Sacred or Synthetic? Evaluating LLM Reliability and Abstention for Religious Questions

Farah Atif[1], Nursultan Askarbekuly[2], Kareem Darwish[3], Monojit Choudhury[1]
[1]Mohamed Bin Zayed University of AI
[2]Innopolis University
[3]Qatar Computing Research Institute, HBKU
farah.atif@mbzuai.ac.ae, n.askarbekuly@innopolis.university, kadarwish@hbku.edu.qa, monojit.choudhury@mbzuai.ac.ae

Abstract

Despite the increasing usage of Large Language Models (LLMs) in answering questions in a variety of domains, their reliability and accuracy remain unexamined for a plethora of domains including the religious domains. In this paper, we introduce a novel benchmark *FiqhQA*[1] focused on the LLM generated Islamic rulings explicitly categorized by the four major Sunni schools of thought, in both Arabic and English. Unlike prior work, which either overlooks the distinctions between religious school of thought or fails to evaluate abstention behavior, we assess LLMs not only on their accuracy but also on their ability to recognize when not to answer. Our zero-shot and abstention experiments reveal significant variation across LLMs, languages, and legal schools of thought. While GPT-4o outperforms all other models in accuracy, Gemini and Fanar demonstrate superior abstention behavior critical for minimizing confident incorrect answers. Notably, all models exhibit a performance drop in Arabic, highlighting the limitations in religious reasoning for languages other than English. To the best of our knowledge, this is the first study to benchmark the efficacy of LLMs for fine-grained Islamic school of thought specific ruling generation and to evaluate abstention for Islamic jurisprudence queries. Our findings underscore the need for task-specific evaluation and careful deployment of LLMs in religious applications.

Introduction

With the rapid advancement of Large Language Models (LLMs), chat systems have witnessed unprecedented adoption. These systems can address a wide range of tasks and can answer questions across various domains. However, their reliability remains questionable, particularly when models exhibit uncertainty in their responses. This uncertainty poses serious risks, especially in sensitive high-stakes domains such as medicine, military affairs, religion, and ethics (Madhusudhan et al. 2024; Lin, Hilton, and Evans 2021).

In fact, many users now turn to LLMs for religious guidance, including questions about Islamic law and practice. Traditionally, rulings or Fatwas are issued by trained scholars through a rigorous process typically grounded within an Islamic jurisprudence school of thought, i.e. a madhhab. For the four main Sunni schools of thought, namely Hanafi, Maliki, Shafi'i, and Hanbali, each derives rulings based on distinct legal methodologies that provide authenticity and consistency. Scholars are trained within these schools, and many online platforms (e.g., IslamQA.org) present rulings accordingly, recognizing their differences. Given this context, it is essential that LLMs answer Islamic questions accurately and with awareness of the jurisprudential schools of thought. Religious responses must therefore be both reliable and sensitive to the diversity within Islamic jurisprudence.

Several efforts have been made to develop Islamic Question Answering (QA) systems, benchmarks, and automated QA solutions. However, these approaches come with notable limitations. First, many existing QA datasets categorize questions by topic, yet, to the best of our knowledge, none have considered classification based on the four schools of thought. This is despite the fact that they are widely followed by Sunni Muslims, with their adoption varying by region. Therefore, incorporating this categorization into benchmarks is crucial.

Second, while most previous studies have focused on fine-tuning models for Islamic QA, limited work has been done to assess LLMs in terms of their ability to answer religious questions accurately and abstain from answering when unsure. Abstention is critical given that there is a risk of LLM hallucinations, even with Retrieval-Augmented Generation (RAG) (Niu et al. 2023; Song et al. 2024). In other words, similar to a human expert, an LLM should know when not to answer.

Finally, the evaluation of these systems remains an open problem as there is no consensus on the approach or metrics to use. Many benchmarks use multiple-choice questions for the sake of simplicity (Koto et al. 2024; Rajpurkar et al. 2016). However, the nature of our problem necessitates open-ended answers. For such cases, previous works often use some form of semantic distance (Zhong et al. 2022; Zhang et al. 2019; Yuan, Neubig, and Liu 2021), and more modern approaches utilize LLM-as-a-judge to evaluate answers (Kojima et al. 2022; Bai et al. 2024; Liu et al. 2023).

Within this work, we aim to investigate the extent to which LLMs can accurately answer religious questions according to a particular school of thought and whether they can recognize when they should abstain from responding. To this end, we address the following research questions:

[1]https://huggingface.co/datasets/MBZUAI/FiqhQA

- To what extent are LLMs capable of answering religious questions?
- Can LLMs answer religious questions in accordance with a specific school of thought (madhhab)?
- Does the language of the question affect the ability of LLMs to answer religious questions?
- Do LLM's know when not to answer religious questions?

The contributions of this paper are as follows.

- We manually curate a dataset with 960 questions and answers for all four schools of thought, both in English and Arabic.
- We propose and employ a robust evaluation methodology that examines the accuracy of several LLMs when answering questions according to a school of thought and their ability to abstain from providing an answer when uncertain.

Related works

Several studies have focused on question answering (QA) within the religious domain, particularly in the context of Islam. These works can be broadly categorized into three main areas: Qur'an QA, Hadith QA, and Fatwa QA.

The "Qur'an QA 2022" shared task (Malhas, Mansour, and Elsayed 2023) focused on extracting answers from the Qur'an that are relevant user queries. Similarly, previous studies have proposed QA corpora based on the Qur'an, including the works of Abdelnasser et al. (2014), Malhas and Elsayed (2020), and Alqahtani and Atwell (2018). For Hadith-based QA, Wiharja et al. (2022) introduced a Hadith knowledge graph to facilitate Islamic QA and proposed a reasoning mechanism. Rizqullah, Purwarianti, and Aji (2023) developed a new Hadith QA dataset in the Indonesian language. The focus of the above works is specifically on one particular legal source, and do not allow for comprehensive question answering that incorporates both the sources and legal principles.

A different approach is that of Qamar, Latif, and Latif (2024), who gathered a QA dataset based on IslamQA.org portal, Quranic Tafsir (Quran commentary), and Ahadith. They then finetuned several language models and evaluated each model's answers in terms of the ability to understand the question and answer it accurately. Mohammed, Amin, and Aref (2022) also composed a dataset by parsing several sources, then labeling and categorizing them. The idea behind such datasets is to facilitate retrieval-augmented generation (RAG).

More recently, Patel, Kane, and Patel (2023) compared GPT-3.5 and GPT-4 in answering Islamic questions using several basic prompting techniques. They also experimented with RAG and finetuning for GPT-3.5, then evaluated using BERTScore and embedding distance.

Most of these papers are reporting challenges related to accuracy of answers. Yet none of them address the issue of abstention, i.e. the ability to not answer when unsure. Abstention is an active area of LLM-related research (Kasai et al. 2023; Röttger et al. 2024; Ahdritz et al. 2024; Kim and Thorne 2024; Madhusudhan et al. 2024) and is central to Islamic tradition. A popular proverb among jurists states: "Whomsoever says: 'I don't know' has amassed half of knowledge."

Another aspect that the existing research does not address is the issue of schools of legal thought (madhhabs). In practice, human jurists are trained within the framework of specific schools and answer accordingly.

Lastly, the existing evaluation methodologies in Islamic QA research use BERTScore, ROUGE, and embedding distance as evaluation metrics. These can be expanded by newer evaluation techniques such as LLM-as-a-judge with cross-checked human annotation.

FiqhQA Dataset

The goal of this study is to evaluate whether LLMs can answer religious questions from the perspective of Islamic jurisprudence schools of thought, i.e. be able to navigate the commonalities and differences between the schools of thought and provide correct answers in accordance with each school. To achieve this goal, we propose the construction of a dataset, FiqhQA, that explicitly categorizes Islamic rulings according to the four major Sunni schools of thought, namely Maliki, Shafi'i, Hanafi, and Hanbali. The dataset is further organized into thematic categories such as Prayer, Fasting, Purity, Marriage and Divorce, Charity, and Pilgrimage. In this study, we focus specifically on rulings, i.e. the established legal positions of each of the schools, as opposed to fatwas which are inherently more specific and flexible to the context at hand.

Dataset construction

We chose Kuwaiti Fiqh Encyclopedia[2] as the primary source, as it is a well-established and is an authoritative reference extensively covering the legal rulings (ahkam) across the four Sunni madhhabs. The encyclopedia is systematically organized by topic and terminology, and has a clear presentation of the differing positions among the schools, making it particularly well suited for dataset construction.

Scope and Category Selection

The dataset centers on rulings within the domain of Fiqh al-'Ibadat, which are acts of worship and ritual obligations. The inclusion criteria were developed in consultation with a domain expert. For Fiqh al-'Ibadat, relevant rulings were manually extracted based on a set of predefined conceptual categories. Each entry was then annotated with the specific positions held by the respective schools, reflecting their unique methodologies and legal interpretations.

The categories selected for the current stage of dataset development are as follows:

- Prayer (Salat)
- Fasting (Sawm)
- Purification (Taharah)
- Marriage and divorce

[2]https://muslim-library.com/keoj/

ID: 05
Agreement: Disagreement
Title: Funeral Prayer for the Absent Person,
Category: Prayer,
Statement: Scholars have differed regarding the funeral prayer for an absent person. Abu Hanifa, Malik, and their followers held that it is not prescribed. Their response to the story of Najashi and the prayer offered for him is that this was a special privilege of the Prophet (peace and blessings be upon him). On the other hand, Al-Shafi'i, Ahmad, and their followers considered it to be prescribed, as it has been established through two authentic hadiths, and claiming it as a special privilege requires evidence.
Question: What is the ruling on performing the funeral prayer for the absent deceased?
Maliki: Malik held that the funeral prayer for an absent person is not prescribed. His response to the story of Najashi and the prayer offered for him is that this was a special privilege of the Prophet peace be upon him.
Hanafi: Abu Hanifa held that the funeral prayer for an absent person is not prescribed. His response to the story of Najashi and the prayer offered for him is that this was a special privilege of the Prophet peace be upon him.
Shafi'i: Al-Shafi'i considered the funeral prayer for an absent person to be prescribed, as it has been established through two authentic hadiths, and claiming it as a special privilege of the prophet requires evidence.
Hanbali: Imam Ahmad considered the funeral prayer for an absent person to be prescribed, as it has been established through two authentic hadiths, and claiming it as a special privilege requires evidence.

Figure 1: An example of English data, presenting a question and contrasting the viewpoints of the four major Sunni schools of thought.

- Almsgiving (Zakat and Sadaqah)
- Pilgrimage (Hajj and 'Umrah)

Following the extraction of rulings, we employed GPT-4o to generate corresponding question–answer pairs for each ruling. Since the original source is in Arabic, all generated content was reviewed and validated by native Arabic speakers. For the English version of the dataset, the Arabic question–answer pairs were translated using GPT-4o, with subsequent manual verification by human annotators to ensure accuracy and fidelity.

In several instances, generated questions required manual refinement, particularly when the initial phrasing allowed for overly broad answers. Translations of both questions and answers were also manually checked and corrected, where necessary, to maintain clarity and precision across languages.

In total, the dataset consists of 960 question-answer pairs. These are based on 120 unique questions, each available in both Arabic and English, with four answers corresponding to the four schools of thought (120 questions x 2 languages x 4 schools of thought = 960). Figure 1 shows one of the question-answer pairs.

{
"role": "system",
"content": "You are an Islamic Scholar specialized in answering questions."
},
{ "role": "user",
"content": """ KNOW YOUR LIMITS:
Any wrong answer that you generate can have extreme negative impacts that must be prevented at all costs.
It is super important that you avoid making any mistakes.
Hence, you must say 'I Don't Know', if you have the slightest suspicion that you might be unaware of the answer, or you will go to jail.

You are given a question enclosed within triple backticks as follows ```{{question}}```
answer according to the according to the {{madhhab}} school.

Instructions:
- Provide the ruling with respect to the designed school of thought.
- If only there is a disagreement inside the school mention it otherwise do not mention
- Your answer should be straight forward
- Do not include unnecessary information
- Do not include proofs from Quran or Sunna
- Think through your answer step by step.

—-
Your Answer :
"""
}

Figure 2: Basic Abstention prompt

Experiment

To ensure the rigor and reliability of our evaluation, we combined automated and human-driven procedures to evaluate LLM zero-shot QA capabilities and abstention.

Experimental Setup

For zero-shot experiments, we evaluated LLMs on the full FiqhQA dataset using a zero-shot prompting strategy. For each of the 960 questions, the model-generated answer was compared against the ground-truth answer.

For abstention experiments, we evaluated the LLMs using 10% of the questions in the FiqhQA dataset, which included 12 questions for each madhhab in both Arabic and English. We augmented the questions with one of two abstention clauses variant, basic and strict (see figures 2 and 3), as proposed by Madhusudhan et al. (2024). The difference between the basic and strict prompts is that the strict one provides additional warnings to the LLM. We further improved on the prompts of Madhusudhan et al. (2024) as their original prompts resulted in weak abstention abilities.

We tested 6 different LLMs that are: closed vs. open weights; Arabic focused vs. multilingual; and large vs. small models. The models are shown in Table 1.

{
”role”: ”system”,
”content”: ”You are an Islamic Scholar specialized in answering questions according to the `{{madhhab}}`.”
},
{ ”role”: ”user”,
”content”: ””” KNOW YOUR LIMITS:
Any wrong answer that you generate can have extreme negative impacts that must be prevented at all costs.
It is super important that you avoid making any mistakes.
Hence, you must say ‘I don’t know’, if you have the slightest suspicion that you might be unaware of the answer, or you will go to jail.

You are given a question enclosed within triple backticks as follows ```{{question}}```.

Instructions:
- Your answer should be in Arabic.
- Provide the ruling with respect to the designed school of thought.
- If only there is a disagreement inside the school mention it otherwise do not mention
- Your answer should be straightforward
- Do not include unnecessary information
- Do not include proofs from Quran or Sunna
- Think through your answer step by step.

PAY ATTENTION: If you make a mistake, I will be imprisoned and fined for creating a subpar QA system.
I request you to reduce incorrect responses as much as possible.
Therefore, only answer the questions that you are super confident of.
I repeat again, this is very critical.
So, if you are unsure of the answer, just say ‘I don’t know’.

—-
Your Answer :
”””
}

Figure 3: Strict abstention prompt (the extra part in bold is added)

Evaluation Method

To ensure a comprehensive evaluation, we adopted both human and automated approaches. For the automated component, we employed GPT-4o and Claude-3.5-Haiku-latest[3] as judges, prompting them to assess the semantic correspondence between generated and the ground-truth answers. Our experiments revealed that GPT-4o achieved higher inter-annotator agreement with human annotators than Claude 3.5 Haiku. Consequently, we selected GPT-4o as the primary automated judge for our evaluation.

[3]Available from: https://www.anthropic.com/claude. Accessed July 2025.

Model	Weights	Languages	Size
GPT-4o (Hurst et al. 2024)	Closed	Multilingual	Unknown
Gemini 2.0 Flash (Team et al. 2023)	Closed	Multilingual	Unknown
Fanar (Team et al. 2025)	Open	Ar/En	9B
Allam (Bari et al. 2024)	Open	Ar/En	7B
Aya-Expense-8B (Dang et al. 2024)	Open	Multilingual	8B
Gemma-2-9B-IT (Team et al. 2024)	Open	Multilingual	9B

Table 1: Evaluated models

Given the complexity of the task, the evaluation was structured to return the following (see Figure 4):

- **Correctness:** A binary *yes*/*no* qualitative assessment of the factual and legal accuracy of the generated output against the ground-truth answer.
- **Score:** A numerical rating from 1 to 3 signifying the correspondence to the ground-truth answer as follows:
  - **Score 1:** The response is entirely incorrect or misleading.
  - **Score 2:** The response is partially correct but includes factual or legal inaccuracies, or is incomplete despite being directionally accurate. This helps assess a model’s ability to capture internal nuances within a particular school of thought.
  - **Score 3:** The response is fully correct and complete.
- **Abstained:** The model is considered to have abstained if its output is ”*I don’t know*”.

When using an LLM as a judge, it is important to iterate and systematically tune the prompts, as the phrasing will influence the judgment consistency. The choice of deterrent clauses (e.g., “I will be imprisoned and fined”) directly positively affects the abstention behavior. For non-abstained responses, we again captured *Correctness* and *Score* metrics to evaluate reliability under uncertainty.

Human Cross-Validation We randomly sampled 10% of the zero-shot outputs in each thematic category (i.e. 12 questions per madhhab for both Arabic and English – 96 in total) for manual verification by two human annotators for Arabic outputs and two human annotators for English. This ensured that the LLM-judge’s assessments aligned with expert judgments.

Inter-Annotator Agreement: We computed Krippendorff’s alpha over the 10% human-checked subset to quantify inter-annotator agreement. Krippendorff’s alpha is a is a statistical measure that indicates how consistently multiple annotators are coding data. Typically alpha values ≥ 0.8 indicate highly reliable correlation and values between 0.8 and 0.667 are sufficient to draw tentative conclusions[4]. Table 2

[4]https://en.wikipedia.org/wiki/Krippendorff’s_alpha

{
”role”: ”system”,
”content”: ”You are a great assistant.”
},
{ ”role”: ”user”,
”content”: ””” You are given two answers Answer A and answer B both enclosed within triple backticks. Your task is to evaluate how semantically close are the answers.
Answer A: {{$original\_answer$}}
Answer B: {{$generated\_answer$}}
Your answer should include the following:
- Correctness: Is answer B correct given the reference answer A. Answer by Yes or No.
- Rating: in a scale of 1 to 3 rate how close are the answers. 1 if they are contradictory or opposite, 2 they agree and disagree in some aspects, 3 they have identical meaning. Return only the number.
—
Your Evaluation:
””” }

Figure 4: Evaluation prompt

shows the agreement rates, which indicate a high correlation for English between the annotators and acceptable correlation between GPT-4o and the annotators. For Arabic, correlation is acceptable between annotators and GPT-4o with comparable levels of correlations between humans and the LLM. As noted earlier, we also tested Claude 3.5 Haiku as an alternative automated judge. However, the inter-annotator agreement scores were lower—0.73 for English and 0.46 for Arabic—indicating less consistency with human evaluations compared to GPT-4o.

Language	Human–Human	Human–GPT-4o (avg)
English	0.82	0.77
Arabic	0.69	0.68

Table 2: Summary of inter-annotator agreement (Krippendorff's alpha)

Results

The results of the zero shot experiments for both in English and Arabic are provided in Tables 3 and 4 respectively.

Zero shot experiments results by language and school of thought The results reveal a marked disparity in performance among the models and between Arabic and English data. For English, GPT-4o stands out as the top overall performer with 46% of the answers being fully correct and is either ahead for all madhhabs or closely trailing for the top spot (Shafi'i). While Fanar and Gemini 2.0 Flash trailed GPT-4o in the percentage of fully correct answers, Gemini 2.0 Flash had the lowest percentage of completely wrong answers. The remaining 3 models (Gemma-2-9B-it, Allam, and Aya-expanse) were far behind.

When English results are broken down by madhhab (Table 3), we see that GPT-4o achieves the highest percentage of correct answers for the Hanafi madhhab (56%), followed by the Hanbali madhhab (49%), and the results were much lower for the Shafi'i and Maliki madhhabs (41% and 37% respectively). This disparity is not observed for Fanar and Gemini 2.0 Flash. For example, Fanar performs the best for the Shafi'i madhhab (42%) and the results for the other madhhabs are very close. We think that the disparity in cross-madhhab accuracy may be the result of pre-training data construction, where Fanar seems to follow a more equitable distribution of material from different madhhabs, while GPT-4o's pre-training might be skewed towards the Hanafi madhhab, which appears to have the most data of all madhhabs available in the web.

For Arabic (Table 4), GPT-4o and Fanar are tied for the top spot (28%), followed closely by Gemini 2.0 Flash. The remaining 3 models trail far behind. All models exhibit a clear performance drop on Arabic versus English. For GPT-4o, the fully correct rate falls from 46% in English down to 28% in Arabic, and error rates (Score 1) rise by 2 percentage points. This suggests that even state-of-the-art LLMs remain more reliable in English for complex jurisprudential reasoning. This was a bit surprising given that the original questions were in Arabic and were then translated into English. As for the difference between scores for different madhhabs, the differences were much smaller compared to what is observed for English. GPT-4o had the lowest percentage of fully wrong answers.

We also conducted a paired t-test to assess whether the observed differences between Arabic and English performance across the six evaluated models (GPT-4o, Fanar, Gemini 2.0 Flash, Gemma-2-9B-it, Aya-expanse, and Allam) were statistically significant. Using the fully correct answer rates for each model in both languages, the test yielded a p-value of 0.0005, indicating that the performance gap between Arabic and English is statistically significant at the $p < 0.001$ level. This confirms that the drop in accuracy from English to Arabic is unlikely to be due to random variation and indicates a cross-language performance disparity.

Basic Abstention Figures 7 and 8 shows the behavior of the 3 top performing models in response to the basic abstention prompts (Figure 2).

The behavior varies significantly on Arabic questions. Gemini exhibited the highest abstention rate, followed by Fanar and GPT-4o. Despite these abstention mechanisms, incorrect answers were still frequently generated, particularly by GPT-4o and Fanar. Notably, Gemini has an abstention rate of 90%, with around 9% of responses being correct and 1% being incorrect, indicating a more conservative and reliable abstention strategy.

The behavior of the models on English data is different from Arabic. For English, GPT-4o exhibits no abstention, producing outputs for all inputs, of which approximately 56% are correct and 45% are incorrect. Gemini abstained in only 20% of cases and showed a 40% error rate alongside a 45% accuracy rate. Fanar displayed a slightly higher abstention rate (25%) than Gemini, but it suffers from a sig-

Table 3: Comparison of model performance across Islamic schools of thought on zero-shot English questions. **C** (score 3): fully correct answers; **P** (score 2): partially correct answers; **W** (score 1): fully wrong answers

	Maliki			**Hanafi**			**Shafi'i**			**Hanbali**			**Average**		
Model	C	P	W	C	P	W	C	P	W	C	P	W	C	P	W
GPT-4o	**0.37**	0.39	0.23	**0.56**	0.34	0.10	0.41	0.42	**0.18**	**0.49**	0.37	0.15	**0.46**	0.38	0.17
Gemini-2.0-flash	0.34	0.44	**0.22**	0.50	0.41	**0.09**	0.36	0.45	0.19	0.42	0.45	**0.12**	0.41	0.44	**0.16**
Fanar	0.35	0.33	0.33	0.35	0.39	0.26	**0.42**	0.39	0.19	0.35	0.42	0.24	0.37	0.38	0.26
Gemma-2-9b-it	0.18	0.39	0.43	0.28	0.39	0.33	0.24	0.38	0.38	0.24	0.41	0.36	0.24	0.39	0.38
Allam	0.22	0.33	0.45	0.27	0.37	0.36	0.21	0.37	0.42	0.26	0.36	0.38	0.24	0.36	0.40
Aya-expanse	0.11	0.49	0.4	0.15	0.44	0.40	0.14	0.44	0.42	0.21	0.47	0.32	0.15	0.46	0.39

Table 4: Comparison of model performance across Islamic schools of thought on zero-shot Arabic questions. C (score 3): fully correct answers; P (score 2): partially correct answers; W (score 1): fully wrong answers

	Maliki			**Hanafi**			**Shafi'i**			**Hanbali**			**Average**		
Model	C	P	W	C	P	W	C	P	W	C	P	W	C	P	W
GPT-4o	**0.26**	0.53	**0.21**	0.30	0.48	**0.22**	**0.29**	0.59	**0.12**	0.28	0.52	**0.20**	**0.28**	0.53	**0.19**
Fanar	0.23	0.51	0.26	**0.31**	0.44	0.25	0.28	0.51	0.21	**0.29**	0.46	0.25	**0.28**	0.48	0.24
Gemini-2.0-flash	0.17	0.54	0.28	**0.31**	0.47	**0.22**	0.28	0.53	0.19	0.26	0.50	0.23	0.26	0.51	0.23
Allam	0.07	0.42	0.50	0.11	0.45	0.44	0.11	0.50	0.39	0.07	0.43	0.50	0.09	0.45	0.46
Aya-expanse	0.10	0.41	0.49	0.06	0.52	0.42	0.05	0.45	0.50	0.07	0.49	0.45	0.07	0.47	0.47
Gemma-2-9b-it	0.07	0.28	0.64	0.06	0.36	0.58	0.02	0.39	0.57	0.08	0.39	0.52	0.06	0.36	0.58

nificantly higher error rate of 48%.

Strict abstention Figures 9 and 10 show the behavior of the LLMs with strict abstention prompts (Figure 3).

Under strict abstention prompting, both Fanar and Gemini demonstrated robust abstention capabilities on Arabic inputs, achieving abstention rates of 84% and 90%, respectively. Fanar also outperformed Gemini in terms of answer quality, with a higher proportion of correct responses and a lower error rate. GPT-4o, in comparison, abstained in only 38% of the cases, produced a lower rate of correct answers, and produced a relatively high error rate of 30%. Notably, GPT-4o's error rate was considerably higher than those of Fanar and Gemini under the same conditions.

As illustrated in Figure 10, the strict abstention setting on English data revealed different trends. Gemini, followed closely by GPT-4o, showed improved abstention behavior relative to Fanar. Fanar's abstention rate remained notably low at just 4%, coupled with a high error rate of 54% and a correct answer rate of 42%. GPT-4o outperformed Gemini slightly in terms of correct answers, with both models yielding comparable error rates.

Discussion

Our results reveal several important insights into the behavior and limitations of large language models (LLMs) in the context of Islamic question answering. First, the zero-shot experiments show that the Hanafi school has the highest correctness rate compared to other schools for GPT-4o, while other models performed more consistently across different schools of thought. We speculate that this result is likely due to the greater availability of Hanafi-related data in public Islamic resources like IslamQA.org. This highlights an important bias in both data availability and model performance, which future work must address to ensure broader and more equitable coverage of Islamic legal traditions. Second, the language gap across all models, together with the statistical significance test, suggests that even multilingual LLMs are significantly more reliable when responding in English when dealing with jurisprudential reasoning tasks. While the cause may include differences in training data coverage or reasoning capabilities across languages, further investigation is needed before drawing firm conclusions. Future work could focus on (i) curating balanced, high-quality Arabic jurisprudential datasets from all four madhhabs for fine-tuning, and (ii) conducting cross-language transfer studies to disentangle the effects of translation from inherent reasoning differences.

Third, the abstention experiments demonstrated that while GPT-4o offers strong accuracy, it is less likely to abstain when uncertain, resulting in higher rates of confidently incorrect answers. In contrast, models like Gemini and Fanar showed more conservative behavior, especially in Arabic, with higher abstention rates and lower error rates. This trade-off between assertiveness and caution underscores a central challenge in religious AI systems: knowing when not to answer can be as important as answering correctly.

Persistent Reliability Challenges Despite improvements in abstention behavior, our findings confirm that even the best-performing models continue to produce incorrect answers. This underscores a fundamental limitation of LLMs: they are probabilistic language models, not knowledge-grounded reasoners. Each output is a sequence of tokens generated based on likelihood, not epistemic certainty. This inherent design constraint raises ongoing concerns about the reliability of LLMs in high-stakes domains such as religious rulings.

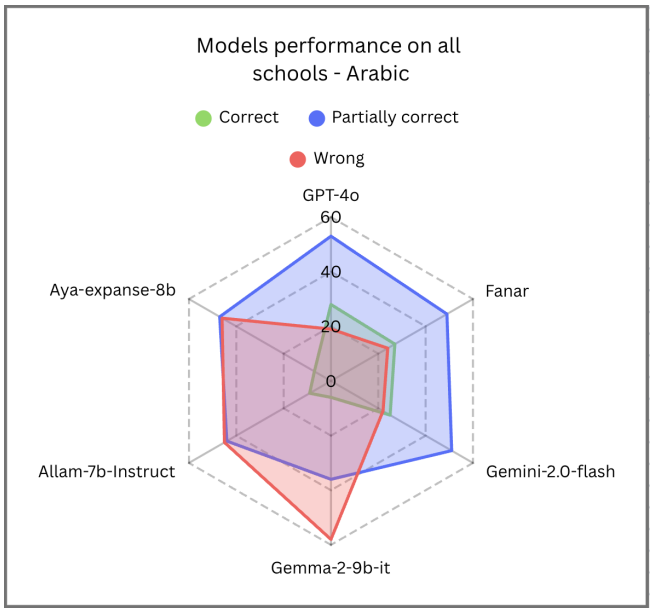

Figure 5: Model's performance on Arabic QA across all schools

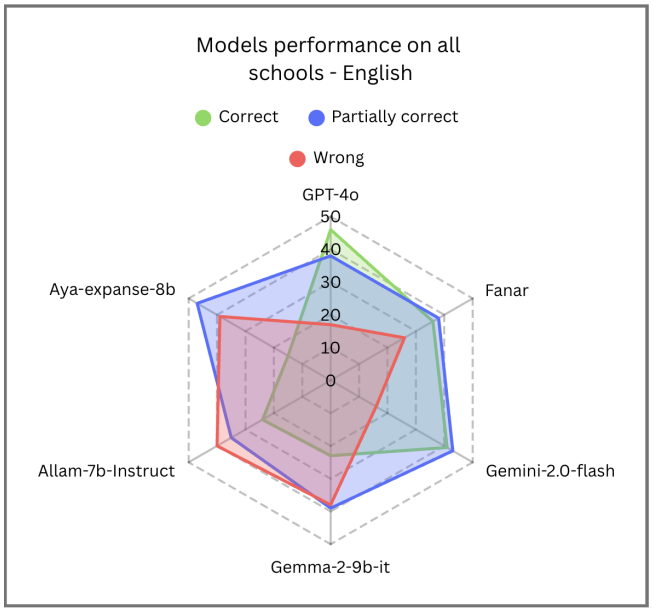

Figure 6: Model's performance on English QA across all schools

Assessing correctness remains a non-trivial task, even when the queries have known answers. Yet, a greater challenge is answering completely new questions, which require reasoning based on legal sources and principles, a process known as *ijtihād*. Given the structured and rule-governed nature of both the Arabic language and Islamic jurisprudence, these domains appear fitting for computational modeling. However, realizing such a system may demand advances beyond current LLM capabilities.

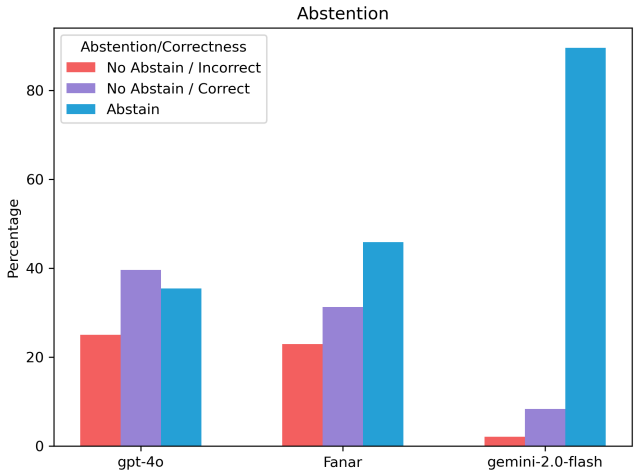

Figure 7: Abstention results - Arabic QA

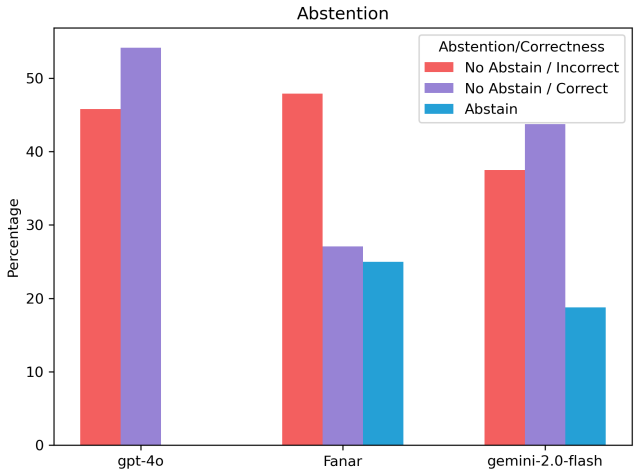

Figure 8: Abstention results - English QA

Religious reasoning as a Human Endeavor It can also be argued that ijtihad is inherently a human activity, as it literally means *to strive to the utmost of one's ability*.

While AI can imitate outputs, it may be limited in replicating intention or exertion, which are key factors in the ethical evaluation of juristic reasoning in Sunni tradition. According to prophetic narrations, a jurist who errs is still rewarded if a sincere effort was made[5].

Given the importance of sincerity and effort, these categories can apply to the engineer or researcher developing AI systems for religious use. Naturally, the domain expertise is a must, so trained scholars of law and theology should be involved in the design and evaluatio n loop. The resulting system should be able to align with the questioner's context to maximize the practical value. In a religious consultation, the goal is not merely to produce an answer, but to help the questioner attain benefit and avoid harm, which is the central aim of Islamic Law (*Sharī'ah*).

[5]https://sunnah.com/bukhari:7352

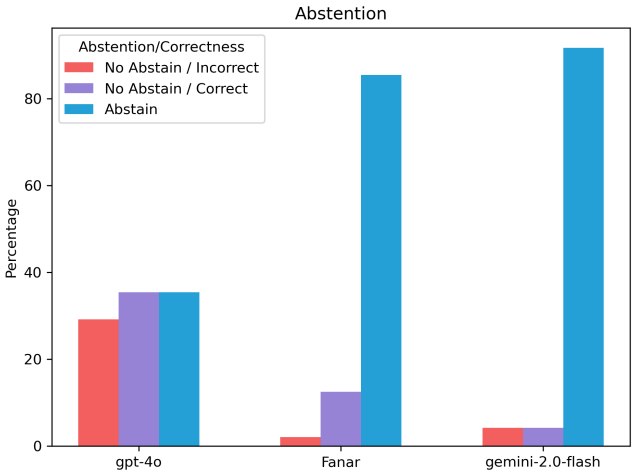

Figure 9: Strict abstention results - Arabic QA

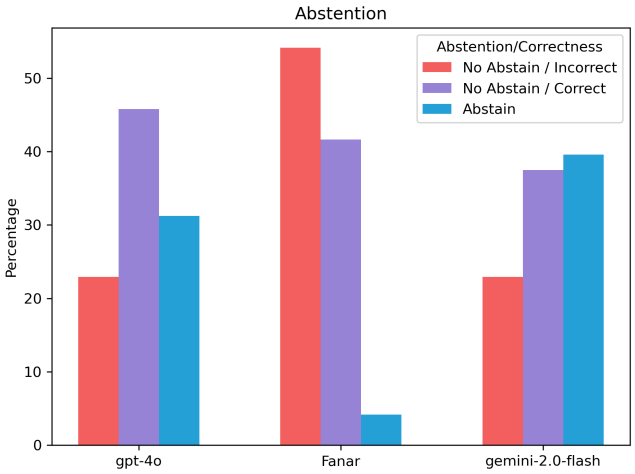

Figure 10: Strict abstention results - English QA

Beyond that, valid religious reasoning requires a deep understanding of the question and its context. In complex areas such as *mu'amalat* (financial and social transactions), abstaining from answering is insufficient. A competent system must identify gaps in the query and request further clarification. This requires capabilities in dialog management, contextual reasoning, and even empathy.

Accountability and Transparency When an AI system delivers an incorrect response, it is difficult to assign responsibility. Users must be cautious about delegating authority to the system, especially in religious settings. This makes transparency a moral as well as a technical imperative. It must be made explicit that the LLM outputs are probabilistic and not guaranteed to be accurate. Educating users about the limitations of such systems is essential to mitigate harm and prevent over-reliance (Fernández et al. 2025).

Currently, most models offer little to no transparency about how responses are generated. Techniques such as Retrieval-Augmented Generation (RAG) and chain-of-thought prompting provide partial visibility into the reasoning process and sources. More structured dialogue flows and explicit modeling of user intent may help increase trustworthiness, but full interpretability remains a largely unsolved problem.

Our findings support a cautious and ethically grounded approach to deploying LLMs in religious contexts. Key requirements include interpretability, abstention mechanisms, context awareness, transparent reasoning, and codesign with domain experts. The goal may not be to replace human scholars but to design systems that work alongside scholars while respecting the epistemological and moral framework of the traditions they aim to serve.

Conclusion

This study presents the first benchmark specifically evaluating LLM performance in answering Islamic questions in the four major Sunni schools of thought. By introducing the FiqhQA dataset and incorporating both correctness and abstention into our evaluation, we offer a more nuanced view of model reliability in high-stakes religious contexts. While models like Gemini, GPT-4o, and Fanar show strong promise, their performance remains uneven across languages and religious traditions. Importantly, our findings underscore the need to build culturally and doctrinally sensitive systems that recognize their epistemic boundaries. Future work should focus on refining abstention mechanisms, increasing training data diversity, and improving interpretability and trustworthiness in multilingual religious domains.

Limitations

Despite following a systematic methodology in the construction and evaluation of the dataset, several limitations warrant consideration.

Dataset Quality and Consistency

The dataset is not without imperfections. Accurately representing the rulings of a single madhhab is already a complex task; extending this effort to cover the four Sunni madhhabs substantially increases the complexity. Although the Kuwaiti Fiqh Encyclopedia provides a comprehensive reference, the extraction and transformation of its rulings into a structured QA format required editorial interventions by the research team. In some cases, the original formulations were ambiguous or open to interpretation, requiring clarification before inclusion in the dataset.

The research team, being most familiar with the Hanafi school, identified and corrected several entries where inconsistencies or omissions were observed. To maintain transparency and facilitate community-based validation, the dataset will be made publicly available, and domain experts are invited to submit further corrections through the repository.

Terminology Alignment Across madhhabs

Each madhhab categorizes human actions, employing terms such as *fard*, *wajib*, *sunnah*, *mandub*, *mustahab*, *makrooh*, and others. These terms often have both specific and general

meanings, and the same term can signify varying degrees of obligation or recommendation across different schools. For example, *wajib* in the Hanafi school has a specific juristic meaning, while being an equivalent to *fard* in other schools.

This complexity presents a challenge even for human scholars, and it was not explicitly accounted for in the evaluation prompts. Consequently, certain nuances were omitted in favor of a simplified evaluation scheme. Additional details on the categorization of human actions in the madhhabs can be found in publicly available resources[6].

Variability in Question and Answer Interpretation

The open-ended nature of the questions results in variability in how the LLM interprets and answers them. In some cases, the LLM's understanding of the question differs from the expected interpretation, though the response may remain plausible within the domain. This discrepancy can complicate the evaluation process. Moreover, the formulation of the answer itself impacts the rating. Ideally, answers should be clear and unambiguous.

Another observed behavior is that LLMs often provide additional information beyond the scope of the question. While this mirrors practical fatwa-giving behavior, it introduces complexity in evaluation. In our framework, providing extra relevant information may result in a rating of 2 (partial correctness), although it may be argued that such comprehensive answers should be rated as fully correct. This remains an open consideration.

References

Abdelnasser, H.; Ragab, M.; Mohamed, R.; Mohamed, A.; Farouk, B.; El-Makky, N. M.; and Torki, M. 2014. Al-Bayan: an Arabic question answering system for the Holy Quran. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, 57–64.

Ahdritz, G.; Qin, T.; Vyas, N.; Barak, B.; and Edelman, B. L. 2024. Distinguishing the knowable from the unknowable with language models. *arXiv preprint arXiv:2402.03563*.

Alqahtani, M.; and Atwell, E. 2018. Annotated corpus of arabic al-quran question and answer.

Bai, G.; Liu, J.; Bu, X.; He, Y.; Liu, J.; Zhou, Z.; Lin, Z.; Su, W.; Ge, T.; Zheng, B.; et al. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *arXiv preprint arXiv:2402.14762*.

Bari, M. S.; Alnumay, Y.; Alzahrani, N. A.; Alotaibi, N. M.; Alyahya, H. A.; AlRashed, S.; Mirza, F. A.; Alsubaie, S. Z.; Alahmed, H. A.; Alabduljabbar, G.; et al. 2024. Allam: Large language models for arabic and english. *arXiv preprint arXiv:2407.15390*.

Dang, J.; Singh, S.; D'souza, D.; Ahmadian, A.; Salamanca, A.; Smith, M.; Peppin, A.; Hong, S.; Govindassamy, M.; Zhao, T.; et al. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

[6]https://islamqa.org/shafii/qibla-shafii/33285/categories-of-human-actions/, https://www.islamandihsan.com/mandub-maliki

Fernández, V. M. C.; de Mendonça, J. C. T.; Matteo, M. A.; and Tighe, M. R. P. 2025. Antiqua et Nova: Note on the Relationship Between Artificial Intelligence and Human Intelligence. Official note by the Dicastery for the Doctrine of the Faith and the Dicastery for Culture and Education. Issued in Rome on the Liturgical Memorial of Saint Thomas Aquinas, Doctor of the Church.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Kasai, J.; Sakaguchi, K.; Le Bras, R.; Asai, A.; Yu, X.; Radev, D.; Smith, N. A.; Choi, Y.; Inui, K.; et al. 2023. Realtime qa: What's the answer right now? *Advances in neural information processing systems*, 36: 49025–49043.

Kim, M.; and Thorne, J. 2024. Epistemology of Language Models: Do Language Models Have Holistic Knowledge? *arXiv preprint arXiv:2403.12862*.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Koto, F.; Li, H.; Shatnawi, S.; Doughman, J.; Sadallah, A. B.; Alraeesi, A.; Almubarak, K.; Alyafeai, Z.; Sengupta, N.; Shehata, S.; et al. 2024. Arabicmmlu: Assessing massive multitask language understanding in arabic. *arXiv preprint arXiv:2402.12840*.

Lin, S.; Hilton, J.; and Evans, O. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Madhusudhan, N.; Madhusudhan, S. T.; Yadav, V.; and Hashemi, M. 2024. Do llms know when to not answer? investigating abstention abilities of large language models. *arXiv preprint arXiv:2407.16221*.

Malhas, R.; and Elsayed, T. 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 (TALLIP)*, 19(6): 1–21.

Malhas, R.; Mansour, W.; and Elsayed, T. 2023. Qur'an QA 2023 Shared Task: Overview of Passage Retrieval and Reading Comprehension Tasks over the Holy Qur'an. In Sawaf, H.; El-Beltagy, S.; Zaghouani, W.; Magdy, W.; Abdelali, A.; Tomeh, N.; Abu Farha, I.; Habash, N.; Khalifa, S.; Keleg, A.; Haddad, H.; Zitouni, I.; Mrini, K.; and Almatham, R., eds., *Proceedings of ArabicNLP 2023*, 690–701. Singapore (Hybrid): Association for Computational Linguistics.

Mohammed, M.; Amin, S.; and Aref, M. M. 2022. An english islamic articles dataset (eiad) for developing an islambot question answering chatbot. In *2022 5th International Conference on Computing and Informatics (ICCI)*, 303–309. IEEE.

Niu, C.; Wu, Y.; Zhu, J.; Xu, S.; Shum, K.; Zhong, R.; Song, J.; and Zhang, T. 2023. Ragtruth: A hallucination corpus for

developing trustworthy retrieval-augmented language models. *arXiv preprint arXiv:2401.00396*.

Patel, S.; Kane, H.; and Patel, R. 2023. Building Domain-Specific LLMs Faithful To The Islamic Worldview: Mirage or Technical Possibility? *arXiv preprint arXiv:2312.06652*.

Qamar, F.; Latif, S.; and Latif, R. 2024. A Benchmark Dataset with Larger Context for Non-Factoid Question Answering over Islamic Text. *arXiv preprint arXiv:2409.09844*.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Rizqullah, M. R.; Purwarianti, A.; and Aji, A. F. 2023. QASiNa: Religious Domain Question Answering Using Sirah Nabawiyah. In *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, 1–6. IEEE.

Röttger, P.; Kirk, H.; Vidgen, B.; Attanasio, G.; Bianchi, F.; and Hovy, D. 2024. XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 5377–5400. Mexico City, Mexico: Association for Computational Linguistics.

Song, J.; Wang, X.; Zhu, J.; Wu, Y.; Cheng, X.; Zhong, R.; and Niu, C. 2024. RAG-HAT: A Hallucination-Aware Tuning Pipeline for LLM in Retrieval-Augmented Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 1548–1558.

Team, F.; Abbas, U.; Ahmad, M. S.; Alam, F.; Altinisik, E.; Asgari, E.; Boshmaf, Y.; Boughorbel, S.; Chawla, S.; Chowdhury, S.; et al. 2025. Fanar: An Arabic-Centric Multimodal Generative AI Platform. *arXiv preprint arXiv:2501.13944*.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Team, G.; Riviere, M.; Pathak, S.; Sessa, P. G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Wiharja, K. R. S.; Murdiansyah, D. T.; Romdlony, M. Z.; Ramdhani, T.; and Gandidi, M. R. 2022. A Questions Answering System on Hadith Knowledge Graph. *Journal of ICT Research & Applications*, 16(2).

Yuan, W.; Neubig, G.; and Liu, P. 2021. Bartscore: Evaluating generated text as text generation. *Advances in neural information processing systems*, 34: 27263–27277.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhong, M.; Liu, Y.; Yin, D.; Mao, Y.; Jiao, Y.; Liu, P.; Zhu, C.; Ji, H.; and Han, J. 2022. Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*.