IRPI PUBLISHER
Institute of Research and Publication Indonesia
IJIRSE: Indonesian Journal of Informatic Research and Software Engineering
Journal Homepage: https://journal.irpi.or.id/index.php/ijirse
Vol. 5. No.1. Maret 2025, pp: 68 - 78
E-ISSN: 2775-5754 | P-ISSN: 2797-2712

Optimization of Feminacare Chatbot Application Using SeaLLM Model

Nurul Nyi Qoniah[1], Dian Ramadhani[2]

[1,2] Dept.of Informatics Engineering, Faculty of Engineering, Universitas Riau, Pekanbaru,Indonesia
E-Mail: [1]nurulqoniah313@gmail.com, [2]dianramadhani@lecturer.unri.ac.id

Papers: Received November 20, 2024; Refereed March 01, 2025; Approved March 10, 2025
Corresponding Author: Dian Ramadhani

Abstract

Chatbots are increasingly used in healthcare to improve access and reduce consultation times. Feminacare, a women's health application, previously used an LSTM-based chatbot with only 61% accuracy and low response relevance. This study aims to optimize the chatbot using the SeaLLM model with a Retrieval-Augmented Generation (RAG) approach to enhance accuracy and relevance. The development process included domain-specific data collection, preprocessing, chunking, embedding, and vector storage using ChromaDB. A hybrid search combining BM25 and vector similarity was implemented to retrieve relevant information, which SeaLLM then used to generate responses. Evaluation using 120 questions across three trials demonstrated improved performance, achieving 87% accuracy, 93% precision, 89% recall, and a 91% F1-score. Compared to the LSTM model, SeaLLM produced more relevant and accurate answers. This study takes a closer look at integrating the SeaLLM large language model with hybrid retrieval techniques to improve performance in domain-specific chatbots. Despite these improvements, limitations remain in handling complex or out-of-distribution queries due to a still-limited knowledge base, suggesting the need for future dataset expansion and model refinement.

Keyword: Chatbot, SeaLLM, Retrieval-Augmented Generation (RAG), Women's Health, Feminacare.

1. INTRODUCTION

In today's digital era, advances in information technology are proliferating. One of the latest information technologies developing is chatbots, a form of artificial intelligence (AI) that allows conversations or interactive communication between users and computer systems through text, voice, and/or visuals automatically [1]. Chatbots are currently being used in various fields, such as in healthcare [2][3][4][5].

Healthcare chatbots are increasingly being used to improve access and efficiency by reducing hospital readmissions, slashing consultation wait times, and reducing unnecessary hospital visits [4]. By automating routine tasks and providing basic medical information, chatbots help lower healthcare costs and improve accessibility [3][5].

Feminacare is a health-based mobile application that aims to provide online women's health consultation services. One of the features owned by this application is a chatbot in the Indonesian language that can be used to ask questions about women's health problems, such as menstruation, pregnancy, and other reproductive health problems. The use of chatbots in the Feminacare application using the LSTM method achieved an accuracy of around 61% with a loss value of 1.65. However, in testing, the chatbot often gives less relevant answers, so performance improvements are needed to increase the validity and accuracy of answers.

Based on this approach, chatbots can be divided into two main categories: retrieval-based and generative-based [6]. The LSTM models used in current chatbots are retrieval-based [7], where responses are limited to pre-trained data. This makes the model less adaptive to questions outside the training dataset's scope [6]. In contrast, generative-based chatbots can generate new responses [8] by utilizing an understanding of language patterns learned from a large dataset of trains [7].

Commonly used generative-based models are currently categorized as Large Language Models (LLM), such as GPT-3.5, GPT-4.0, LLaMA2-7B, and LLaMA2-13B [9]. There are LLMs trained in the Indonesian language, such as IndoBART, IndoGPT, Merak, SeaLLM, SEA-LION, and Komodo [10]. The study by Koto et al. (2024) showed that the SeaLLM model gave the best performance in multiple-choice questions (MCQ) tasks related to general knowledge and Indonesian culture [10]. In addition, in testing on various professional fields, SeaLLM recorded the highest accuracy in the health category [11].

Integration of domain-specific data into LLM can be done through two approaches, namely Retrieval-Augmented Generation (RAG) and Fine-Tuning (FT) [12]. The RAG method incorporates external data into the answer generation process by LLM [2], while FT adapts the pre-trained model with specific data for a particular task [12]. Research by [13] shows that the RAG approach is more effective than FT in building LLM-based knowledge systems. Through a search process based on embedded vectors from external databases [14], RAG can generate relevant responses with a lower degree of hallucination [15].
RAG also overcomes the main limitation of the LSTM model, which relies only on static training data. With this approach, the chatbot can still answer questions beyond the scope of the training data as information is dynamically retrieved from indexed external sources.

This study aims to develop an RAG-based generative chatbot by utilizing LLM models trained in Indonesian. This approach is expected to improve the relevance and accuracy of chatbot responses to questions about women's health.

2. RESEARCH MATERIALS AND METHODS

Figure 1 shows the research stages, which consist of seven main stages in the development of the Feminacare chatbot using SeaLLM with the RAG approach.

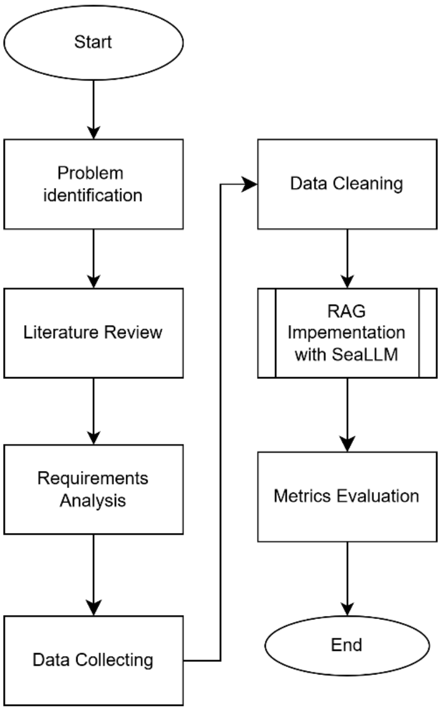


Figure 1. Research stages

2.1 Problem Identification

Previous testing of the LSTM model on a chatbot showed a low accuracy of 61% and a loss value of 1.65, and it failed to provide relevant answers. The limited dataset is also a cause of this model's limitations. This is the basis for the need for a new LLM-based approach with a retrieval mechanism.

2.2 Literature Review

A review of recent NLP developments, particularly the LLM and RAG methods, was conducted. The study also covers the understanding of women's health and the application of chatbots in the healthcare field.

2.3 Requirements Analysis

To improve the performance and relevance of Feminacare chatbot responses, an LLM model integrated with the RAG method is used to reduce hallucinations and improve answer accuracy, especially in women's health topics. The SeaLLM model was chosen because it has been optimized for Southeast Asian languages, including Indonesian [10][11]. In addition, expanding the dataset through the conversion of articles and reference books to PDF and CSV formats will enrich the chatbot's knowledge to provide more specific and relevant answers according to user needs.

2.4 Data Collection

Data for the Feminacare chatbot was collected from two main sources: articles from the Hello Sehat platform and the textbook *Kesehatan Reproduksi Remaja dan Lansia*, which serves as an additional medical reference. Table 1 shows an example of scraping results on the Hello platform.

Table 1. Example of data collection

Title	Content
Metroragia	... Metroragia ormetrorrhagiais a medical term used to describe irregular or abnormal uterine bleeding outside of normal menstrual periods. Normally in one cycle, menstruation will last for 4â€“7 days and this menstrual cycle occurs every 21â€“35 days ...

2.5 Data Cleaning

The data cleaning process ensures the quality of the data before it is used in analysis or model development. Data is cleaned by removing duplicate, typos and irrelevant entries. The final dataset is formatted in CSV and prepared for use in the model. An example of data cleaning is shown in Table 2.

Table 2. Example of data cleaning

Content	Content cleaned
... Metroragia **ormetrorrhagiais** a medical term used to describe irregular or abnormal uterine bleeding outside of normal menstrual periods. Normally in one cycle, menstruation will last for **4â€“7** days and this menstrual cycle occurs every **21â€“35** days ...	... Metroragia or metrorrhagia is a medical term used to describe irregular or abnormal uterine bleeding outside of normal menstrual periods. Normally in one cycle, menstruation will last for 4-7 days and this menstrual cycle occurs every 21-35 days ...

2.6 RAG Implementation with SeaLLM

The RAG implementation was conducted in the following seven stages, as shown in Figure 2.

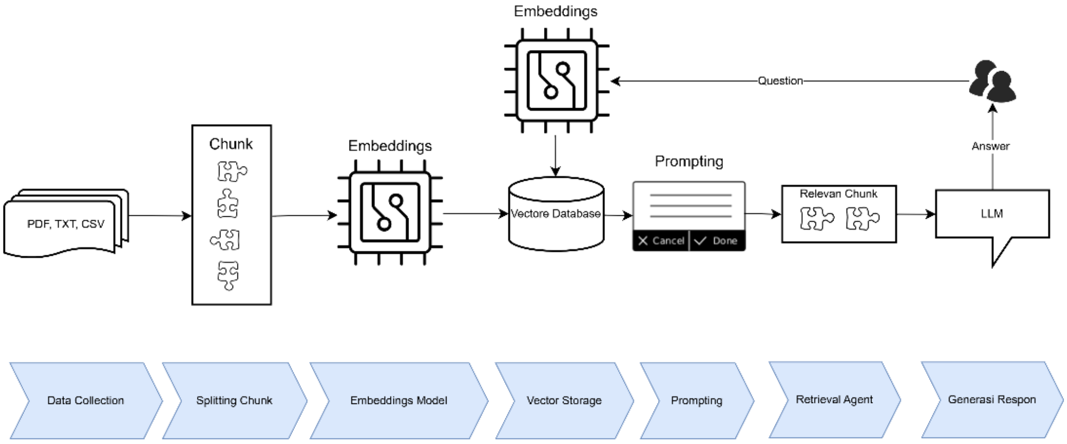


Figure 2. RAG implementation stages

1. Data Collection
   Data that have been collected from various sources related to women's health and cleaned of duplicates, irrelevant entries, and typos were stored in a single folder in CSV and PDF formats.
2. Splitting Chunk
   The documents that have been collected are broken down into small parts called chunks to facilitate information retrieval by the chatbot model as shown in Table 3. Chunking is important because the LLM model limits the amount of text it can process at once, so breaking the documents into semantically relevant chunks will improve search efficiency and response accuracy. This process must

consider the size of the chunk so that it is neither too small to lose context nor too large to mix different ideas. Some commonly used chunking techniques include splitting based on word count, sentence boundaries, and Recursive Chunking methods that utilize a sequential list of separators. In its implementation, such as with the Langchain library, the RecursiveCharacterTextSplitter class is used to break up documents with parameters such as chunk_size, chunk_overlap, and separators so that the resulting chunk remains meaningfully intact and in context.

Table 3. Example of splitting chunk

Content cleaned	Chunk
The following is a complete explanation of the symptoms of menopause in women... Basically, complaints and symptoms or characteristics of menopause in women... Conclusion Some symptoms of menopause in women include: changes in the menstrual cycle... Hot flashes is a condition where you experience a sensation of heat, either in the upper part of the body or... Obstructive sleep apnea (OSA) is a sleep disorder in which there are pauses in breathing....	Here's a full explanation of menopause symptoms in women...
	Basically, complaints and symptoms or characteristics of menopause in women are...
	Conclusion Some symptoms of menopause in women include:changes in menstrual cycle...
	Hot flashes are when you experience a sensation of heat, either in the upper body or even all over...
	Obstructive sleep apnea (OSA) is a sleep disorder characterized by pauses in breathing....

3. Embedding Model

Each chunk is converted into a vector representation using embedding, which represents the semantic meaning of the text as in Figure 3 and Table 4.

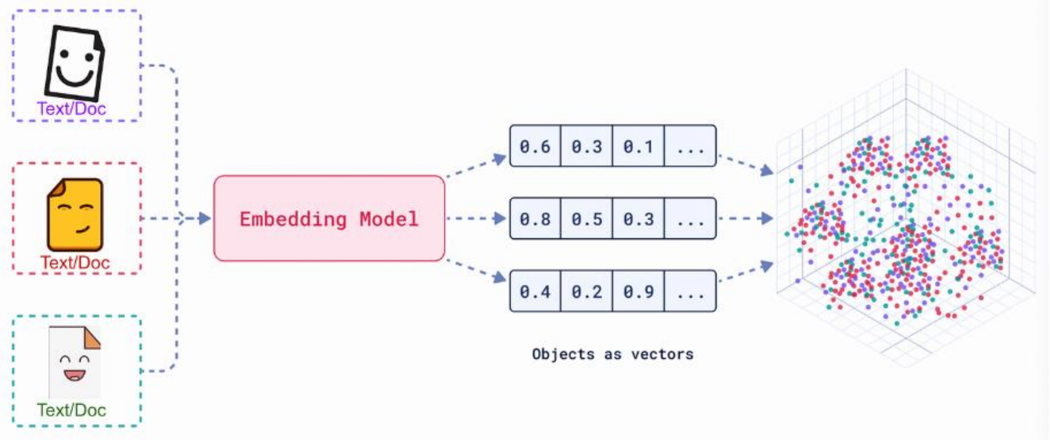


Figure 3. Embedding model

Table 4. Example of embedding model

Chunk	Vector chunk
Here's a full explanation of menopause symptoms in women...	c65df215-feab-4d45-8fb0-af2802314eb4
Basically, complaints and symptoms or characteristics of menopause in women are...	7980bd33-a0f0-45dd-813c-e1d7c9841b36
Conclusion Some symptoms of menopause in women include:changes in menstrual cycle...	4913e68a-ed5f-4999-bd38-6fbb2021901c
Hot flashes are when you experience a sensation of heat, either in the upper body or even all over...	d10b1680-c7fd-42e4-bfb2-011638a16784

Obstructive sleep apnea (OSA) is a sleep disorder characterized by pauses in breathing....	c325214e-3819-43df-a4c7-c603c8a283f5

4. Vector storage
   The embedded vectors are stored in a vector database using ChromaDB, which stores vector databases [16]. ChromaDB makes it easy to search vectors based on semantic similarity.
5. Prompting
   Prompts are used to structure instructions to the LLM and generate relevant and contextualized answers based on search results [17].
6. Retrieval Agent
   User queries are converted into vectors, and retrieval finds the most relevant chunk from the database [9] using techniques such as cosine similarity or *Euclidean* distance [2]. In this stage, retrieving relevant documents using two approaches, hybrid search and similarity search. Hybrid search is a search method that combines two main techniques, namely sparse search (BM25) and dense search (vector similarity) [18]. Sparse search works by explicitly matching keywords, while dense search uses a vector representation of the text to capture deeper semantic meaning. By combining the two, hybrid search can capture both keyword similarity and meaning context, thus improving search accuracy. Meanwhile, similarity search only relies on semantic proximity between embedding vectors.
7. Response Generation with SeaLLM
   The SeaLLM model generates responses based on relevant documents found, with parameter settings such as temperature to reduce hallucinations [9].

2.7 Metrics Evaluation and Testing

Evaluation is done using accuracy, precision, recall, and F1-score metrics and confusion matrix visualization to understand the distribution of correct and incorrect predictions with scenarios, as shown in Table 5. In addition, BERTScore was used to assess the semantic congruence between the chatbot responses and the information retrieved from the database.

Table 5. Metric evaluation scenario

Conditions	Chatbot Answers	Categories
Questions according to the dataset	Correct answer	True Positive (TP)
Question not in the dataset	Answering ignorance	True Negative (TN)
Question not in the dataset	Did not answer due to ignorance	False Positive (FP)
Questions according to the dataset	Gives wrong answer	False Negative (FN)

Accuracy, precision, recall, and F1-score metrics are calculated by [19]

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{1}$$

$$Precision = \frac{TP}{(TP+FP)} \tag{1}$$

$$Recall = \frac{TP}{(TP+FN)} \tag{1}$$

$$F1 - score = \frac{2\times(Precision \times Recall)}{(Precision+Recall)} \tag{1}$$

Nine testers with different backgrounds conducted the test, each asking five random questions about women's health. The results of the chatbot's answers were assessed based on accuracy, relevance, and conformity to the topics in the knowledge base.

3. RESULTS AND DISCUSSION

In this section, we will discuss the results of the research carried out on the application of RAG with the SeaLLM model for Feminacare chatbot optimization through several stages.

3.1 Data Collection and Cleaning

Data was collected through two primary sources: scraping from the Hello Sehat website and extraction from the book "*Kesehatan Reproduksi Remaja Dan Lansia*." the scraping resulted in 99 women's health topics, while 169 were added from previous research. All data was cleaned to remove duplicates and irrelevant formats and saved in csv format, as shown in table 6. Data from the book was converted into pdf format and covered six important topics related to women's health, such as the reproductive cycle, women's nutrition, and sexually transmitted infections.

.

Table 6. CSV data format

Content cleaned
Sanitary pads are menstrual products used by women to absorb menstrual blood. Usually, pads are made of absorbent material that is attached to a woman's underwear and used to keep menstrual blood from seeping or soiling clothes.

3.2 Splitting and Chunking

After cleaning the dataset, splitting is performed using LangChain's RecursiveCharacterTextSplitter method. Text splitting is done with separators ["\n\n", "."], where the text is first split based on two newlines, then a complete stop if it is still too long. Each chunk is limited to a maximum of 700 characters with a chunk_overlap of 0 (no overlap). From this process, a total of 787 chunks were generated.

3.3 Embedding and Storage in Vector Database

After the chunking process, each chunk is converted into a vector using the sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 embedding model, which supports multiple languages, including Bahasa Indonesia. The embedded vectors are stored in the ChromaDB vector database and visualized using PCA, as shown in Figure 4.

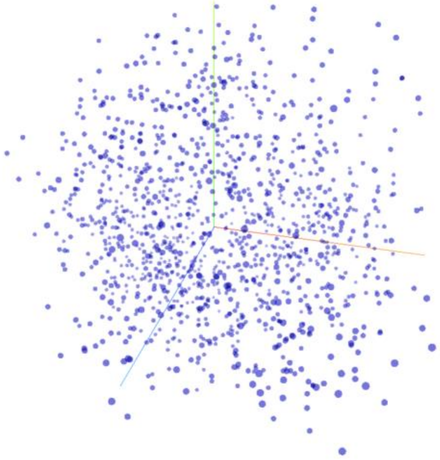


Figure 4. Vector distribution of the dataset

3.4 Retrieval

After the embedding is stored in ChromaDB, the next stage is retrieving relevant documents using hybrid search and similarity search approaches. In this research, a value of k=6 is used to retrieve the top six most relevant document chunks for each query or question from user.

3.5 Generative Model

After acquiring the relevant chunks, the next step was to generate responses using the SeaLLMs-v3-1.5B-Chat model. This model was downloaded via Hugging Face and set to respond only based on the available context using a prompt. The prompting technique was used by constructing a prompt template emphasizing that responses should refer to the context.

Prompt template

prompt_template *= """Answer the following query exclusively based on the information provided in the Context. *
If the information is not found in the context, please say "Maaf, saya tidak memiliki pemahaman tentang hal itu" instead of making up facts!

Context*: {context}*
Question*: {question}*

```
"""
system_message = """You are a humble and helpful women's health assistant. Your name is Feminacare.
You must give answers about women's health according to the information you get from the context.
If the Question and Context are different topics or NOT ENOUGH information is available, say "Maaf, saya tidak memiliki pemahaman
tentang hal itu".
"""
```

Generative response using SeaLLM

```
tokenizer = AutoTokenizer.from_pretrained(
    "SeaLLMs/SeaLLMs-v3-1.5B-Chat",
    use_auth_token="********* "
)

model = AutoModelForCausalLM.from_pretrained(
    "SeaLLMs/SeaLLMs-v3-1.5B-Chat",
    device_map="auto",
    torch_dtype=torch.float16,
    use_auth_token="*********",
)
```

An example of generative response query results using SeaLLM is shown in Table 7.

Table 7. Example answers of questions

No	Question	Answer
1	How is post-COVID syndrome diagnosed and treated?	Sorry, I have no understanding of post-COVID syndrome...
2	What causes brown menstrual blood, and is this normal?	In most cases, brown menstrual blood is normal. However, if there are other symptoms such as changes in menstrual volume, menstrual period length, or feelings of pain or cramping in the abdomen, it is advisable to consult a doctor...

3.6 Chatbot Evaluation

The evaluation was conducted using 120 questions, consisting of 80 questions that matched the dataset content and 40 questions that did not. Each question was tested three times. An example of the test results can be seen in Table 8. The chatbot responses from each experiment were then classified into True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) categories, and evaluation metrics such as accuracy, precision, recall, and F1-score were calculated. An evaluation was also conducted using BERTScore to measure the semantic similarity between the chatbot answers and the information in the dataset, which strengthens the analysis of the model's performance in terms of content quality.

Table 8. Example test question result

No	Question	Dataset	E 1	E 2	E 3
1	*How is post-COVID syndrome diagnosed and treated?*	0	0	0	0
2	*How to distinguish COVID-19 from influenza clinically?*	0	0	0	0
3	*What causes brown menstrual blood, and is this normal?*	1	1	1	1
4	*What are the causes of left breast pain that need to be considered?*	1	1	1	1

Evaluation was conducted on two retrieval methods: hybrid search and similarity search. The test results using hybrid search are shown in Table 9, while the test results using similarity search are shown in Table 10. Each method was tested with 120 queries, and each query was tested three times to measure the consistency of the model performance.

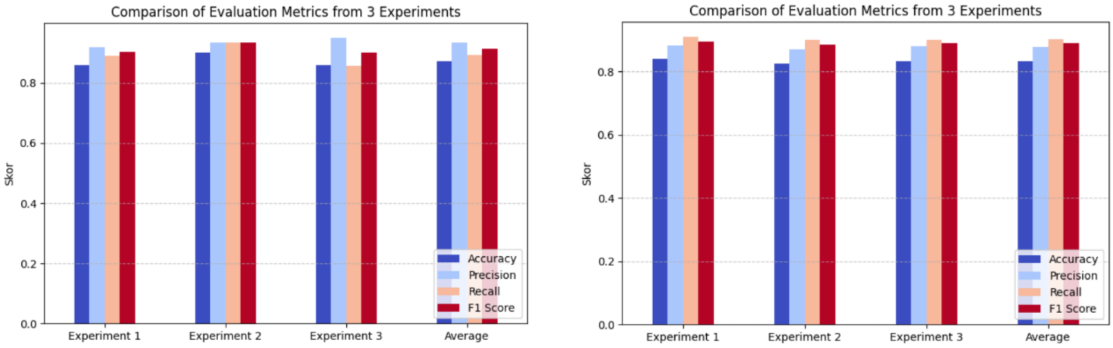
Table 9. Test results with hybrid search

Experiment	Accuracy	Precision	Recall	F1-Score	Bert Precision	Bert Recall	Bert F1-S
1	0.86	0.92	0.89	0.90	0.7595	0.6974	0.7229
2	0.90	0.93	0.93	0.93	0.7721	0.6905	0.7251
3	0.86	0.95	0.86	0.90	0.7682	0.6941	0.7252
Average	0.87	0.93	0.89	0.91	0.7666	0.6940	0.7244

Table 10. Test results with similarity search

Experiment	Accuracy	Precision	Recall	F1-Score	Bert Precision	Bert Recall	Bert F1-S
1	0.84	0.88	0.91	0.90	0.7172	0.6787	0.6937
2	0.83	0.87	0.90	0.89	0.7137	0.6836	0.6941
3	0.83	0.88	0.90	0.89	0.7156	0.6883	0.6981
Average	0.83	0.87	0.90	0.89	0.7155	0.6835	0.6835

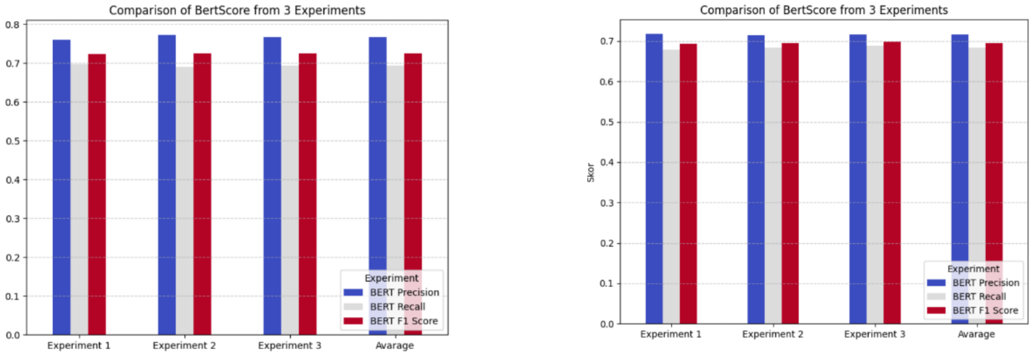
A comparative visualization of the evaluation results of the three experiments is shown in Figure 5.



(a) (b)

Figure 5. (a) Comparison of evaluation metrics from 3 experiments with hybrid search method **(b)** Comparison of evaluation metrics from 3 experiments with the similarity search method

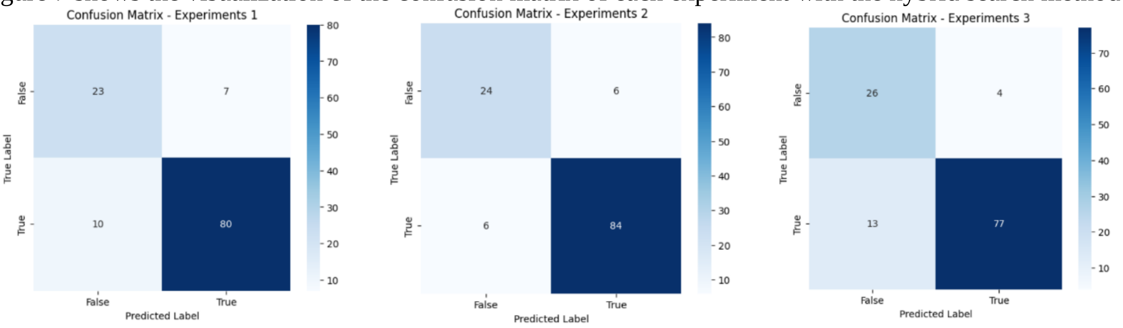
In addition, an evaluation was also conducted using BERTScore to measure the semantic similarity between the chatbot answers and the information in the dataset. The BERTScore results are visualized in Figure 6.



(a) (b)

Figure 6. (a) BERTScore comparison of 3 trials with hybrid search method **(b)** BERTScore comparison of 3 trials with similarity search method

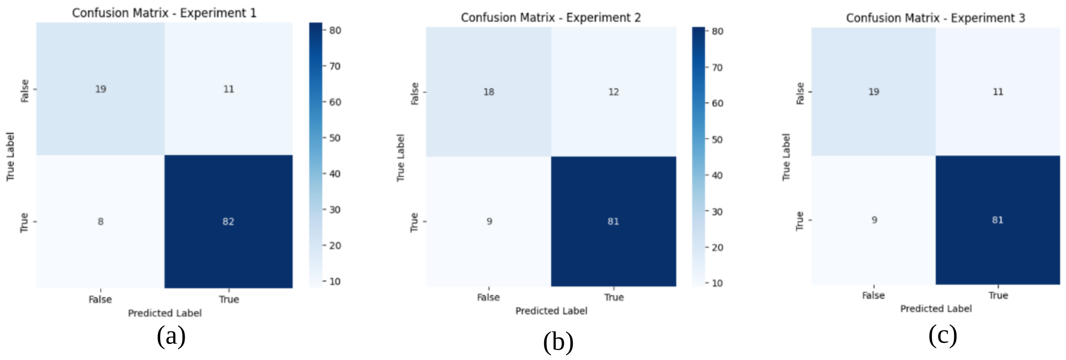
Figure 7 shows the visualization of the confusion matrix of each experiment with the hybrid search method.



(a) (b) (c)

Figure 7. Confusion matrix with hybrid search method **(a)** experiment 1 **(b)** experiment 2 **(c)** experiment 3

In addition, the confusion matrix visualization for the similarity search method is shown in Figure 8.



(a) (b) (c)

Figure 8. Confusion matrix with similarity search method**(a)** experiment 1 **(b)** experiment 2 **(c)** experiment 3

The evaluation results show that the hybrid search method consistently provides higher performance than similarity search, both in terms of accuracy, precision, recall, F1-score metrics, and semantic similarity according to BERTScore.

3.7 Comparison with LSTM Model

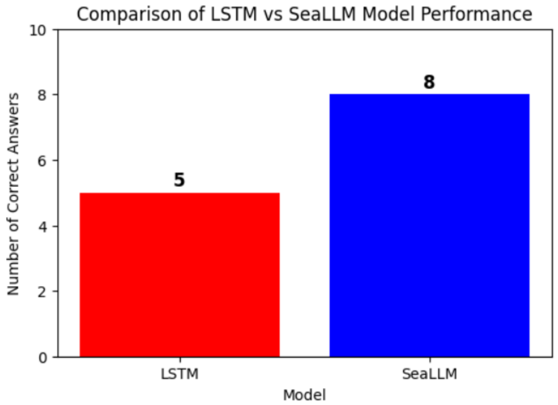
In this research, response comparison is required to evaluate the quality of answers provided by the previous model using LSTM and the current model using LLM SeaLLM with RAG using hybrid search method. The comparison of response results can be seen in Table 11.

Table 11. Comparison of response results of LSTM and LLM SeaLLM

Question	LSTM Model Response	LLM SEALLM Model Response	Response Comparison
What is menstrual pain?	Menstrual pain, also known as dysmenorrhea, is...	Menstruation is a natural process that occurs in...	Both models gave fairly good answers.
What causes menstrual distress or pain?	Some complications of menstrual disorders can affect a woman's ability to conceive...	Untreated menstrual disorders can lead to a number of complications, one of which is painful...	The LSTM model does not answer the cause of pain while SeaLLM provides a more precise answer
When does monoupouse usually occur in women?	Cutting your nails during your period shouldn't have any negative impact on your body...	Sorry, I have no understanding of that.	The LSTM model provides irrelevant answers, while SeaLLM fails to provide valid responses.
iud side effects	After IUD insertion, some women may experience mild side effects...	After IUD insertion, some women may experience side effects...	Both models gave fairly good answers.
I use a mentrual cup during my period, how do I clean it?	Eating pineapple during menstruation generally has no particular effect on the body. Pineapple is a fruit that can be eaten all the time, including during menstruation...	To clean the menstrual cup after use, wash your hands first, then take out the menstrual cup, pour running water into it, leave it until the water becomes clear, then drain it and put it back into the vagina....	The LSTM model provides irrelevant answers, while SeaLLM provides valid responses according to the dataset.
Is exercise safe to do during menstruation?	Exercise is safe to do during menstruation...	Yes, it is safe to exercise during menstruation...	Both models gave fairly good answers.
Which is better to use during menstruation, tampons, menstrual cups, or pads?	Clotted menstrual blood is usually not a sign of a serious health problem...	Menstrual cups are more environmentally friendly and economical in the long run compared to tampons and pads....	The LSTM model provides irrelevant answers, while SeaLLM provides valid responses according to the dataset.

What is the difference between menstrual blood and implantation blood?	Normal menstrual blood is usually liquid and bright red...	Menstrual blood and implantation blood have some similarities...	Both models gave fairly good answers.
Can stress during menstruation affect menstruation?	Soft drinks or carbonated drinks have no direct relationship with menstruation...	There is not enough information in context to give the right answer...	The LSTM model provides irrelevant answers, while SeaLLM fails to provide valid responses.
How long do iud side effects last?	Most side effects are mild after IUD insertion...	After IUD insertion, side effects such as mild abdominal pain...	Both models gave fairly good answers.

Based on Table 6, the SeaLLM model proved superior to LSTM in providing relevant and contextually appropriate responses. SeaLLM could answer 8 out of 10 questions, while LSTM only 5 (Figure 9). SeaLLM is more recommended for a women's health consultation chatbot as it provides more accurate and helpful answers.



.

Figure 9. Comparison of LSTM and SeaLLM test results

3.8 Testing

The testing was conducted with 9 participants from diverse backgrounds, including students, employees, and housewives. These testers had no prior knowledge of the internal logic of the chatbot system. Each participant was asked to submit 5 random questions covering a variety of topics related to women's health. The results showed that 22 out of 45 questions were not found in the dataset, suggesting that the dataset used is not yet comprehensive enough to cover the full scope of women's health topics.

3.9 Discussion

The previous LSTM-based Feminacare chatbot only achieved 61% accuracy with a loss value of 1.65 and often gave irrelevant responses. The performance improved after being optimized using the RAG approach with the SeaLLM model. Evaluation of 120 questions showed that the hybrid search method combining BM25 and vector similarity achieved 87% accuracy, 93% precision, 89% recall, and 91% F1-score, while similarity search only achieved 83% accuracy, 87% precision, 90% recall, and 80% F1-score. This shows that a hybrid search is more effective in generating relevant responses than a similarity search. Assessment using BERTScore also shows good semantic quality of answers, although there is room for improvement in handling complex questions.

Additional testing of 10 questions showed that SeaLLM could answer 8 questions correctly, while LSTM only 5, indicating SeaLLM's superiority in context and accuracy. Overall, SeaLLM with RAG improves the reliability of chatbots in women's health consultations, although there is still a need to expand the dataset to include questions beyond the current scope.

4. CONCLUSION

Based on this research, applying the SeaLLM model with the RAG method has successfully optimized the Feminacare chatbot in providing women's health information in a more relevant and contextual manner than the LSTM model. Evaluation of 120 questions with the hybrid search retrieval method showed good performance with an average accuracy of 87%, precision of 93%, recall of 89%, and F1-score of 91%, as well as BERTScore F1 of 0.7244 which reflects the high semantic quality of the responses.

REFERENCES

[1] M. Muliyono and S. Sumijan, "Identifikasi Chatbot dalam Meningkatkan Pelayanan Online Menggunakan Metode Natural Language Processing," *J. Inform. Ekon. Bisnis*, vol. 3, pp. 142–147, 2021, doi: 10.37034/infeb.v3i4.102.

[2] L. Xu, L. Lu, M. Liu, C. Song, and L. Wu, "Nanjing Yunjin intelligent question-answering system based on knowledge graphs and retrieval augmented generation technology," *Herit. Sci.*, vol. 12, no. 1, pp. 1–23, 2024, doi: 10.1186/s40494-024-01231-3.

[3] L. Athota, V. K. Shukla, N. Pandey, and A. Rana, "Chatbot for Healthcare System Using Artificial Intelligence," *ICRITO 2020 - IEEE 8th Int. Conf. Reliab. Infocom Technol. Optim. (Trends Futur. Dir.*, pp. 619–622, 2020, doi: 10.1109/ICRITO48877.2020.9197833.

[4] J. N. K. Wah, "Revolutionizing e-health: the transformative role of AI-powered hybrid chatbots in healthcare solutions," *Front. Public Heal.*, vol. 13, no. February, pp. 1–14, 2025, doi: 10.3389/fpubh.2025.1530799.

[5] M. Laymouna, Y. Ma, D. Lessard, T. Schuster, K. Engler, and B. Lebouché, "Roles, Users, Benefits, and Limitations of Chatbots in Health Care: Rapid Review," *J. Med. Internet Res.*, vol. 26, pp. 1–28, 2024, doi: 10.2196/56930.

[6] L. Li, "Studies advanced in chatbots based on deep learning," *Appl. Comput. Eng.*, vol. 6, no. 1, pp. 678–683, 2023, doi: 10.54254/2755-2721/6/20230921.

[7] S. Pandey and S. Sharma, "A comparative study of retrieval-based and generative-based chatbots using Deep Learning and Machine Learning," *Healthc. Anal.*, vol. 3, no. May, p. 100198, 2023, doi: 10.1016/j.health.2023.100198.

[8] Y. Chang *et al.*, "A Survey on Evaluation of Large Language Models," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, 2024, doi: 10.1145/3641289.

[9] Y. Ke *et al.*, "Development and Testing of Retrieval Augmented Generation in Large Language Models -- A Case Study Report," 2024, [Online]. Available: http://arxiv.org/abs/2402.01733

[10] F. Koto, R. Mahendra, N. Aisyah, and T. Baldwin, "IndoCulture: Exploring Geographically-Influenced Cultural Commonsense Reasoning Across Eleven Indonesian Provinces," 2024, [Online]. Available: http://arxiv.org/abs/2404.01854

[11] F. Koto, "Cracking the Code: Multi-domain LLM Evaluation on Real-World Professional Exams in Indonesia," 2024, [Online]. Available: http://arxiv.org/abs/2409.08564

[12] A. Balaguer *et al.*, "RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture," 2024, [Online]. Available: http://arxiv.org/abs/2401.08406

[13] R. Lakatos, P. Pollner, A. Hajdu, and T. Joo, "Investigating the performance of Retrieval-Augmented Generation and fine-tuning for the development of AI-driven knowledge-based systems," pp. 1–17, 2024, [Online]. Available: http://arxiv.org/abs/2403.09727

[14] M. Fatehkia, J. K. Lucas, and S. Chawla, "T-RAG: Lessons from the LLM Trenches," pp. 1–22, 2024, [Online]. Available: http://arxiv.org/abs/2402.07483

[15] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval Augmentation Reduces Hallucination in Conversation," *Find. Assoc. Comput. Linguist. Find. ACL EMNLP 2021*, pp. 3784–3803, 2021, doi: 10.18653/v1/2021.findings-emnlp.320.

[16] M. Șorecău and E. Șorecău, "An Alternative Application to CHATGPT that Uses Reliable Sources to Enhance the Learning Process," *Int. Conf. KNOWLEDGE-BASED Organ.*, vol. 29, no. 3, pp. 113–119, 2023, doi: 10.2478/kbo-2023-0084.

[17] G. F. Febrian and G. Figueredo, "KemenkeuGPT: Leveraging a Large Language Model on Indonesia's Government Financial Data and Regulations to Enhance Decision Making," 2024, [Online]. Available: http://arxiv.org/abs/2407.21459

[18] P. Zhao *et al.*, "Retrieval-Augmented Generation for AI-Generated Content: A Survey," no. March, 2024, [Online]. Available: http://arxiv.org/abs/2402.19473

[19] D. A. Ramadhan, D. Ramadhani, and U. Riau, "Classification of Riau Batik Motifs Using the Convolutional Neural Network (CNN) Algorithm," vol. 07, no. 03, pp. 201–211, 2024.