A Knowledge Graph Based Diagnostic Framework for Analyzing Hallucinations in Arabic Machine Reading Comprehension

Najwa AlGhamdi[1,2], Sadam Al-Azani[2], Kwabena Nuamah[1], Alan Bundy[1]

[1]School of Informatics, University of Edinburgh, UK
[2]SDAIA–KFUPM Joint Research Center for Artificial Intelligence,
Information and Computer Science Department,
King Fahd University of Petroleum & Minerals, Saudi Arabia

s2448091@ed.ac.uk, sadam.azani@kfupm.edu.sa, k.nuamah@ed.ac.uk, a.bundy@ed.ac.uk

Abstract

Large Language Models (LLMs) frequently generate answers that are fluent but not fully grounded in the provided context, a phenomenon commonly referred to as hallucination. While recent work has explored hallucination detection primarily in English and open domain settings, comparatively little attention has been given to Arabic machine reading comprehension (MRC), particularly in culturally sensitive domains such as Qur'anic texts. In this paper, we present a knowledge graph based diagnostic framework for analyzing hallucinations and question misalignment in Arabic MRC. Rather than proposing a new detection model or metric, the framework provides an interpretable, triple level analysis of model generated answers by comparing subject-relation-object representations derived from the passage, the question, and the answer. The approach incorporates question-aware filtering and operates under weak supervision, combining automatic analysis with targeted human adjudication to handle annotation gaps and semantic ambiguity. We apply the framework to the Qur'anic Reading Comprehension Dataset (QRCD) and demonstrate how it exposes systematic hallucination patterns that are difficult to capture using surface level similarity metrics alone, particularly for questions requiring justification or abstract interpretation. The results highlight the value of structured, transparent diagnostic evaluation for understanding LLM behavior in low resource and high stakes Arabic NLP settings.

1 Introduction

Large language models (LLMs) are now widely used across a range of natural language processing tasks, including dialogue, summarization, translation, and open ended text generation (Zhao et al., 2023). Despite their strong empirical performance, their outputs cannot always be assumed to be fully reliable. In particular, LLMs may generate responses that are linguistically fluent yet insufficiently supported by the source text or underlying facts. This behavior, commonly referred to as hallucination has been documented across tasks and model architectures (Huang et al., 2025; Ji et al., 2023; Maynez et al., 2020). In settings where answers are expected to be directly grounded in a given context, such behavior poses challenges for trustworthiness, especially in knowledge intensive and high stakes domains (Liu et al., 2023; McIntosh et al., 2023).

Analyzing hallucinations is therefore a necessary but non-trivial step toward understanding and mitigating them. Hallucinated content is often context dependent and may not be detectable through surface level similarity measures alone. For this reason, recent work has explored structured representations, such as knowledge graphs, to make factual relationships explicit. By modeling information in terms of entities and relations, knowledge graphs enable more direct comparison between generated outputs and trusted sources, revealing inconsistencies that may otherwise remain unnoticed (Sansford et al., 2024). Beyond evaluation, such representations have also been incorporated into generation pipelines, where they can constrain model behavior and reduce unsupported claims (Agrawal et al., 2023; Guan et al., 2024).

In this paper, we explore these ideas in the context of Arabic machine reading comprehension (MRC), with a particular focus on Qur'anic passages. Qur'anic Arabic poses distinctive challenges for automatic processing: the language is Classical rather than modern, meaning is often conveyed implicitly, and even small inaccuracies can have disproportionate interpretive consequences. While hallucination detection has been studied extensively in English and in open-domain settings, comparatively little attention has been given to low resource and culturally sensitive contexts (Mubarak et al., 2024), and even less to Arabic religious texts. In that study, we examined model outputs using a combination of

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 413–421
March 28, 2026. ©2026 Association for Computational Linguistics

automatic metrics and manual inspection, identifying recurring error patterns such as missed entities, unsupported answers, over quotation of verses, and failures to respect the semantic focus of the question. While this analysis offered insight into the types of errors produced by LLMs and their variation across prompting strategies, it did not provide a scalable way to identify such issues in a systematic manner. The present paper addresses this limitation by introducing a structured diagnostic framework that captures these error patterns explicitly at the level of subject–relation–object triples.

We evaluate the framework in a weakly supervised setting, assuming access only to the passage, the question, and a reference (gold) answer, without relying on explicit hallucination annotations. Model generated answers are transformed into triples and compared against knowledge representations derived from the passage, allowing hallucinated, missing, off-focus, and supported content to be identified. Human adjudication is reserved for genuinely ambiguous cases. The framework is designed to balance automation with interpretability, reflecting both the limitations of existing gold annotations and the linguistic complexity of Classical Arabic. Rather than producing a single correctness score, it serves as a diagnostic tool that supports fine-grained inspection of model behavior.

Overall, our approach addresses two practical challenges: first, the scarcity of datasets annotated specifically for hallucination detection in Arabic MRC, which limits the applicability of fully supervised methods; and second, the need for interpretable and linguistically grounded evaluation techniques that are suitable for low resource and culturally sensitive domains.

The main contributions of this work can be summarized as follows:

- We introduce a knowledge graph based diagnostic framework for triple level analysis of hallucinations and question misalignment in Arabic machine reading comprehension.
- We propose a question-aware error taxonomy that distinguishes between supported, missing, hallucinated, off-focus, and gold-only content under weak supervision.
- We present an empirical diagnostic analysis of LLM behavior on Qur'anic reading comprehension, highlighting systematic hallucination patterns that are not captured by surface level evaluation metrics.

While this work focuses on model-generated answers, our goal is not to rank or compare models in terms of overall performance. Instead, we treat generated answers as diagnostic artifacts that allow us to examine how hallucinations arise during the generation process. By analyzing model outputs at a fine-grained, triple-based level, the framework aims to expose patterns of failure that remain opaque under standard correctness-oriented evaluation metrics.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the proposed framework. Section 4 presents the experimental results and analysis. Finally, Section 5 concludes the paper and highlights directions for future research.

2 Related Work

LLMs have demonstrated strong performance on open domain question answering and machine reading comprehension. At the same time, a growing body of evidence shows that these models frequently produce hallucinated responses that are fluent but not grounded in the source text or supported by external knowledge (Ji et al., 2023; Maynez et al., 2020). This vulnerability has prompted extensive research on how to evaluate and mitigate factual inconsistency in generated outputs.

Early evaluation methods have largely relied on reference based metrics such as Accuracy, F1, ROUGE, and BLEU (Mündler et al., 2023; Zhang et al., 2023a). Although effective at capturing surface level overlap, these metrics often fail to distinguish between answers that are genuinely supported and those that merely resemble a reference. To address this limitation, prior work has explored alternative signals, including token level confidence and probability based approaches (Manakul et al., 2023), attribution and preservation oriented metrics (Chen et al., 2023; Zhang et al., 2023b), and correlation based measures such as Pearson and Kendall's τ (Elaraby et al., 2023). Although these techniques provide complementary perspectives, many remain sensitive to phrasing variation and do not explicitly model factual structure.

More recently, knowledge graph based approaches have been proposed as a way to make factual relationships explicit and interpretable. (Sansford et al., 2024) introduced GraphEval, which represents both reference content and model outputs as sets of triples and evaluates hallucinations by comparing their graph structures. This formulation allows errors to be localized at the level of individual relations rather than entire answers.

Similarly, (Fang et al., 2025) proposed a zero resource hallucination detection method that models context and generated text as knowledge graphs and aligns extracted triples to identify unsupported claims. Related efforts include GraphHallucination (Zhang et al., 2023c), which provides a benchmark for graph based factual consistency evaluation in summarization, and FactScore (Min et al., 2023), which decomposes generated text into atomic facts and verifies them against evidence. Unlike scoring-based approaches, our framework is diagnostic and focuses on identifying supported, missing, and hallucinated content at the triple level.

While these graph and claim based methods have advanced hallucination analysis in English and open domain settings, their applicability to Arabic—and especially to religious texts remains limited. Most existing frameworks assume linguistic resources, normalization pipelines, and annotation standards that do not readily transfer to Arabic, let alone to Classical or Qur'anic Arabic. In addition, prior work typically evaluates freely generated text rather than tightly constrained reading comprehension tasks where answers must be grounded in a fixed passage.

Arabic NLP presents additional challenges arising from rich morphology, orthographic variation, and the limited availability of labeled data and domain-specific tools (Alyafeai et al., 2021; Antoun et al., 2020; Mohamed and Al-Azani, 2025). These issues are amplified in the context of sacred texts such as the Qur'an, where meaning is often implicit and even minor inaccuracies can carry significant interpretive consequences (Alqahtani et al., 2022; Al-Azani et al., 2025). Despite this, relatively little work has focused on systematic, symbolic evaluation of LLM outputs for Arabic religious texts, and most existing studies rely on surface similarity metrics or manual inspection.

Our work builds on prior research in knowledge graph based hallucination evaluation while addressing these gaps. Rather than proposing a new detection algorithm, we focus on the adaptation and operationalization of symbolic, triple level evaluation for Arabic machine reading comprehension over Qur'anic passages. The framework operates entirely in Arabic, incorporates question-aware filtering, and is designed to function under weak supervision with targeted human adjudication. This positioning distinguishes our approach from earlier graph based methods and makes it suitable for low resource, culturally sensitive, and high stakes evaluation settings.

Finally, this paper should be read in conjunction with our earlier empirical study on Qur'anic reading comprehension (AlGhamdi et al., 2025), which analyzed LLM outputs using similarity based metrics and manual inspection to characterize common error patterns. While that work focused on diagnosing model behavior and identifying recurring hallucination types, it did not propose a structured or scalable evaluation mechanism. The present work complements it by operationalizing these observations through a symbolic, knowledge graph based diagnostic framework.

3 Knowledge Graph Based Diagnostic Framework

This section describes the diagnostic evaluation framework used to analyze hallucination and question relevance in LLM generated answers for Arabic machine reading comprehension (MRC). An overview of the framework is shown in Figure 1. Rather than treating model outputs as unstructured text, the framework represents both passages and answers using lightweight knowledge graphs and compares them at the level of extracted triples. The design combines automatic processing with targeted human adjudication, allowing ambiguous cases and incomplete gold annotations to be handled in a principled way.

3.1 Data Preparation

All experiments are conducted on the Qur'anic Reading Comprehension Dataset (QRCD) (Malhas et al., 2023). The dataset contains 992 passage-question-answer instances spanning 603 unique Qur'anic passages, with an average of approximately 1.64 questions per passage. Each instance consists of a short Qur'anic passage (typically two to five verses), a natural language question, and a gold standard answer annotated by humans. The dataset is particularly suitable for studying grounding and hallucination, as answers are expected to be derived from the passage rather than freely generated.

To support our diagnostic framework, we augment the original dataset with model generated an-

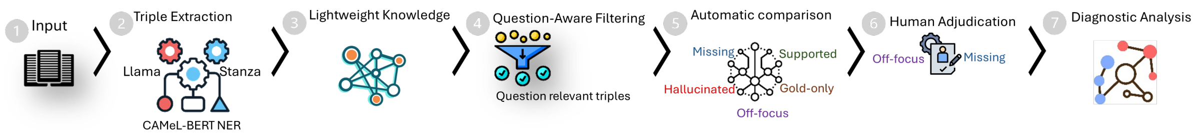
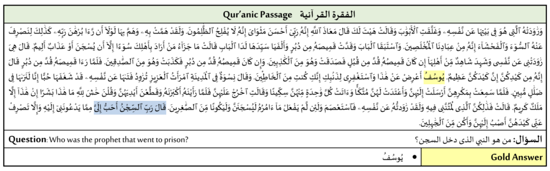


Figure 1: Knowledge graph based diagnostic framework.



Qur'anic Passage الفقرة القرآنية

وَرَٰوَدَتْهُ ٱلَّتِى هُوَ فِى بَيْتِهَا عَن نَّفْسِهِۦ وَغَلَّقَتِ ٱلْأَبْوَٰبَ وَقَالَتْ هَيْتَ لَكَ قَالَ مَعَاذَ ٱللَّهِ إِنَّهُۥ رَبِّىٓ أَحْسَنَ مَثْوَاىَ إِنَّهُۥ لَا يُفْلِحُ ٱلظَّٰلِمُونَ. وَلَقَدْ هَمَّتْ بِهِۦ وَهَمَّ بِهَا لَوْلَآ أَن رَّءَا بُرْهَٰنَ رَبِّهِۦ كَذَٰلِكَ لِنَصْرِفَ
عَنْهُ ٱلسُّوٓءَ وَٱلْفَحْشَآءَ إِنَّهُۥ مِنْ عِبَادِنَا ٱلْمُخْلَصِينَ. وَٱسْتَبَقَا ٱلْبَابَ وَقَدَّتْ قَمِيصَهُۥ مِن دُبُرٍ وَأَلْفَيَا سَيِّدَهَا لَدَا ٱلْبَابِ قَالَتْ مَا جَزَآءُ مَنْ أَرَادَ بِأَهْلِكَ سُوٓءًا إِلَّآ أَن يُسْجَنَ أَوْ عَذَابٌ أَلِيمٌ. قَالَ هِىَ
رَٰوَدَتْنِى عَن نَّفْسِى وَشَهِدَ شَاهِدٌ مِّنْ أَهْلِهَآ إِن كَانَ قَمِيصُهُۥ قُدَّ مِن قُبُلٍ فَصَدَقَتْ وَهُوَ مِنَ ٱلْكَٰذِبِينَ. وَإِن كَانَ قَمِيصُهُۥ قُدَّ مِن دُبُرٍ فَكَذَبَتْ وَهُوَ مِنَ ٱلصَّٰدِقِينَ. فَلَمَّا رَءَا قَمِيصَهُۥ قُدَّ مِن دُبُرٍ قَالَ
إِنَّهُۥ مِن كَيْدِكُنَّ إِنَّ كَيْدَكُنَّ عَظِيمٌ. يُوسُفُ أَعْرِضْ عَنْ هَٰذَا وَٱسْتَغْفِرِى لِذَنۢبِكِ إِنَّكِ كُنتِ مِنَ ٱلْخَاطِـِٔينَ. وَقَالَ نِسْوَةٌ فِى ٱلْمَدِينَةِ ٱمْرَأَتُ ٱلْعَزِيزِ تُرَٰوِدُ فَتَىٰهَا عَن نَّفْسِهِۦ قَدْ شَغَفَهَا حُبًّا إِنَّا لَنَرَىٰهَا فِى
ضَلَٰلٍ مُّبِينٍ. فَلَمَّا سَمِعَتْ بِمَكْرِهِنَّ أَرْسَلَتْ إِلَيْهِنَّ وَأَعْتَدَتْ لَهُنَّ مُتَّكَـًٔا وَءَاتَتْ كُلَّ وَٰحِدَةٍ مِّنْهُنَّ سِكِّينًا وَقَالَتِ ٱخْرُجْ عَلَيْهِنَّ فَلَمَّا رَأَيْنَهُۥٓ أَكْبَرْنَهُۥ وَقَطَّعْنَ أَيْدِيَهُنَّ وَقُلْنَ حَٰشَ لِلَّهِ مَا هَٰذَا بَشَرًا إِنْ هَٰذَآ إِلَّا
مَلَكٌ كَرِيمٌ. قَالَتْ فَذَٰلِكُنَّ ٱلَّذِى لُمْتُنَّنِى فِيهِ وَلَقَدْ رَٰوَدتُّهُۥ عَن نَّفْسِهِۦ فَٱسْتَعْصَمَ وَلَئِن لَّمْ يَفْعَلْ مَآ ءَامُرُهُۥ لَيُسْجَنَنَّ وَلَيَكُونًا مِّنَ ٱلصَّٰغِرِينَ. قَالَ رَبِّ ٱلسِّجْنُ أَحَبُّ إِلَىَّ مِمَّا يَدْعُونَنِىٓ إِلَيْهِ وَإِلَّا تَصْرِفْ
عَنِّى كَيْدَهُنَّ أَصْبُ إِلَيْهِنَّ وَأَكُن مِّنَ ٱلْجَٰهِلِينَ.

Question: Who was the prophet that went to prison? **السؤال:** من هو النبي الذى دخل السجن؟

Gold Answer • يُوسُفُ

Figure 2: Example from the QRCD dataset.

Model	Recall	Precision	F1
ALLaM	76.24%	72.71%	74.16%
GPT-4	79.15%	77.04%	77.89%
LLaMA-3	77.54%	80.04%	78.60%

Table 1: Overall performance of the evaluated LLMs on QRCD.

swers and derived annotations required for knowledge graph comparison. Figure 2 shows a representative example, where the gold answer ("Joseph") is highlighted and the supporting verse is marked within the passage.

In such cases, the gold answer is an entity mention rather than a relational statement. In our framework, single-entity answers are handled through an entity-centric representation: the entity is kept as a node and aligned to the question predicate during comparison, allowing these instances to be evaluated within the same triple-based pipeline.All text is processed in Arabic without translation.

We generate answers using three large language models: GPT-4o-mini (OpenAI, 2023), LLaMA-3 (Grattafiori et al., 2024), and ALLaM (Bari et al., 2025). These outputs were originally produced during an earlier exploratory study focused on characterizing model behavior rather than structured evaluation. For the experiments reported here, we select LLaMA-3 as the primary model for pipeline assessment, as it achieved the strongest balance between precision and recall across standard metrics, as shown in Table 1. This makes it a suitable candidate for stress testing the proposed diagnostic framework.

Each evaluation instance therefore includes four components: the passage, the question, the gold answer, and the model generated answer. These form the inputs to the subsequent stages of the framework.

Question Type	Question Type in English	Count
ما	what	444
هل	is / does	151
من	who	132
لماذا	why	52
متى	when	19
كيف	how	17
ما هي	what is / are	211
ما الدلائل	what are the indications	144
من هم	who are	88
هل هناك	is there / are there	39

Table 2: Distribution of common question starters in the QRCD dataset.

3.2 Dataset and Question Focus

We evaluate the proposed framework on 992 Arabic Qur'anic reading comprehension questions drawn from the QRCD dataset. The questions exhibit substantial variation in both form and intent, reflecting the linguistic and semantic diversity of Qur'anic inquiry. Most questions begin with a small set of common interrogative forms, either as single words or short phrases. The most frequent single word question starters are shown in Table 2.

To support question-aware evaluation, we automatically categorize questions by their expected semantic focus, NER-related questions. Using this categorization, 132 questions are identified as person focused, 249 as reason focused, 25 as animal related, 19 as time related, and 7 as location related, with the remaining 549 questions falling into a broader other category. This distribution highlights the dataset's heterogeneity and underscores the need for a diagnostic framework capable of handling both entity centric and reasoning questions.

3.3 Knowledge Triple Extraction

To enable structured comparison, passages and answers are converted into sets of subject–relation–object triples using an automatic Arabic information extraction pipeline. We first apply the Stanza toolkit (Qi et al., 2020) to extract noun phrases, which serve as candidate concepts. Named entities are identified using CAMeL-BERT (Inoue et al., 2021), a transformer based Arabic NER model.

To capture semantic relations, we prompt ALLaM and GPT-4o-mini with a small number of in context examples to extract triples from the text. These models were selected based on pilot experiments demonstrating reliable instruction following behavior in Arabic. The extracted triples are converted into a consistent (subject, relation, object) format, lightly normalized using simple rule-based canonicalization, and deduplicated using exact matching before comparison. Triples are stored separately for the passage, the gold answer, and the model generated answer.

Although this extraction process is automated and reproducible given fixed inputs and prompts, minor variation may arise due to the stochastic nature of LLM based extraction. To minimize extraction bias, we use identical prompt templates and extraction settings for all text sources, including the passage, the gold answer, and the model generated answer. As a result, differences observed in the extracted triples reflect differences in the underlying content rather than artifacts of the extraction process.

3.4 Knowledge Graph Construction

From the extracted triples, we build a reference knowledge graph for each passage. Nodes correspond to entities or concepts, and edges encode semantic relations expressed in the text. The resulting graph is intentionally lightweight and does not depend on an external ontology; instead, its structure mirrors the factual content of the passage. Equivalent graphs are constructed for the gold answer and the model generated answer, enabling systematic comparison at the level of individual triples across all components.

3.5 Question-Aware Filtering

Because factual correctness alone does not guarantee a valid answer in reading comprehension, we explicitly model the semantic focus of each question. Questions are grouped into coarse answer categories (e.g., entity, reason, time, location) using a lightweight LLM based classifier, supplemented with simple rules when needed.

This categorization guides which triples are treated as relevant during evaluation. When multiple entities of the same type appear in the passage, relevance is not decided by type alone; it is assessed through alignment with the relation implied by the question, and genuinely ambiguous cases are deferred to the human adjudication stage.

Label	Interpretation (after question-aware filtering)
Supported	The triple is grounded in the passage and is also present in the gold answer.
Missing	The triple is supported by the passage and included in the gold answer, but is omitted from the model generated answer.
Hallucinated	The triple is introduced by the model generated answer but lacks support from the passage and is absent from the gold answer.
Gold-only	The triple appears in the gold answer despite not being explicitly supported by the passage, typically reflecting annotation inconsistencies.
Off-focus	The triple is supported by the passage but does not align with the semantic intent of the question.

Table 3: Interpretation of automatic triple labels used in the framework.

We note that these rules may not generalize to less common or highly implicit formulations. To evaluate their reliability, we manually examined a representative subset of questions and observed an accuracy of approximately 93% in question type classification. Although this result is encouraging, it also points to the need for more flexible and context aware approaches to question understanding in future work.

3.6 Comparison, Detection, and Error Typing

After question-aware filtering, we compare triples extracted from the model generated answer with those derived from the passage and the gold answer. This comparison is performed automatically and forms the basis for identifying different types of model behavior. At this stage, we focus on assigning coarse labels that reflect factual support and question relevance, without making any final judgment about answer quality.

These labels distinguish unsupported content (hallucinations), omitted information, and factually correct but question irrelevant triples. The labeling step is purely automatic and does not yet account for alternative valid answers or annotation limitations.

3.7 Human Adjudication

Triples labeled as Off-focus, supported but absent from the gold answer, or appearing only in the gold answer are flagged for manual review; all other cases are handled automatically.

Label	Count	Interpretation
Supported	8,628	Passage supported and aligned with the question focus.
Missing	5,150	Supported by the passage and present in the gold answer, but omitted by the model.
Hallucinated	3,377	Present only in the model answer and unsupported by the passage.
Gold-only	5,202	Present in the gold answer but not supported by the passage.
Off-focus	565	Passage supported but not aligned with the semantic intent of the question.

Table 4: Distribution of triple-level labels assigned by the knowledge graph based diagnostic framework.

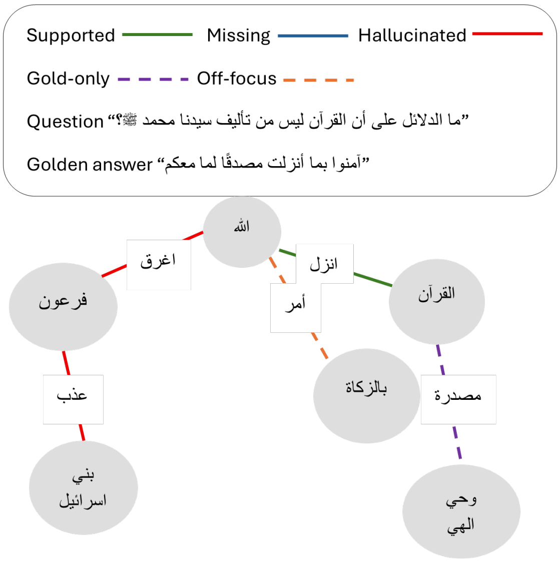

Supported
Missing
Hallucinated
Gold-only
Off-focus
Question “ما الدلائل على أن القرآن ليس من تأليف سيدنا محمد ﷺ؟”
Golden answer “آمنوا بما أنزلت مصدقًا لما معكم”
الله
اغرق
فرعون
انزل
أمر
القرآن
عذب
بالزكاة
مصدرة
بني اسرائيل
وحي الهي

Figure 3: Sample output from the knowledge graph based diagnostic framework. Edge colors indicate triple labels: supported, missing, hallucinated, gold-only, and off-focus.

Human annotators determine whether such cases reflect genuine omissions, incomplete gold annotations, or valid but alternative passage grounded answers. This step ensures that evaluation remains fair and does not penalize models for producing correct information that falls outside narrow reference answers.

4 Results and Analysis

4.1 Knowledge Graph Triple Analysis

Applying the proposed framework to model generated answers resulted in a total of 23,388 extracted triples from both gold standard and model outputs. Of these, 6,233 triples were flagged for further inspection due to uncertainty arising from question relevance or discrepancies between passage support and gold annotations. Table 4 summarizes the distribution of triple-level labels assigned by the framework, including both automatically accepted and adjudicated cases.

Figure 3 shows a representative knowledge graph produced by the framework, where edges are color coded according to their assigned labels. This visualization provides an intuitive overview of how different error types are distributed within a single model response and supports rapid inspection by human adjudicators.

To better understand where models struggle most, Table 5 lists a subset of questions that triggered the highest number of flagged triples. These examples are shown in both Arabic and English for clarity.

A clear pattern emerges across these examples: questions that require justification, evidence, or abstract interpretation tend to generate substantially more flagged triples than fact based questions. Such questions often cannot be answered by a single explicit fact in the passage and instead require synthesis or reasoning, increasing the risk of unsupported or off-focus content.

With respect to extraction quality, the triple extraction component achieved a precision of 74.6% and a recall of 57.7% when evaluated against passage derived reference graphs. These results suggest that many relevant relations are successfully recovered, while also revealing limitations in coverage and faithfulness that influence downstream hallucination analysis.

Closer manual inspection sheds light on recurring failure modes. Hallucinated triples often took the form of plausible yet unsupported statements, such as generic religious assertions not explicitly stated in the passage. Missing triples, in contrast, frequently reflected omitted secondary entities or relations that were expressed implicitly rather than stated directly. We also observed a substantial number of gold-only triples, indicating that some gold answers contain implicit or interpretive information that is not explicitly stated in the passage, and/or that certain relations confirmed by annotators are difficult to recover reliably through automatic extraction in Classical Qur'anic Arabic.

Taken together, these results highlight the value of structured, question-aware evaluation in exposing hallucination patterns that surface level similarity metrics often fail to capture.

Question in English	Question in Arabic	Flagged Triples
What evidence shows that the Qur'an is not written by our master Muhammad (PBUH)?	ما الدلائل على أن القرآن ليس من تأليف سيدنا محمد ﷺ؟	1011
What are the verses that discuss the topic of the will?	ما هي الآيات التي تتحدث عن موضوع الوصية؟	82
What is the evidence that the Qur'an is valid for all times and places?	ما الدليل على أن القرآن صالح لكل زمان ومكان؟	50

Table 5: Examples of questions associated with the highest number of flagged triples.

5 Conclusion

This paper introduced a knowledge graph–based diagnostic evaluation framework for analyzing hallucinations in LLM generated answers for Arabic machine reading comprehension, with a particular emphasis on Qur'anic passages. Through symbolic, triple level comparison and question-aware filtering, the framework supports systematic analysis of hallucinations and semantic misalignment in a high stakes, low resource setting. Our findings indicate that questions requiring justification, evidence, or interpretation remain especially challenging for current models, revealing limitations that are not readily captured by surface level similarity metrics.

The analysis also highlights both the strengths and limitations of the framework. While it is effective at identifying unsupported, missing, and off-focus content, it remains limited in its ability to capture deeper forms of reasoning and implicit semantic relations. These challenges are especially pronounced in cases where gold answers themselves involve interpretation or external contextual knowledge.

Looking ahead, an important direction for future work is the integration of symbolic reasoning techniques, such as rule based inference and multi-hop reasoning over knowledge graphs, to better capture complex dependencies and higher level semantic inconsistencies. We also expect that the methodology introduced here can be adapted to other Arabic tasks and low resource domains where transparency, interpretability, and factual reliability are critical.

Limitations

While the proposed framework supports structured and interpretable analysis of hallucinations in Arabic machine reading comprehension, several limitations remain.

First, the framework relies on LLM based triple extraction. Although identical prompts and settings are used across passages, gold answers, and model outputs, the stochastic nature of LLMs can introduce minor variability that affects downstream comparison.

Second, the framework operates under weak supervision and depends on the quality of gold answers. Some gold answers include implicit or interpretive information that is not explicitly stated in the passage, resulting in gold-only triples. Targeted human adjudication helps address such cases but limits full automation.

Third, question-aware filtering relies on coarse semantic categories and may fail for highly implicit or uncommon question formulations, particularly in Classical Qur'anic Arabic. In addition, the framework focuses on explicit factual relations and does not fully capture deeper implicit reasoning or theological interpretation.

Finally, the effectiveness of the framework depends on the accuracy of upstream components such as triple extraction and question classification. Improving these components through more robust semantic representations and symbolic reasoning is left for future work.

Acknowledgments

The authors would like to acknowledge the support provided by Saudi Data and AI Authority (SDAIA) and King Fahd University of Petroleum and Minerals (KFUPM) under SDAIA-KFUPM Joint Research Center for Artificial Intelligence grant No. JRC-UCG-09.

References

Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. 2023. Can knowledge graphs reduce hallucinations in llms?: A survey. *arXiv preprint arXiv:2311.07914*.

Sadam Al-Azani, Maad Alowaifeer, Alhanoof Alhunief, and Ahmed Abdelali. 2025. Ontologyrag-q: Resource development and benchmarking for retrieval-augmented question answering in qur'anic tafsir. In *Proceedings of the 30th Conference on Empirical*

Methods in Natural Language Processing, pages 15551–15569.

Najwa AlGhamdi, Sadam Al-Azani, Kwabena Nuamah, and Alan Bundy. 2025. Evaluating llms on arabic reading comprehension: Errors, hallucinations, and factual inconsistencies. In *Proceedings of the 7th International Conference on AI in Computational Linguistics (ACLing 2025)*, Procedia Computer Science, Dubai, United Arab Emirates. Accepted.

Abdulaziz Alqahtani, Mohammed Alothman, Tamer Elsayed, Abdulmohsen Al-Thubaity, and Wael Shalaby. 2022. Quranqa: A qur'an question answering dataset for reading comprehension and beyond. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 4484–4494.

Zaid Alyafeai, Hamza Alkaoud, and Arabic BERT. 2021. Araelectra: Pre-training text discriminators for arabic language understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 133–143.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools*, pages 9–15.

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, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan, Dr. Majed Alrubaian, Ali Alammari, Zaki Alawami, and 7 others. 2025. ALLam: Large language models for arabic and english. In *The Thirteenth International Conference on Learning Representations*.

Anthony Chen, Panupong Pasupat, Sameer Singh, Hongrae Lee, and Kelvin Guu. 2023. Purr: Efficiently editing language model hallucinations by denoising language model corruptions. *arXiv preprint arXiv:2305.14908*.

Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, Shizhu Liu, Pingchuan Tian, Yuping Wang, and Yuxuan Wang. 2023. Halo: Estimation and reduction of hallucinations in open-source weak large language models. *arXiv preprint arXiv:2308.11764*.

Xinyue Fang, Zhen Huang, Zhiliang Tian, Minghui Fang, Ziyi Pan, Quntian Fang, Zhihua Wen, Hengyue Pan, and Dongsheng Li. 2025. Zero-resource hallucination detection for text generation via graph-based contextual knowledge triples modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23868–23877.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2024. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18126–18134.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Zijian Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Yifan Sun, Pascale Fung, Soroush Vosoughi, and Jiwei Wen. 2023. Survey of hallucination in natural language generation. volume 55, pages 1–38.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

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. Association for Computational Linguistics (ACL).

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*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1906–1919.

Timothy R McIntosh, Tong Liu, Teo Susnjak, Paul Watters, Alex Ng, and Malka N Halgamuge. 2023. A culturally sensitive test to evaluate nuanced gpt hallucination. *IEEE Transactions on Artificial Intelligence*, 5(6):2739–2751.

Sewon Min, Taehwan Jung, Qian Lin, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained evaluation of factual consistency in abstractive summarization.

In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 10498–10512.

Mohanad Mohamed and Sadam Al-Azani. 2025. Enhancing arabic nlp tasks through character-level models and data augmentation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2744–2757.

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.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*.

OpenAI. 2023. Chatgpt: Gpt-4 (mar 14 version). https://chat.openai.com/chat. Accessed: 2025-07-07.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Hannah Sansford, Nicholas Richardson, Hermina Petric Maretic, and Juba Nait Saada. 2024. Grapheval: A knowledge-graph based llm hallucination evaluation framework. *arXiv preprint arXiv:2407.10793*.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023a. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. 2023b. Mitigating language model hallucination with interactive question-knowledge alignment. *arXiv preprint arXiv:2305.13669*, 3.

Wenxuan Zhang, Kun Song, Sebastian Riedel, and Antoine Bosselut. 2023c. Benchmarking graph-based hallucination detection for factual consistency in summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4403–4418.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).