IslamicEval 2025: The First Shared Task of Capturing LLMs Hallucination in Islamic Content

Hamdy Mubarak[1] , Rana Malhas[2], Watheq Mansour[3], Abubakr Mohamed[1], Mahmoud Fawzi[4], Majd Hawasly[1], Tamer Elsayed[2], Kareem Darwish[1], Walid Magdy[4]
[1] QCRI, HBKU, Qatar
[2] Qatar University, Qatar
[3] The University of Queensland, Australia
[4] School of Informatics, The University of Edinburgh, UK
hmubarak@hbku.edu.qa, telsayed@qu.edu.qa, wmagdy@inf.ed.ac.uk

Abstract

Hallucination in Large Language Models (LLMs) remains a significant challenge and continues to draw substantial research attention. The problem becomes especially critical when hallucinations arise in sensitive domains, such as religious discourse. To address this gap, we introduce IslamicEval 2025—the first shared task specifically focused on evaluating and detecting hallucinations in Islamic content. The task consists of two subtasks: (1) Hallucination Detection and Correction of quoted verses (Ayahs) from the Holy Quran and quoted Hadiths; and (2) Qur'an and Hadith Question Answering, which assesses retrieval models and LLMs by requiring answers to be retrieved from grounded, authoritative sources. Thirteen teams participated in the final phase of the shared task, employing a range of pipelines and frameworks. Their diverse approaches underscore both the complexity of the task and the importance of effectively managing hallucinations in Islamic discourse.

1 Introduction

Large Language Models (LLMs) are becoming an integral part of natural language processing applications in Arabic. Recent advancements have produced several Arabic-focused and multilingual LLMs, such as Jais (Sengupta et al., 2023), Allam (Bari et al., 2024), and Fanar (Fanar Team et al., 2025), which have shown promising results across a variety of tasks, from open-domain question answering to content generation. However, alongside these advances, a critical challenge remains unresolved, namely hallucination, i.e. the generation of text that appears plausible but is factually incorrect or fabricated (Rawte et al., 2023).

This issue is particularly sensitive in domains where accuracy and authenticity are paramount, such as religion. In the Arabic-speaking world, religious topics are not only culturally central but also frequently searched, discussed, and queried online and in social media (Abokhodair et al., 2020; Fawzi et al., 2025), often driven by a deep sense of learning, curiosity, and at times, skepticism. This has made religious discourse, particularly question answering, among the most common applications of Arabic LLMs, both directly and indirectly.

Among religious sources, the Qur'an and Hadith literature stand out due to their sacred status and the high expectations of precision when they are quoted or referenced. The Qur'an, regarded as the ultimate and divine word of Allah, serves as the foundation of Islamic teachings. In tandem, Hadith encompasses the sayings, deeds, and implied approvals of the Prophet Muhammad (Peace Be Upon Him), serving in part as a practical illustration of Qur'anic teachings (Musallam, 2022). Given this sanctity, LLM hallucination in Islamic content poses serious risks: it can lead to misattributions, paraphrased verses falsely labeled as genuine, or entirely fabricated Hadiths (Fawzi et al., 2026), raising serious ethical, theological, and social concerns. Such hallucinations can unintentionally propagate misinformation or be exploited for disinformation, undermining trust in AI technologies and amplifying harm.

To address this gap, we have organized the IslamicEval 2025 shared task at ArabicNLP 2025,[1] which consists of two subtasks: (1) Hallucination Detection and Correction, and (2) Qur'an and Hadith Question Answering (QA). The first subtask focuses on detecting and correcting hallucinations in Qur'an and Hadith content within Arabic LLM-generated text. To our knowledge, it is the first task of its kind to target semantic and source-faithful evaluation of generated religious text. The second subtask is primarily intended to provide authentic QA benchmarks and standardized evalu-

[1] https://sites.google.com/view/islameval-2025/home

Proceedings of The Third Arabic Natural Language Processing Conference, pages 480–493
November 8-9, 2025 ©2025 Association for Computational Linguistics

ation testbeds for question answering models and systems on the Holy Qur'an and Hadith. Such benchmarks and testbeds are of paramount importance in the era of Generative AI, as they constitute a first line of defense against hallucination.

To this end, we aim to bring the Arabic NLP community together to develop robust systems for hallucination detection, localization, verification, and correction, as well as question answering on the Qur'an and Hadith:

Detection Determine whether a generated Arabic text contains a claimed Qur'anic verse (Ayah) or a Hadith. This involves building systems capable of semantic matching against canonical sources, accounting for variations in phrasing / paraphrasing.

Span Identification : Identify the exact span within the text corresponding to the claimed verse or Hadith. This requires models to accurately delimit religious content from surrounding context, often under noisy or stylistically varied conditions.

Verification Assess whether the detected quote is accurate—i.e., whether it exists in the authentic sources (e.g., Qur'an text or recognized Hadith collections) and is correctly cited. This step combines information retrieval with textual entailment techniques.

Correction If a quote is found to be inaccurate or hallucinated, reproduce the correct version, being the closest matching verse or Hadith if it exists, or indicate it is fabricated if no close match is found.

Passage Retrieval Given a free-text question in Modern Standard Arabic (MSA), a collection of Qur'anic passages covering the Holy Qur'an, and a collection of Hadiths from Sahih Al-Bukhari, the system must retrieve a ranked list of up to 20 answer-bearing passages—Qur'anic passages or Hadiths—that may contain one or more answers to the question, drawn from both collections.

This task raises unique NLP challenges:

- Fuzzy matching and paraphrase detection for verses and Hadiths expressed in non-standard forms;
- Robustness to stylistic variation and dialectal influence in generated text;
- Semantic grounding in authoritative religious corpora;
- Trust-sensitive evaluation, where false positives and false negatives have different and context-dependent implications.

We believe this shared task will catalyze research in faithful generation, hallucination detection, and knowledge-grounded NLP—not only for Arabic but as a reference for similar tasks in other languages and sensitive domains. It also supports the broader goal of responsible AI, promoting the development of LLMs that are not only fluent but also accurate, culturally aware, and ethically aligned.

2 IslamicEval Task 1: Hallucination Detection and Correction

Task 1 of the IslamicEval shared task addresses the detection and correction of hallucinations in LLM outputs that reference Qur'anic verses and Prophetic Hadiths. It is organized into three subtasks: identifying the intended references, validating their correctness against authoritative sources, and providing corrected versions when errors are found. The following subsections present the task setup, datasets, annotation guidelines, evaluation metrics, and results of participating systems.

2.1 Task Description

1. Subtask A - Identification of Intended verses (Qur'anic Ayahs) and Hadiths (Prophetic sayings) Given an LLM-generated response, participants will determine the spans of the "intended", since they might be inaccurate, verses and Hadiths in the text. Spans are represented by the character indexes, e.g. from character 0 to character 72 (inclusive). Evaluation is based on span precision and recall (macro-averaged F1 score). References to verse number and Hadith narrators and punctuations are ignored in this version.

2. Subtask B - Validation of content accuracy For each identified verse and Hadith, participants will categorize them as correct or incorrect based on established Islamic references. Evaluation is based on accuracy. Incorrect diacritics will be considered as mistakes.

3. Subtask C - Correction of Erroneous Content. Participants will provide corrected versions for any incorrectly quoted verse or Hadith, ensuring fidelity to the original sources. Evaluation is based on accuracy. Note that complete verse(s) from the Qur'an and complete Hadiths are expected. Writing and diacritics should be obtained from the shared Qur'an and Hadith sources.

2.2 Dataset

Starting from Qur'an QA 2023 dataset (n=251) that covers a broad range of topics including Fiqh, Tafsir, and Islamic teachings, a training (174), validation (25), and test (52) sets were created.

Six LLMs were prompted with the questions, with the prompt explicitly asking the models to cite Qur'anic and Hadith evidence in their responses (see Appendix B for the prompt). The question–output pairs, along with anonymized model IDs, were stored in XML format. The models used could be seen in Table 1. The LLM choice aimed to balance Arabic-focused models with state-of-the-art multilingual ones.

2.3 Annotation Setup and Guidelines

The generated answers were manually annotated by domain experts using the Label Studio platform (Tkachenko et al., 2020-2025). A separate annotation task was created for each question–response pair. Annotators were instructed to highlight every span containing an intended Qur'anic verse or Hadith and assign it one of four labels: Correct Qur'an, Incorrect Qur'an, Correct Hadith, or Incorrect Hadith. For each span marked as incorrect, annotators were required to either provide the corrected text or write "خطأ" (Wrong) if no valid correction existed. Figure 1 shows an example of an annotated output.

All annotators were experts in Islamic studies to ensure accuracy and reliability. Qur'anic references were standardized to the Uthmani script, while Hadith references were cross-checked against the six authoritative collections (الكتب الستة) including Sahih al-Bukhari and Sahih Muslim. The annotation guidelines emphasized precise span boundary selection and careful evaluation of correctness. The full annotation guidelines are available in Appendix C.

2.4 Evaluation Measures

Each subtask in Task 1 was evaluated using metrics suited to its specific objectives:

Subtask A (Identification) Performance was measured using the **macro-averaged F1** score, computed at the character level by classifying each character in the response string as belonging to a Qur'anic verse, a Hadith, or neither. Macro-averaged F1 is well-suited for this subtask because the data is highly imbalanced, with far fewer Ayah and Hadith spans than "neither", so accuracy alone would be misleading. Character-level F1 ensures that partial matches and boundary errors are fairly captured, while macro-averaging gives equal weight to each class rather than letting the dominant class overwhelm the results.

Subtask B (Validation) **Accuracy** was used as the evaluation metric, defined as the proportion of correctly assigned labels (Correct or Incorrect) over the total number of labels.

Subtask C (Correction) **Accuracy** was used, defined as the proportion of corrected outputs that exactly matched the corresponding ground truth over the total number of corrections. Strict accuracy was adopted for this subtask because even minor deviations - such as omitted words or altered diacritics - can substantially affect the meaning of a Qur'anic verse or Hadith. To avoid penalizing superficial formatting inconsistencies, both reference and system outputs were preprocessed prior to evaluation by removing default diacritics (e.g., sukun).

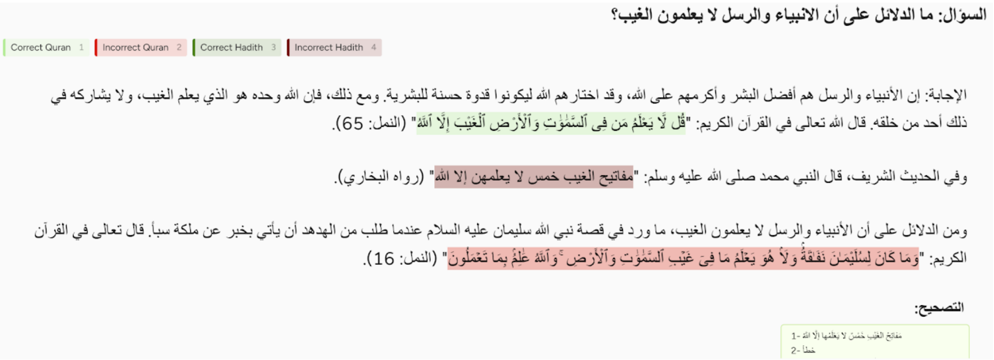
2.5 Task Setup

The dataset comprises 1,506 annotated answers (251 questions × 6 models). The development set corresponds to the original Qur'an QA 2023 dev set, consisting of 10% of the generations (n=150), yielding 50 annotated answers per subtask A, B and C. Similarly, the test set corresponds to the Qur'an QA 2023 test set, consisting of 20% of the questions (n=312), yielding 104 annotated answers per subtask. All annotations for development and test sets were manually reviewed and refined through multiple iterations (with the help of validation scripts) to ensure accuracy before release. A revised version of the training set (n=1,044) will be released in the future.

To facilitate participation, we hosted three competitions on CodaBench[2]. The development sets, along with the Qur'an and Hadith texts in JSON format (see Appendix D for a sample), were made publicly available. Participants were required to rely exclusively on the provided data.

The competition was launched on June 16, with test sets released on July 29, and final submissions closed on August 8. The shared task drew strong engagement, with 20 participants in Subtask 1A (87 submissions), 16 participants in Subtask 1B (41 submissions), and 15 participants in Subtask

[2] https://www.codabench.org/competitions/9820/, https://www.codabench.org/competitions/9822/, and https://www.codabench.org/competitions/9824/



السؤال: ما الدلائل على أن الانبياء والرسل لا يعلمون الغيب؟

الإجابة: إن الأنبياء والرسل هم أفضل البشر وأكرمهم على الله، وقد اختارهم الله ليكونوا قدوة حسنة للبشرية. ومع ذلك، فإن الله وحده هو الذي يعلم الغيب، ولا يشاركه في ذلك أحد من خلقه. قال الله تعالى في القرآن الكريم: "قُل لَّا يَعْلَمُ مَن فِى ٱلسَّمَٰوَٰتِ وَٱلْأَرْضِ ٱلْغَيْبَ إِلَّا ٱللَّهُ" (النمل: 65).

وفي الحديث الشريف، قال النبي محمد صلى الله عليه وسلم: "مفاتيح الغيب خمس لا يعلمهن إلا الله" (رواه البخاري).

ومن الدلائل على أن الأنبياء والرسل لا يعلمون الغيب، ما ورد في قصة نبي الله سليمان عليه السلام عندما طلب من الهدهد أن يأتي بخبر عن ملكة سبأ. قال تعالى في القرآن الكريم: "وَمَا كَانَ لِسُلَيْمَٰنَ نَفَقَةٌ وَلَا هُوَ يَعْلَمُ مَا فِى غَيْبِ ٱلسَّمَٰوَٰتِ وَٱلْأَرْضِ ۚ وَٱللَّهُ عَلِمٌ بِمَا تَعْمَلُونَ" (النمل: 16).

التصحيح:

1- مفاتح الغيب خمس لا يعلمها إلا الله
2- خطأ

Figure 1: Example of an annotated LLM response. Question translation: “What is the evidence that the prophets and messengers do not know the unseen?”. Spans highlighted in light green and dark green represent correct Qur’anic verses and Hadiths, respectively. Spans highlighted in light red and dark red represent incorrect Qur’anic verses and Hadiths. Corrections for each incorrect span are listed in the box at the bottom.

Model	#Answers	Avg Word Len	#Ayahs	Correct%	#Hadiths	Correct%
ALLaM-7B-Instruct-preview	251	297	1104	**84.06**	654	59.33
In-house fine-tuned Gemma-2-9b	251	153	548	65.33	372	38.17
In-house fine-tuned Gemma-2-9b + RAG*	246	742	1634	82.01	467	**63.17**
Jais-13B-Chat	251	46	151	41.72	83	26.51
Qwen3-8B	251	202	379	6.86	55	1.82
Llama-3.1-8B-Instruct	251	230	797	4.77	564	0.53

Table 1: Performance of models during dataset curation where LLM responses were annotated by experts. The model families include ALLam (Bari et al., 2024), Jais (Sengupta et al., 2023), Llama-3 (Grattafiori et al., 2024), Qwen3 (Qwen Team, 2025), in addition to fine-tuned versions of Gemma-2 (Gemma Team, 2024) developed in-house by the Fanar team (Fanar Team et al., 2025). Model marked with * failed to give answer to some questions. Best results in generating correct verses and Hadiths are written in bold.

1C (59 submissions). Since some teams submitted under multiple individual accounts, this amounted to five distinct teams overall, listed in Table 2.

Teams were allowed to submit an unlimited number of runs; however, only their most recent three submissions were considered for evaluation. Results were provided to participants on these final three runs, and they were requested to describe them in their system description papers.

2.5.1 Participating Teams

Burhan AI (Al Adel et al., 2025): For Subtask 1A, the authors fine-tuned a domain-adapted LLM (gpt-4.1-mini) for hallucination span detection, incorporating synthetic augmentation, diacritic variation, and morphological normalization to enhance robustness (F1 = 87.10%). In addition, they explored an agentic approach with specialized tools (OpenAI’s code interpreter), achieving an **F1 of 90.06% (Best in subtask 1A)**. For Subtasks 1B (Accuracy = 88.60%) and 1C (Accuracy = 66.56%), they developed a multistage hierarchical correction pipeline that combined exact, normalized, fuzzy, and semantic matching with prompt-driven repair to ensure canonical alignment and diacritic fidelity.

HUMAIN (Omayrah et al., 2025): HUMAIN addressed Subtask 1 using a three-stage LLM-based pipeline grounded in the TANL framework (Paolini et al., 2021). For Subtask 1A, they modeled span detection as sequence-to-sequence annotation with bracket-based tags aligned via dynamic programming, with an alternative guided decoding setup through vLLM producing structured JSON. This achieved a strong 87.20% F1 on the test set. In Subtask 1B, validation combined retrieval-based similarity with strict substring matching, using higher thresholds for Qur’anic verses and exact substring logic for Hadith, yielding 86.14% accuracy. Finally, Subtask 1C correction employed multi-stage matching - exact, LCS alignment, and semantic reranking with bge-reranker-v2-m3 - reaching 68.18% accuracy, though rare Hadiths and implicit references remained challenging.

TCE (ElKoumy et al., 2025): The TCE team tackled Subtasks 1A and 1B of IslamicEval 2025 using few-shot prompting with state-of-the-art LLMs such as Qwen-235B (MoE) and GPT-4o. For span detection (1A), they used prompts enriched with trigger words and citation patterns, as well as chunking, and fuzzy matching to identify Qur'anic and Hadith content, achieving a macro-F1 of 86.11% on the test set. For 1B, they designed a retrieval-augmented pipeline: Qur'anic spans were retrieved with word-level fuzzy voting and Hadith with character n-gram TF-IDF, then verified by LLMs with strict word-for-word rules for Qur'an and lenient matching for Hadith. This hybrid system, enhanced with an efficient early-exit strategy, scored **89.82% accuracy (Best in Subtask 1B)** on the test set, with GPT-4o outperforming Gemma variants and showing improved performance when diacritics were preserved.

Isnad AI (Elden, 2025): The authors proposed a rule-based preprocessing and augmentation pipeline that systematically transforms raw religious texts into a large-scale, high-quality training corpus. The pipeline embeds processed Qur'anic verses and Hadiths into contextual templates. A set of common prefixes (eg. قال الله تعالى "God Almighty said") and suffixes (eg. رواه البخاري "Narrated by Al-Bukhari") was applied, and each unique instance was expanded into multiple training examples by randomly combining it with different prefixes, suffixes, and neutral connecting sentences. The authors reported that synthetic data generation using AraGPT was less effective.

2.5.2 Task 1 Results

Table 2 shows the results for Task 1 across the three subtasks. Participating systems employed a wide range of approaches to detect the intended Qur'anic verses and Hadiths, including LLMs such as GPT-4 and Qwen, as well as fuzzy matching with search engines and rule-based techniques. Our evaluation shows a significant performance gap: the rule-based approach (e.g. Isnad AI) lag considerably behind LLM-based systems, highlighting the inherent difficulty of this task. Lists of rules and patterns are insufficient to capture the diverse styles and degrees of distortion found in LLM generations.

We also observe that detecting the textual boundaries of verses and Hadiths is substantially easier than correcting them, underscoring the fact that hallucinations in LLM outputs are often non-trivial to repair. Recovery from hallucinated references remains highly challenging, suggesting that hallucination prevention should occur during generation, e.g. via RAG to constrain outputs to authentic sources, instead of post-hoc correction.

Finally, we find that models perform consistently better on Qur'anic verses than on Hadiths (either by the participating teams or the LLMs in Table 1). This can be attributed to the relative size and structure of the corpora: the Qur'an is comparatively compact and standardized, whereas Hadith collections (e.g., the six authoritative books) are far larger and more variable, making hallucination detection and correction more complex.

3 IslamicEval Task 2: Qur'an and Hadith Question Answering

In this section, we define Task 2, its dataset, annotation and evaluation setup, and the measures used to rank systems. Results are presented and discussed before concluding with an overview of the approaches adopted by the systems of participating teams (with accepted description papers).

3.1 Task Description

The Qur'an and Hadith QA subtask is a continuation of the Qur'an QA 2022[3] (Malhas et al., 2022) and Qur'an QA 2023[4] (Malhas et al., 2023) Shared Tasks. This year's subtask introduces Hadith as an additional Islamic resource for answering questions, marking the first such inclusion in the task's history. We define the task as follows: Given a free-text question in Modern Standard Arabic (MSA), a collection of Qur'anic passages covering the Holy Qur'an, and a collection of Hadiths from Sahih Al-Bukhari, systems are required to retrieve a ranked list of up to 20 answer-bearing Qur'anic passages or Hadiths (i.e., that potentially contain the answer(s) to the given question) drawn from these two collections. Questions may be factoid or non-factoid. Example questions with answer-bearing Qur'anic passages and Hadith *matns* are exhibited in Figures 2 and 3, respectively. The *matn* refers to the core text of the Hadith itself, while the *isnad* outlines the chain of narrators who convey and authenticate the *matn* (Azmi et al., 2019).

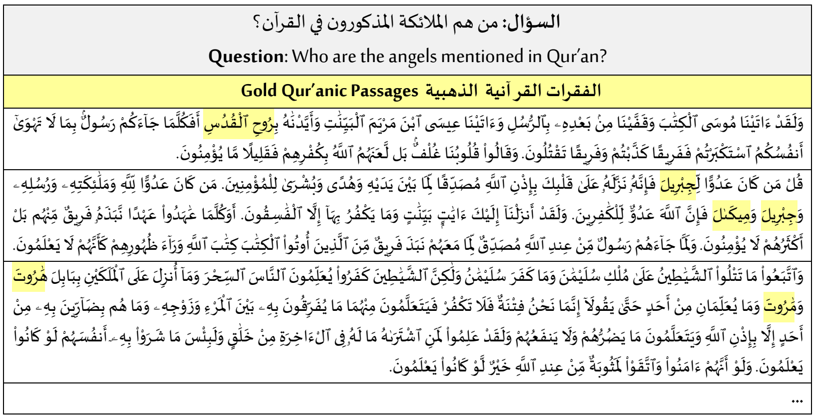
To better reflect real-world conditions and make the task more challenging, we included questions that lack answers in the Qur'an and/or Sahih Al-Bukhari. We label a question *zero-answer* only when neither source contains an answer. For such questions, the ideal system returns no result; otherwise, it should output a ranked list of up to 20 answer-bearing Qur'anic passages or Hadith *matns*.

[3] https://sites.google.com/view/quran-qa-2022

[4] https://sites.google.com/view/quran-qa-2023

	Subtask 1A				Subtask 1B				Subtask 1C			
Team Name	**F1%**	**F1-Q**	**F1-H**	**Rank**	**Acc%**	**Acc-Q**	**Acc-H**	**Rank**	**Acc%**	**Acc-Q**	**Acc-H**	**Rank**
Burhan AI	**90.06**	89.47	86.99	**1**	88.60	89.45	86.63	2	66.56	65.70	67.65	2
HUMAIN	87.20	86.61	85.11	2	86.14	90.20	76.74	3	**68.18**	62.21	75.74	**1**
TCE	86.11	86.60	80.51	3	**89.82**	91.21	86.63	**1**	-	-	-	-
Isnad AI	66.97	72.39	48.94	4	-	-	-	-	-	-	-	-
mucAI*	44.88	46.24	29.80	5	-	-	-	-	-	-	-	-

Table 2: Task 1 results across subtasks. Teams are ranked per subtask. The majority baseline in Subtask 1A is **36.17%** Macro-Avg. F1 (assuming no Ayah or Hadith), in Subtask 1B is **70.00%** (assuming all Ayahs and Hadiths are correct), and in Subtask 1C is **67.52%** (assuming all errors are not correctable). For Subtask 1A we report the overall Macro-averaged F1 (F1) and for Qur'an (F1-Q) and Hadith (F1-H) individually. Similarly, for Subtasks 1B and 1C we report the accuracies Acc, Acc-Q, and Acc-H. Teams marked with * did not submit a system paper.



السؤال: من هم الملائكة المذكورون في القرآن؟

Question: Who are the angels mentioned in Qur'an?

Gold Qur'anic Passages الفقرات القرآنية الذهبية

وَلَقَدْ ءَاتَيْنَا مُوسَى ٱلْكِتَٰبَ وَقَفَّيْنَا مِنۢ بَعْدِهِۦ بِٱلرُّسُلِ وَءَاتَيْنَا عِيسَى ٱبْنَ مَرْيَمَ ٱلْبَيِّنَٰتِ وَأَيَّدْنَٰهُ بِرُوحِ ٱلْقُدُسِ أَفَكُلَّمَا جَآءَكُمْ رَسُولٌۢ بِمَا لَا تَهْوَىٰٓ
أَنفُسُكُمُ ٱسْتَكْبَرْتُمْ فَفَرِيقًا كَذَّبْتُمْ وَفَرِيقًا تَقْتُلُونَ. وَقَالُوا۟ قُلُوبُنَا غُلْفٌۢ بَل لَّعَنَهُمُ ٱللَّهُ بِكُفْرِهِمْ فَقَلِيلًا مَّا يُؤْمِنُونَ.

قُلْ مَن كَانَ عَدُوًّا لِّجِبْرِيلَ فَإِنَّهُۥ نَزَّلَهُۥ عَلَىٰ قَلْبِكَ بِإِذْنِ ٱللَّهِ مُصَدِّقًا لِّمَا بَيْنَ يَدَيْهِ وَهُدًى وَبُشْرَىٰ لِلْمُؤْمِنِينَ. مَن كَانَ عَدُوًّا لِّلَّهِ وَمَلَٰٓئِكَتِهِۦ وَرُسُلِهِۦ
وَجِبْرِيلَ وَمِيكَىٰلَ فَإِنَّ ٱللَّهَ عَدُوٌّ لِّلْكَٰفِرِينَ. وَلَقَدْ أَنزَلْنَآ إِلَيْكَ ءَايَٰتٍۭ بَيِّنَٰتٍ وَمَا يَكْفُرُ بِهَآ إِلَّا ٱلْفَٰسِقُونَ. أَوَكُلَّمَا عَٰهَدُوا۟ عَهْدًا نَّبَذَهُۥ فَرِيقٌ مِّنْهُم بَلْ
أَكْثَرُهُمْ لَا يُؤْمِنُونَ. وَلَمَّا جَآءَهُمْ رَسُولٌ مِّنْ عِندِ ٱللَّهِ مُصَدِّقٌ لِّمَا مَعَهُمْ نَبَذَ فَرِيقٌ مِّنَ ٱلَّذِينَ أُوتُوا۟ ٱلْكِتَٰبَ كِتَٰبَ ٱللَّهِ وَرَآءَ ظُهُورِهِمْ كَأَنَّهُمْ لَا يَعْلَمُونَ.

وَٱتَّبَعُوا۟ مَا تَتْلُوا۟ ٱلشَّيَٰطِينُ عَلَىٰ مُلْكِ سُلَيْمَٰنَ وَمَا كَفَرَ سُلَيْمَٰنُ وَلَٰكِنَّ ٱلشَّيَٰطِينَ كَفَرُوا۟ يُعَلِّمُونَ ٱلنَّاسَ ٱلسِّحْرَ وَمَآ أُنزِلَ عَلَى ٱلْمَلَكَيْنِ بِبَابِلَ هَٰرُوتَ
وَمَٰرُوتَ وَمَا يُعَلِّمَانِ مِنْ أَحَدٍ حَتَّىٰ يَقُولَآ إِنَّمَا نَحْنُ فِتْنَةٌ فَلَا تَكْفُرْ فَيَتَعَلَّمُونَ مِنْهُمَا مَا يُفَرِّقُونَ بِهِۦ بَيْنَ ٱلْمَرْءِ وَزَوْجِهِۦ وَمَا هُم بِضَآرِّينَ بِهِۦ مِنْ
أَحَدٍ إِلَّا بِإِذْنِ ٱللَّهِ وَيَتَعَلَّمُونَ مَا يَضُرُّهُمْ وَلَا يَنفَعُهُمْ وَلَقَدْ عَلِمُوا۟ لَمَنِ ٱشْتَرَىٰهُ مَا لَهُۥ فِى ٱلْءَاخِرَةِ مِنْ خَلَٰقٍ وَلَبِئْسَ مَا شَرَوْا۟ بِهِۦٓ أَنفُسَهُمْ لَوْ كَانُوا۟
يَعْلَمُونَ. وَلَوْ أَنَّهُمْ ءَامَنُوا۟ وَٱتَّقَوْا۟ لَمَثُوبَةٌ مِّنْ عِندِ ٱللَّهِ خَيْرٌ لَّوْ كَانُوا۟ يَعْلَمُونَ.

...

Figure 2: An example question with some of its gold (answer-bearing) Qur'anic passages. Answers are highlighted.

3.2 Dataset

In this section, we introduce the test collections used for the Qur'an-Hadith QA subtask (or QH-QA for short). In information retrieval, a *test collection* consists of a document collection[5] (here, the Holy Qur'an and Sahih al-Bukhari), a set of queries (questions), and their relevance judgments (Lin and Katz, 2006) (i.e., the gold answers or, in our case, the passages that contain them).

The document collections used for this subtask comprise the Qur'anic Passage collection (QPC) (Swar, 2007; Malhas, 2023), and Sahih Al-Bukhari collection. QPC was developed by topically segmenting the 114 Qur'anic chapters using the Thematic Holy Qur'an (Swar, 2007)[6], a printed edition that clusters the chapter verses into topics. This segmentation resulted in a total of 1,266 passages. For the Sahih Al-Bukhari collection, we used the Tajreed Sarih version (Al-Zubaidi, 2009) that comprises 2,254 Hadiths, from which redundant Hadiths, Arabic commentary, and chain of narrators (except the last) have been excluded. However, Al-Zubaidi may repeat a Hadith if there was a beneficial addition in a later occurrence. Moreover, only authenticated Hadiths with a continuous chain of narrators are included in this collection. The digital version of this book[7] is available on `shamela.ws`, a project for collecting classical Arabic books. We contacted an Islamic scholar who provided us with an offline version of the book, which we parsed later to generate the final JSON lines *(.jsonl)* format[8].

For the questions, we used the 250 questions of *AyaTEC v1.2* dataset (Malhas and Elsayed, 2020; Malhas et al., 2023), split into training (84%) and development (16%) sets. The relevance judgments for these questions are provided over the QPC **only**.

For the test dataset, we developed a new set of 71 questions, 23 of which are paraphrased versions of natural user prompts drawn from usage logs of the Fanar Arabic LLM (Fanar Team et al., 2025). Only 51 questions were used to evaluate the systems of participating teams. The relevance judgments for all 71 questions over the Qur'anic Passage col-

[5] The term "document collection" or "collection" refers to a corpus or dataset (Lin et al., 2021); we use these terms interchangeably.

[6] `https://archive.org/details/Quran27/page/n13/mode/2up`

[7] `shamela.ws/book/96283/`

[8] `https://gitlab.com/bigirqu/quran-hadith-qa-2025`

السؤال: ما هي علامات يوم القيامة؟ **Question**: What are the signs of the Day of Judgment?
Gold Hadith Matns (Texts) متون (نصوص) الأحاديث الذهبية
أَتَيْتُ النَّبِيَّ صَلَّى اللهُ عليه وسلَّمَ في غَزْوَةِ تَبُوكَ وهو في قُبَّةٍ مِن أَدَمٍ، فَقالَ: اعْدُدْ سِتًّا بيْنَ يَدَيِ السَّاعَةِ: مَوْتِي، ثُمَّ فَتْحُ بَيْتِ المَقْدِسِ، ثُمَّ مُوتَانٌ يَأْخُذُ فِيكُمْ كَقُعَاصِ الغَنَمِ، ثُمَّ اسْتِفَاضَةُ المَالِ حتَّى يُعْطَى الرَّجُلُ مِئَةَ دِينَارٍ فَيَظَلُّ سَاخِطًا، ثُمَّ فِتْنَةٌ لا يَبْقَى بَيْتٌ مِنَ العَرَبِ إلَّا دَخَلَتْهُ، ثُمَّ هُدْنَةٌ تَكُونُ بيْنَكُمْ وبيْنَ بَنِي الأصْفَرِ، فَيَغْدِرُونَ فَيَأْتُونَكُمْ تَحْتَ ثَمَانِينَ غَايَةً، تَحْتَ كُلِّ غَايَةٍ اثْنَا عَشَرَ أَلْفًا.
لَا تَقُومُ السَّاعَةُ حتَّى يُقْبَضَ العِلْمُ، وتَكْثُرَ الزَّلَازِلُ، ويَتَقَارَبَ الزَّمَانُ، وتَظْهَرَ الفِتَنُ، ويَكْثُرَ الهَرْجُ - وهو القَتْلُ القَتْلُ - حتَّى يَكْثُرَ فِيكُمُ المَالُ فَيَفِيضَ.
لَأُحَدِّثَنَّكُمْ حَدِيثًا سَمِعْتُهُ مِن رَسولِ اللَّهِ صلَّى اللهُ عليه وسلَّم، لا يُحَدِّثُكُمْ به أحَدٌ غيرِي: سَمِعْتُ رَسولَ اللَّهِ صلَّى اللهُ عليه وسلَّم يقولُ: إنَّ مِن أشْرَاطِ السَّاعَةِ أنْ يُرْفَعَ العِلْمُ، ويَكْثُرَ الجَهْلُ، ويَكْثُرَ الزِّنَا، ويَكْثُرَ شُرْبُ الخَمْرِ، ويَقِلَّ الرِّجَالُ، ويَكْثُرَ النِّسَاءُ حتَّى يَكونَ لِخَمْسِينَ امْرَأَةً القَيِّمُ الوَاحِدُ.
...

Figure 3: An example question with some of its gold (answer-bearing) Hadith *matns* from Sahih Al-Bukhari.

lection and the Sahih Al-Bukhari collection were conducted by Qur'an and Hadith specialists, as described in the next section.

We note that the relevance judgments for the test dataset will not be released. Nevertheless, future run submissions for evaluation on this dataset may be obtained by contacting one of the organizers. All datasets and test collections are publicly available in the official Qur'an-Hadith QA repository.[9]

3.3 Annotation Setup and Guidelines

Two annotation guidelines and rubrics, with illustrative examples, were meticulously developed for the Qur'an and Hadith specialists, labeling potential answer-bearing Quranic passages and Hadith *matns*. Each candidate passage or *matn* was annotated as either having a *direct answer*, an *indirect answer*, *relevant but no answer*, or *irrelevant* to a given question. The Arabic definitions for these labels are in Figures 7 and 8 (Appendix E).

Moreover, Arabic web-based GUIs were developed in line with these guidelines and rubrics to streamline annotation and gather specialist-suggested passages and *matns* potentially containing *direct* or *indirect answers* to the given question.

Retrieval and pooling: We constructed a pooled candidate set per question by taking the deduplicated union of top-k results from multiple retrieval models. The pooled candidates were re-ranked using GPT-4.1 and GPT-4.1-mini. We applied a cutoff at the top-20 items after re-ranking to define the Round 1 candidate set presented to annotators.

Annotation rounds and coverage. In Round 1, specialists annotated the re-ranked top-20 candidates per question (across both collections). In Round 2, they annotated additional candidates that they had proposed during Round 1. Round 3 took place after the test-set submission phase closed, during which specialists annotated a pooled candidate set per question, formed as the deduplicated union of the top-k responses from the best submitted run of each team, after excluding candidates with a frequency less than 2.

Each candidate passage/matn in Round 1 was independently labeled by three Qur'an specialists (for Qur'anic passages) or three Hadith specialists (for Hadith *matns*). Additional candidates in Rounds 2 and 3 were likewise independently labeled by three domain specialists.

Aggregation and agreement: We applied majority voting across the three domain specialists; ties were resolved by a fourth. Despite our careful design and piloting of the annotation rubrics, inter-annotator agreement was fair: Fleiss' kappa was 0.283 among Qur'an specialists and 0.235 for Hadith.

Label normalization: Consistent with the training and development sets, final test-set relevance judgments over both collections were binarized: only passages/matns containing a *direct answer* received a positive label (1); all others received 0.

3.4 Evaluation Setup

We chose Codabench[10] as a platform for hosting our subtask, similar to Task 1. We used trec_eval tool[11] to compute the evaluation metrics. We made our training and development sets available during the development phase and allowed each team to run 100 submissions on the development set and receive scores from the system. Our evaluation script was also made available for local evaluation. During the testing phase, we allowed teams to submit 13 submissions; however, we stated that only the last 3 submitted runs would be considered for evaluation.

[9] https://gitlab.com/bigirqu/quran-hadith-qa-2025

[10] codabench.org/competitions/9939/

[11] github.com/usnistgov/trec_eval

Team	MAP@10	MAP_Q@5	MAP_H@5
Burhan	**0.3351**	**0.3389**	**0.3876**
BurhanAI	0.2807	0.3257	0.2386
ThinkDrill	0.2296	0.2623	0.215
NUR	0.1809	0.2334	0.1923
BayaNet*	0.1504	0.2064	0.224
MSA*	0.1185	0.1674	0.0685
Maged*	0.0332	0.0887	0.0457
CISRG*	0.0116	0.0294	0.0128

Table 3: Results of Task 2 showing the best run per team ranked by MAP@10. Teams with * did not submit a system paper.

3.4.1 Evaluation Measures

For the classical ranked retrieval formulation of the task, MAP (Mean Average Precision) serves as the primary official evaluation metric. The no-answer cases are handled simply by giving full credit to "no answers” system output and zero otherwise. We report three measures: **MAP@10** computed over the top 10 ranked answers, **MAP_Q@5** computed over the top 5 ranked Qur'anic passages (after discarding all ranked Hadiths), and **MAP_H@5** computed over the top 5 ranked Hadiths (after discarding all ranked Qur'anic passages).

3.4.2 Participating Teams and Results

While 30 teams registered in Task 2, eight teams submitted runs during the test phase. The evaluation of the best run per team is shown in Table 3. For the full evaluation results, see Table 4 in Appendix. Only four out of eight participating teams submitted papers describing their work, namely Burhan (Basheer et al., 2025), ThinkDrill (Elrefai et al., 2025), Nur (Amin et al., 2025), and BurhanAI (Al Adel et al., 2025). It is evident that the task of this year is quite challenging since the top MAP@10 score is 0.3351 achieved by Burhan.

3.5 Methods and Analysis

The main observation in all participants is the reliance on LLMs in their systems. We categorize the discussion of adopted methods by techniques.

Augmentation The top team (Burhan) utilized LLMs to extract facts and relationships from Qur'an and Hadith passages and then augmented the extracted text with the corresponding passages. ThinkDrill team extended hadith question-answer pairs from HAQA dataset, and employed GPT-4 to extract relevant keywords from questions, and then apply fuzzy string matching to determine the relevance score. NUR team augmented the provided dataset with the Arabic portion of the TyDi dataset, the Jalalayn Tafseer of the Qur'an, and the QuQA and HaQA datasets. They also embedded negative samples to increase their models' sensitivity to zero-answer questions. BurhanAI team employed iterative semantic search, expanding the query with the initial results.

Reranking Burhan and ThinkDrill adopted an LLM as a reranker, leading to remarkable improvements as reported by Burhan team. NUR team used a fine-tuned cross-encoder or Gemini for reranking and identification of zero-answer questions.

Embedding Toward building sematic-based retrieval pipelines, multiple teams focused on the choice of the encoder embedding model. Burhan team experimented with multiple embedding models to identify the best model in *zero-shot* setup. However, ThinkDrill *fine-tuned* a multilingual embedding model using triplet loss on augmented data of Qur'an and Hadith. NUR team has compared a large set of publicly available Arabic sentence embedding models on the development set (Qur'an-only) to select the backbone encoder for their retrieval and reranking pipeline. On the other hand, BurhanAI team employed OpenAI's `file_search` directly as the backbone for semantic search.

Paraphrasing Burhan team was the only team that worked on improving the query representation. In particular, they utilize LLMs to paraphrase the questions or append synonyms to them. The paraphrasing component revealed clear benefits.

Zero-answer Questions Handling the zero-answer questions differed across teams. Burhan team employed an LLM to judge whether a passage provides an answer to a given question on a binary basis. ThinkDrill adopted a thresholding mechanism to detect such questions, i.e., if the relevance score is above s certain threshold, the question then has an answer. Similarly, NUR team adopted the thresholding-based approach with fine-tuned cross-encoders, in addition to directly prompting Gemini LLM to identify such questions.

4 Conclusion

We introduced IslamicEval, the first shared task dedicated to addressing hallucination in Islamic contexts. The challenges posed by this task aim to significantly advance the reliability of LLMs in generating accurate Islamic content. Moreover, it supports broader efforts to uphold the integrity of religious information in the digital age.

5 Limitations

Labeling religious data is an exhaustive sensitive task. As a result, the number of records in our datasets is not big. We plan in the future to extend our datasets by labeling more samples.

Our study only considers Qur'an and Hadith in the Arabic language; however, there are hundreds of millions of people worldwide who communicate Hadith in other languages like Turkish, Farsi, Malay, and Urdu (Fawzi et al., 2026). Since these languages have their own customized LLMs, it is very likely that they will produce different variants of religious hallucinations. In addition, each LLM output in Subtask 1 was annotated by a single annotator, which may introduce annotation errors. We evaluated answers from six LLMs (Arabic-centric and multilingual), each with distinct styles of responding to Islamic questions, which may not generalize to other models. The test set is relatively small (312 question–answer pairs), and model performance could vary on larger or thematically different test sets. Furthermore, annotation was limited to assessing the correctness of Qur'anic verses and Hadiths, without considering whether the overall answer was accurate or relevant to the input question. A more comprehensive evaluation of LLMs in this domain should therefore extend beyond text correction to include additional dimensions of answer quality.

Since this is the first edition of the Qur'an–Hadith QA task to incorporate Hadith as an additional Islamic resource for answering questions, we limited the Hadith collection to Sahih al-Bukhari. We plan to include other Hadith collections in future versions of the task.

Unlike the AyaTEC and QRCD datasets used in prior versions of Subtask 2, the annotation phase for the current test set may not have exhaustively identified all answer-bearing candidates. Consequently, evaluation is subject to the usual risk that some relevant results may not be rewarded.

6 Ethical Considerations

Subtask 1 involves questions and answers generated by LLMs, which were manually annotated to correct errors in cited Qur'anic verses and Hadiths. Given the religious sensitivity of the content, we took care to ensure accuracy and respect: annotations were carried out by three qualified linguists from Egypt with expertise in Arabic language and Islamic studies, and all annotators were compensated fairly for their work. The dataset is released strictly for research purposes, with the intention of improving the reliability and safety of LLMs in handling religious material. We explicitly caution against any misuse of this resource in contexts that could distort, misrepresent, or disrespect Islamic teachings.

7 Acknowledgments

We gratefully acknowledge the Qur'an and Hadith specialists who helped fine-tune test-question language, evaluated and annotated answers from the Qur'an and Sahih al-Bukhari, and advised on Hadith resources for Subtask 2—especially the professors of the Department of Qur'an and Sunnah, College of Sharia and Islamic Studies, Qatar University: Prof. Ahmad Shukri, Dr. Abdulhamid Alsis, and Dr. Abada Tahhan. The work of Watheq Mansour would not have been possible without the support of resources from Dr. Joel Mackenzie.

References

Samir Abdaljalil, Hasan Kurban, and Erchin Serpedin. 2025. Halluverse25: Fine-grained multilingual benchmark dataset for llm hallucinations. *arXiv preprint arXiv:2503.07833*.

Norah Abokhodair, AbdelRahim Elmadany, and Walid Magdy. 2020. Holy tweets: Exploring the sharing of the quran on twitter. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–32.

Basem H. Ahmed, Motaz K. Saad, and Eshrag A. Refaee. 2022. QQATeam at Quran QA 2022: Fine-Tuning Arabic QA Models for Quran QA Task. In *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022)*.

Arij Al Adel, Abu Bakr Soliman, Mohamed Sakher Sawan, Rahaf Al-Najjar, and Sameh Amin. 2025. Combating hallucinations in llms for islamic content: Evaluation, correction, and retrieval-based solution. In *Proceedings of the Third Arabic Natural Language Processing Conference*.

Hend Al-Khalifa, Tamer Elsayed, Hamdy Mubarak, Abdulmohsen Al-Thubaity, Walid Magdy, and Kareem Darwish. 2022. Proceedinsg of the 5th workshop on osact with shared tasks on qur'an qa and fine-grained hate speech detection. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*.

Z. A. B. Al-Zubaidi. 2009. *Al-Tajreed Al-Sareeh of Collective Sahih Hadith*. Resalah Publishers. Author died 893 AH/1488 CE.

Hayfa A Aleid and Aqil M Azmi. 2025. Hajj-fqa: A benchmark arabic dataset for developing question-answering systems on hajj fatwas: H. aleid and a. azmi. *Journal of King Saud University Computer and Information Sciences*, 37(6):135.

Serag Amin, Ranwa Aly, Yara Allam, Yomna Eid, and Ensaf Hussein Mohamed. 2025. Nur at islamiceval 2025 shared task: Retrieval-augmented llms for qur'an and hadith qa. In *Proceedings of the Third Arabic Natural Language Processing Conference*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. Araelectra: Pre-training text discriminators for arabic language understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 191–195. Association for Computational Linguistics.

Aqil M Azmi, Abdulaziz O Al-Qabbany, and Amir Hussain. 2019. Computational and natural language processing based studies of hadith literature: a survey. *Artificial Intelligence Review*, 52:1369–1414.

M Saiful Bari, Yazeed Alnumay, Norah A Alzahrani, Nouf M Alotaibi, Hisham A Alyahya, Sultan AlRashed, Faisal A Mirza, Shaykhah Z Alsubaie, Hassan A Alahmed, Ghadah Alabduljabbar, et al. 2024. ALLam: Large language models for Arabic and English. *arXiv preprint arXiv:2407.15390*.

Mohammad Basheer, Watheq Mansour, Abdulhamid Touma, and Ahmad Qadeib Alban. 2025. Burhan at islamiceval: Fact-augmented and llm-driven retrieval for islamic qa. In *Proceedings of the Third Arabic Natural Language Processing Conference*.

Amina El Ganadi, Sania Aftar, Luca Gagliardelli, Federico Ruozzi, et al. 2025. Generative ai for islamic texts: The eman framework for mitigating gpt hallucinations. In *roceedings of the 17th International Conference on Agents and Artificial Intelligence-ICAART*, volume 3, pages 1221–1228.

Fatimah Mohamed Emad Elden. 2025. Isnad ai at islamiceval 2025: A rule-based system for identifying religious texts in llm outputs. In *Proceedings of the Third Arabic Natural Language Processing Conference*.

Mohammed ElKomy and Amany M. Sarhan. 2022. TCE at Qur'an QA 2022: Arabic Language Question Answering Over Holy Qur'an Using a Post-Processed Ensemble of BERT-based Models. In *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022)*.

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.

Mohammed ElKoumy, Mohamed Ibrahim Alqablawi, Ahmad Tamer, and Khalid Allam. 2025. Tce at islamiceval 2025: Retrieval-augmented llms for quranic and hadith content identification and verification. In *Proceedings of the Third Arabic Natural Language Processing Conference*.

Eman Elrefai, Toka Khaled, and Ahmed Soliman. 2025. Thinkdrill at islamiceval 2025: Llm hybrid approach for quran and hadith question answering. In *Proceedings of the Third Arabic Natural Language Processing Conference*.

Fanar 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*.

Mahmoud Fawzi, Walid Magdy, and Björn Ross. 2025. 'the prophet said so!': On exploring hadith presence on arabic social media. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–23.

Mahmoud Fawzi, Björn Ross, and Walid Magdy. 2026. Fabricating holiness: Characterizing religious misinformation circulators on arabic social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 20.

Gemma Team. 2024. Gemma.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and Arun Rao et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Jimmy Lin and Boris Katz. 2006. Building a reusable test collection for question answering. *Journal of the American Society for Information Science and Technology*, 57(7):851–861.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021. Pretrained transformers for text ranking: Bert and beyond. *Synthesis Lectures on Human Language Technologies*, 14(4):1–325.

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, Watheq Mansour, and Tamer Elsayed. 2022. Qur'an QA 2022: Overview of the First Shared Task on Question Answering over the Holy Qur'an. In *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022)*, pages 79–87.

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 ArabicNLP 2023*, pages 690–701, Singapore (Hybrid). Association for Computational Linguistics.

Rana R Malhas. 2023. *ARABIC QUESTION ANSWERING ON THE HOLY QUR'AN*. Ph.D. thesis, Qatar University.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. *arXiv preprint arXiv:2401.06855*.

Hamdy Mubarak, Hend Al-Khalifa, and Khaloud Suliman Alkhalefah. 2024. Halwasa: Quantify and analyze hallucinations in large language models: Arabic as a case study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8008–8015.

Al-Tahir A.R. Musallam. 2022. Prophetic interpretation of the quran: Between quantity and quality. *Rehan Journal for Scientific Publishing*, 26(1):51–76.

Arwa Omayrah, Sakhar Alkhereyf, Ahmed Abdelali, Abdulmohsen Al-Thubaity, Jeril Kuriakose, and Ibrahim AbdulMajeed. 2025. Humain at islamiceval 2025 shared task 1: A three-stage llm-based pipeline for detecting and correcting hallucinations in quran and hadith. In *Proceedings of the Third Arabic Natural Language Processing Conference*.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. *arXiv preprint arXiv:2101.05779*.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Marwan N. Swar. 2007. *Mushaf Al-Tafseel Al-Mawdoo'ee*. Dar Al-Fajr Al-Islami, Damascus.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Raúl Vázquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sánchez-Vega, Alessandro Raganato, Jindřich Libovickỳ, et al. 2025. Semeval-2025 task 3: Mu-shroom, the multilingual shared task on hallucinations and related observable overgeneration mistakes. *arXiv preprint arXiv:2504.11975*.

Abdulrezzak Zekiye and Fadi Amroush. 2023. Al-jawaab at Qur'an QA 2023 shared task: Exploring Embeddings and GPT Models for Passage Retrieval and Reading Comprehension. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore.

A Related Work

A.1 Hallucination Detection

Hallucination detection methods can be grouped into uncertainty-based predictors (Manakul et al., 2023), entailment or consistency checks against retrieved evidence (Ji et al., 2023), and span-level labeling frameworks (Mishra et al., 2024). Recent work emphasizes span-level detectors for interpretability, with SemEval-2025 introducing a shared task that explicitly included Arabic (Vázquez et al., 2025).

For Arabic hallucination detection, resources remain limited. The OSACT-6 Hallucination Shared Task "Halwasa" (Mubarak et al., 2024) released the first Arabic data set (10K sentences generated by GPT and manually annotated for factuality), with baselines that highlight challenges due to morphological richness. HalluVerse25 (Abdaljalil et al., 2025) is a multilingual benchmark that categorizes fine-grained hallucinations in English, Arabic, and Turkish. The authors used GPT-4 to inject hallucinations into factual biographical sentences extracted from Wikipedia.

In religious domains, hallucination risks are amplified by doctrinal sensitivity. Qur'an QA (Malhas et al., 2022, 2023) established benchmarks for comprehension and passage retrieval, while (Aleid and Azmi, 2025) supports research on fatwa related to Hajj (Muslim pilgrimage). Most approaches mitigate hallucinations through retrieval-augmented generation (RAG) (Lewis et al., 2020), conservative prompting, and reranking rather than explicit detectors. Recent frameworks such as EMAN (El Ganadi et al., 2025) stress governance and cultural alignment when deploying LLMs on Islamic texts.

Overall, prior work shows progress but also gaps: (i) reliance on mitigation rather than calibrated detectors in high-stakes religious contexts, and (ii) lack of standardized evaluation for detecting misquotations or unsupported doctrinal claims. Our work builds on these efforts by extending hallucination detection to Arabic religious texts with domain-grounded and span-level evaluation.

A.2 Qur'an QA 2022 and 2023

With Qur'an and Hadith QA being a continuation of Qur'an QA 2022[12] (Malhas et al., 2022) and Qur'an QA 2023[13] (Malhas et al., 2023) shared tasks, we provide an overview of those two editions.

The Qur'an QA shared task in its first round (2022) comprised a single machine reading comprehension (MRC) task over the Holy Qur'an: given a passage of consecutive verses from one Surah and an MSA question about that passage, systems had to extract *any* correct answer *span*. The main measure used in evaluation was partial Reciprocal Rank (pRR) (Malhas and Elsayed, 2020). The task attracted 30 teams, 13 of which submitted 30 runs in the test phase. Ten system description papers were published in OSACT 2022 (Al-Khalifa et al., 2022), and the best-performing systems achieved pRR=0.586, underscoring the difficulty of the MRC task. Leading systems (ElKomy and Sarhan, 2022; Ahmed et al., 2022) mainly used fine-tuned encoder-only BERT-based models, notably AraELECTRA (Antoun et al., 2021) and AraBERT (Antoun et al., 2020).

Qur'an QA 2023 introduced a more challenging MRC task and a new Qur'anic Passage Retrieval (QPR) task, which parallels the Qur'an QA component of Subtask 2 in the present shared task. The primary goal of QPR is to retrieve *all* Qur'anic passages that contain potential answers to a question posed in MSA. A total of 38 and 29 teams registered for QPR and MRC, respectively, and 10 teams submitted 39 runs in the test phase across the two tasks. The evaluation results revealed the inherent difficulty of the tasks: the top team achieved $pRR = 0.571$ on MRC and $MAP@10 = 0.251$ on QPR. For MRC, fine-tuned AraELECTRA and AraBERT models remained leading performers for the top team that employed them. Notably, the second-place team was the only one to adopt a GPT-4 model in a zero-shot prompt setting (Zekiye and Amroush, 2023). For QPC, the top-performing approach ensembled dual- and cross-encoder BERT-based models with staged fine-tuning on Arabic QA and domain-specific datasets (Elkomy and Sarhan, 2023). Attempts to use LLMs as embedding models or re-rankers were modest and did not feature among the top systems.

[12] https://sites.google.com/view/quran-qa-2022

[13] https://gitlab.com/bigirqu/quran-qa-2023

B Prompt for Generating Responses with Qur’anic and Hadith Evidence

Prompt:

أجب عن السؤال التالي واستشهد بآيات من القرآن الكريم وأحاديث شريفة.

السؤال:

Translation:

Provide an answer to the following question, citing evidence from the Qur’an and Prophetic Hadiths.
Question:

C Annotation and Correction Guidelines

1. Incomplete texts: Any incomplete Qur’anic verse (Ayah) or incomplete Hadith is considered an error.

2. Diacritization: Incorrect diacritization is marked as an error, whereas partially correct diacritization or the absence of diacritics is not treated as an error.

3. Error granularity: A single error in an Ayah or Hadith suffices to label the span as erroneous.

4. Reference verification: In this version, verification of metadata such as chapter or Hadith reference numbers is not required.

5. Span boundaries: Annotated spans exclude outer punctuation marks, if present.

6. Multiple Ayahs: If more than one Ayah appears in the same span, the entire span is selected even if an internal verse number appears in the middle.

7. Sources: Corrected Qur’anic text must be copied from `https://quran.ksu.edu.sa/`, and corrected Hadith from `https://dorar.net/hadith`.

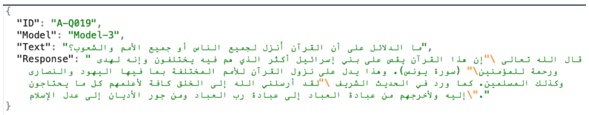
8. Correction task: Annotators predict the intended Ayah or Hadith and copy the exact corrected and complete text from the designated sources. If no valid correction can be determined, they write “Wrong” in the correction field.

9. Consistency in length: Corrections must preserve the number of intended Ayahs. For instance, if the erroneous text contains two Ayahs, the corrected version should also contain two.

10. Output assessment: In this version, we focus solely on the verification and correction of Qur’anic verses (Ayahs) and Hadiths, without assessing the completeness of the answers or their relevance to the given question. We leave this for future releases.

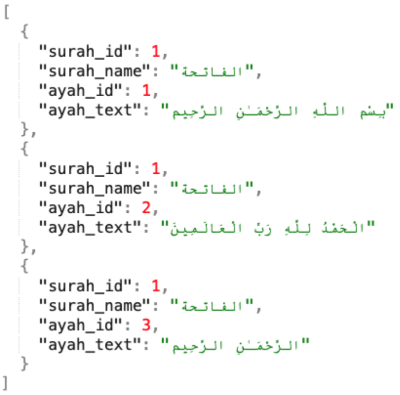
11. Correction formatting: For quality control, annotators were instructed to prepend a serial number to each correction in the text area, reflecting its order in the list of erroneous Ayahs or Hadiths. Each correction was required to be written on a separate line.

D Sample Data Files Provided to Participants



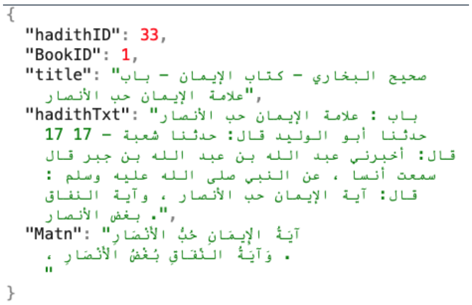
```
{
  "ID": "A-Q019",
  "Model": "Model-3",
  "Text": "ما الدلائل على أن القرآن أنزل لجميع الناس أو جميع الأمم والشعوب؟",
  "Response": " قال الله تعالى \"إن هذا القرآن يقص على بني إسرائيل أكثر الذي هم فيه يختلفون وإنه لهدى
    ورحمة للمؤمنين\" (سورة يونس). وهذا يدل على نزول القرآن للأمم المختلفة بما فيها اليهود والنصارى
    وكذلك المسلمين. كما ورد في الحديث الشريف \"لقد أرسلني الله إلى الخلق كافة لأعلمهم كل ما يحتاجون
    إليه ولأخرجهم من عبادة العباد إلى عبادة رب العباد ومن جور الأديان إلى عدل الإسلام\"."
}
```

Figure 4: Example entry from the development set showing a question, model ID, and model-generated response.



```
[
  {
    "surah_id": 1,
    "surah_name": "الفاتحة",
    "ayah_id": 1,
    "ayah_text": "بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ"
  },
  {
    "surah_id": 1,
    "surah_name": "الفاتحة",
    "ayah_id": 2,
    "ayah_text": "الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ"
  },
  {
    "surah_id": 1,
    "surah_name": "الفاتحة",
    "ayah_id": 3,
    "ayah_text": "الرَّحْمَٰنِ الرَّحِيمِ"
  }
]
```

Figure 5: Sample JSON entries from the Qur’an reference collection, showing Surah name, Ayah ID, and Ayah text.



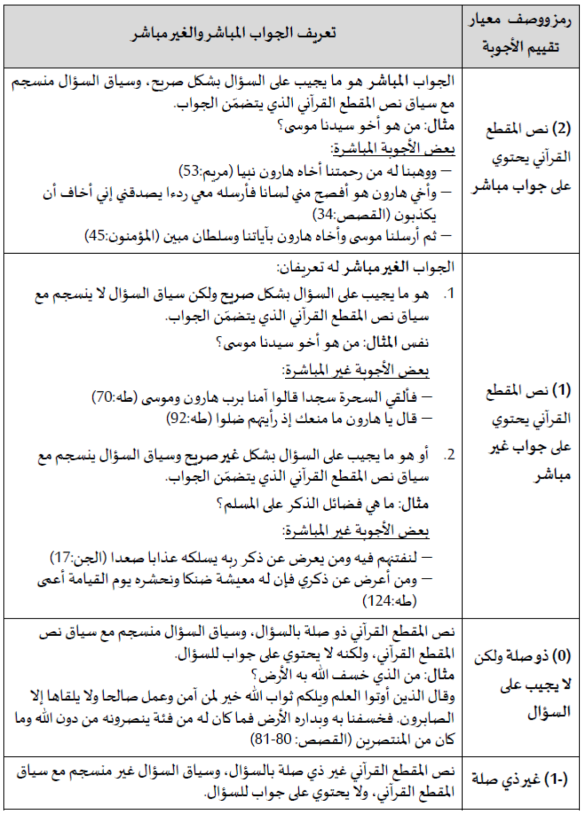
```
{
  "hadithID": 33,
  "BookID": 1,
  "title": "صحيح البخاري - كتاب الإيمان - باب
    علامة الإيمان حب الأنصار",
  "hadithTxt": "باب : علامة الإيمان حب الأنصار
    17 17 - حدثنا أبو الوليد قال: حدثنا شعبة
    قال: أخبرني عبد الله بن عبد الله بن جبر قال
    سمعت أنسا ، عن النبي صلى الله عليه وسلم :
    قال: آية الإيمان حب الأنصار ، وآية النفاق
    بغض الأنصار .",
  "Matn": "آيَةُ الإِيمَانِ حُبُّ الأَنْصَارِ
    ، وَآيَةُ النِّفَاقِ بُغْضُ الأَنْصَارِ .
    "
}
```

Figure 6: Sample JSON entry from the Hadith reference collection, including metadata (Book ID, title) and Hadith text.

E Annotation Rubrics for Qur’an Hadith QA Subtask

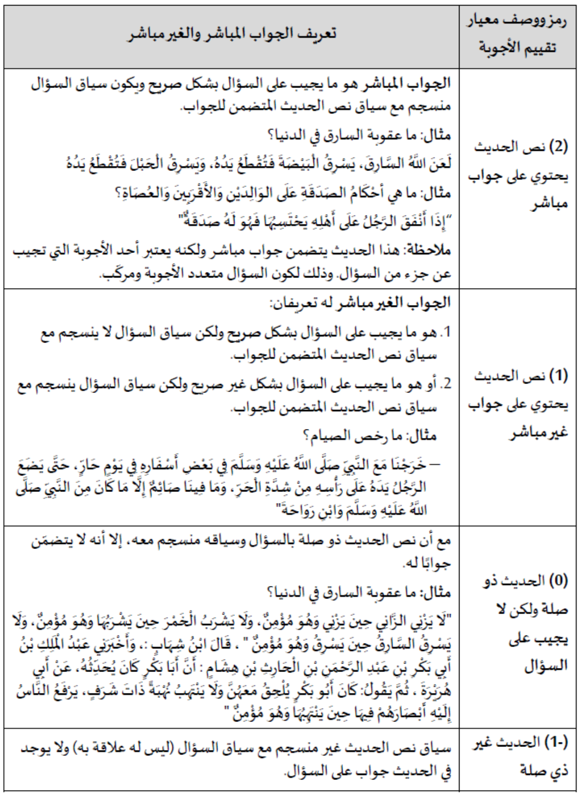
Team Name	Run	MAP@10	MAP_Q@5	MAP_H@5
Burhan	351588_Burhan_PQQFHF	0.3351	0.3389	0.3876
Burhan	351587_Burhan_QFHF	0.3021	0.3091	0.3461
Burhan	351586_Burhan_QFH	0.2916	0.3130	0.2936
BurhanAI	351568_burhanai_task_2_RAG_gpt5high	0.2807	0.3257	0.2386
ThinkDrill	351792_run_sample	0.2296	0.2623	0.2150
NUR	351549_nur_run01	0.1809	0.2334	0.1923
NUR	351550_nur_run02	0.1804	0.2257	0.1961
BayaNet	351272_BayaNet_run02mod	0.1504	0.2064	0.2240
NUR	351551_nur_run03	0.1257	0.1438	0.1569
MSA	350916_MSA_02	0.1185	0.1674	0.0685
MSA	351316_MSA_04	0.1185	0.1674	0.0685
MSA	351275_MSA_03	0.1185	0.1674	0.0685
ThinkDrill	351585_run_sample	0.0509	0.0977	0.0841
Maged	351633_run_sample	0.0332	0.0887	0.0457
ThinkDrill	351580_run_sample	0.0226	0.0482	0.1569
BayaNet	351263_BayaNet_b6453eb4	0.0157	0.0205	0.0067
CISRG	350176_CISRG_r25	0.0116	0.0294	0.0128
Maged	351629_run_sample	0.0000	0.1569	0.1961
Maged	351462_run_sample	0.0000	0.0588	0.0196

Table 4: The evaluation results of the last three runs submitted to Subtask 2 ranked by MAP@10. Teams with * did not submit a system paper. The run name is formatted as CodaBenchSubmissionID_RunName.



تعريف الجواب المباشر والغير مباشر	رمز ووصف معيار تقييم الأجوبة
الجواب المباشر هو ما يجيب على السؤال بشكل صريح، وسياق السؤال منسجم مع سياق نص المقطع القرآني الذي يتضمَن الجواب. مثال: من هو أخو سيدنا موسى؟ بعض الأجوبة المباشرة: – ووهبنا له من رحمتنا أخاه هارون نبيا (مريم:53) – وأخي هارون هو أفصح مني لسانا فأرسله معي ردءا يصدقني إني أخاف أن يكذبون (القصص:34) – ثم أرسلنا موسى وأخاه هارون بآياتنا وسلطان مبين (المؤمنون:45)	(2) نص المقطع القرآني يحتوي على جواب مباشر
الجواب الغير مباشر له تعريفان: 1. هو ما يجيب على السؤال بشكل صريح ولكن سياق السؤال لا ينسجم مع سياق نص المقطع القرآني الذي يتضمَن الجواب. نفس المثال: من هو أخو سيدنا موسى؟ بعض الأجوبة غير المباشرة: – فألقي السحرة سجدا قالوا آمنا برب هارون وموسى (طه:70) – قال يا هارون ما منعك إذ رأيتهم ضلوا (طه:92) 2. أو هو ما يجيب على السؤال بشكل غير صريح وسياق السؤال ينسجم مع سياق نص المقطع القرآني الذي يتضمَن الجواب. مثال: ما هي فضائل الذكر على المسلم؟ بعض الأجوبة غير المباشرة: – لنفتنهم فيه ومن يعرض عن ذكر ربه يسلكه عذابا صعدا (الجن:17) – ومن أعرض عن ذكري فإن له معيشة ضنكا ونحشره يوم القيامة أعمى (طه:124)	(1) نص المقطع القرآني يحتوي على جواب غير مباشر
نص المقطع القرآني ذو صلة بالسؤال، وسياق السؤال منسجم مع سياق نص المقطع القرآني، ولكنه لا يحتوي على جواب للسؤال. مثال: من الذي خسف الله به الأرض؟ وقال الذين أوتوا العلم ويلكم ثواب الله خير لمن آمن وعمل صالحا ولا يلقاها إلا الصابرون. فخسفنا به وبداره الأرض فما كان له من فئة ينصرونه من دون الله وما كان من المنتصرين (القصص: 80-81)	(0) ذو صلة ولكن لا يجيب على السؤال
نص المقطع القرآني غير ذي صلة بالسؤال، وسياق السؤال غير منسجم مع سياق المقطع القرآني، ولا يحتوي على جواب للسؤال.	(-1) غير ذي صلة

Figure 7: Rubric for annotating potential answer-bearing Qur'anic passages to a given question.



تعريف الجواب المباشر والغير مباشر	رمز ووصف معيار تقييم الأجوبة
الجواب المباشر هو ما يجيب على السؤال بشكل صريح ويكون سياق السؤال منسجم مع سياق نص الحديث المتضمن للجواب. مثال: ما عقوبة السارق في الدنيا؟ لَعَنَ اللَّهُ السَّارِقَ، يَسْرِقُ الْبَيْضَةَ فَتُقْطَعُ يَدُهُ، وَيَسْرِقُ الْحَبْلَ فَتُقْطَعُ يَدُهُ مثال: ما هي أحْكَامُ الصَدَقَةِ عَلَى الوَالِدَيْنِ وَالأَقْرَبِينَ وَالعُصَاةِ؟ "إِذَا أَنْفَقَ الرَّجُلُ عَلَى أَهْلِهِ يَحْتَسِبُهَا فَهُوَ لَهُ صَدَقَةٌ" ملاحظة: هذا الحديث يتضمن جواب مباشر ولكنه يعتبر أحد الأجوبة التي تجيب عن جزء من السؤال. وذلك لكون السؤال متعدد الأجوبة ومركَب.	(2) نص الحديث يحتوي على جواب مباشر
الجواب الغير مباشر له تعريفان: 1. هو ما يجيب على السؤال بشكل صريح ولكن سياق السؤال لا ينسجم مع سياق نص الحديث المتضمن للجواب. 2. أو هو ما يجيب على السؤال بشكل غير صريح ولكن سياق السؤال ينسجم مع سياق نص الحديث المتضمن للجواب. مثال: ما رخص الصيام؟ – خَرَجْنَا مَعَ النَّبِيِّ صَلَّى اللَّهُ عَلَيْهِ وَسَلَّمَ فِي بَعْضِ أَسْفَارِهِ فِي يَوْمٍ حَارٍّ، حَتَّى يَضَعَ الرَّجُلُ يَدَهُ عَلَى رَأْسِهِ مِنْ شِدَّةِ الْحَرِّ، وَمَا فِينَا صَائِمٌ إِلَّا مَا كَانَ مِنَ النَّبِيِّ صَلَّى اللَّهُ عَلَيْهِ وَسَلَّمَ وَابْنِ رَوَاحَةَ"	(1) نص الحديث يحتوي على جواب غير مباشر
مع أن نص الحديث ذو صلة بالسؤال وسياقه منسجم معه، إلا أنه لا يتضمَن جوابًا له. مثال: ما عقوبة السارق في الدنيا؟ "لَا يَزْنِي الزَّانِي حِينَ يَزْنِي وَهُوَ مُؤْمِنٌ، وَلَا يَشْرَبُ الْخَمْرَ حِينَ يَشْرَبُهَا وَهُوَ مُؤْمِنٌ، وَلَا يَسْرِقُ السَّارِقُ حِينَ يَسْرِقُ وَهُوَ مُؤْمِنٌ" . قَالَ ابْنُ شِهَابٍ :، وَأَخْبَرَنِي عَبْدُ الْمَلِكِ بْنُ أَبِي بَكْرِ بْنِ عَبْدِ الرَّحْمَنِ بْنِ الْحَارِثِ بْنِ هِشَامٍ : أَنَّ أَبَا بَكْرٍ كَانَ يُحَدِّثُهُ، عَنْ أَبِي هُرَيْرَةَ ، ثُمَّ يَقُولُ: كَانَ أَبُو بَكْرٍ يُلْحِقُ مَعَهُنَّ وَلَا يَنْتَهِبُ نُهْبَةً ذَاتَ شَرَفٍ، يَرْفَعُ النَّاسُ إِلَيْهِ أَبْصَارَهُمْ فِيهَا حِينَ يَنْتَهِبُهَا وَهُوَ مُؤْمِنٌ "	(0) الحديث ذو صلة ولكن لا يجيب على السؤال
سياق نص الحديث غير منسجم مع سياق السؤال (ليس له علاقة به) ولا يوجد في الحديث جواب على السؤال.	(-1) الحديث غير ذي صلة

Figure 8: Rubric for annotating potential answer-bearing Hadith *matns* to a given question.