

## CONFERENCE PAPER • KICE 2026 • YOGYAKARTA, INDONESIA

## Question Answering System to Educate Woman's *Fiqh* using Indonesia Bidirectional Encoder Representations from Transformers

Nurdila Farha Kamila<sup>1</sup>, Dian Sa'adillah Maylawati<sup>2</sup>, Eva Nurlatifah<sup>3</sup>, Yana Aditia Gerhana<sup>4</sup>, Ichsan Taufik<sup>5</sup>

<sup>1,2,3,4,5</sup>Department of Informatics, Faculty of Science and Technology, UIN Sunan Gunung Djati, Bandung, Indonesia

<sup>2</sup>Email: diansm@uinsgd.ac.id

*Proceedings of the Kalijaga International Conference on Education 2026*

**Abstract.** Women frequently seek guidance on these matters because they directly affect daily religious practices, yet reliable information can be difficult to access quickly. To address this need, the study implements an intelligent chatbot capable of providing accurate and relevant responses. Therefore, The development of an automated question-answering system designed to support learning and understanding of women's Islamic jurisprudence (*fiqh*), particularly the topics of menstruation, postpartum bleeding, and *istihādah*. The system utilizes the Indonesian Bidirectional Encoder Representations from Transformers (IndoBERT) model as a natural language processing engine and is integrated into the Telegram platform for ease of user access. This model facilitates the acquisition of information related to women's *fiqh* through an automated question answering system. There are two IndoBERT models were used in the fine-tuning process, including Rifky/Indobert-QA and Wikidepia/Indobert-lite-squad. The results of this study indicate that the Rifky/IndoBERT-QA model provides better performance in the question-and-answer system for the menstruation, postpartum bleeding, and *istihadah* domains. Evaluation using BERTScore yielded a 73.7% of precision, 73.4% of recall, and an 73.3%. of f1-score. The user performance evaluation used a Mean Opinion Score (MOS) with an average value of 4,09, indicating that the chatbot was considered good. Overall, this research contributes valuable insights into the application of Transformer-based language models for religious knowledge dissemination and demonstrates the potential of IndoBERT as an effective solution for specialized Islamic jurisprudence domains for women's need.

**Keywords:** *Fiqh*; IndoBERT; Natural Language Processing; Question Answering System;

### INTRODUCTION

*Fiqh* is a structured body of Islamic knowledge that guides daily conduct and worship practices. Within this field, women's *fiqh* has an important educational role because it addresses matters that directly affect the validity and timing of worship, including menstruation, postpartum bleeding, and *istihādah* [1], [2], [3], [4]. These topics become especially important for adolescent girls and young women because the onset of menstruation often occurs during school age. Prior studies on Islamic education have reported that many students still have limited understanding of women's *fiqh*, particularly when they are required to distinguish between different types of blood and their implications for purification and prayer [2], [5]. Although information about women's *fiqh* is available through books, religious teachers, and online sources, learners may face difficulty in finding answers quickly and verifying whether a source is credible. Digital-native learners increasingly depend on online information, yet information searching can still require substantial effort because users must formulate queries, compare sources, and filter irrelevant results [6]. In an educational context, this situation indicates a need for a learning medium that can provide concise, credible, and contextually relevant answers. A question answering system (QAS) can help meet this need because it accepts natural language questions and returns direct answers rather than a list of documents [7], [8].

The development of QAS has advanced rapidly with Transformer-based models. Bidirectional Encoder Representations from Transformers (BERT) introduced contextual language understanding by reading text bidirectionally, which has made it effective for extractive QA tasks [9], [10]. For Indonesian language processing, IndoBERT and related benchmark resources have enabled more accurate contextual representation for local language structures [11]. Several studies have applied IndoBERT or BERT-based models to Indonesian QA and chatbot systems, including website content evaluation, campus orientation guides, academic services, nutrition information, and child consultation services [12], [13], [14]. These studies indicate that fine-tuned Transformer models can answer domain-specific questions with promising performance.

In the Islamic knowledge domain, recent works have explored expert systems, semantic retrieval, and large language model approaches for fiqh or Islamic chatbots [15], [16], [17]. Previous research has demonstrated the success of various methods and models in developing Question Answering Systems (QAS). However, no study has yet been conducted to develop an automated question-answering system specifically for the domain of women's *fiqh* (Islamic jurisprudence), comprehensively covering menstruation (*haid*), postpartum bleeding (*nifas*), and irregular bleeding (*istihadah*). Understanding the *fiqh* regarding women's bleeding is essential for every Muslim woman, yet it requires easy access to information. Therefore, this study employs the IndoBERT model, which, based on prior research, has proven effective for QAS in the Indonesian language. By focusing on this specific domain of women's *fiqh*, this research aims to facilitate information retrieval through a QAS built using IndoBERT. Therefore, this study integrates model evaluation with user evaluation through Mean Opinion Score (MOS).

This study aims to develop and evaluate an IndoBERT-based QAS for women's fiqh education. The system is designed as a Telegram chatbot to make access easier for learners. The specific objectives are to construct a domain-specific QA dataset, fine-tune and compare two IndoBERT-based models, deploy the best model through a chatbot interface, and evaluate both answer quality and user perception. The contribution of this study is the application of Indonesian Transformer-based QA to a specialized Islamic education domain, supported by empirical evaluation of model performance and learner-facing usability.

## METHODS

### Study Design

This study is applied research that develops a closed-domain extractive QAS for women's fiqh. The research process followed Cross-Industry Standard Process for Data Mining (CRISP-DM) as shown in Figure 1, which consists of business understanding, data understanding, data preparation, modeling, evaluation, and deployment [18]. The educational problem was defined as limited access to concise and credible answers about menstruation, postpartum bleeding, and istihādah. The technical objective was to fine-tune an Indonesian Transformer model that can extract appropriate answer spans from a trusted fiqh context.

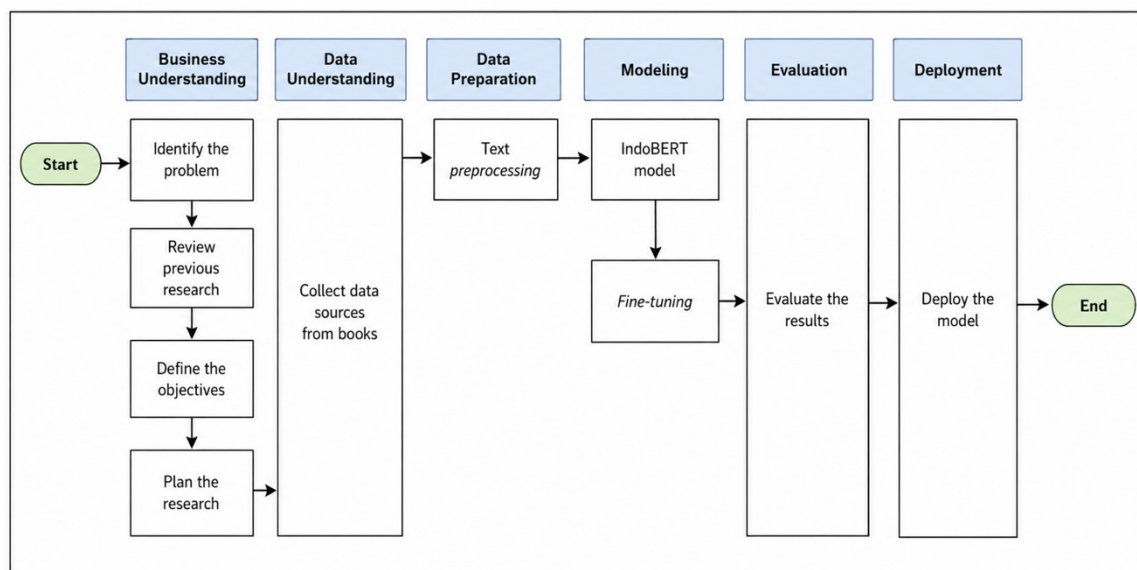


Figure 1. CRISP-DM Adoption

### Dataset Construction and Preprocessing

The knowledge source was the Indonesian book "*Kupas Tuntas Haid, Nifas, dan Istihadhoh*", a translation of *Al-Ibanah wal Ifadah* [17], [19]. The printed book was scanned and converted into digital text using optical character recognition. The extracted material was organized into 35 contexts and transformed into 504 question-answer pairs. The dataset was then cleaned through case folding, removal of excessive whitespace, and regular expression-based correction of irrelevant characters as shown in Figure 2. The question and answer spans were annotated using cdQA-annotator so that the data could be represented in a QA-compatible format. Table 1 summarizes the dataset and pre-processing in this research.

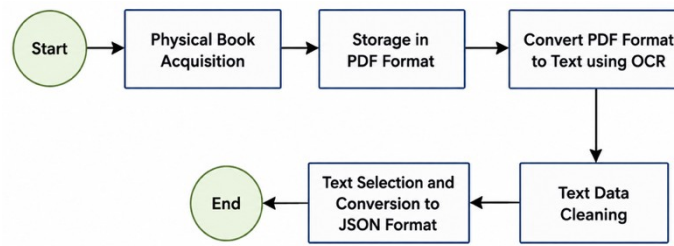


Figure 2. Pre-processing stage

Table 1. Dataset and preprocessing summary.

| Component             | Description                                                                                   |
|-----------------------|-----------------------------------------------------------------------------------------------|
| Knowledge domain      | Women's fiqh: menstruation, postpartum bleeding, and istihādah                                |
| Source                | Indonesian translation of Al-Ibanah wal Ifadah                                                |
| Contexts              | 35 contexts                                                                                   |
| Question-answer pairs | 504 pairs                                                                                     |
| Preprocessing         | OCR extraction, case folding, whitespace cleaning, regex cleaning, and answer-span annotation |
| Application domain    | Closed-domain extractive question answering for educational chatbot use                       |

## Modeling and Deployment

Two IndoBERT-based models were compared: Rifky/Indobert-QA and Wikipedia/Indobert-lite-squad. Both models were selected because they had been prepared for Indonesian QA and could be further fine-tuned on the women's fiqh dataset. Model 1 is Rifky/Indobert-QA, a fine-tuned version of IndoBERT Base-Uncased developed by IndoLEM. Model 2 is Wikipedia/Indobert-lite-squad, a fine-tuned version of IndoBERT-Lite Base developed by IndoBenchmark. Both models have been fine-tuned on the SQuAD 2.0 dataset and possess a foundational understanding of the Indonesian QA format, allowing for further fine-tuning on the specific domains of menstruation (*haid*), postpartum bleeding (*nifas*), and irregular uterine bleeding (*istihadah*). Each model was trained to predict the start and end positions of answer spans within a context. During deployment, the system used a retriever-reader design. TF-IDF retrieved the most relevant context for a user question, and the fine-tuned IndoBERT model served as the reader that extracted the answer span [20]. The best-performing model was integrated into Telegram using the pyTelegramBotAPI library, enabling users to select topics and submit questions through a familiar mobile messaging interface. The process of developing a QA model for the topics of menstruation, postpartum, *istihadah* is carried out through the following stages as in Figure 3.

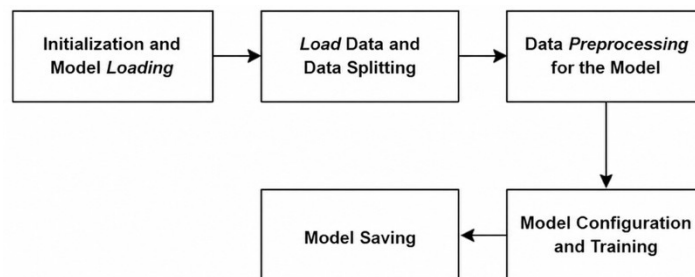


Figure 3. Modeling process

## Evaluation

Model performance was evaluated using BERTScore, which measures contextual similarity between the model answer and the reference answer through precision, recall, and F1-score [21]. In addition to automatic evaluation, prompt-based testing was conducted by using 12 natural questions representing different contexts. Human judgement classified answers as suitable or unsuitable. User evaluation used Mean Opinion Score (MOS) on a five-point scale to assess perceived usefulness, understanding, ease of interaction, answer clarity, and response speed. Twenty-one female respondents participated, consisting of seven junior high school students, seven senior high school students, and seven university students. Questionnaire-based evaluation is appropriate for assessing perceived usability and quality in user-facing digital systems [22], [23]. Figure 4 shows the overall evaluation process in this research.

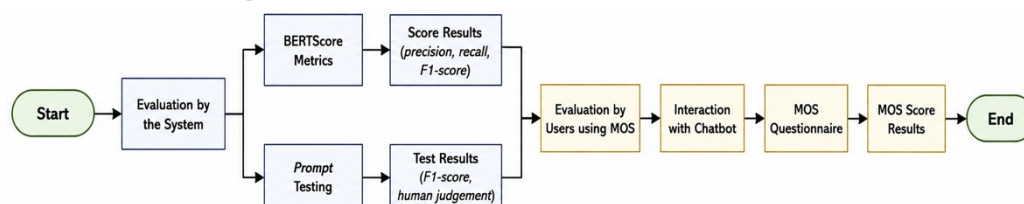


Figure 3. Evaluation process

## RESULTS AND DISCUSSION

### Data Preparation Result

The raw data acquisition process involves scanning physical books and saving the content in PDF format. Subsequently, text is extracted from the PDF files using a web-based OCR (Optical Character Recognition) tool, allowing the text to be detected, anoted, and converted into JSON format, as shown in Figure 4. Data sourced from the translated book *Al-Ibanah wal Ifadah* is utilized at this stage as a question-answering dataset, comprising 504 question-answer pairs derived from a total of 35 contexts.

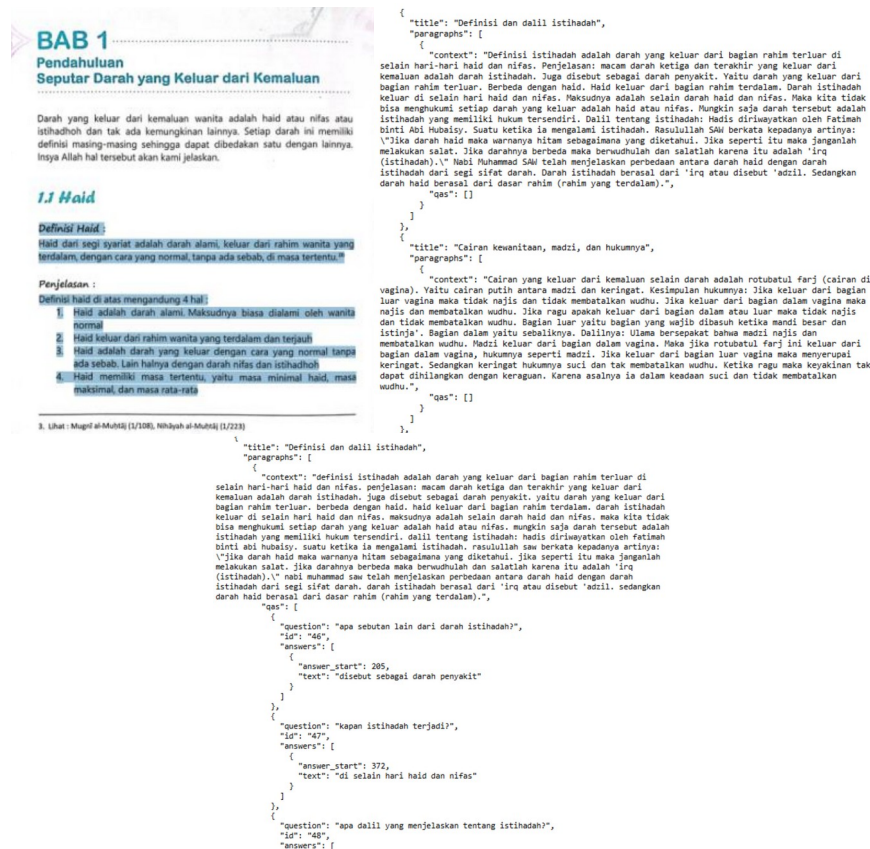


Figure 4. The example of data preparation

### Fine-Tuned IndoBERT Performance

An evaluation was conducted to measure the model's performance in performing question-answering tasks within the domain of Islamic jurisprudence regarding women (menstruation, postpartum bleeding, and irregular uterine bleeding). The evaluation encompassed an analysis of loss curves, semantic similarity measurement using BERTScore, an assessment of answers derived from user prompt testing, and an evaluation of system quality via user questionnaires. Figure 5 displays the training and validation loss graphs for the first and second models: Rifky/Indobert-QA and Wikidepia/Indobert-lite-squad. The loss graph illustrates the performance of the Indobert-QA model during the training process for the question-answering task covering topics related to menstruation, postpartum bleeding, and irregular bleeding. The training loss curve for the first model shows a significant decline, dropping from 2.95 in the first epoch to 0.56 by the eighth epoch. This indicates that the model successfully learned patterns from the training data in a gradual and consistent manner. Meanwhile, the validation loss also decreased, falling from 2.67 to 1.79 by the sixth epoch, before beginning to rise again in epochs 7 and 8. This increase suggests the onset of overfitting, where the model adapted too closely to the training data, resulting in a decline in performance on the validation data. Overall, the graph demonstrates that the Rifky/Indobert-QA model was capable of learning effectively from the data, despite indications of overfitting toward the end of training. These results show that Rifky/Indobert-QA maintained reasonable stability during training, notwithstanding fluctuations in validation loss and a gap of 1.4 between the training and validation loss values.

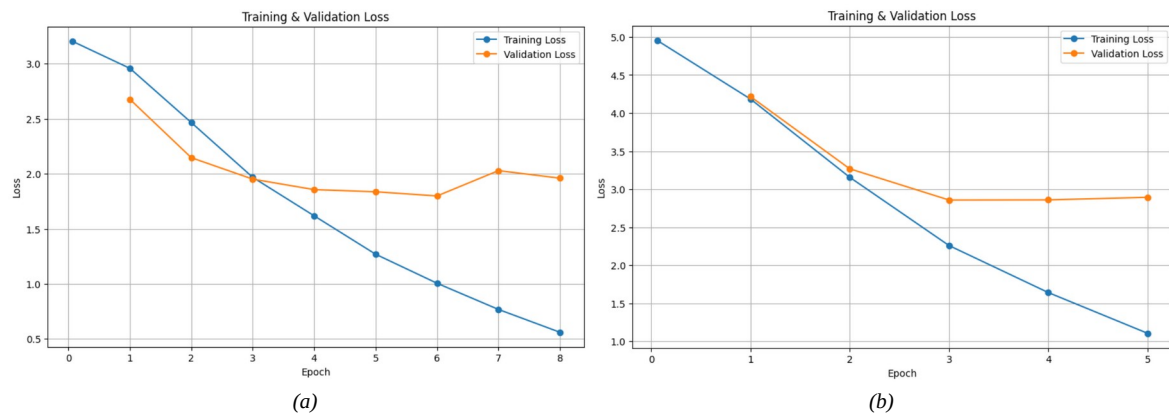


**Figure 5.** Curve of Training and Validation Loss: (a) first model, (b) second model

In contrast, the curve for the second model shows that training loss started at a high level, 4.18, but consistently decreased to reach 1.10 by the fifth epoch. This indicates that the model was able to learn patterns from the training data effectively. Meanwhile, the validation loss also showed an initial decline, dropping from 4.21 to 2.85, but stabilized at that level during the fourth and fifth epochs. This suggests the model had reached a point where further learning yielded no significant improvement regarding the validation data. The substantial gap of 1.8 between training loss and validation loss at the end of training indicates clearer signs of overfitting compared to the previous model. The persistently high loss values suggest that the model still struggled to generalize to data beyond the training set.

BERTScore evaluation is conducted to assess the similarity between predicted answers and reference answers. This metric evaluates answer quality based on three components: precision, recall, and F1-score. Precision reflects the relevance of words in the predicted answer to the reference answer, recall indicates the amount of information from the reference answer successfully captured by the system, and the F1-score provides the harmonic mean of precision and recall. Figure 6 displays the BERTScore evaluation results of both models.

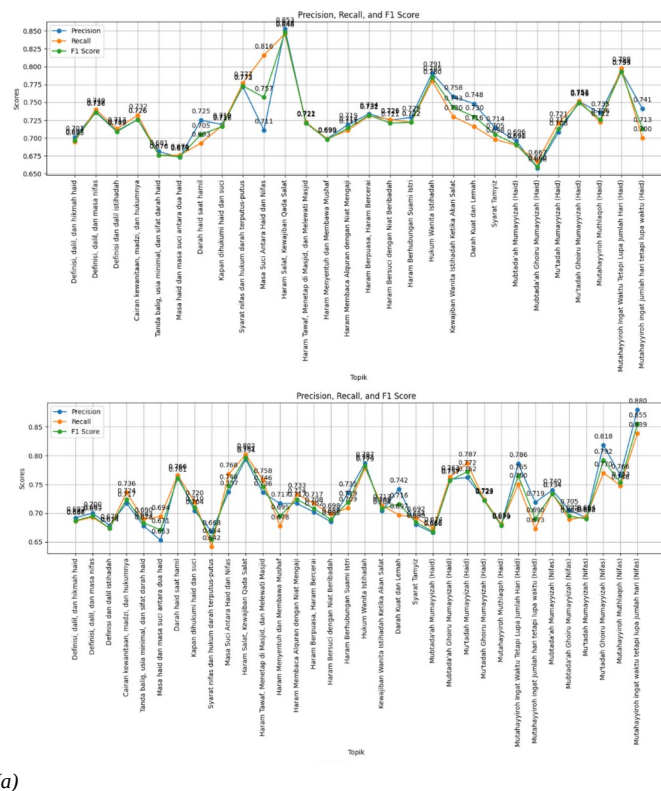


**Figure 6.** Curve of BERTScore based on topics: (a) first model, (b) second model

Table 2 summarizes the performance comparison between the two IndoBERT-based models. Rifky/Indobert-QA obtained the highest BERTScore values, with 73.7% precision, 73.4% recall, and 73.3% F1-score. It also achieved stronger performance in the prompt-based test, with an 85.4% F1-score and 83.3% answer suitability based on human judgement. Wikidpedia/Indobert-lite-squad produced a lower prompt F1-score of 61.7% and 66.6% answer suitability, although its BERTScore was relatively close to the first model.

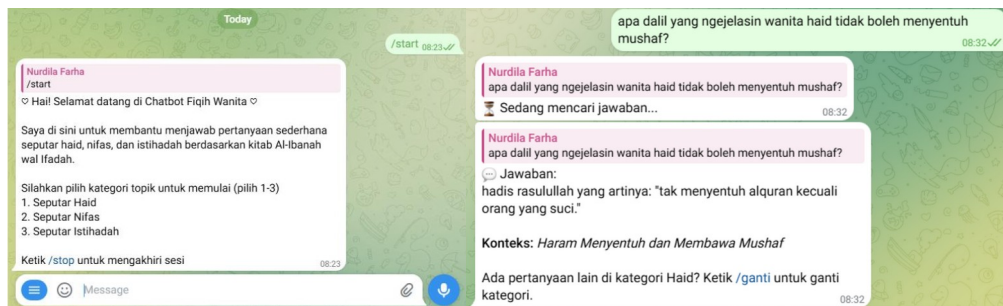
**Table 2. Evaluation results of fine-tuned IndoBERT models.**

| Model                          | BERTScore Precision | BERTScore Recall | BERTScore F1 | Prompt F1 | Suitable Answers |
|--------------------------------|---------------------|------------------|--------------|-----------|------------------|
| Rifky/Indobert-QA              | 73.7%               | 73.4%            | 73.3%        | 85.4%     | 83.3%            |
| Wikidpedia/Indobert-lite-squad | 73.0%               | 72.0%            | 72.5%        | 61.7%     | 66.6%            |

The difference between the two models became more visible in prompt testing than in BERTScore. This finding suggests that automatic semantic similarity is useful for measuring general answer closeness, but natural user questions may expose limitations in vocabulary, question phrasing, and Arabic-derived religious terminology. Rifky/Indobert-QA produced answers that were more complete and closer to the reference in most prompt cases. This result is consistent with previous Indonesian QA studies showing that IndoBERT-based models perform well when fine-tuned for domain-specific tasks [12]–[14].

### Telegram Chatbot Implementation

The deployment stage implemented the best model in a Telegram chatbot. The chatbot begins with the /start command and presents three main categories: menstruation, postpartum bleeding, and istihādah. After a category is selected, users can ask questions within the selected topic. The /ganti command allows users to change the category, while /stop ends the interaction. The retriever-reader design made the interaction more flexible because the system first selected a relevant context before the IndoBERT reader extracted the answer. This design follows the general logic of Indonesian document QA systems that combine retrieval with Transformer-based reading [19]. Figure 7 show the example of Telegram implementation with Indonesian language.

**Figure 7. Example of Telegram Deployment**

### User Evaluation with Mean Opinion Score

Table 3 presents the MOS results across three education levels. The overall MOS was 4.09, indicating that users generally perceived the chatbot as good. The junior high school group produced the lowest MOS at 3.8, while the senior high school and university groups both achieved 4.2. This pattern indicates that the chatbot was more easily accepted by users with higher educational levels, while younger learners may require simpler phrasing or additional explanation.

**Table 3. User evaluation using Mean Opinion Score.**

| Respondent Group   | Number of Respondents | MOS  | Interpretation      |
|--------------------|-----------------------|------|---------------------|
| Junior high school | 7                     | 3.8  | Neutral toward good |
| Senior high school | 7                     | 4.2  | Good                |
| University         | 7                     | 4.2  | Good                |
| Overall            | 21                    | 4.09 | Good                |

The item-level MOS analysis showed that users gave high scores to fast response and information access, but the lowest average score was associated with answer clarity and suitability. This result is educationally meaningful: speed and accessibility are important, but learners also need explanations that are precise, readable, and age-appropriate. Compared with Islamic chatbot studies that use semantic retrieval or retrieval-augmented generation [15]–[17], this study shows that extractive IndoBERT QA can be effective for trusted-source answers, but it is constrained by the wording and coverage of the source context. The system is therefore suitable as a learning aid, while complex personal cases should still be referred to qualified religious educators.

### Limitations

This study has several limitations. First, the dataset was derived from a single translated book, so the knowledge coverage remains limited. Second, the model was not fully robust in handling Arabic terminology and informal question variations. Third, the system used extractive QA, which means that it could answer only when the relevant answer span existed in the context. Questions requiring reasoning, case comparison, or synthesis beyond the source text were outside the scope of this study.

## CONCLUSION

This study developed an IndoBERT-based question answering system for women's fiqh education, focusing on menstruation, postpartum bleeding, and istihādah. The system used a domain-specific dataset of 504 question-answer pairs from 35 contexts and was implemented as a Telegram chatbot through a retriever-reader architecture. Rifky/Indobert-QA outperformed Wikipedia/Indobert-lite-squad, achieving 73.3% BERTScore F1, 85.4% prompt F1, and 83.3% answer suitability. User evaluation also showed a good overall MOS of 4.09, indicating that the chatbot was useful and accessible for women's fiqh learning. The main limitations are the restricted dataset size, limited coverage of Arabic terminology, and the extractive nature of the system. Future research should expand the data sources, simplify answer delivery for younger learners, and explore generative or retrieval-augmented approaches to provide more natural and explanatory responses.

## ACKNOWLEDGMENTS

The authors thank the Department of Informatics, Faculty of Science and Technology, UIN Sunan Gunung Djati Bandung, for academic support during the completion of this research.

## REFERENCES

- [1] M. Zaki, "Fikih, Ushul Fikih dan Qawaid Al-Fiqhiyyah dalam Lintasan Sejarah," *NUR EL-ISLAM : Jurnal Pendidikan dan Sosial Keagamaan*, vol. 9, no. 2, pp. 1–16, Sep. 2022, doi: 10.51311/nuris.v9i2.521.
- [2] Sudarsri Lestari, Endhang Suhilmiati, and Erisy Syawiril Ammah, "Kajian Fiqih Wanita Tentang Taharah Haid Pada Siswa Mi Al Ihsan Banyuwangi," *Dedikasi Nusantara: Jurnal Pengabdian Masyarakat Pendidikan Dasar*, vol. 1, no. 2, pp. 86–91, Dec. 2021, doi: 10.29407/dedikasi.v1i2.17031.
- [3] N. Sa'adah and A. A. Zafi, "Hukum Seputar Darah Perempuan Dalam Islam," *Martabat: Jurnal Perempuan dan Anak*, vol. 4, no. 1, pp. 155–174, 2020.
- [4] A. U. Fadhillah, "Women's Fiqih (Haid, Nifas, And Istihadhah) on Reproductive Health," in *Proceeding International Conference on Religion, Science and Education*, 2023, pp. 1029–1034.
- [5] F. Kustina, "Fikih Wanita dan Pemahaman Remaja Putri di Pondok Pesantren Sunan Drajat," *Multidisciplinary Journal of Education, Economic and Culture*, vol. 1, no. 1, pp. 41–51, 2023.
- [6] F. Handayani and F. Fauzi, "Kendala-kendala yang Dihadapi Digital Native dalam Pencarian Informasi," *Shaut Al-Maktabah: Jurnal Perpustakaan, Arsip dan Dokumentasi*, vol. 15, no. 1, pp. 31–39, Jun. 2023, doi: 10.37108/shaut.v15i1.766.
- [7] M. A. Calijorne Soares and F. S. Parreiras, "A literature review on question answering techniques, paradigms and systems," *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 6, pp. 635–646, Jul. 2020, doi: 10.1016/j.jksuci.2018.08.005.
- [8] A. F. Hanifah and R. Kusumaningrum, "Non-Factoid Answer Selection in Indonesian Science Question Answering System using Long Short-Term Memory (LSTM)," *Procedia Comput. Sci.*, vol. 179, pp. 736–746, 2021, doi: 10.1016/j.procs.2021.01.062.
- [9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Oct. 2018.
- [10] M. V. Koroteyev, "BERT: A Review of Applications in Natural Language Processing and Understanding," Mar. 2021.
- [11] B. Wilie *et al.*, "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," *arXiv preprint arXiv:2009.05387*, 2020.
- [12] S. U. Masrurroh, K. H. Putri, K. Hulliyah, I. M. Shofi, and V. Amrizal, "Development Of A Question Answering Model Using The Indobert Algorithm To Evaluate Website Content In Indonesian Language," in *2023 11th International Conference on Cyber and IT Service Management (CITSM)*, IEEE, Nov. 2023, pp. 1–6. doi: 10.1109/CITSM60085.2023.10455383.
- [13] F. Dartiko, M. Yusa, A. Erlansari, and S. A. Basha, "Comparative analysis of transformer-based method in a question answering system for campus orientation guides," *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi*, vol. 8, no. 1, pp. 126–143, 2024.
- [14] B. Richardson and A. Wicaksana, "Comparison of indobert-lite and roberta in text mining for Indonesian language question answering application," *Int. J. Innov. Comput. Inf. Control*, vol. 18, no. 6, pp. 1719–1734, 2022.
- [15] S. Rahayu, N. S. Harahap, S. Agustian, and P. Pizaini, "Penerapan Teknologi LangChain pada Question Answering System Fikih Empat Madzhab: Application of Langchain Technology to the Fiqh Question Answering System of Four Madhhab," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 3, pp. 974–983, 2024.
- [16] T. R. Masuzzahra, K. Umam, H. Mustofa, and M. R. Handayani, "Hana: An ai chatbot for islamic jurisprudence on menstruation using sbert and tf-idf," *Journal of Applied Informatics and Computing*, vol. 9, no. 3, pp. 1013–1024, 2025.
- [17] D. A. bin A. Assegaf, "Terjemah Al-Ibanah Wal Ifadhah Kupas Tuntas Hukum Haid, Nifas Dan Istihadhah," 2013, *Terjemahan oleh Fakhry Emil Habib*.
- [18] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Comput. Sci.*, vol. 181, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.
- [19] M. F. N. Zamri, A. M. Zabidi, N. Siamil, and Z. Hussin, "Penerokaan Kaedah Fiqh dalam Perbahasan Fiqh Darah Haid: Kajian Terhadap Buku al-Ibanah wa al-Ifadhah fi Ahkam al-Haid wa al-Nifas wa al-Istihadhah 'Ala Mazhab al-Imam

- al-Syafi'i: Islamic Legal Maxim Exploration in Fiqh Debate for Menstrual Blood: A Study of the Scripture al-Ibanah wa al-Ifadhah fi Ahkam al-Haid wa al-Nifas wa al-Istihadhah 'Ala Mazhab al-Imam al-Syafi'i," *Journal of Muwafaqat*, vol. 5, no. 1, pp. 71–87, 2022.
- [20] G. Nur Ahmad and A. Romadhony, "End-to-End Question Answering System for Indonesian Documents Using TF-IDF and IndoBERT," in *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, IEEE, Oct. 2023, pp. 1–6. doi: 10.1109/ICAICTA59291.2023.10390111.
- [21] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," Feb. 2020.
- [22] S. Hajesmaeel-Gohari, F. Khordastan, F. Fatehi, H. Samzadeh, and K. Bahaadinbeigy, "The most used questionnaires for evaluating satisfaction, usability, acceptance, and quality outcomes of mobile health," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, p. 22, Dec. 2022, doi: 10.1186/s12911-022-01764-2.
- [23] S. Syaifullah, I. G. P. S. Wijaya, and A. Y. Husodo, "Satisfaction Information System of Academic Administration Services Based on IPA (Importance Performance Analysis) Study Case in Faculty of Engineering, Mataram University," *Journal of Computer Science and Informatics Engineering (J-Cosine)*, vol. 2, no. 1, Aug. 2018, doi: 10.29303/jcosine.v2i1.50.