucel Saygin. 2024. Sg-rag: Multi-hop question answering with large language models through knowledge graphs. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 439–448.

Abdul-Baquee M Sharaf and Eric Atwell. 2012. Qurana: Corpus of the Quran annotated with Pronominal Anaphora. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 130–137, Istanbul, Turkey. European Language Resources Association (ELRA).

Mohamed Ahmed Sherif and Axel-Cyrille Ngonga Ngomo. 2015. Semantic quran. *Semantic Web*, 6(4):339–345.

Marco Siino, Mariana Falco, Daniele Croce, and Paolo Rosso. 2025. Exploring llms applications in law: A literature review on current legal nlp approaches. *IEEE Access*.

Azman Ta'a, Mohd Syazwan Abdullah, Abdul Bashah Mat Ali, and Muhammad Ahmad. 2014. Themes-based classification for Al-Quran knowledge ontology. In *2014 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 89–94. IEEE.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 6.

Zheng Wang, Shu Xian Teo, Jieer Ouyang, Yongjun Xu, and Wei Shi. 2024. M-rag: Reinforcing large language model performance through retrieval-augmented generation with multiple partitions. *arXiv preprint arXiv:2405.16420*.

Qingsong Wen, Jing Liang, Carles Sierra, Rose Luckin, Richard Tong, Zitao Liu, Peng Cui, and Jiliang Tang. 2024. Ai for education (ai4edu): Advancing personalized education with llm and adaptive learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6743–6744.

Wajdi Zaghouani, Abdelati Hawwari, and Mona Diab. 2012. A pilot propbank annotation for quranic Arabic. In *Proceedings of the NAACL-HLT 2012 workshop on computational linguistics for literature*, pages 78–83.

Kuo Zhang, Xiangbin Meng, Xiangyu Yan, Jiaming Ji, Jingqian Liu, Hua Xu, Heng Zhang, Da Liu, Jingjia Wang, Xuliang Wang, et al. 2025. Revolutionizing health care: The transformative impact of large language models in medicine. *Journal of Medical Internet Research*, 27:e59069.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

## A Additional details of resources construction and annotation

This section provides more details on the developed resources, including the description of Tafsir books in Table 6, Tafaseer dataset attributes in Table 7, and QA dataset attributes in Table 8. Aysar at-Tafaseer and Tafsir Al-Muyassar were selected for building QA dataset. Aysar at-Tafaseer was chosen as it is a contemporary Tafsir, which means it can be utilized by readers without requiring deep knowledge such as that needed for the Major Tafsir books. In addition, it contains more knowledge than nearly all other contemporary Tafsir books, as shown in Table 6, where it has 3 volumes while nearly all others have just one volume. Moreover, it provides vocabulary, Tafsir (exegesis), and Hedayat for each verse or sequence of verses, not just vocabularies. Tafsir Al-Muyassar was selected as it explicitly provides the Surah Objectives (Maqasid) and the reasoning behind the naming of each Surah (chapter).

A taxonomy of the labels, classes, and the Q start attribute in the QA dataset is presented in Figure 12. In addition, the distribution of the dataset across these labels and attributes is provided in Table 10. Figures 13 (a) and (b) present word clouds that illustrate the most frequent terms appearing in the questions and answers of the QA dataset, respectively.

## B Error analysis

Figure 7 presents examples where the proposed approach returns responses that reflect an "I don't know" outcome. We refer to these as context-dependent questions. This type of task is particularly challenging because it involves two distinct subtasks: information retrieval and machine reading comprehension. Unlike standard reading comprehension, where the answer is typically found within a given passage, context-dependent questions require retrieving relevant information from external sources before reasoning can occur. Figure 8 showcases two cases where domain experts assigned a score of 5, noting that the generated responses outperformed the ground truth.

| Tafsir Name | Tafsir Name 2 | Tafsir Category | DS-I samples | DS-II samples | Size (Volumes) |
|---|---|---|---|---|---|
| الجامع لأحكام القران : Al-Jami' li Ahkam al-Quran | al-Qurtubi | Major | 6236 | 3316 | 24 |
| المختصر : Al-Mukhtasar | Tafsir Center for Quranic Studies | Contemporary | 6236 | 6169 | 1 |
| الميسر : Al-Muyassar | King Fahd Complex | Contemporary | 6236 | 6197 | 1 |
| النكت والعيون : An-Nukat wa al-'Uyoon | al-Mawardi | Compilation of Statements | 6236 | 1709 | 6 |
| التحرير التنوير : At-Tahreer wa at-Tanweer | Ibn Ashour | Language and Rhetoric | 6236 | 3969 | 24 |
| التسهيل لعلوم التنزيل : At-Tahseel li 'Uloom at-Tanzeel | Ibn Juzay | General | 6236 | 3377 | 3 |
| أيسر التفاسير : Aysar at-Tafaseer | Abu Bakr al-Jazairi | Contemporary | 6236 | 1290 | 3 |
| فتح البيان : Fath al-Bayan | Siddiq Hassan Khan | General | 6236 | 6228 | 12 |
| فتح القدير : Fath al-Qadeer | al-Shawkani | General | 6236 | 845 | 11 |
| جامع البيان : Jami' al-Bayan | Ibn Jarir al-Tabari | Major | 6236 | 3687 | 28 |
| مفاتيح الغيب : Mafatih al-Ghayb | Fakhr al-Din al-Razi | Encyclopedic Works | 6236 | 2969 | 24 |
| روح المعاني : Rooh al-Ma'ani | al-Alusi | Encyclopedic Works | 6236 | 5629 | 28 |
| تفسير القران العظيم : Tafsir al-Quran al-Azim | Ibn Kathir | Major | 6236 | 1992 | 19 |
| تيسير الكريم الرحمن : Tayseer al-Kareem ar-Rahman | al-Saadi | Contemporary | 6236 | 5910 | 4 |
| زاد المسير : Zad al-Maseer | Ibn al-Jawzi | Compilation of Statements | 6236 | 2184 | 5 |
| | | Total | 93540 | 55471 | 193 |

Table 6: Tafsir books details

| | |
|---|---|
| SURA_num | Num of SURA (num 1-114) |
| AYA_num | Number of Ayah/verse in Surah/Chapter |
| Q_num | Number from 1 to 6236 which equals to the number of Ayat/verses Qura'an |
| Question | The Question itself for example "ما هو تفسير (من شر الوسواس الخناس ) من كتاب التحرير التنوير |
| Ayah | The Ayah itself as in the Qura'an |
| Tafsir | The exegesis of Ayah in the specific Tafsir |
| Related | When an ayah's exegesis is closely linked to subsequent verses, many Tafsir books interpret a sequence of consecutive verses together rather than individually. |

Table 7: Tafaseer dataset attributes

| | |
|---|---|
| Q_num | Refers to the Question number |
| Sura_ID | Refers to the Num of SURA (num 1-114) |
| SURA_name | Refers to the Sura Name |
| Verse_Number_start | Refers to the verse start number |
| Verse_Number_End | Refers to the verse end number |
| Ayah_text | Refers to the Ayah text or a sequence of Ayat |
| Tafsir_text | Refers to the tafsir text |
| Question_text | Refers to the Question Text |
| Answer_text | Refers to the Answer text |
| Task/Type | Refers to the question is standalone or related to context (Direct or Indirect Question) |
| Remark | Additional comments |
| Source_name | The name of the main resource |
| Q_ID_in_the_Tafsir | Refers to the Question ID in the Tafaseer dataset |
| Q_Type I | For questions' Type I classifications: Descriptive, Confirm, Fatoid |
| Q_Type II | For questions' Type II classifications: refers to the answer Intent or Tafsir category |
| Q_Start_Keywords | TO DO: refer to the question keyword |

Table 8: QA dataset attributes

| التصنيف | Category (English) | Definition (English + Arabic) |
|---|---|---|
| مفردات | Vocabulary | The meaning of a Qur'anic word in its original linguistic root and within its Qur'anic context.<br>معنى اللفظة القرآنية في أصل اللغة وفي سياقها القرآني |
| تفسير | Tafsir | Explanation of the verse's meaning, either through transmitted tradition (ma'thur) or scholarly reasoning within human capability.<br>بيان معنى الآية بالمأثور أو الاجتهاد حسب الطاقة البشرية |
| أحكام | Rulings | Legal rulings in the verse (obligatory, recommended, permissible, prohibited, disliked).<br>الأحكام التكليفية في الآية (واجب، مستحب، مباح، حرام، مكروه) |
| العلوم العربية | Arabic Linguistic Sciences | Grammar & Morphology: related to word structure and syntactic function. Qur'anic Rhetoric: rhetorical style including eloquence, literary devices, and meanings.<br>النحو والصرف: كل ما تعلق بالكلمة من حيث الإعراب والبناء.<br>البلاغة القرآنية: كل ما تعلق بالآية من البيان والبديع والمعاني |
| أسباب النزول | Reasons for Revelation | Incidents or questions that triggered the revelation of a specific verse at a particular time.<br>ما نزل بشأنه قرآن وقت وقوعه كحادثة أو سؤال |
| هدايات | Guidance | Subtle or indirect meanings and insights that can be inferred from the verse.<br>ما يستنبط من الآية من معاني دقيقة وغير مباشرة |
| أسماء وأعلام | Names and Proper Nouns (Entities) | Mentions of people, places, or times.<br>الإطلاقات على الأشخاص والأماكن والأزمان |
| اعجاز | Miraculous Nature (I'jaz) | Any miraculous aspect found in the verse—linguistic, scientific, prophetic (unseen), legal, etc.<br>ما وقع في الآية من معجزة بيانية أو علمية أو غيبية أو تشريعية وغيرها |
| قصص القرآن | Qur'anic Stories | Historical events, whether from the distant past or during the time of Prophet Muhammad.<br>الأحداث التاريخية الغابرة والمعاصرة للنبي محمد صلى الله عليه وسلم |
| أمثال القرآن | Parables of the Qur'an | Analogies or comparisons made in the Qur'an between two things.<br>تشبيهات القرآن لشيء بشيء |
| مناسبات القرآن | Qur'anic Context | The relationship between verses and between surahs (chapters).<br>العلاقة بين الآيات وبين السور |
| مقاصد السور | Surah Objectives (Maqasid) | The aims and thematic goals of a surah.<br>غايات السور وأهدافها |
| أقسام القرآن | Divine Oaths | Instances where Allah swears by Himself or by His creation.<br>إقسام الله بنفسه أو بمخلوقاته |
| أخرى | Other | Abrogation: The later ruling that overrides an earlier one in the same context. Readings: Mutawatir readings within a verse. Problematic Verses: Verses that may appear contradictory on the surface.<br>الناسخ والمنسوخ: الناسخ الحكم الأخير المعمول به والمنسوخ الحكم الأول المتروك العمل به.<br>القراءات: القراءات القرآنية المتواترة.<br>مشكل القرآن: الآيات التي ظاهرها التعارض والتناقض |

Table 9: Proposed Qur'anic Categories with English and Arabic Definitions

| |
|---|
| ما هو معنى كلمة "مثلهم" في قوله تعالى "مثلهم كمثل الذي استوقد نارا"؟<br>What is the meaning of the word “mathaluhum” (“their example”) in His saying “Their example is like that of one who kindled a fire...”? |
| ما معنى "وقودها" في قوله تعالى "فإن لم تفعلوا ولن تفعلوا فاتقوا النار التي وقودها الناس والحجارة أعدت للكافرين"؟<br>What is the meaning of “waqūdahā” (“its fuel”) in His saying “But if you do not - and you will never be able to - then fear the Fire, whose fuel is men and stones , prepared for the disbelievers” |
| ماذا تعني كلمة "أعدت" في قوله تعالى "فإن لم تفعلوا ولن تفعلوا فاتقوا النار التي وقودها الناس والحجارة أعدت للكافرين"؟<br>What does “uʿiddat” (“prepared”) mean in His saying “But if you do not - and you will never be able to - then fear the Fire, whose fuel is men and stones , prepared for the disbelievers” |
| ما هو "المن والسلوى" في قوله تعالى "وأنزلنا عليكم المن والسلوى"؟<br>What are “al-mann wa-al-salwā” (“manna and quails”) in His saying “And We provided you with manna and quails.” |
| ماذا يعني استسقى في قوله تعالى "وإذ استسقى موسى لقومه"؟<br>What is the meaning of “is'tasqā” (“prayed for water”) in His saying “And [remember] when Moses prayed for water for his people...” |
| ماذا تعني كلمة "مشربهم" في قوله تعالى "قد علم كل أناس مشربهم"؟<br>What does the word “mashrabahum” (“their drinking-place”) mean in His saying “Each group of people knew its drinking-place.” |
| ما هو المقصود بكلمة "البقل" في قوله تعالى "فادع لنا ربك يخرج لنا مما تنبت الأرض من بقلها وقثائها وفومها وعدسها وبصلها"؟<br>What is meant by the word “al-baql” (“herbs”) in His saying “Pray for us to your Lord that He may bring forth for us of what the earth grows – of its herbs, its cucumbers, its garlic, its lentils, and its onions.” |

Figure 7: Examples of context-dependent questions. The context is the text used to extract the question.

| |
|---|
| السياق (Context):<br>شرح الكلمات: اسجدوا: السجود هو وضع الجبهة والأنف على الأرض، وقد يكون باتحناء الرأس دون وضعه على الأرض لكن مع تذلل وخضوع. إبليس: قيل كان اسمه الحارث ولما تكبر عن طاعة الله أبلسه الله أي أيأسه من كل خير ومسخه شيطانا. أبى: امتنع ورفض السجود لآدم. استكبر: تعاظم في نفسه فمنعه الإستكبار والحسد من الطاعة بالسجود لآدم. الكافرين: جمع كافر. من كذب بالله تعالى أو كذب بشيء من آياته أو بواحد من رسله أو أنكر طاعته. معنى الآية: يذكر تعالى عباده بعلمه وحكمته وإفضاله عليهم بقوله: (وإذ قلنا للملئكة أسجدوا لأدم...) سجود تحية وإكرام فسجدوا إلا إبليس تعاظم في نفسه وامتنع عن السجود الذي هو طاعة الله، وتحية آدم. تكبرا وحسدا لأدم في شرفه فكان بامتناعه عن طاعة الله من الكافرين الفاسقين عن أمر الله، الأمر الذي استوجب إبلاسه وطرده. |
| السؤال (Question): من هو إبليس وما الذي حدث له؟<br>Question: Who is Iblis, and what happened to him? |
| الجواب من السياق (Answer from context): قيل كان اسمه الحارث ولما تكبر عن طاعة الله أبلسه الله أي أيأسه من كل خير ومسخه شيطانا.<br>Before, his name was Al-Harith. When he disobeyed Allah and became arrogant, Allah made him despair of all good and transformed him into a devil. |
| الجواب من النظام (Answer from the system): إبليس هو أحد الجن، وقد أبى أن يسجد لآدم عندما أمره الله بذلك. قال الله تعالى: "وإذ قلنا للملائكة اسجدوا لآدم فسجدوا إلا إبليس أبى واستكبر وكان من الكافرين" (البقرة: 34). بسبب كبره ورفضه السجود، طُرد من الجنة وأصبح عدوًا لبني آدم. قال الله تعالى: "فقلنا يا آدم إن هذا عدو لك ولزوجك فلا يخرجنكما من الجنة فتشقى" (طه: 117).<br>Iblis is one of the jinn. He refused to prostrate to Adam when Allah commanded him. Allah Almighty said: “And [mention] when We said to the angels, ‘Prostrate before Adam’; so they prostrated, except for Iblees. He refused and was arrogant and became of the disbelievers.” (Al-Baqarah 2:34) And Allah said: “So We said, ‘O Adam, indeed this is an enemy to you and to your wife. Then let him not remove you from Paradise so you would suffer.’” (Ta-Ha 20:117) |

| |
|---|
| السياق (Context):<br>شرح الكلمات: اهبطوا منها جميعا: إنزلوا من الجنة إلى الأرض لتعيشوا فيها متعادين. فإما يأتينكم مني هدى: إن يجئكم من ربكم هدى: شرع ضمنه كتاب وبينه رسول. فمن اتبع هداي: أخذ بشرعي فلم يخالفه ولم يحد عنه. فلا خوف عليهم ولا هم يحزنون: جواب شرط فمن اتبع هداي، ومعناه إتباع الهدى يفضي بالعبد إلى أن لا يخاف ولا يحزن لا في الدنيا ولا في الآخرة. كفروا وكذبوا: جحدوا شرع الله، وكذبوا رسوله. أصحاب النار: أهلها الذين لا يفارقونها بحيث لا يخرجون منها. معنى الآيتين: يخبر تعالى أنه أمر آدم وحواء وإبليس بالهبوط إلى الأرض بعد أن وسوس الشيطان لهما فأكلا من الشجرة، وأعلمهم أنه إن أتاهم منه هدى فاتبعوه ولم يحيدوا عنه يأمنوا ويسعدوا فلن يخافوا ولن يحزنوا، وتوعد من كفر به وكذب رسوله فلم يؤمن ولم يعمل صالحا بالخلود في النار. |
| السؤال (Question): ما هو معنى "فلا خوف عليهم ولا هم يحزنون"؟<br>Question: What is the meaning of ‘no fear will be upon them, nor will they grieve’? |
| الجواب من السياق (Answer from context): هو جواب شرط، ومعناه إتباع الهدى يفضي بالعبد إلى أن لا يخاف ولا يحزن لا في الدنيا ولا في الآخرة.<br>It is the result of the condition: whoever follows the guidance of Allah will be free from fear and grief, in this world and the Hereafter. |
| الجواب من النظام (Answer from the system): "فلا خوف عليهم ولا هم يحزنون" تعني أن من اتبع هدى الله لن يخاف من عذاب الله في الدنيا والآخرة، ولن يحزن على ما فاته من أمور الدنيا، بل سيعيش في أمن وسعادة.<br>The meaning of “there will be no fear concerning them, nor will they grieve” is that whoever follows Allah’s guidance will not fear punishment nor grieve over worldly losses, but will live in safety and happiness. |

Figure 8: Two cases where the generated answers are better than the ground truth

| Label | Q_Type I | | | Q_Type II | | | | | | | | | | | Task | | | Q start | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Class | Confirmation | Descriptive | Factual | Guidance | Tafsir | Vocabulary | Maqasid | Stories | Rules | Entities | Ie'jaz | Language | Reasons | Parables | Oaths | Direct set I | Direct set II | Indirect | What (ما) | What (ماذا) | Who | How | Why | Is/Do | How many/much | When | Where | Others |
| Count | 60 | 3804 | 335 | 1465 | 1116 | 950 | 228 | 191 | 105 | 47 | 44 | 22 | 15 | 13 | 3 | 1178 | 2350 | 1849 | 2313 | 1330 | 215 | 198 | 84 | 18 | 17 | 4 | 4 | 16 |
| Total | 4199 | | | 4199 | | | | | | | | | | | 4199 | | | 4199 | | | | | | | | | |

Table 10: Distribution of the QA dataset across different labels and attributes. Direct Set II is an extended version of Direct Set I.

السؤال (Question): ماذا دعا نوح عليه السلام على الظالمين بعد يأسه من إيمانهم؟
**Question:** What did Noah (peace be upon him) pray against the wrongdoers after despairing of their belief?

الجواب من السياق (Ground truth answer): دعا نوح عليه السلام قائلا "ولا تزد ٱلظالمين إلا ضلالا".
Noah (peace be upon him) prayed, saying: “And, [my Lord], do not increase the wrongdoers except in error.”

الجواب من النظام (Answer from the system): قال نوح عليه السلام: "رب لا تذر على الأرض من الكافرين ديارا إنك إن تذرهم يضلوا عبادك ولا يلدوا إلا فاجرا كفارا".
Noah (peace be upon him) said: “My Lord, do not leave upon the earth from among the disbelievers an inhabitant... Indeed, if You leave them, they will mislead Your servants...”

تعليق خبير المجال (Scholar comment): لا يكفي إيراد الآية فقط (verse is not enough)

---

السؤال (Question): ماذا فعل قوم ثمود بالصخور في واديهم؟
**Question:** What did the people of Thamūd do with the rocks in their valley?

الجواب من السياق (Ground truth answer): قوم ثمود قطعوا الصخور وجعلوا منها بيوتا ومنازل لهم.
The people of Thamūd cut the rocks and made houses and dwellings out of them.

الجواب من النظام (Answer from the system): تنحتون من الجبال بيوتا فارهين.
“And you carve out of the mountains, homes, with skill.”

تعليق خبير المجال (Scholar comment): الجواب صحيح لكن من غير السورة (Accurate response but from different chapter in the same source)

---

السؤال (Question): ما هي القبلة التي أراد الله أن يحول إليها رسول الله ﷺ؟
**Question:** What is the Qiblah that Allah wanted His Messenger (peace and blessings be upon him) to turn toward?

الجواب من السياق (Ground truth answer): القبلة التي أراد الله أن يحول إليها رسول الله ﷺ هي الكعبة.
The Qiblah that Allah wanted His Messenger (peace and blessings be upon him) to turn toward is the Kaʿbah (the Sacred House).

الجواب من النظام (Answer from the system): فول وجهك شطر المسجد الحرام
“So turn your face toward al-Masjid al-Haram.”

تعليق خبير المجال (Scholar comment):
None

Figure 9: Cases rated with a score of 4 by the domain scholar indicate that the responses are correct.

السؤال (Question): كيف يُعرّف الصبر في قوله "الذين إذا أصابتهم مصيبة قالوا إنا لله وإنا إليه راجعون"؟
**Question:** How is patience defined in His saying “[Those] who, when disaster strikes them, say, ‘Indeed we belong to Allah, and indeed to Him we will return.’”?

الجواب من السياق (Answer from context): الصبر هو حمل النفس على المكروه وتوطينها على احتمال المكاره.
Patience is to force oneself to bear what is disliked and accustom it to endure hardships.

الجواب من النظام (Answer from the system): :
الصبر في قوله "الذين إذا أصابتهم مصيبة قالوا إنا لله وإنا إليه راجعون" يُعرّف بأنه تسليم لحكم الله ورضا بقضائه وقدره، حيث يعبر المؤمنون عن إيمانهم بأن ما أصابهم هو من الله، ويظهرون الاسترجاع والاعتراف بأنهم ملك لله وأنهم راجعون إليه.
Patience – in His saying “Who, when disaster strikes them, say, ‘Indeed we belong to Allah, and indeed to Him we will return.’” It is defined as submission to God's judgment and acceptance of His decree and destiny, whereby believers express their belief that what has befallen them is from God, and they show their acceptance and recognition that they belong to God and that they will return to Him.

تعليق خبير المجال (Scholar comment): جواب المصدر أدق (The source's answer is more accurate)

---

السؤال (Question): ماذا ينكر الحق سبحانه وتعالى على اليهود في قوله تعالى "أوكلما عاهدوا عهدا نبذه فريق منهم بل أكثرهم لا يؤمنون"؟
**Question:** What does the Almighty reproach the Jews for in His saying “Is it not [true] that every time they made a covenant, a party of them threw it away?”

الجواب من السياق (Answer from context): ينكر كفرهم ونبذهم للعهود والمواثيق، ويشير إلى عدم إيمان أكثرهم.
He condemns their disbelief and their casting aside of covenants and pledges, alluding to their lack of faith and trustworthiness.

الجواب من النظام (Answer from the system): ينكر الحق سبحانه وتعالى على اليهود في قوله تعالى "أوكلما عاهدوا عهدا نبذه فريق منهم بل أكثرهم لا يؤمنون" عدم وفائهم بالعهد والميثاق، حيث يبين أنهم ينقضون العهود التي عاهدوا عليها، ويظهر عدم إيمان أكثرهم.
The Almighty reproaches the Jews in His saying “...whenever they made a covenant, a party of them threw it away...” and the covenant, where He shows that they break the covenants they made, and shows the lack of faith of most of them.

تعليق خبير المجال (Scholar comment): الجواب صحيحا لكن ليس دقيقاً (The answer is correct, but not accurate)

---

السؤال (Question): ماذا أمر الله تعالى اليهود في شأن القتيل؟
**Question:** What did Allah — exalted be He — command the Jews to do regarding the murdered person?

الجواب من السياق (Answer from context): أمرهم أن يضربوا القتيل ببعض أجزاء البقرة فيحيا ويخبر عن قاتله.
He commanded them to strike the slain man with a part of the cow, so that he would revive and reveal his killer.

الجواب من النظام (Answer from the system): أمر الله تعالى اليهود في شأن القتيل بقوله: "فقلنا اضربوه ببعضها كذلك يحيي الله الموتى ويريكم آياته لعلكم تعقلون".
Allah’s command to the Jews about the slain person is in His saying: “So We said, ‘Strike him with a part of it.’...”

تعليق خبير المجال (Scholar comment): المفروض التوضيح وعدم الاكتفاء بذكر الآية (Mentioning the verse is not enough)

---

السؤال (Question): ما هو الحلال؟
**Question:** What is al-ḥalāl (the lawful)?

الجواب من السياق (Answer from context): الحلال هو ما انحلت عقدة الحظر عنه وهو ما أذن الله تعالى فيه.
Al-ḥalāl (the lawful) is that from which the knot of prohibition has been untied – in other words, whatever Allah has permitted.

الجواب من النظام (Answer from the system): الحلال هو ما أحله الله تعالى من الأطعمة والشراب، كما جاء في قوله: "يا أيها الذين آمنوا كلوا من طيبات ما رزقناكم واشكروا لله إن كنتم إياه تعبدون" (البقرة: 172).
Al-ḥalāl is whatever Allah has made lawful of food and drink, as mentioned “O you who have believed, eat from the good things which We have provided for you and be grateful to Allah if it is [indeed] Him that you worship “

تعليق خبير المجال (Scholar comment): الجواب من غير السياق ومن سورة أخرى (The answer is out of context and from another Surah)

Figure 10: Cases rated with a score of 3 by the domain scholar indicate that the responses, while not fully optimal, remain correct in terms of content and relevance.

السؤال (Question): ما هي "البعوضة" في قوله تعالى "إن الله لا يستحيي أن يضرب مثلا ما بعوضة فما فوقها"؟

**Question:** What is “the mosquito” in His saying “Indeed, Allah is not ashamed to set forth an example – [that of] a mosquito or something even smaller”?

الجواب من السياق (Answer from context):
البعوضة هي واحدة من البعوض، وهي صغار البق، وتستخدم كمثال في النص.
Al-baʻūdhah (the mosquito) is a type of gnat – a very small bug – used as an example (in this verse).

الجواب من النظام (Answer from the system): "البعوضة" هي شيء صغير يُضرب به المثل، كما جاء في قوله تعالى: "إن الله لا يستحيي أن يضرب مثلا ما بعوضة فما فوقها".
“Al-baʻūdhah” is a tiny creature given as an example, as mentioned in His saying “Indeed, Allah is not ashamed to set forth an example – [that of] a mosquito or something even smaller”.

تعليق خبير المجال (Scholar comment):
None

---

السؤال (Question): ماذا فعل آدم بعد أن علمه الله الأسماء كلها؟

**Question:** What did Adam do after Allah taught him all the names?

الجواب من السياق (Answer from context):
أنبأ آدم الملائكة بأسماء تلك المخلوقات المعروضة واحدا واحدا.
Adam informed the angels of the names of those presented objects, one by one.

الجواب من النظام (Answer from the system):
قال الله تعالى: "قال يا آدم أنبئهم بأسمائهم فلما أنبأهم بأسمائهم قال ألم أقل لكم إني أعلم غيب السماوات والأرض وأعلم ما تبدون وما كنتم تكتمون" (البقرة: 33).
Allah the Exalted said: “He said, "O Adam, inform them of their names." And when he had informed them of their names, He said, "Did I not tell you that I know the unseen [aspects] of the heavens and the earth? And I know what you reveal and what you have concealed. “

تعليق خبير المجال (Scholar comment):
المفروض التوضيح وعدم الاكتفاء بذكر الآية (Mentioning the verse is not enough)

---

السؤال (Question): ماذا يسأل الناس عن الساعة؟

**Question:** What do people ask about the Hour (Day of Judgment)?

الجواب من السياق (Answer from context):
يسألون عن الساعة أيان مرساها، أي متى وقوعها وقيامها.
They ask about the time of the Hour’s establishment – in other words, when it will occur and come to pass.

الجواب من النظام (Answer from the system):
يسأل الناس عن الساعة: (يسألك الناس عن الساعة قل إنما علمها عند الله وما يدريك لعل الساعة تكون قريبا) (الأحزاب: 63).
People ask about the Hour: “People ask you concerning the Hour. Say," Knowledge of it is only with Allah. And what may make you perceive? Perhaps the Hour is near " “

تعليق خبير المجال (Scholar comment):
الجواب من غير السياق ومن سورة أخرى (The answer is out of context and from another Surah)

Figure 11: Cases rated with a score of 2 by the domain scholar indicate that the responses, while not fully optimal, remain correct in terms of content and relevance.

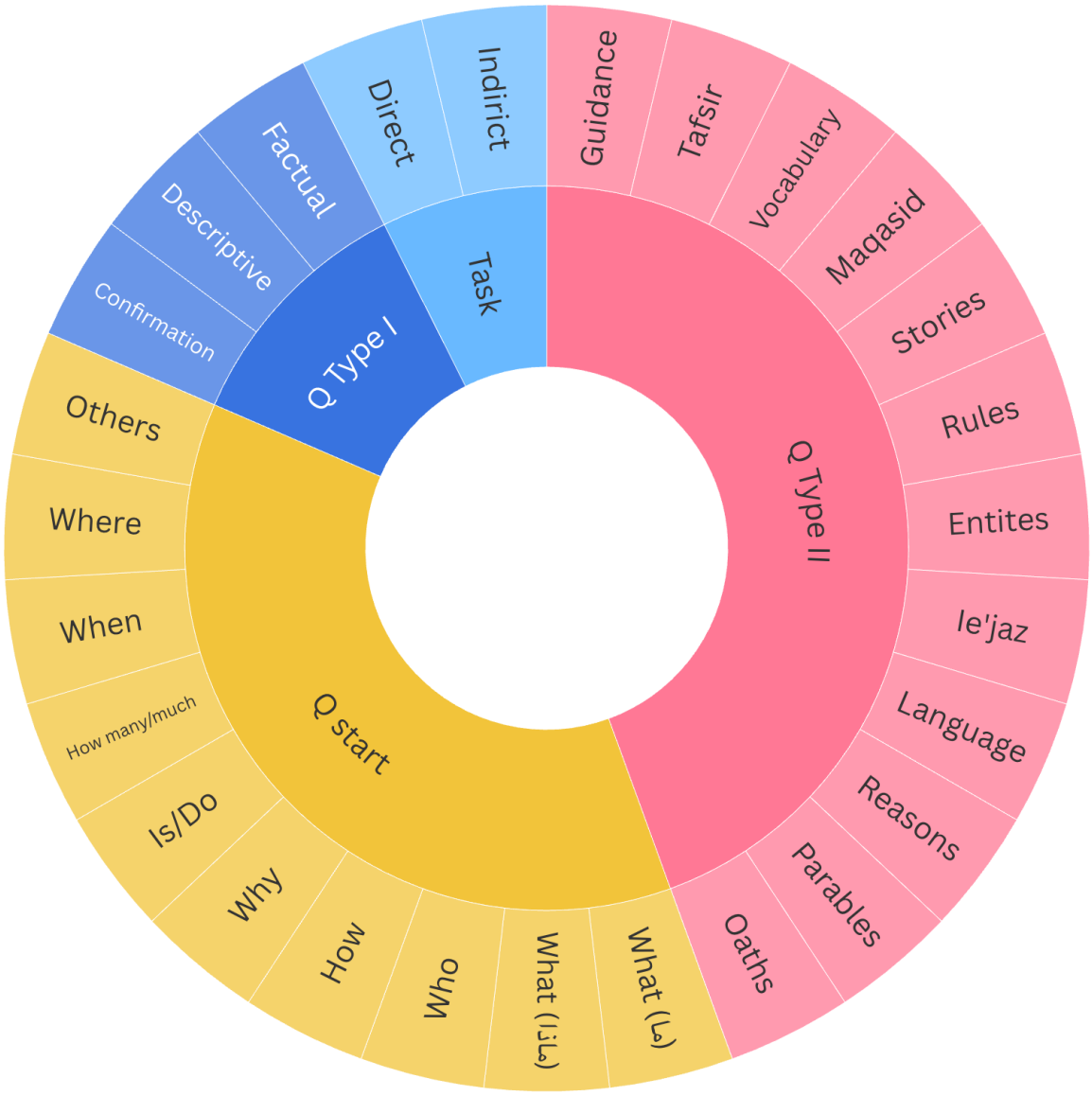


Figure 12: Taxonomy of the QA dataset showing the considered labels, classes, and the Q start attribute.

(a) Most frequent words in the Questions

(b) Most frequent words in the Answers

Figure 13: Most frequent words in the dataset: (a) Questions and (b) Answers.