LLMs and Islamic Fiqh: A Reliability Study Grounded in Maliki Jurisprudential Principles

Alaa Alharbi[1], **Nada Almarwani**[2], **Samah Aloufi**[2]
[1]Department of Information Systems, [2]Department of AI and Data Science
College of Computer Science and Engineering, Taibah University
`{alaharbi, nmarwani, slhebi}@taibahu.edu.sa`

Abstract

In recent years, large language models have become increasingly prevalent in knowledge-based domains, including religion. However, their reliability in domain-specific religious questions remains underexplored. To address this gap, this study evaluates GPT-4o and ALLaM on Islamic jurisprudence (Fiqh) questions based on the Maliki school. We construct a dataset from Maliki sources and test the models across three domains. Results show that GPT-4o consistently outperformed ALLaM; however, both models exhibited significant limitations that affected their reliability in answering domain-specific questions. The models struggled with nuanced rulings requiring deep contextual understanding and showed sensitivity to prompt phrasing. These findings highlight the challenges of applying general-purpose LLMs in religious domains and underscore the need for domain adaptation or retrieval-based enhancements.

1 Introduction

Recently, Large Language Models (LLMs) emerged as a powerful search platform, capable of addressing straightforward questions and tackling complex interactive tasks. Due to their remarkable capabilities, their adoption is rising across tasks in multiple domains, such as translation, summarization, and questions-answering. As their influence grows, more diverse users rely on them for human-like interactions.

Consequently, their use extends into sensitive, complex domains of knowledge and belief. Certain LLM applications, especially in religion, raise major ethical concerns. This is especially relevant as LLMs are increasingly used for religious information, an area tied to cultural heritage and ethics. Thus, ensuring fairness, inclusivity, and cultural sensitivity requires evaluating these models' accuracy, consistency, and context. For instance, accurately interpreting religions such as Islam is essential to developing LLMs that respect and reflect diverse cultural contexts.

In the Islamic system, jurisprudence, known as *Fiqh*, refers to understanding Islamic law (Sharia), based on the Quran, Sunnah, and scholarly consensus. Fiqh is defined by Imam al-Shāfi'i as "the knowledge of the practical Sharia rulings derived from detailed evidence" (Al-Zuhayli, 1985). There are four main Sunni schools of Fiqh (*madhāhib*): *Hanafī*, *Mālikī*, *Shāfiī*, and *Hanbalī*. Each school developed unique methodologies for interpreting texts, applying legal principles, and analogical reasoning for new cases. Beyond being repositories of legal opinions, these schools embody complete intellectual frameworks with structured methods of deriving and applying legal judgments.

Given the methodological differences among Islamic Fiqh schools and the fact that questioners often seek answers aligned with their followed school, evaluating LLMs' ability to generate responses consistent with a specific jurisprudential school, such as the Maliki school, is highly relevant. The Maliki school, dominant in North and West Africa, emphasizes the practices of the early Muslim

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The 5th Muslims in ML Workshop at NeurIPS 2025.

community in Medina, relies on the four classical sources of Islamic law (Quran, Sunnah, consensus and analogical reasoning), and incorporates additional principles such as juridical preference, public interest, custom, and blocking harmful means (Nabit, 2020). These sources collectively shape its distinct legal reasoning framework and motivate our focused analysis: rather than evaluating the general Islamic knowledge of LLMs, we specifically examine their capacity to adhere to the Maliki school's unique legal reasoning methodology. The main contributions of this paper are as follows: 1) A construction of specialized dataset for evaluating Maliki Fiqh across diverse topics; 2) A systematic evaluation of LLMs' response fidelity to Maliki jurisprudential principles; 3) Analyzing prompt and role impact on LLM performance in religious QA.

2 Related work

While LLMs have emerged as knowledge-based platforms with impressive understanding in diverse fields like math, science, and history, they are still prone to generating unreliable, inaccurate, and low-quality answers (Gupta et al., 2025; Chernyshev et al., 2024; Satpute et al., 2024; Hauser et al., 2024; Feng et al., 2024).

LLMs' ability to understand and represent different religions is a crucial aspect that has been explored in several studies (Alnefaie et al., 2023a; Khalila et al., 2025; Plaza-del Arco et al., 2024; Trepczyński, 2023). For example, Plaza-del Arco et al. (2024) found that LLMs depict Western religions, like Christianity, with nuance, but misrepresent Eastern religions, including Islam, through stereotypes and stigmatization. Liu et al. (2024) further showed that LLMs encode differing spiritual values across religions, but such biases can be reduced through targeted religious exposure in training.

In the context of Islam, Alnefaie et al. (2023b) introduced two Arabic QA corpora on Islamic texts: HAQA (1,598 Hadith pairs) and QUQA (3,382 Quran pairs), providing benchmarks for Arabic QA. Using QUQA, Alnefaie et al. (2023a) showed GPT-4o performed poorly (pAP=0.23, EM=0.19), reflecting limits with Classical Arabic and religious contexts. Similarly, Qamar et al. (2024) released a 73,000-pair benchmark from Quranic Tafsir and Hadith, where fine-tuned transformers still fell short, revealing persistent challenges in complex religious QA.

To reduce hallucinations in religious QA, Khalila et al. (2025) applied RAG to 13 open-source LLMs on Quranic data, demonstrating substantial improvements in factuality and contextual relevance, with LLaMA 3 70B with RAG achieving the best performance. Similarly, Alnefaie et al. (2024) reported notable accuracy gains for GPT-4o on the QUQA dataset when combined with RAG, while Alan et al. (2024) introduced MufassirQAS, a RAG-based system using Turkish Islamic texts that yielded more reliable answers than ChatGPT.

For Islamic fatwa generation, Mohammed et al. (2025) proposed Aftina, a two-stage RAG system with a Flash re-ranker, evaluated on 18,407 Dar Al Ifta fatwa QA pairs. Testing three LLMs across base, RAG, and RAG with re-ranker settings, they found the full Aftina setup substantially reduced hallucinations and improved reliability. While previous studies have explored Quranic and Hadith QA, LLM evaluation on school-specific Islamic Fiqh remains limited. We benchmark GPT-4o and ALLaM on Maliki Fiqh using 550 curated questions across three domains, highlighting gaps in adherence to established Fiqhi rules. This is the first systematic study of school-specific Islamic legal reasoning in Arabic, addressing a key challenge in NLP and Islamic AI.

3 FiqhAlign-Maliki: jurisprudential LLM alignment dataset

To evaluate Arabic-capable LLMs in Islamic Fiqh, we created the FiqhAlign-Maliki dataset, based on the Maliki school. It contains 550 questions across three core Fiqh domains: purification (Tahārah), marital jurisprudence (Fiqh Al-Nikāh), and financial transactions (Buyū). These areas were selected to represent a broad range of Islamic legal topics, from personal obligations to complex economic dealings.

The majority of questions (about 90%) in the Fiqh Al-Nikāh and Buyū sections were sourced from the official website of *Dr. Walid Shawish*.[1] The remaining questions, including all those in the purification section, were compiled from various reliable Maliki sources, including textbooks, classroom materials,

[1]A recognized Maliki scholar and professor of Islamic Fiqh at The World Islamic Sciences and Education University. `https://walidshawish.com`

Table 1: Summary of the Islamic Fiqh dataset.

Question Type	Tahārah	Fiqh al-Nikāh	Buyū	**Total**
True/False	150	150	150	450
MCQ	50	50	–	100
Total	200	200	150	550

and contributions from instructors specializing in the Maliki Fiqh. Table 1 summarizes the dataset. A Maliki jurisprudence researcher reviewed the questions, all written in MSA for clarity consistency.

Covering a wide range of aspects within each topic, the questions vary in difficulty to reflect both foundational concepts and nuanced legal rulings. Each question is formatted as either a True/False (binary) item or a multiple-choice question (MCQ) with four options and a single correct answer. Examples of True/False questions along with their correct answers are presented in Appendix A, Table 3.

4 Experimental setup

In this study, we evaluated the performance of two Arabic-capable Large Language Models (LLMs) in answering Fiqh-related questions: GPT-4o (Brown et al., 2020) and ALLaM-7B-Instruct (Bari et al., 2024). GPT-4o is a proprietary model from OpenAI, while ALLaM is an open-source model from the Saudi Data and AI Authority (SDAIA). All experiments were conducted via API calls without local training or fine-tuning, thus requiring no specialized computing resources. For GPT-4o, we used the OpenAI API with the temperature set to 0 for deterministic outputs. For ALLaM, sampling was disabled to encourage focused completions. Since ALLaM does not support chat roles, the system instruction was prepended to the prompt. All prompts for both models were in MSA.

To examine the effect of role-specific instructions on Fiqh reasoning, we used prompt engineering to guide the model from general reasoning to domain-specific expertise. For MCQs, the model selected one option (A–D); for True/False tasks, it responded with True or False. Both GPT-4o and ALLaM were tested under four system-role settings (Appendix B, Table 4):

1. **Baseline:** No instructions; model responded freely.
2. **General Fiqh Expert:** Act as a general Islamic jurisprudence expert.
3. **Maliki Expert-1:** Follow Maliki school principles.
4. **Maliki Expert-2 (strict):** Rely solely on authoritative Maliki sources.

The incremental design assessed how increased prompt specificity affects adherence to Maliki Fiqh. Comparing general and school-specific instructions evaluates the model's ability to reproduce distinctive Islamic jurisprudential reasoning. Experiments used a zero-shot paradigm, generating responses solely from Arabic instructions without examples.

5 Results

Table 2 highlights GPT-4o and ALLaM performance across different prompts in three Islamic Fiqh domains. Overall, GPT-4o outperformed ALLaM across all domains, showing strong Arabic reasoning. ALLaM remained competitive in some cases, notably in Fiqh Al-Nikāh, indicating that well-crafted prompts can boost its performance.

Accross the three domains, both models performed worst on the Tahārah dataset, with ALLaM showing the greatest decline in accuracy. This suggests that purification-related questions are either more complex or underrepresented in the models' training data. The drop was most pronounced in multiple-choice questions, which are harder than True/False due to lower guess probability, emphasizing the need for deeper topic understanding. With a 25% chance of guessing correctly versus 50% for True/False, multiple-choice questions are more demanding, highlighting the need for deeper topic understanding.

Table 2: Accuracy of GPT-4o and ALLaM under different prompt settings across three fiqh domains.

		Baseline		General Fiqh		Maliki Expert 1		Maliki Expert 2	
Domain	**Model**	**T/F**	**MCQ**	**T/F**	**MCQ**	**T/F**	**MCQ**	**T/F**	**MCQ**
Tahārah	GPT-4o	0.63	0.52	0.67	0.52	**0.69**	0.52	0.67	**0.58**
	ALLaM	0.57	0.36	0.54	0.42	0.55	0.40	0.59	0.30
Nikāh	GPT-4o	0.64	**0.72**	0.61	0.66	0.68	**0.72**	**0.70**	0.66
	ALLaM	0.68	0.66	0.68	0.64	0.66	0.62	0.67	0.62
Buyū	GPT-4o	0.72	–	0.70	–	**0.73**	–	0.72	–
	ALLaM	0.61	–	0.62	–	0.63	–	0.65	–

Prompt engineering significantly boosted GPT-4o's performance. Instructing the model to adopt a Maliki Fiqh expert role, particularly under the Maliki Expert-1 and Expert-2 (strict) settings, achieved the highest accuracies, demonstrating that aligning the model with a school-specific perspective strengthens its jurisprudential reasoning and the reliability of its answers. GPT-4o's Maliki Expert-2 (strict) setting achieved the highest accuracy in Fiqh Al-Nikāh (0.70) and Tahārah MCQs (0.58), and nearly matched its best score in Buyū (0.72 vs. 0.73), showing that a strict, school-specific identity and reliance on authoritative Maliki sources improve reasoning by narrowing answers to school-aligned rules.

In contrast, ALLaM was more variable. Its best Buyū score (0.65) occurred under Maliki Expert-2, but top Fiqh Al-Nikāh (T/F) performance (0.68) came from Baseline and General Fiqh Expert prompts, indicating that expert-role prompts do not consistently boost ALLaM, likely due to limited domain-specific alignment. Both models often underperformed with general Fiqh expert prompts, likely because broader opinions reduced alignment with the target school. In contrast, strict Maliki-focused prompts improved accuracy, especially for mufradāt-rich topics like purification.

6 Limitation and future directions

These results reveal a notable limitation in Arabic LLMs when handling Maliki-specific Fiqh content. A plausible explanation is the underrepresentation of such materials in the models' pretraining data. Although our dataset was carefully compiled from authoritative Maliki sources, these sources may be scarce in the corpora used to train GPT-4o and ALLaM, whose exact pretraining data remain undisclosed. Prior research shows that Arabic online content is geographically and ideologically uneven, with countries such as Saudi Arabia, Egypt, and the Gulf states dominating the discourse (Warf and Vincent, 2007), which may further limit the models' exposure to Maliki jurisprudence.

Dominance of Shāf'ī and Hanbali content may bias models against Maliki jurisprudence. With pretraining largely on general-domain corpora and limited inclusion of classical Fiqh texts, the lack of doctrinal granularity likely restricts school-specific reasoning, contributing to the observed performance gaps.

Building on the observed limitations, future work should adapt LLMs to Islamic Fiqh via fine-tuning or Retrieval-Augmented Generation with explicit evaluation of jurisprudential alignment. Large-scale Fiqh datasets remain challenging due to school variations and multiple opinions. FiqhAlign-Maliki offers a preliminary evaluation using closed-type questions from authoritative classical sources.

7 Conclusion

Using a curated dataset of 550 Maliki Fiqh questions, this study compared GPT-4o and ALLaM, showing that performance varied by question type and prompt design. GPT-4o generally outperformed ALLaM, but both had clear limitations in handling complex religious queries, highlighting their potential in education and research while cautioning against use in sensitive contexts. Future directions include developing a RAG system to improve accuracy, expanding the dataset with additional chapters and formats, and benchmarking more Arabic-capable LLMs to better map their strengths and weaknesses.

Appendix

A Examples of questions from FiqhAlign-Maliki dataset

Table 3: Examples of True/False Fiqh questions

Domain	Question (with English translation)	Correct answer
Tahārah	يجب تخليل أصابع القدمين في الوضوء. (It is obligatory to wash between the toes during wudhu.)	FALSE
	يجوز للمتيمم أن يصلي فرضا آخر بنفس التيمم. (It is permissible for the one who performed tayammum to pray another obligatory prayer with the same tayammum.)	FALSE
Fiqh Al-Nikāh	ذكر الصداق عند العقد شرط في صحة عقد الزواج. (Mentioning the dowry in the marriage contract is a condition for the validity of the contract.)	FALSE
	يندب أن يكون المهر كله معجلاً. (It is recommended that the entire dowry be paid in advance.)	TRUE
Buyū	يجوز اجتماع البيع والسلف من غير شرط على المعتمد. (According to the reliable opinion in the school, it is permissible to combine a sale and a loan without stipulation.)	TRUE
	يصح بيع العنب بالزبيب متماثلاً. (It is valid to sell fresh grapes for an equal amount of raisins.)	FALSE

B Prompt settings used in the experiments

Table 4: Prompt settings used in the experiments

Setting	System Prompt (Arabic with English translation)
Baseline	None (no instruction provided)
General Fiqh Expert	أنت خبير في الفقه الإسلامي (You are an expert in Islamic Fiqh.)
Maliki Expert-1	أنت خبير في المذهب المالكي، وتجيب على الأسئلة بناءً على أصول وقواعد هذا المذهب (You are an expert in the Maliki school, and you answer questions based on its principles and rules.)
Maliki Expert-2 (strict)	أنت فقيه متخصص في المذهب المالكي، ولا تعتمد في إجاباتك إلا على ما هو معتمد في كتب المذهب المالكي مثل أقرب المسالك والشرح الصغير والكبير. لا تستخدم آراء المذاهب الأخرى، ولا تقارن بينها، ولا تجيب إلا بما هو مشهور ومعتمد في المذهب المالكي فقط. إذا ورد سؤال يختلف فيه رأي المذهب المالكي عن غيره، فاختر الجواب الذي يمثل المذهب المالكي فقط، ولو خالف ما هو شائع في الفتاوى العامة أو المذاهب الأخرى. إذا لم يكن الجواب واضحاً في كتب المذهب فقل: لا يوجد نص صريح في المذهب، ولا تُخمّن من نفسك (You are an expert in the Maliki school, and you answer questions based on based on what is established in the authoritative Maliki sources, such as Aqrab al-Masālik, al-Sharh al-Saghir and al-Sharh al-Kabir. Do not use opinions from other schools, do not compare between them, and only mention what is well-known and adopted within the Maliki school. If a question involves an issue where the Maliki opinion differs from others, select only the answer that represents the Maliki view, even if it contradicts widely known fatwas or the positions of other. If the answer is not explicitly found in Maliki sources, state: "There is no explicit text in the school," and do not guess.)

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction explicitly state the paper's scope and contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA] .

Justification: This paper does not include theoretical results, theorems, or formal proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides full details on dataset and experimental setup (models evaluated, prompts, temperature, evaluation metrics).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The dataset and experimental code can be shared upon request from the corresponding author to support reproducibility.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies all necessary experimental details. See Section 4

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The paper reports single-run accuracy results for GPT-4o and ALLaM across different prompt settings and fiqh domains (Table 2). While these results support the main claims, error bars or significance tests were not included, since the experiments involved deterministic model outputs under fixed prompts.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Thr experiments were conducted via API calls to LLMs without local training or fine-tuning. The experiments therefore did not require specialized compute resources such as GPUs or large memory.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research follows the NeurIPS Code of Ethics. The dataset was constructed exclusively from publicly available religious and legal texts, with no use of private or personally identifiable information. All experiments were conducted in compliance with ethical standards.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Discussion section addresses both potential benefits and risks of applying LLMs in Fiqh domain.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not release high-risk models or sensitive datasets.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All external assets used in this study are properly credited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets; the dataset described is available upon request but not released as an open resource at submission time.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve human subjects and therefore does not require IRB approval.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs were used only to assist with proofreading and language clarity; all scientific content was authored and verified by the authors.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

References

Wahbah Al-Zuhayli. الفقه الإسلامي وأدلته *[Islamic Jurisprudence and its Proofs]*. Dar Al-Fikr, Damascus, Syria, 1985.

Ahmet Yusuf Alan, Enis Karaarslan, and Ömer Aydin. A rag-based question answering system proposal for understanding islam: Mufassirqas llm. *arXiv preprint arXiv:2401.15378*, 2024.

Sarah Alnefaie, Eric Atwell, and Mohammad Ammar Alsalka. Is GPT-4 a good islamic expert for answering Quran questions? In Jheng-Long Wu and Ming-Hsiang Su, editors, *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 124–133, Taipei City, Taiwan, October 2023a. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP). URL `https://aclanthology.org/2023.rocling-1.15/`.

Sarah Alnefaie, Eric Atwell, and Mohammad Ammar Alsalka. HAQA and QUQA: Constructing two Arabic question-answering corpora for the Quran and Hadith. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 90–97, Varna, Bulgaria, September 2023b. INCOMA Ltd., Shoumen, Bulgaria. URL `https://aclanthology.org/2023.ranlp-1.10/`.

Sarah Alnefaie, Eric Atwell, and Mohammed Ammar Alsalka. Using the retrieval-augmented generation technique to improve the performance of gpt-4 in answering quran questions. In *2024 6th International Conference on Natural Language Processing (ICNLP)*, pages 377–381, 2024. doi: 10.1109/ICNLP60986.2024.10692797.

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. Allam: Large language models for arabic and english. *arXiv preprint arXiv:2407.15390*, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Konstantin Chernyshev, Vitaliy Polshkov, Ekaterina Artemova, Alex Myasnikov, Vlad Stepanov, Alexei Miasnikov, and Sergei Tilga. U-math: A university-level benchmark for evaluating mathematical skills in llms. *arXiv preprint arXiv:2412.03205*, 2024.

Kehua Feng, Keyan Ding, Weijie Wang, Xiang Zhuang, Zeyuan Wang, Ming Qin, Yu Zhao, Jianhua Yao, Qiang Zhang, and Huajun Chen. Sciknoweval: Evaluating multi-level scientific knowledge of large language models. *arXiv preprint arXiv:2406.09098*, 2024.

Adit Gupta, Jennifer Reddig, Tommaso Calo, Daniel Weitekamp, and Christopher J MacLellan. Beyond final answers: Evaluating large language models for math tutoring. *arXiv preprint arXiv:2503.16460*, 2025.

Jakob Hauser, Daniel Kondor, Jenny Reddish, Majid Benam, Enrico Cioni, Federica Villa, James Bennett, Daniel Hoyer, Pieter Francois, Peter Turchin, et al. Large language models' expert-level global history knowledge benchmark (hist-llm). *Advances in Neural Information Processing Systems*, 37:32336–32369, 2024.

Zahra Khalila, Arbi Haza Nasution, Winda Monika, Aytug Onan, Yohei Murakami, Yasir Bin Ismail Radi, and Noor Mohammad Osmani. Investigating retrieval-augmented generation in quranic studies: A study of 13 open-source large language models. *International Journal of Advanced Computer Science and Applications*, 16(2), 2025. ISSN 2158-107X. doi: 10.14569/ijacsa.2025.01602134. URL http://dx.doi.org/10.14569/IJACSA.2025.01602134.

Songyuan Liu, Ziyang Zhang, Runze Yan, Wei Wu, Carl Yang, and Jiaying Lu. Measuring spiritual values and bias of large language models. *arXiv preprint arXiv:2410.11647*, 2024.

Marryam Yahya Mohammed, Sama Ayman Ali, Salma Khaled Ali, Ayad Abdul Majeed, and Ensaf Hussein Mohamed. Aftina: enhancing stability and preventing hallucination in ai-based islamic fatwa generation using llms and rag. *Neural Computing and Applications*, pages 1–26, 2025.

Abdulrahman Al Nabit. تقريب المدارك في أصول الإمام مالك *[Facilitating the Comprehension of the Principles of Imam Mālik]*. Elaf, Hawalli, Kuwait, 2020.

Flor Miriam Plaza-del Arco, Amanda Cercas Curry, Susanna Paoli, Alba Curry, and Dirk Hovy. Divine llamas: Bias, stereotypes, stigmatization, and emotion representation of religion in large language models. *arXiv preprint arXiv:2407.06908*, 2024.

Faiza Qamar, Seemab Latif, and Rabia Latif. A benchmark dataset with larger context for non-factoid question answering over islamic text, 2024. URL https://arxiv.org/abs/2409.09844.

Ankit Satpute, Noah Gießing, André Greiner-Petter, Moritz Schubotz, Olaf Teschke, Akiko Aizawa, and Bela Gipp. Can llms master math? investigating large language models on math stack exchange. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 2316–2320, 2024.

Marcin Trepczyński. Religion, theology, and philosophical skills of llm–powered chatbots. *Disputatio Philosophica: International Journal on Philosophy and Religion*, 25(1):19–36, 2023.

Barney Warf and Peter Vincent. Multiple geographies of the arab internet. *Area*, 39(1):83–96, 2007.