Jurnal Asy-Syukriyyah
Vol. 27 | Nomor 1 | Januari – Juni 2026

P-ISSN 1693-136X
E-ISSN 2715-6753
DOI : https://doi.org/10.36769/asy.v27i1.1822

EVALUATION OF THE QUALITY OF LARGE LANGUAGE MODELS' ANSWERS IN ANSWERING RELIGIOUS QUESTIONS FOR DIGITAL DA'WAH

Munirah
Faculty of Engineering, Universitas Muhammadiyah Ponorogo, Indonesia
munirah.mt@umpo.ac.id

Aslan Alwi
Faculty of Engineering, Universitas Muhammadiyah Ponorogo, Indonesia
aslan.alwi@umpo.ac.id

Abdul Karim
Hallym University, Chunceon, Gangwon, South Korea
abdulkarim@korea.ac.kr

Abstract: *The development of Large Language Models (LLMs) has opened new opportunities for utilizing artificial intelligence for digital da'wah. However, the use of LLMs in the religious domain presents challenges related to accuracy, potential errors (hallucinations), and conformity with Islamic values. This study aims to evaluate the quality of LLM responses in answering religious questions as part of digital da'wah. The method used is a descriptive evaluation of 30 religious questions covering aspects of worship, faith, morals, contemporary issues, and ambiguous questions. LLM responses were analyzed based on three main indicators: accuracy, potential hallucinations, and conformity with Islamic values. The results show that LLMs have a fairly good level of accuracy on basic religious questions, but the potential for hallucinations is still found, especially in ambiguous and contemporary questions. These findings indicate that LLMs have potential as a digital da'wah tool, but their use requires a verification mechanism to maintain accuracy and conformity of values.*

Keywords: *Large Language Models, Digital Da'wah, AI Evaluation, Hallucination, AI Ethics*.

INTRODUCTION

Recent advances in Artificial Intelligence (AI) have accelerated innovation in the field of Natural Language Processing (NLP). Among the most significant breakthroughs is the development of Large Language Models (LLMs), which are trained on extensive datasets and possess the capability to comprehend, generate, and respond to textual information in a natural and human-like manner.[1] These capabilities have positioned LLMs as a widely adopted

[1] Thomas B. Brown et al., "Language Models are Few-Shot Learners," Advances in Neural Information Processing Systems (NeurIPS), 2020; S.K. Dam, C.S. Hong, Y. Qiao, dan C. Zhang, "A Complete Survey on LLM-based AI Chatbots," arXiv, 2024; A. Ehrlich-Sommer, "ForestGPT: Domain-Specific LLM," Electronics, 2025; Y. Geifman dan R. El-Yaniv, "SelectiveNet: A Deep Neural Network with Reject Option," Proceedings of the International Conference on Machine Learning (ICML), 2019; S. Kadavath et al., "Uncertainty Estimation for Language Model Predictions," 2022; Patrick Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,"

P-ISSN 1693-136X
E-ISSN 2715-6753
DOI : https://doi.org/10.36769/asy.v27i1.1822

technology across various sectors, including education, healthcare, public services, and religious applications.

In the religious domain, the adoption of LLMs offers new opportunities for the dissemination of religious knowledge and the advancement of digital da'wah. LLM-based systems enable individuals to obtain answers to religious inquiries quickly, interactively, and conveniently, without the constraints of time and geographical boundaries. Several studies have demonstrated that modern language models have been utilized to support religious scholarship, including Qur'anic studies and Islamic knowledge-based question-answering systems.[2] The emergence of this technology has the potential to broaden public access to Islamic knowledge and support the transformation of da'wah in the digital era. Digital transformation has not only changed the way individuals access information but has also reshaped learning processes and the dissemination of religious values. Studies on Islamic education in the digital age indicate that the effective and purposeful use of technology can enhance access to Islamic knowledge while simultaneously strengthening the development of religious character. Therefore, the advancement of artificial intelligence–based technologies, including Large Language Models (LLMs), can be viewed as part of the ongoing evolution of Islamic educational and da'wah media, warranting further investigation into their effectiveness and the quality of the information they generate. [3]

Despite their potential benefits, the use of LLMs in religious contexts raises several challenges. Among the most significant is the phenomenon known as hallucination, whereby a model produces responses that appear credible and authoritative despite lacking support from valid, trustworthy, or verifiable sources.[4] Within religious settings, inaccuracies of this nature may result in misunderstandings of religious doctrines, the spread of misleading information, and potential distortions in public perceptions of Islamic principles. Moreover, previous studies have highlighted the potential for religious bias in large language models, which may compromise the quality, fairness, and objectivity of the generated responses.[5]

NeurIPS, 2020; Long Ouyang et al., "Training Language Models to Follow Instructions with Human Feedback," NeurIPS, 2022.

[2] G. Bhatia, "Advances in AI Systems on Islamic Knowledge Capabilities," 2026; Z. Khalila, H. Khaled, dan M. Elmahdy, "Investigating Retrieval-Augmented Generation in Quranic Studies: A Study of 13 Open-Source Large Language Models," arXiv, 2025.

[3] Muin, M.T. (2024). *Konsep Pendidikan Anak di Era Digital dalam Perspektif Al-Qur'an*. Tarqiyatuna.

[4] L. Huang, "A Survey on Hallucination in Large Language Models," ACM Transactions on Information Systems, 2025.

[5] A.Y. Alan, "Improving LLM Reliability with RAG in Religious QA," Turkish Journal of Engineering, 2025; Luciano Floridi, "Establishing the Rules for Building Trustworthy AI," Nature Machine Intelligence, 2019; A.

P-ISSN 1693-136X
E-ISSN 2715-6753
DOI : https://doi.org/10.36769/asy.v27i1.1822

Beyond technical considerations, the use of AI in religious contexts also raises a range of ethical concerns. Papakostas et all [6] argue that the integration of AI into educational and religious practices requires careful attention to ethical considerations, pedagogical effectiveness, and the preservation of religious scholarly authority to ensure the responsible and appropriate use of such technologies.[7] A similar view was expressed by Vaughan, who argued that the use of AI in spiritual and religious contexts requires appropriate oversight mechanisms to ensure that the information provided remains consistent with established religious values and principles.[8]

Beyond hallucination-related issues, recent research has demonstrated that Large Language Models (LLMs) may contain inherent biases toward certain religious identities and communities. Studies examining religious representation in LLMs have found that some religions are depicted with greater depth and nuance, while others are more likely to be simplified, stereotyped, or even subject to stigmatization. These findings highlight that evaluating da'wah chatbots requires more than assessing response accuracy; it also necessitates examining potential biases that could undermine the objectivity, fairness, and quality of the religious information provided to users.

In addition to hallucination-related concerns, emerging research has demonstrated that Large Language Models (LLMs) may contain inherent biases toward particular religious identities and groups. Studies examining religious representation within LLMs have found that some religions are represented in a more comprehensive and nuanced manner, while others are more likely to be subject to oversimplification, stereotyping, or stigmatization. These findings underscore the need for a broader evaluation framework for da'wah chatbots, one that not only assesses response accuracy but also examines potential biases that may compromise the objectivity, fairness, and reliability of the religious information provided to users.[9]

Sharma dan M. Gupta, "Quantifying Religious Bias in Open LLMs through Demographic Attributes," arXiv, 2025; D. Simbeck dan M. Mahran, "Mechanistic Interpretability with SAEs: Probing Religion, Violence, and Geography in Large Language Models," arXiv, 2025; L. Wang, "Survey on LLM-based Autonomous Agents," 2023; F.M. Plaza-del-Arco et al., "Divine LLaMAs: Bias, Stereotypes, Stigmatization, and Emotion Representation of Religion in Large Language Models," EMNLP Findings, 2024.

[6] A. Papakostas, G.D. Kallergis, dan D. Politis, "Artificial Intelligence in Religious Education: Ethical, Pedagogical, and Theological Perspectives," Religions, 2025.

[7] A. Papakostas, G.D. Kallergis, dan D. Politis, "Artificial Intelligence in Religious Education: Ethical, Pedagogical, and Theological Perspectives," Religions, Vol. 16, No. 5, 2025.

[8] G. Vaughan, "Wisdom of the Heart: A Review of Religion and AI," Religions, 2025.

[9] Plaza-del-Arco, F.M., Curry, A.C., Paoli, S., Curry, A., & Hovy, D. (2024). *Divine LLaMAs: Bias, Stereotypes, Stigmatization, and Emotion Representation of Religion in Large Language Models*. arXiv.

P-ISSN 1693-136X
E-ISSN 2715-6753
DOI : https://doi.org/10.36769/asy.v27i1.1822

In Muhammadiyah's digital da'wah ecosystem, the integration of AI technologies presents promising opportunities for enhancing the effectiveness of religious communication and knowledge dissemination. However, before such systems are adopted on a broader scale, it is crucial to evaluate the quality of their outputs, particularly in terms of response accuracy, susceptibility to hallucinations, and compliance with Islamic values and teachings. This evaluation is crucial, as the quality of generated responses directly affects users trust and confidence in the system. In addition to technical accuracy, the adoption of digital technologies in religious domains necessitates careful attention to ethical considerations and the preservation of Islamic values. Contemporary research in Islamic education emphasizes that the integration of technology into educational activities and the dissemination of religious knowledge should be implemented with caution to ensure that digital innovation does not diminish the authenticity of Islamic teachings or undermine Islamic identity. Accordingly, the evaluation of da'wah chatbot responses should extend beyond technical performance to include their compliance with religious values and ethical standards.[10]

A growing body of research has investigated the effectiveness of Large Language Models (LLMs) in answering religious questions, while also identifying potential risks associated with misinformation, factual inaccuracies, and other forms of unreliable content generated by these systems.[11] Despite the growing body of literature on LLMs in religious question-answering, studies that comprehensively evaluate LLM-based da'wah chatbots using multiple dimensions, including response accuracy, hallucination risk, and adherence to Islamic values, are still limited. This gap is particularly evident in the Indonesian digital da'wah context, where empirical evidence regarding the reliability and appropriateness of such systems remains insufficient.

Recent research investigating the reliability of Large Language Models (LLMs) in religious question-answering has demonstrated that model performance remains highly dependent on the complexity of the religious issues being addressed. Through the FiqihQA benchmark, which incorporates perspectives from multiple schools of Islamic jurisprudence, it was found that LLMs frequently generate responses that appear credible and persuasive, yet they are not always consistent with established Islamic scholarly interpretations. These findings

[10] Guci, A. (2024). *Tantangan Pendidikan Islam Zaman Modern*. Tarqiyatuna.
[11] Atif, F., Askarbekuly, N., Darwish, K., & Choudhury, M. (2025). Sacred or Synthetic? Evaluating LLM Reliability and Abstention for Religious Questions. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.

P-ISSN 1693-136X
E-ISSN 2715-6753
DOI : https://doi.org/10.36769/asy.v27i1.1822

underscore the necessity of conducting rigorous evaluations of LLM-based systems prior to their adoption as tools for religious consultation, guidance, or digital da'wah.[12]

In the context of Society 5.0, technological innovations, including artificial intelligence, are assessed not solely in terms of their technical performance but also in relation to their contribution to human-centered development and ethical values. Research on the integration of Islamic education with technological progress highlights that technological adoption should be guided by the principles of public welfare (maslahah), social responsibility, and the reinforcement of spiritual values within society. This perspective is highly relevant to the development of LLM-based da'wah chatbots, as the quality of these systems depends not only on the correctness of their responses but also on their capacity to preserve and promote Islamic values throughout their interactions with users.[13]

Despite the growing body of literature on the reliability of Large Language Models (LLMs) in religious question-answering, existing studies have primarily concentrated on international benchmark evaluations, hallucination-related issues, abstention mechanisms, and the identification of religious biases. Comparatively little attention has been given to the comprehensive evaluation of LLM-based da'wah chatbots that simultaneously considers response accuracy, hallucination susceptibility, and adherence to Islamic values, particularly within the Indonesian digital da'wah context. This limitation reveals an important research gap that warrants further investigation to ensure that the deployment of LLMs in digital da'wah is not only technically reliable but also consistent with Islamic teachings and ethical principles.

The primary novelty of this study resides in its comprehensive evaluation framework for LLM-based da'wah chatbots, which simultaneously assesses three critical dimensions: response accuracy, hallucination risk, and adherence to Islamic values. Furthermore, the study utilizes an Indonesian-language religious question dataset encompassing five key categories, namely worship (ibadah), creed (aqidah), morality (akhlaq), contemporary issues, and ambiguous questions. By integrating these dimensions within a single evaluation framework, this study aims to provide a more holistic assessment of the readiness and suitability of LLM technology as a platform for supporting digital da'wah initiatives.

[12] Atif, F., Agrawal, A., Awadallah, A.H., Caruana, R., & Ribeiro, M.T. (2025). *Sacred or Synthetic? Evaluating LLM Reliability and Abstention for Religious Questions*. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.
[13] Muzakki, Z. (2023). *Integrasi Ilmu Ekonomi Islam dan Pendidikan Agama Islam dalam Era Society 5.0*. I-BEST: Islamic Banking & Economic Law Studies.

P-ISSN 1693-136X
E-ISSN 2715-6753
DOI : https://doi.org/10.36769/asy.v27i1.1822

Drawing upon the identified research gap, this study seeks to evaluate the quality of responses generated by an LLM-based da'wah chatbot using an Indonesian-language dataset of religious questions covering five key categories: worship (ibadah), creed (aqidah), morality (akhlaq), contemporary issues, and ambiguous questions. The evaluation framework incorporates three principal dimensions: response accuracy, hallucination susceptibility, and adherence to Islamic values. It is anticipated that the findings will provide valuable insights for the development of more accurate, trustworthy, and Islamically compliant digital da'wah systems, thereby supporting the responsible adoption of AI technologies in religious contexts.

RESEARCH METHODOLOGY

1. Research Design

This study was conducted at Universitas Muhammadiyah Ponorogo using an LLM-based da'wah chatbot prototype developed by the researchers as the primary object of evaluation. The research dataset consisted of 125 religious questions manually compiled from topics commonly encountered in digital da'wah activities. The questions were categorized into five domains: worship (ibadah), creed (aqidah), morality (akhlaq), contemporary issues, and ambiguous questions. The responses generated by the chatbot were initially evaluated by the researchers and subsequently validated by an Islamic studies expert to ensure response accuracy, identify potential hallucinations, and assess conformity with Islamic values.

This study employed a descriptive evaluative approach to assess the quality of responses generated by the LLM-based da'wah chatbot prototype. The evaluation focused on three key dimensions: response accuracy, hallucination risk, and adherence to Islamic values.

The research procedure was conducted systematically, beginning with the development of the chatbot prototype, followed by the construction of the religious question dataset, system testing, and response quality evaluation. The evaluation process was performed using the three predefined indicators: accuracy, hallucination, and conformity with Islamic values. The overall research workflow adopted in this study is illustrated in Figure 1.

The study commenced with the development of an LLM-based da'wah chatbot prototype, which served as the evaluation object. Subsequently, a dataset comprising 125 religious questions was prepared, consisting of 25 questions for each category: worship (ibadah), creed (aqidah), morality (akhlaq), contemporary issues, and ambiguous questions. Each question was submitted to the chatbot system, and the generated responses were evaluated according to the three evaluation indicators. The evaluation results were then analyzed

P-ISSN 1693-136X
E-ISSN 2715-6753
DOI : https://doi.org/10.36769/asy.v27i1.1822

descriptively to identify the chatbot's performance across different categories of religious questions.

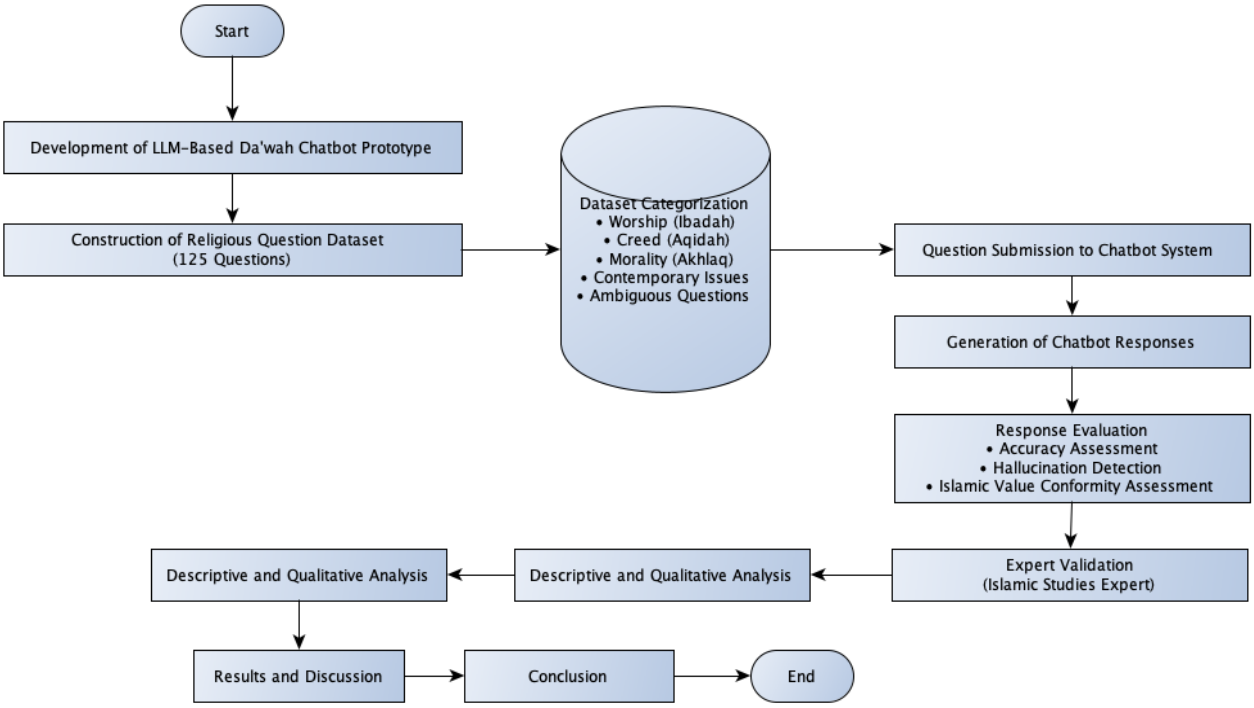

FIgure 1. Research workflow

2. Development of the LLM-Based Da'wah Chatbot Prototype

An LLM-based da'wah chatbot prototype was developed as the primary object of evaluation in this study. The prototype was designed as a text-based conversational system capable of automatically responding to religious inquiries by leveraging the capabilities of Large Language Models (LLMs) :

1) User submits a religious question
2) The system processes the input using a language model
3) The system automatically generates a response

The developed prototype functioned as the main evaluation object in this study, providing the basis for assessing the quality of chatbot-generated responses. To support transparency and reproducibility, the prototype has been deployed online and is available for public access at: https://elangbijak4-risetmu-umpo.hf.space/landing.php.

P-ISSN 1693-136X
E-ISSN 2715-6753
DOI : https://doi.org/10.36769/asy.v27i1.1822

3. Research Dataset

The dataset used in this study consisted of 125 religious questions manually compiled by the researchers and categorized into five main groups:

1. Worship (Ibadah)
2. Creed (Aqidah)
3. Morality (Akhlaq)
4. Contemporary Issues
5. Ambiguous Questions

Each category contained 25 questions, resulting in a total of 125 questions. The dataset was designed to represent a diverse range of religious inquiries commonly encountered in digital da'wah activities and online religious consultation platforms.

The inclusion of these categories was intended to capture varying levels of complexity and contextual interpretation, thereby enabling a comprehensive evaluation of the chatbot's ability to generate accurate, reliable, and Islamically appropriate responses across different types of religious questions.

The dataset was developed manually to ensure the relevance and contextual appropriateness of the questions within the Indonesian Islamic context. The selected categories encompass both well-defined religious topics and more complex issues that may require contextual reasoning or scholarly interpretation, allowing for a more robust assessment of the LLM-based da'wah chatbot.

Table 1. Distribution of the Religious Question Dataset

No	Category	Number of Questions
1	Worship (*Ibadah*)	25
2	Creed (*Aqidah*)	25
3	Morality (*Akhlaq*)	25
4	Contemporary Issues	25
5	Ambiguous Questions	25
	Total	**125**

4. Data Collection Procedure

Data collection was carried out by systematically submitting each question in the religious question dataset to the LLM-based da'wah chatbot prototype. For every input question, the corresponding response generated by the system was recorded and stored for evaluation purposes. Subsequently, all responses were organized in a structured evaluation matrix to

P-ISSN 1693-136X
E-ISSN 2715-6753
DOI : https://doi.org/10.36769/asy.v27i1.1822

facilitate the assessment of response accuracy, hallucination risk, and conformity with Islamic values.

To ensure consistency, the same evaluation procedure was applied across all five question categories, namely worship (ibadah), creed (aqidah), morality (akhlaq), contemporary issues, and ambiguous questions. The collected responses constituted the primary data used for the evaluation and analysis stages of this study.

5. Evaluation Metrics

To assess the quality of the responses generated by the LLM-based da'wah chatbot, three evaluation metrics were employed: accuracy, hallucination, and conformity with Islamic values. These metrics were selected to capture not only the technical reliability of the system but also its suitability for use in religious contexts.

1) Accuracy evaluates the correctness of chatbot responses by comparing them against authoritative Islamic sources, such as the Qur'an, Hadith, and established scholarly references.
2) Hallucination assesses the presence of unsupported, speculative, or fabricated information generated by the model that cannot be verified through reliable Islamic sources.
3) Conformity with Islamic Values measures the extent to which chatbot responses adhere to Islamic teachings, ethical principles, and religious norms, ensuring that the information provided remains appropriate for digital da'wah applications.

6. Evaluation Procedure

The evaluation process was carried out through manual assessment of the responses generated by the LLM-based da'wah chatbot. Each response was evaluated according to the three predefined metrics: accuracy, hallucination, and conformity with Islamic values. A binary classification scheme was employed for each metric to ensure consistency in the assessment process.

For the accuracy metric, responses were classified as either accurate or inaccurate based on their consistency with authoritative Islamic sources. For hallucination, responses were categorized as either hallucination present or hallucination absent, depending on whether unsupported or unverifiable information was identified. For conformity with Islamic values, responses were classified as either conforming or non-conforming according to their alignment with Islamic teachings, ethical principles, and religious norms.

P-ISSN 1693-136X
E-ISSN 2715-6753
DOI : https://doi.org/10.36769/asy.v27i1.1822

To improve the credibility of the evaluation, the assessment results were validated by an expert in Islamic studies. The expert independently reviewed the chatbot responses to verify their correctness, identify potential hallucinations, and determine their compliance with Islamic values.

Whenever discrepancies arose between the researcher's assessment and the expert's evaluation, discussions were conducted to reach a consensus. This process helped establish inter-rater agreement and enhanced the reliability of the evaluation outcomes. The final evaluation results were determined based on the consensus reached between the researcher and the Islamic studies expert.

7. Data Analysis

The collected responses were analyzed using a combination of quantitative and qualitative approaches to obtain a comprehensive understanding of the chatbot's performance.

First, descriptive analysis was employed to calculate the percentages of accurate responses, hallucination occurrences, and responses conforming to Islamic values. These percentages were used to provide an overall performance overview of the chatbot.

Second, category-based analysis was performed to compare the chatbot's performance across five categories of religious questions: worship (ibadah), creed (aqidah), morality (akhlaq), contemporary issues, and ambiguous questions. This analysis enabled the identification of categories in which the chatbot performed well and those that required further improvement.

Finally, qualitative analysis was conducted by reviewing selected response examples to explore patterns of correct and incorrect answers, identify instances of hallucination, evaluate conformity with Islamic values, and examine the strengths and limitations of the proposed chatbot. The findings from these analyses were then synthesized to assess the overall suitability of LLM-based chatbots for digital da'wah applications. All analyses were performed at the category and overall dataset levels to provide both detailed and aggregate evaluations of chatbot performance.

RESULT AND DISCUSSION

This study evaluated the quality of responses generated by an LLM-based da'wah chatbot using a dataset of 125 religious questions covering five categories: worship (ibadah), creed (aqidah), morality (akhlaq), contemporary issues, and ambiguous questions. The

P-ISSN 1693-136X
E-ISSN 2715-6753
DOI : https://doi.org/10.36769/asy.v27i1.1822

evaluation was conducted based on three key dimensions: response accuracy, hallucination, and conformity with Islamic values. The findings provide insights into both the technical reliability and the religious appropriateness of LLM-based chatbots for digital da'wah applications.

1. Accuration Result

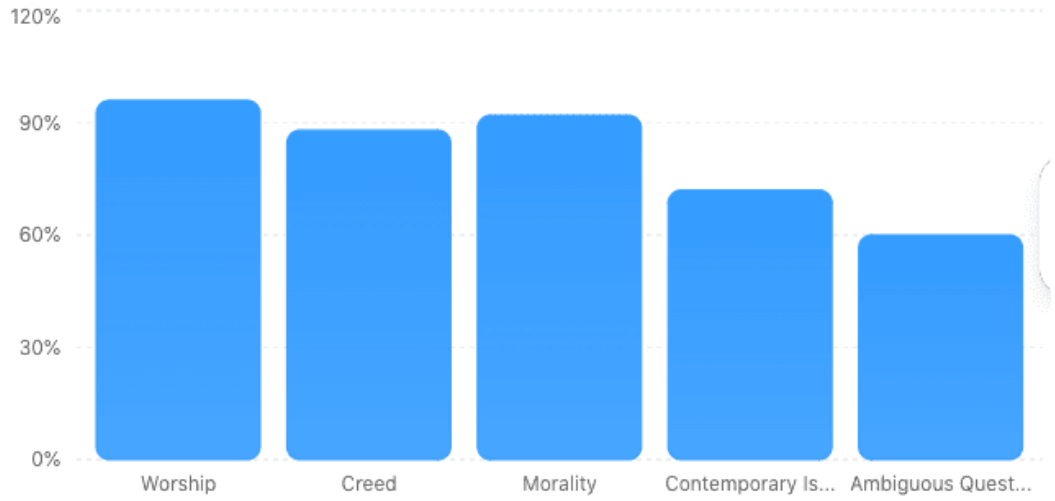

Figure 2. Accuracy Performance across Religious Question Categories

The highest accuracy was achieved in the worship category (96%), followed by morality (92%) and creed (88%). In contrast, the chatbot demonstrated lower performance when responding to contemporary issues (72%) and ambiguous questions (60%), indicating greater difficulty in handling questions requiring contextual interpretation and complex reasoning.

2. Hallucination Result

P-ISSN 1693-136X
E-ISSN 2715-6753
DOI : https://doi.org/10.36769/asy.v27i1.1822

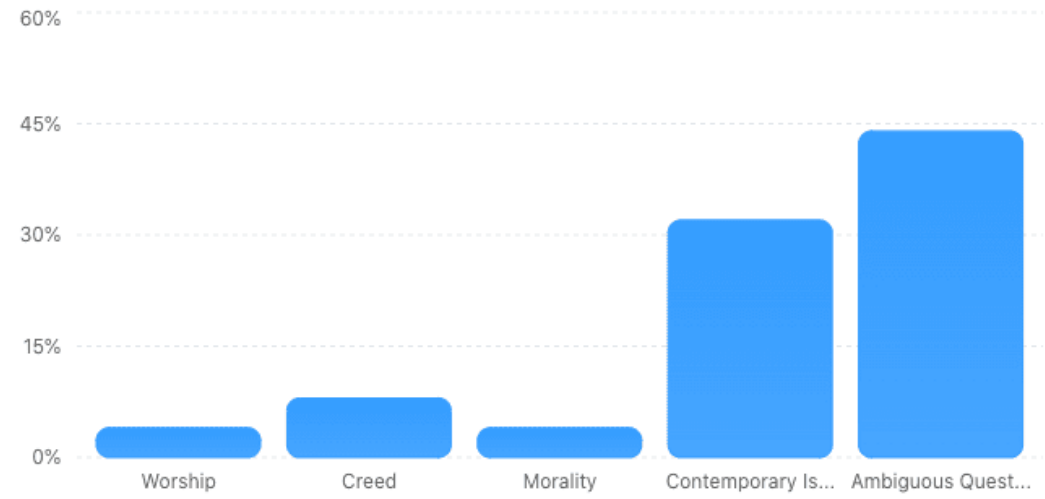

Figure 3. Hallucination Rates across Religious Question Categories

Hallucination occurrences were relatively low in worship, creed, and morality questions. However, the hallucination rate increased substantially for contemporary issues (32%) and ambiguous questions (44%), suggesting that the model struggles when addressing topics with limited explicit references or multiple possible interpretations.

3. Islamic Value Conformity Result

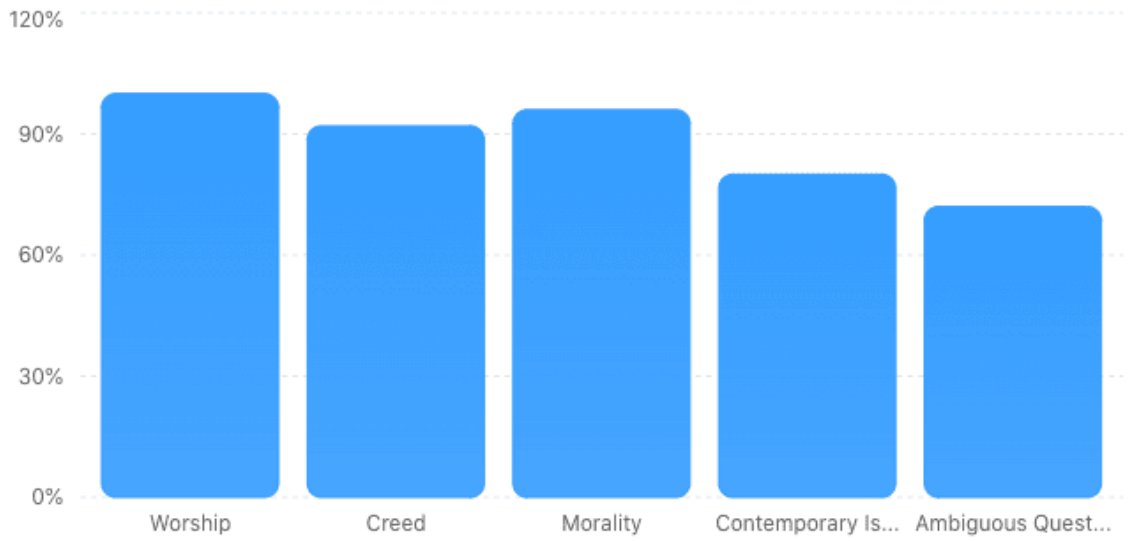

Figure 4. Conformity of Chatbot Responses with Islamic Values

P-ISSN 1693-136X
E-ISSN 2715-6753
DOI : https://doi.org/10.36769/asy.v27i1.1822

The chatbot demonstrated a high level of conformity with Islamic values, particularly in worship-related questions (100%). Nevertheless, lower conformity rates were observed in contemporary issues (80%) and ambiguous questions (72%), highlighting the need for additional validation mechanisms when addressing complex religious topics.

4. Comparative Analysis

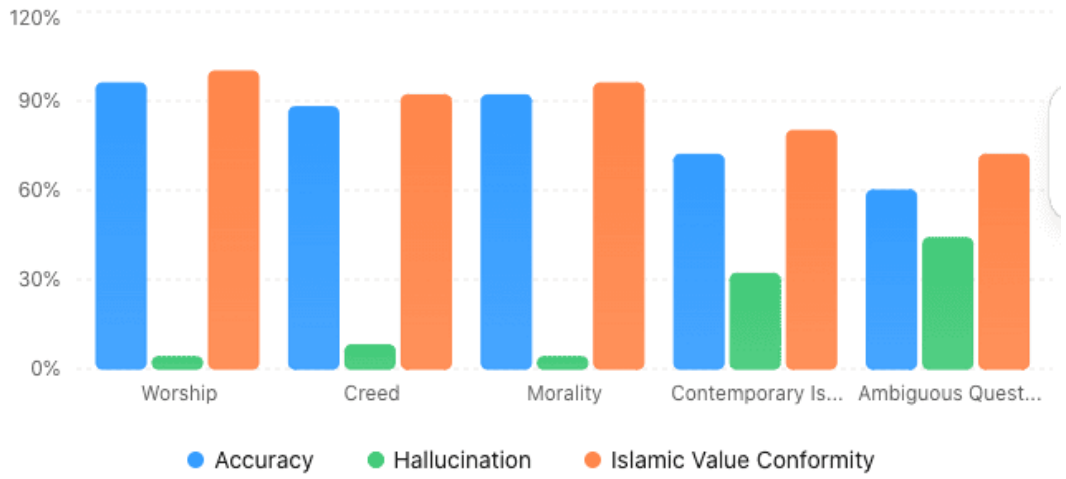

Figure 5. Comparative Analysis of Accuracy, Hallucination, and Islamic Value Conformity Across Religious Question Categories

Figure 5 presents a comparative analysis of the chatbot's performance across the five categories of religious questions. The results reveal a clear relationship between response accuracy, hallucination rate, and conformity with Islamic values. The worship and morality categories achieved the highest accuracy scores (96% and 92%, respectively) while maintaining very low hallucination rates (4%). These categories also exhibited the highest levels of conformity with Islamic values, reaching 100% and 96%, respectively.

In contrast, the contemporary issues and ambiguous question categories demonstrated lower accuracy levels (72% and 60%) and substantially higher hallucination rates (32% and 44%). A corresponding decline in conformity with Islamic values was also observed, with scores of 80% and 72%, respectively. These findings suggest that as the complexity and interpretative nature of religious questions increase, the likelihood of hallucination also increases, which may negatively affect both response accuracy and adherence to Islamic principles.

P-ISSN 1693-136X
E-ISSN 2715-6753
DOI : https://doi.org/10.36769/asy.v27i1.1822

Overall, the results indicate that LLM-based da'wah chatbots perform effectively when addressing religious topics with well-established references and consensus among scholars. However, additional validation mechanisms are required for contemporary and ambiguous religious issues to minimize hallucinations and ensure alignment with Islamic teachings.

5. Integrated Discussion

These results further support the argument that evaluating religious chatbots should extend beyond technical performance metrics. As highlighted in previous studies on religious bias in LLMs, response quality should also be assessed in terms of value alignment, fairness, and theological appropriateness. A technically accurate response may still be problematic if it conflicts with accepted Islamic principles or fails to adequately represent established scholarly perspectives.

These findings suggest that LLMs perform more effectively when answering questions with well-established religious references. Topics related to worship and morality are extensively represented in both religious literature and publicly available textual resources, enabling the model to generate more accurate and consistent responses. Conversely, questions involving contemporary issues and ambiguous religious matters often require contextual interpretation, nuanced reasoning, and consideration of differing scholarly opinions, making them more challenging for the model.

The findings are consistent with the study conducted by Atif et al.[14] which reported that the reliability of LLMs in religious question-answering varies according to the complexity of the questions posed. Their research demonstrated that language models tend to perform better on questions grounded in explicit religious knowledge than on questions requiring deeper interpretation or jurisprudential reasoning.

Regarding hallucination, the evaluation revealed that 18.4% of the generated responses contained unsupported, speculative, or unverifiable information. The highest hallucination rates were identified in ambiguous questions (44%) and contemporary issues (32%). These results indicate that the model still faces difficulties when dealing with topics that lack clear textual references or involve evolving social and technological developments.

This observation aligns with previous studies on hallucination in Large Language Models, which have identified hallucination as one of the most persistent challenges in LLM

[14] F. Atif, N. Askarbekuly, K. Darwish, dan M. Choudhury, "Sacred or Synthetic? Evaluating LLM Reliability and Abstention for Religious Questions," Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2025.

P-ISSN 1693-136X
E-ISSN 2715-6753
DOI : https://doi.org/10.36769/asy.v27i1.1822

deployment. In religious contexts, hallucinated responses may be particularly problematic because users often perceive chatbot-generated information as authoritative. Consequently, inaccurate religious information may contribute to misunderstandings of Islamic teachings and potentially reduce trust in digital da'wah platforms.

The evaluation of conformity with Islamic values showed a relatively high compliance rate of 88%, with worship-related questions achieving full conformity (100%). This finding indicates that the chatbot generally produces responses that are consistent with Islamic ethical principles and religious norms. Nevertheless, lower conformity scores in contemporary and ambiguous questions suggest that some responses may require additional validation by religious scholars before being presented to users.

These results further support the argument that evaluating religious chatbots should extend beyond technical performance metrics. As highlighted in previous studies on religious bias in LLMs, response quality should also be assessed in terms of value alignment, fairness, and theological appropriateness. A technically accurate response may still be problematic if it conflicts with accepted Islamic principles or fails to adequately represent established scholarly perspectives.

From a practical perspective, the findings demonstrate that LLM-based chatbots have considerable potential as supporting tools for digital da'wah. Their ability to provide rapid and interactive responses can enhance public access to religious information, particularly for fundamental religious inquiries. However, the presence of hallucinations and occasional inconsistencies with Islamic values indicates that such systems should not operate without human oversight.

These findings also support previous survey-based studies Huang[15] that hallucination remains one of the most persistent challenges in the development of Large Language Models (LLMs). The likelihood of hallucination increases when models are required to address questions involving contextual interpretation, complex reasoning processes, or information that is not explicitly available in their training corpus. This tendency was particularly evident in the contemporary issues and ambiguous question categories, which recorded significantly higher hallucination rates compared with the worship, creed, and morality categories.

[15] L. Huang, "A Survey on Hallucination in Large Language Models," ACM Transactions on Information Systems, 2025.

P-ISSN 1693-136X
E-ISSN 2715-6753
DOI : https://doi.org/10.36769/asy.v27i1.1822

In addition to hallucination-related issues, the results of this study underscore the importance of addressing potential biases in LLM-based systems used for religious applications. This observation is consistent with the findings of Plaza-del-Arco et al.,[16] who reported that Large Language Models may contain inherent biases in the representation of different religious identities and groups. Such biases may affect not only the neutrality and objectivity of generated responses but also the quality and inclusiveness of religious information provided to users. Therefore, evaluating LLM-based da'wah chatbots should extend beyond technical performance metrics to include assessments of fairness, value alignment, and theological appropriateness.

Furthermore, the results of the Islamic value conformity assessment indicate that most of the chatbot-generated responses were generally consistent with Islamic ethical principles and norms. The chatbot tended to produce responses using respectful, educational, and non-confrontational language, reflecting an appropriate communication style for digital da'wah applications. However, responses addressing sensitive or controversial religious issues may still require review and validation by qualified Islamic scholars to ensure theological accuracy and to minimize the risk of misunderstanding within the broader community.

From the perspective of AI ethics, the findings of this study reinforce the view of Papakostas et al.[17] who argue that the integration of artificial intelligence into educational and religious contexts must be guided by ethical considerations, pedagogical principles, and respect for established religious authority.

Overall, the findings demonstrate that LLMs have considerable potential to support digital da'wah by providing accessible, rapid, and interactive religious information. However, the observed hallucination risks, potential biases, and occasional inconsistencies in addressing sensitive religious issues indicate that AI systems should complement rather than replace human religious expertise. Therefore, the effective deployment of LLM-based da'wah chatbots requires continuous human oversight, expert validation, and appropriate governance mechanisms, particularly when addressing contemporary fiqh issues and questions involving multiple scholarly interpretations.

[16] F.M. Plaza-del-Arco et al., "Divine LLaMAs: Bias, Stereotypes, Stigmatization, and Emotion Representation of Religion in Large Language Models," Findings of EMNLP, 2024.
[17] A. Papakostas, G.D. Kallergis, dan D. Politis, "Artificial Intelligence in Religious Education: Ethical, Pedagogical, and Theological Perspectives," Religions, 2025.

P-ISSN 1693-136X
E-ISSN 2715-6753
DOI : https://doi.org/10.36769/asy.v27i1.1822

CONCLUSION

This study evaluated the quality of responses generated by an LLM-based da'wah chatbot using three evaluation dimensions: accuracy, hallucination, and conformity with Islamic values. The results indicate that the chatbot achieved relatively high accuracy and generally produced responses that were consistent with Islamic values. However, hallucinations were still observed, particularly in contemporary and ambiguous religious questions that require contextual interpretation and complex reasoning.

Overall, the findings demonstrate that LLMs have considerable potential to support digital da'wah by providing rapid and interactive access to religious information. Nevertheless, human oversight and expert validation remain necessary, especially for sensitive religious issues and questions involving multiple scholarly interpretations. Future research may focus on integrating reliable Islamic knowledge sources to further improve response quality and reduce hallucination risks.

REFERENCES

Alan, A. Y. (2025). Improving LLM Reliability with RAG in Religious QA. *Turkish Journal of Engineering*.

Atif, F., Agrawal, A., Awadallah, A. H., Caruana, R., & Ribeiro, M. T. (2025). Sacred or Synthetic? Evaluating LLM Reliability and Abstention for Religious Questions. *arXiv*. https://arxiv.org/abs/2508.08287

Atif, F., Askarbekuly, N., Darwish, K., & Choudhury, M. (2025). Sacred or Synthetic? Evaluating LLM Reliability and Abstention for Religious Questions. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 8(1), 217–226. https://doi.org/10.1609/aies.v8i1.36543

Bhatia, G. (2026). *Advances in AI Systems on Islamic Knowledge Capabilities*.

Brown, T. B. (2020). *Language Models are Few-Shot Learners*. NeurIPS.

Dam, S. K., Hong, C. S., Qiao, Y., & Zhang, C. (2024). A Complete Survey on LLM-based AI Chatbots. *arXiv*.

Ehrlich-Sommer, A. (2025). ForestGPT: Domain-Specific LLM. *Electronics*.

Floridi, L. (2019). Establishing the Rules for Building Trustworthy AI. *Nature Machine Intelligence*.

Geifman, Y., & El-Yaniv, R. (2019). *SelectiveNet: A Deep Neural Network with Reject Option*. ICML.

Guci, A. (2024). Tantangan Pendidikan Islam Zaman Modern. Tarqiyatuna.

Huang, L. (2025). A Survey on Hallucination in Large Language Models. *ACM Transactions on Information Systems*.

Kadavath, S. (2022). *Uncertainty Estimation for Language Model Predictions*.

P-ISSN 1693-136X
E-ISSN 2715-6753
DOI : https://doi.org/10.36769/asy.v27i1.1822

Khalila, Z., Khaled, H., & Elmahdy, M. (2025). Investigating Retrieval-Augmented Generation in Quranic Studies: A Study of 13 Open-Source Large Language Models. arXiv. https://arxiv.org/abs/2503.16581

Kirichenko, P. (2025). *AbstentionBench: Evaluating LLM Abstention*.

Lewis, P. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS.

Muin, M. T. (2024). Konsep Pendidikan Anak di Era Digital dalam Perspektif Al-Qur'an. Tarqiyatuna.

Muzakki, Z. (2023). Integrasi Ilmu Ekonomi Islam dan Pendidikan Agama Islam dalam Era Society 5.0. I-BEST: Islamic Banking and Economic Law Studies.

Ouyang, L. (2022). *Training Language Models to Follow Instructions with Human Feedback*. NeurIPS.

Papakostas, A., Kallergis, G. D., & Politis, D. (2025). Artificial Intelligence in Religious Education: Ethical, Pedagogical, and Theological Perspectives. *Religions*, *16*(5), 563.

Plaza-del-Arco, F. M., Cercas Curry, A., Paoli, S., Cercas Curry, A., & Hovy, D. (2024). Divine LLaMAs: Bias, Stereotypes, Stigmatization, and Emotion Representation of Religion in Large Language Models. Findings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), 4346–4366.

Sharma, A., & Gupta, M. (2025). Quantifying Religious Bias in Open LLMs through Demographic Attributes. *arXiv*. https://arxiv.org/html/2503.07510v1

Simbeck, D., & Mahran, M. (2025). Mechanistic Interpretability with SAEs: Probing Religion, Violence, and Geography in Large Language Models. *arXiv*. https://arxiv.org/abs/2509.17665

Vaughan, G. (2025). Wisdom of the Heart: A Review of Religion and AI. *Religions*.

Wang, L. (2023). *Survey on LLM-based Autonomous Agents*.

Wen, B. (2025). Know Your Limits: A Survey of Abstention in LLMs. *TACL*.