

## Integrating Retrieval-Augmented Generation with Large Language Model Mistral 7b for Indonesian Medical Herb

Diash Firdaus <sup>(1)\*</sup>, Idi Sumardi <sup>(2)</sup>, Yuni Kulsum <sup>(3)</sup>

<sup>1</sup> Informatika, Fakultas Teknologi Industri, Institut Teknologi Nasional, Bandung, Indonesia

<sup>2</sup> Teknik Informatika, STMIK Jawa Barat, Bandung, Indonesia

<sup>3</sup> Laboratorium Terpadu, UIN Sunan Gunung Djati, Bandung, Indonesia

e-mail : diash@itenas.ac.id, idis@stmikjabar.ac.id, yunikulsum05@gmail.com.

\* Penulis korespondensi.

Artikel ini diajukan 11 Mei 2024, direvisi 28 Agustus 2024, diterima 5 September 2024, dan dipublikasikan 25 September 2024.

### Abstract

Large Language Models (LLMs) are advanced artificial intelligence systems that use deep learning, particularly transformer architectures, to process and generate text. One such model, Mistral 7b, featuring 7 billion parameters, is optimized for high performance and efficiency in natural language processing tasks. It outperforms similar models, such as LLaMa2 7b and LLaMa 1, across various benchmarks, especially in reasoning, mathematics, and coding. LLMs have also demonstrated significant advancements in addressing medical queries. This research leverages Indonesia's rich biodiversity, which includes approximately 9,600 medicinal plant species out of the 30,000 known species. The study is motivated by the observation that LLMs, like ChatGPT and Gemini, often rely on internet data of uncertain validity and frequently provide generic answers without mentioning specific herbal plants found in Indonesia. To address this, the dataset for pre-training the model is derived from academic journals focusing on Indonesian medicinal herbal plants. The research process involves collecting these journals, preprocessing them using Langchain, embedding models with sentence transformers, and employing Faiss CPU for efficient searching and similarity matching. Subsequently, the Retrieval-Augmented Generation (RAG) process is applied to Mistral 7b, allowing it to provide accurate, dataset-driven responses to user queries. The model's performance is evaluated using both human evaluation and ROUGE metrics, which assess recall, precision, F1 measure, and METEOR scores. The results show that the RAG Mistral 7b model achieved a METEOR score of 0.22%, outperforming the LLaMa2 7b model, which scored 0.14%.

**Keywords:** LLM, Generative AI, LLAMA2, Retrieval-Augmented Generation, Deep Learning

### Abstrak

Large Language Models (LLM) adalah sistem kecerdasan buatan canggih yang menggunakan pembelajaran mendalam, khususnya arsitektur transformator, untuk memproses dan menghasilkan teks. Salah satu model tersebut, Mistral 7b, yang memiliki 7 miliar parameter, dioptimalkan untuk kinerja tinggi dan efisiensi dalam tugas pemrosesan bahasa alami. Model ini mengungguli model serupa, seperti LLaMa2 7b dan LLaMa 1, di berbagai tolok ukur, terutama dalam penalaran, matematika, dan pengkodean. LLM juga telah menunjukkan kemajuan yang signifikan dalam menjawab pertanyaan-pertanyaan medis. Penelitian ini memanfaatkan keanekaragaman hayati Indonesia yang kaya, yang mencakup sekitar 9.600 spesies tanaman obat dari 30.000 spesies yang diketahui. Penelitian ini dilatarbelakangi oleh pengamatan bahwa LLM, seperti ChatGPT dan Gemini, sering kali mengandalkan data internet yang validitasnya tidak pasti dan sering kali memberikan jawaban umum tanpa menyebutkan tanaman herbal tertentu yang ditemukan di Indonesia. Untuk mengatasi hal ini, *dataset* untuk pra-pelatihan model ini berasal dari jurnal-jurnal akademis yang berfokus pada tanaman herbal obat Indonesia. Proses penelitian melibatkan pengumpulan jurnal-jurnal ini, *preprocessing* menggunakan Langchain, menanamkan model dengan *transformer* kalimat, dan menggunakan CPU Faiss untuk pencarian dan pencocokan kemiripan yang efisien. Selanjutnya, proses Retrieval-Augmented Generation (RAG) diterapkan pada Mistral 7b, yang memungkinkannya untuk memberikan respons yang akurat dan berbasis *dataset* terhadap pertanyaan pengguna. Kinerja model dievaluasi dengan menggunakan evaluasi manusia dan metrik ROUGE, yang menilai



*recall*, presisi, F1 *measure*, dan skor METEOR. Hasilnya menunjukkan bahwa model RAG Mistral 7b mencapai skor METEOR 0,22%, mengungguli model LLaMa2 7b, yang mendapat skor 0,14%.

**Kata Kunci:** LLM, AI Generatif, LLAMA2, Retrieval-Augmented Generation, Deep Learning

## 1. INTRODUCTION

The domain of Natural Language Processing (NLP) and Artificial Intelligence (AI) has witnessed a remarkable breakthrough with the introduction of Large Language Models (LLMs). This innovative development has dramatically amplified the capacity of machines to comprehend and produce human-like language, thereby revolutionizing the field of language processing and AI research (Hadi et al., 2024; Jain et al., 2023; Kaddour et al., 2023). LLMs have consistently demonstrated outstanding performance in a wide range of tasks. However, their exceptional abilities present notable challenges due to their large-scale and intensive computational demands (Zhu et al., 2023).

Some frequently used Large Language Model (LLM) models are LLaMa (Touvron, Lavril, et al., 2023), GPT-4 (OpenAI et al., 2023), LLaMa2 (Touvron, Martin, et al., 2023), PANGU- $\Sigma$  (Ren et al., 2023) and so on. Considering the remarkable achievements of pre-training methods in general LLM, researchers have begun to explore the potential application of these techniques in the biomedical and health domain. However, it has been observed that simply applying these models directly to the biomedical domain does not yield satisfactory results. The underwhelming performance can be attributed to the significant differences and distinct characteristics between the general and specialized biomedical domains, known as domain shift (K. Zhang et al., 2024). Several studies have used the LLM model to be implemented in the field of health and medicine, such as Med-PaLM (Singhal, Azizi, et al., 2023), BioGPT (Luo et al., 2022), Med-PaLM 2 (Singhal, Tu, et al., 2023), Flan-PaLM (Chung et al., 2022), and so on.

The utilization of medicinal plants is of utmost importance in acting as the primary healthcare system, particularly in underserved regions of developing nations. These areas often rely solely on herbal remedies as the sole medication accessible to them (Fathir et al., 2021). Indonesia is known to be a country with biodiversity resources spread from Sabang to Merauke Region, where there are around 30,000 plant species, of which 9600 are medicinal plants. The utilization of herbal medicine, commonly referred to as "jamu" within the Indonesian culture, holds significant importance in healthcare for a staggering majority of the population, approximately 80%. Derived from the Javanese tribal language, the term "jamu" embodies the concept of traditional herbal treatments. Essentially, jamu encompasses the utilization of plant-based substances that are intricately crafted to serve medicinal purposes. Over time, the term "jamu" has seamlessly integrated itself into the Indonesian language, bearing a resemblance to its original definition (Elfahmi et al., 2014; Sianipar, 2021). The application of NLP in the field of Herbal Medicine has been carried out by several previous researchers, such as (T. Zhang et al., 2022). However, it is mentioned that in developing NLP herbal medicine, a collaboration between many experts, such as biologists and computer experts, is required. The next issue is that the currently available models, such as ChatGPT, Google Gemini, Claude, and others, provide answers based on internet data whose validity is uncertain (Ray, 2023). Often, the responses include herbs that are not available in Indonesia.

Mistral 7b is a language model equipped with 7 billion parameters, engineered for superior performance and efficiency in tasks associated with natural language processing. It surpasses larger models such as LLaMa2 7b, which has 13 billion parameters, and LLaMa 1, with 34 billion parameters, across multiple benchmarks, particularly excelling in reasoning, mathematics, and code generation tasks. (Jiang et al., 2023). RAG, or Retrieval-Augmented Generation, is a technique used in natural language processing that combines a retrieval system with a generative model. This approach enhances the generative model's ability to produce relevant and accurate responses by first retrieving relevant information from a large dataset or knowledge base before generating an answer. In the RAG setup, when a question or input is presented to the system,



the retrieval component first searches through a database to find relevant documents or pieces of text. These retrieved texts are then provided as additional context to the generative model, which uses this context to generate a more informed and precise answer. This method helps in improving the quality of the output by grounding the responses in factual content retrieved from the dataset (Radeva et al., 2024).

In this study, we propose a specific Retrieval-Augmented Generation for Herbal Medicine text generation and mining. Herbal Medicine LLM follows the backbone of the Sentence-Transformer language model and Mistral 7b for the Large Language Model. For QA, we used nine international journals on herbal medicine in Indonesia. At the time of evaluation, we used the results issued by RAG Mistral, which were compared with the answers from the experts.

## 2. RESEARCH METHOD

Our research introduces Herbal Medicine using RAG Mistral 7b, an innovative and specialized language model specifically tailored for various Herbal medical applications. With a unique approach, we have successfully trained this model using a diverse and extensive collection of Journal Medical Herb in Indonesia. This abundant dataset enables the infusion of domain-specific knowledge into the model, enhancing its efficacy. Moreover, we have meticulously devised a comprehensive evaluation framework that encompasses question-answering.

### 2.1 Research Stage

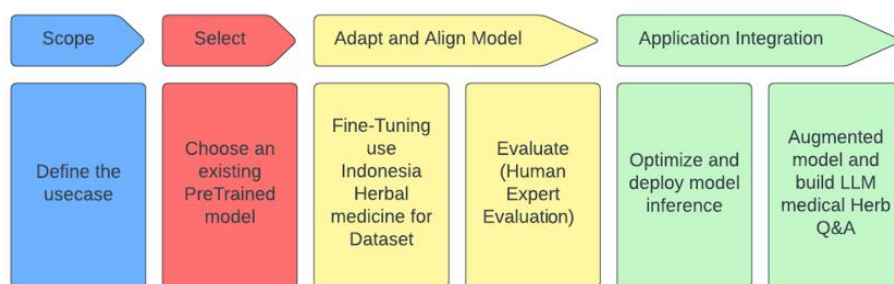


Figure 1 Research Stage of Medical Herb

The research stages of medical herbs using Mistral 7b are shown in Figure 1. In the first stage, we identified the problem, followed by the process of selecting an LLM model. In this study, we chose the Mistral 7b model because previous research has shown that the RAG on the LLM Mistral 7b model provides better results in terms of evaluation time, loading duration, prompt evaluation count, total duration, and tokens per second compared to the LLM LLaMa2 7b and Orca2 7b models. The LLM Mistral 7b RAG model also delivered the best results for METEOR, ROUGE 1 Recall, ROUGE Precision, and BLUE Score when compared to LLaMa2 7b and Orca2 7b (Radeva et al., 2024).

The process depicted in the diagram involves several stages to develop a language model tailored specifically for medical herb Q&A using Indonesian herbal medicine data. First, define the use case clearly to establish the specific application or problem the model will address, focusing on the context of Indonesian herbs. Next, select an appropriate pre-trained model as a foundation for further customization. Then, the model can be adapted by training it with a specialized dataset containing information about Indonesian herbs, ensuring the model learns domain-relevant knowledge. Evaluate the fine-tuned model with the help of human experts, such as ROUGE and METEOR, to assess its performance. Afterward, optimize and deploy the model for real-world application by ensuring efficient inference and accessibility. Finally, the enhanced model will be integrated into an application framework designed for Q&A, focusing on medical herb-related queries. This process ensures that the language model is not only accurate and relevant but also effectively integrated into practical applications. When finished, each answer to the question



posed will be evaluated by an expert. On this occasion, our expert is assisted by an expert in the field of ethnobotany from Sunan Gunung Djati University, Bandung, namely Dr. Tri Cahyanto, M.Si. After the evaluation process is complete, the model will be deployed with the final goal of a web-based application.

## 2.2 RAG LLM Herb Medicine Architecture

Figure 2 is the architecture of LLM Herbal Medicine using Mistral 7b, where, at the initial stage, we will extract data and then preprocess, including quality filtering, de-duplication, and so on. The next process is to divide the data into chunks, where each chunk will be embedded, and a semantics index will be built. Then, the data will be stored in the DB vector data in the form of Knowledge-based Data. In the last stage, the user will give questions to the Knowledge Database, and the Knowledge Database will reply to user questions based on the answers ranking.

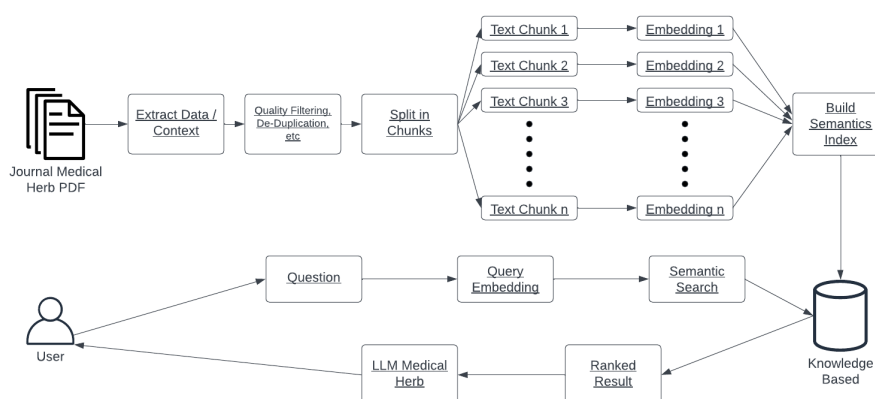


Figure 2 Architecture of RAG LLM Mistral 7b Herbal Medicine

In the dataset section, we utilized nine journal sources focused on herbal medicine in Indonesia, which are detailed in Table 1. These journals were specifically selected for their relevance and suitability concerning various herbal plants found in the region. This selection provides a comprehensive foundation for our research on the efficacy of these plants.

Table 1 Journal Medical Herb Dataset

Journal	Title
(Sianipar, 2021)	The Potential of Indonesian Traditional Herbal Medicine as Immunomodulatory Agents: A Review
(Putri et al., 1970)	Ethnobotanical study of herbal medicine in Ranggawulung Urban Forest, Subang District, West Java, Indonesia
(Fathir et al., 2021)	Ethnobotanical study of medicinal plants used for maintaining stamina in Madura ethnic, East Java, Indonesia
(Sholikhah, 2016)	Indonesian medicinal plants as sources of secondary metabolites for pharmaceutical industry
(Ardiyanto et al., 2021)	The use of hyperuricemia herbs at “Hortus Medicus” herbal medicine clinic Tawangmangu
(Elfahmi et al., 2014)	Jamu: Indonesian traditional herbal medicine towards rational phytopharmacological use
(Arozal et al., 2020)	Selected Indonesian Medicinal Plants for the Management of Metabolic Syndrome: Molecular Basis and Recent Studies
(Kartini et al., 2019)	Standardization of Some Indonesian Medicinal Plants Used in “Scientific Jamu”
(Sumarni et al., 2019)	The scientification of jamu: a study of Indonesian’s traditional medicine



### 2.3 Environment Setup

In this step, the necessary environment for the project is established. This involves configuring tools, software, hardware, and other resources essential for conducting research or development activities. The details of the Environment Setup are provided in Table 2.

**Table 2 Environment Setup**

No.	Name	Version
1	Operating System	Windows 11
2	Programming	Python 11.2
3	Supporting Tools	Library Chainlit Library Huggingface Library Langchain
4	Hardware	CPU AMD Ryzen 7 5800 RAM 16 GB VGA RADEON RX 5500M
5	Model	Mistral 7b LLaMa 7b 500 Chunk 512 Tokens 0.5 Temperatur Mistral 7b

### 2.4 Metric Evaluation

For Metric Evaluation, we use several evaluation models such as METEOR (Metric for Evaluation of Translation with Explicit Ordering), ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Radeva et al., 2024), and human evaluation (Wang et al., 2023)

#### 2.4.1 METEOR (Metric for Evaluation of Translation with Explicit Ordering)

The METEOR score is a metric used to evaluate the quality of machine-generated text by comparing it to one or more reference texts. The calculation involves several steps:

- 1) Word Alignment: Align words between candidate and reference translations based on exact matches, stems, synonyms, and paraphrases, ensuring each word in the candidate and reference sentences is used only once to maximize the overall match.
- 2) Precision and Recall Calculation:
  - a) Precision measures the proportion of matched words in the candidate compared to the total number of words in the candidate.
  - b) Recall measures the proportion of matched words in the candidate against the total number of words in the reference.

$$P = \frac{m}{w_c}, R = \frac{m}{w_r} \quad (1)$$

The formulas for precision and recall are presented in Equation (1). In this context,  $m$  represents the number of unigrams in the candidate translation that match those in the reference,  $w_c$  denotes the total number of unigrams in the candidate translation, and  $w_r$  indicates the total number of unigrams in the reference translation(s).

- 3) The penalty accounts for chunkiness, which refers to the arrangement and fluency of matched chunks, represented by the formula in Equation (2), where  $c$  defines the number of contiguous matched unigrams.

$$Penalty = 0.5 \left(\frac{c}{m}\right)^3 \quad (2)$$





- 4) The final METEOR score is computed using the harmonic mean of precision and recall, adjusted by the penalty factor. The formula is shown in Equation (3).

$$M_{Score} = F_{Mean} \times (1 - Penalty) \quad (3)$$

Where

$$F_{Mean} = \frac{10PR}{R + 9P}$$

The calculations for Equations (1) to (3) are implemented using the NLTK library's *single\_meteor\_score* function, specifically at line 58 in the Python script. This pseudocode outlines the process of dividing two texts into individual words and computing the METEOR score for them. In the context of RAG models, the METEOR score is used to assess the quality of generated responses. A high METEOR score signifies that the generated response closely aligns with the reference text, indicating the model's effectiveness in accurately retrieving and generating responses. On the other hand, a low METEOR score may highlight areas where the model's performance could be enhanced.

#### 2.4.2 ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE is a collection of metrics used to assess automatic summarization and machine translation. It evaluates by comparing a machine-generated summary or translation with one or more reference summaries (typically human-generated). ROUGE includes different variants, including ROUGE-N, ROUGE-L, and ROUGE-W.

ROUGE-N measures the overlap of n-grams (sequences of n words) between the system-generated summary and the reference summaries. It is calculated using recall, precision, and F1 score:

- 1) **Recall for ROUGE-N** is the ratio of overlapping n-grams between the system summary and the reference summaries to the total n-grams in the reference summaries. The formula is presented in Equation (4).

$$Recall_{ROUGE-N} = \frac{\sum_{S \in \{Reference\ Summaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{Reference\ Summaries\}} \sum_{gram_n \in S} Count(gram_n)} \quad (4)$$

- 2) **Precision for ROUGE-N** is the ratio of overlapping n-grams in the system summary to the total n-grams in the system summary itself. The formula is presented in Equation (5).

$$Precision_{ROUGE-N} = \frac{\sum_{S \in \{System\ Summaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{System\ Summaries\}} \sum_{gram_n \in S} Count(gram_n)} \quad (5)$$

- 3) **F1 Score for ROUGE-N** is the harmonic mean of precision and recall. The formula is presented in Equation (6).

$$F1_{ROUGE-N} = 2 \frac{Precision_{ROUGE-N} \times Recall_{ROUGE-N}}{Precision_{ROUGE-N} + Recall_{ROUGE-N}} \quad (6)$$

ROUGE-L emphasizes the longest common subsequence (LCS) between the generated summary and reference summaries. The LCS is the longest sequence of words that appear in both texts in the same order, but not necessarily consecutively. The parameters for ROUGE-L include:

- 1) **Recall for ROUGE-L** is calculated by dividing the length of the LCS by the total number of words in the reference summary. This evaluates how well the generated summary captures the content of the reference summaries. The formula is presented in Equation (7).



$$Recall_{ROUGE-L} = \frac{LCS(\text{System Summary}, \text{Reference Summary})}{\text{Length of Reference Summary}} \quad (7)$$

- 2) **Precision for ROUGE-L** is determined by dividing the length of the LCS by the total number of words in the generated summary. This assesses how much of the generated summary's content appears in the reference summaries. The formula is presented in Equation (8).

$$Precision_{ROUGE-L} = \frac{LCS(\text{System Summary}, \text{Reference Summary})}{\text{Length of System Summary}} \quad (8)$$

- 3) **F1 for ROUGE-L** is the harmonic mean of LCS-based precision and recall. The formula is presented in Equation (9).

$$F1_{ROUGE-L} = 2 \frac{Precision_{ROUGE-L} \times Recall_{ROUGE-L}}{Precision_{ROUGE-L} + Recall_{ROUGE-L}} \quad (6)$$

ROUGE-W extends ROUGE-L by assigning more weight to longer sequences of matching words. However, in this context, ROUGE-W is not utilized.

The calculations for Equations (4) to (9) are implemented using the rouge library, specifically the `rouge.get_scores` function, located on line 65 of the Python script. The following pseudocode outlines the steps to initialize a ROUGE object and compute ROUGE scores between two texts:

- 1) Assign 'hypotheses' to the reference text and 'ref' to the candidate text.
- 2) Create a ROUGE object.
- 3) Use the `get_scores` method of the ROUGE object to compute ROUGE scores between 'hypotheses' and 'ref.'

### 3. RESULT AND DISCUSSION

In this section, we present the experimental design and implementation details of LLM Medical Herb, highlighting its superior performance evaluated by Dr. Tri Cahyanto, M.Si., an expert in ethnobotany. We intentionally selected datasets from reputable journals to demonstrate the promising generalization and validity of our approach. This thorough evaluation underscores the effectiveness of our method in assessing herbal medicine applications.

#### 3.1 Implementation Details

At the time of implementation, the chunk size is 500 with a maximum number of tokens of 512, along with a temperature of 0.5. The training model used is Mistral 7b, where this model has 7 billion parameters. In the application implementation stage, we use a web-based application that utilizes Chainlit. Chainlit is an open-source package that has tremendous speed when it comes to developing and distributing LLM applications. It provides a seamless experience by allowing users to incorporate the Chainlit API into their existing code. We can create a user interface similar to ChatGPT.

The testing process will be conducted under the guidance of expert Dr. Tri Cahyanto, M.Si. Tables 3 to 5 present six questions posed to the Herb Medical Bot, along with the corresponding answers provided by experts. This setup facilitates a comprehensive evaluation of the bot's performance and effectiveness in delivering accurate information.

Table 6 presents the results of the evaluation matrix, comparing the responses generated by the RAG model with those provided by the experts. This comparison highlights the effectiveness of the RAG model in generating accurate and relevant answers in relation to expert opinions, providing insights into its performance and areas for improvement.



Figure 3 presents a question posed to the Mistral 7b model, asking, "What's an herb for diabetes?" The model responds as shown in the Q2 Mistral 7b Answer. Figure 4 displays the results of a question directed to the RAG Medical Chatbot utilizing the LLaMa2 7b model. In this case, the question asked was, "What's an herb for diabetes?" The model responded that *Guazuma ulmifolia* leaves and *Trigonella foenum-graceum* seeds are known for their potential to manage diabetes mellitus through their anti-diabetic effects.

**Table 3 Human Evaluation Question and Answer**

No.	Question Prompt	Answer		
		LLaMa2 7b	Mistral 7b	Expert
Q1	Herb for headache	Cinnamon is an herb that can be used to treat headaches, including migraines. It has been shown to have anti-inflammatory properties and can help to reduce inflammation in the body, which may contribute to headache pain.	Herbs are a plant or part of a plant that is used to make medicine. They can also be eaten as food, such as cinnamon. Some herbs have been used medicinally for thousands of years. The properties of some herbs may be beneficial in relieving pain and muscle tension, but there is limited clinical evidence about how effective they are at treating headaches. However, it is thought that these herbs can help to reduce the symptoms of a headache if you use them as part of a regular treatment plan alongside other lifestyle changes. There are many different types of herbal remedies available on the market today, but only a few have been tested scientifically for effectiveness against headache relief. Some may contain harmful ingredients, so it's important to check with your doctor before taking any natural supplement or medication. A common home remedy for headaches is peppermint oil, which is applied topically around the temple area where most people feel their pain when they experience this condition. This method works well because peppermint contains menthol, which has been shown in studies to have anti-inflammatory properties that help reduce swelling and inflammation associated with headaches caused by stress or allergies.	Cinnamon is a spice that has anti-inflammatory and neuroprotective properties. Researchers were therefore interested in studying whether cinnamon could help reduce migraine attacks and inflammation. For example, this journal describes about that (Zareie et al., 2020).
Q2	Herb For Diabetes	<i>Guazuma ulmifolia</i> leaves and <i>Trigonella foenum-graceum</i> seeds are known to have the potential to manage diabetes mellitus through their anti-diabetic effects.	The most popular herbs in the treatment of diabetes are bitter melon, fenugreek, bilberry, cinnamon, ginger, chromium, bitter melon, gymnema, zinc, psyllium husk, aloe vera, magnesium, ginseng, and garlic. Bitter melon ( <i>Momordica charantia</i> ), which is also known as bitter melon or bitter melon, has been used in Chinese and Ayurvedic medicine for centuries to treat diabetes mellitus (12). In recent years, laboratory studies have demonstrated that the plant extracts lower blood sugar levels. The anti-diabetic effects of bitter melon are related to its ability to increase insulin production by pancreatic beta cells and reduce glucose absorption in the intestine. The hypoglycemic effect of bitter melon is also associated with increased insulin sensitivity at the cellular level, which results in more efficient uptake of glucose by muscle and fat tissues (13).	<i>Trigonella foenum-graecum</i> is one of the important medicinal plants in the management of diabetes mellitus. Several studies, such as (Geberemeskel et al., 2019), have investigated the effect of <i>Trigonella foenum-graecum</i> seed powder on the lipid profile of newly diagnosed type II diabetic patients.







Table 5 Human Evaluation Question and Answer (Continued)

No.	Question Prompt	Answer		
		LLaMa2 7b	Mistral 7b	Expert
Q5	Medical herb for rheumatism	Cynamaldehyde from Cinnamomum verum can enhance the immune response in Rheumatoid Arthritis (RA) by inhibiting the proteins involved in RA progression and ins in RA.	<p>Curcumin is a chemical compound found in turmeric, which has been shown to have anti-inflammatory and antioxidant properties. It is thought that curcumin may help reduce pain and stiffness associated with rheumatism. However, more research is needed to confirm its effectiveness as a treatment for rheumatism.</p> <p>Several medical herbs have been traditionally used in the treatment of rheumatism. These include turmeric, ginger, and willow bark. Turmeric has anti-inflammatory and antioxidant properties, while ginger can help reduce pain and stiffness. Willow bark contains salicylic acid, which is a compound that is similar to aspirin and can also help reduce inflammation and pain.</p> <p>Curcumin has been shown to have anti-inflammatory and antioxidant properties. It may help reduce pain and stiffness associated with rheumatism. However, more research is needed to confirm its effectiveness as a treatment for rheumatism.</p>	Rheumatoid arthritis is one part of rheumatic disease. In Indonesia, some herbs that are often used for rheumatism are jambe, jackfruit, and several other examples, such as cinnamon, curcumin, African tree, and so on.
Q6	Medical herb for Hartburn	The herb commonly used to treat heartburn is ginger. Ginger has natural anti-inflammatory properties that can help reduce inflammation in the esophagus and alleviate symptoms of heartburn.	Medicinal plants have been used in Ayurveda, traditional Chinese medicine, the Unani system of medicine, folk medicine, and home remedies since ancient times. In India, about 80 percent of people still depend on herbal medicines for their primary healthcare needs. In the past two decades, research has been focused on identifying and testing the active ingredients in traditional medicinal plants. This research is aimed at developing safe and effective drug products based on natural compounds that are able to treat specific diseases. According to a 2008 report, approximately 75 percent of the population in rural areas of India rely solely on traditional medicines for their primary healthcare needs. Traditional medicine systems, including Ayurveda, Siddha, Unani, and naturopathy, were included in the Indian Constitution in 1970 under Schedule H of the Drugs and Cosmetic Act.	Some sources have been researched, such as Harvard Medical School, which states that ginger root is a popular herbal remedy for heartburn. It has been used for centuries to relieve the symptoms of heartburn, such as a burning sensation in the chest.

Table 6 Metric Evaluation

Question	Model	ROUGE Precision	ROUGE Recall	ROUGE F-Measure	METEOR
Q1	Mistral 7b	0.09	0.29	0.13	0.23
	LLaMa2 7b	0.31	0.17	0.22	0.13
Q2	Mistral 7b	0.05	0.33	0.09	0.26
	LLaMa2 7b	0.17	0.31	0.22	0.24
Q3	Mistral 7b	0.09	0.20	0.12	0.24
	LLaMa2 7b	0.18	0.06	0.09	0.06
Q4	Mistral 7b	0.09	0.10	0.10	0.16
	LLaMa2 7b	0.06	0.12	0.08	0.11
Q5	Mistral 7b	0.05	0.22	0.09	0.21
	LLaMa2 7b	0.16	0.11	0.13	0.08
Q6	Mistral 7b	0.07	0.25	0.11	0.22
	LLaMa2 7b	0.21	0.15	0.17	0.24
Average	Mistral 7b	0.07	0.23	0.10	0.22
	LLaMa2 7b	0.18	0.15	0.15	0.14



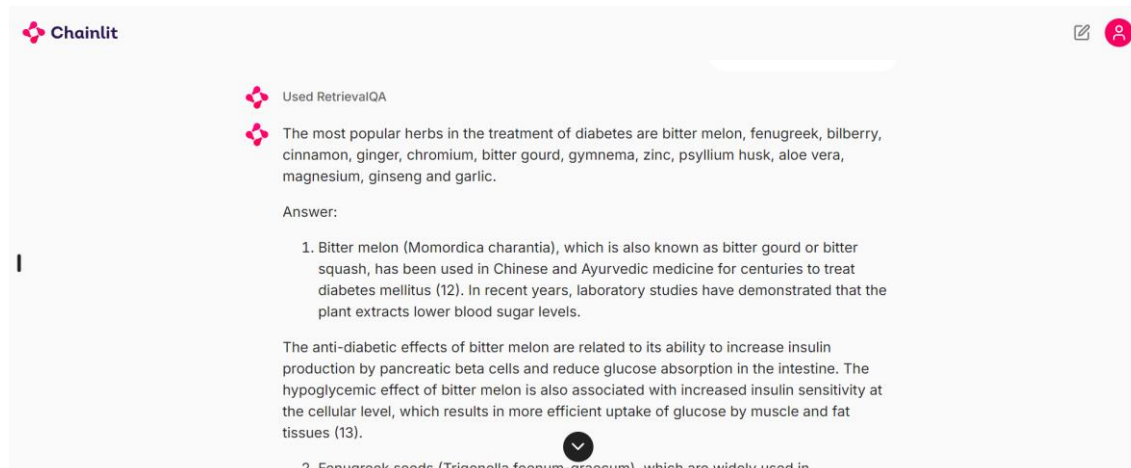


Figure 3 Question and Answer Medical Herb with Mistral 7b

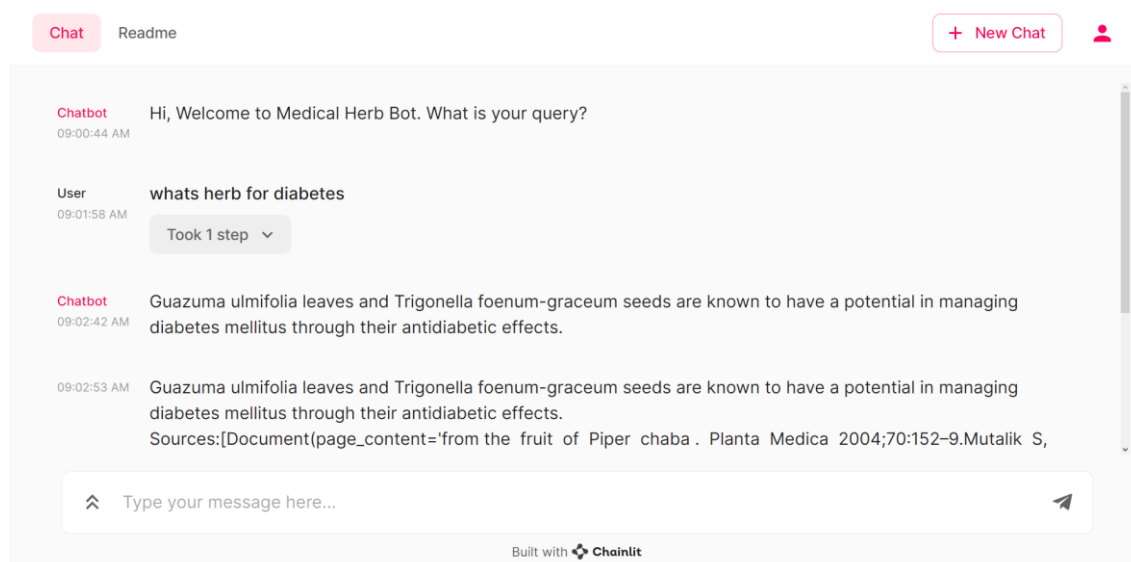


Figure 4 Question and Answer Medical Herb with LLaMa2 7b

### 3.2 Discuss

In the Metric Evaluation at Precision section, Mistral 7b has a low percentage because the output generated by Mistral 7b exhibits a high level of creativity, even with the same temperature setting of 0.5. However, when evaluated using METEOR, Mistral 7b shows a high percentage, making the Mistral model arguably better than the LLaMa2 7b model. However, if you prefer a model with higher precision in text, you might choose LLaMa2 7b because it has higher text precision.

## 4. CONCLUSION

Indonesia has a large and abundant number of plants, but their utilization is still derived from customs only. Several studies have been conducted to explore the potential of plants in Indonesia, and they have been stored in accredited journals, which will be used as datasets. With RAG, we can obtain answers where each response is valid based on journals and contains data on herbal plants from Indonesia. Based on the results of the table above, the use of RAG Mistral 7b as an LLM model for Question Answering Medical Herb can be stated quite well. When viewed from six questions, the answers from the medical herb LLM and the solutions from the experts are by the



average Score from METEOR is 0.22% where the score given is higher than LLaMa2 7b, which only received 0.14%.

Furthermore, the precision score for Mistral 7b is only 0.07%, compared to 0.18% for LLaMa2 7b. This is because the Mistral 7b model answers questions creatively, as seen in Table 1. The creative responses from Mistral 7b reduce the precision score compared to expert answers. Next, one of the answers from LLaMa2 7b, specifically for question Q4, contains repetitive content, indicating that the model did not perform RAG correctly and effectively. Meanwhile, the answer from Mistral 7b does not contain repetitive content.

For further research, tests should be conducted with several experts and several questions related to herbal medicines for several diseases. Furthermore, journal sources should be added to the dataset so that the answers from LLM Medical Herb are more valid and qualified. This approach would not only improve the model's reliability but also contribute to a deeper understanding of the therapeutic potential of various herbs.

## REFERENCES

- Ardiyanto, D., Triyono, A., Nisa, U., Fitriani, U., Astana, P. R., Novianto, F., & Zulkarnain, Z. (2021). The use of hyperuricemia herbs at "Hortus Medicus" herbal medicine clinic Tawangmangu. *Jurnal Kedokteran Dan Kesehatan Indonesia*. <https://doi.org/10.20885/JKKI.Vol12.Iss2.art9>
- Arozal, W., Louisa, M., & Soetikno, V. (2020). Selected Indonesian Medicinal Plants for the Management of Metabolic Syndrome: Molecular Basis and Recent Studies. *Frontiers in Cardiovascular Medicine*, 7. <https://doi.org/10.3389/fcvm.2020.00082>
- Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., ... Wei, J. (2022). *Scaling Instruction-Finetuned Language Models: Vol. 1754* (H. W. Chung, S. Longpre, B. Zoph, A. Castro-ros, A. Yu, & A. Dai, Eds.). <http://arxiv.org/abs/2210.11416>
- Elfahmi, Woerdenbag, H. J., & Kayser, O. (2014). Jamu: Indonesian traditional herbal medicine towards rational phytopharmacological use. *Journal of Herbal Medicine*, 4(2), 51–73. <https://doi.org/10.1016/j.hermed.2014.01.002>
- Fathir, A., HAIKAL, MOCH., & Wahyudi, D. (2021). Ethnobotanical study of medicinal plants used for maintaining stamina in Madura ethnic, East Java, Indonesia. *Biodiversitas Journal of Biological Diversity*, 22(1), 386–392. <https://doi.org/10.13057/biodiv/d220147>
- Geberemeskel, G. A., Debebe, Y. G., & Nguse, N. A. (2019). Antidiabetic Effect of Fenugreek Seed Powder Solution ( *Trigonella foenum-graecum L.* ) on Hyperlipidemia in Diabetic Patients. *Journal of Diabetes Research*, 2019, 1–8. <https://doi.org/10.1155/2019/8507453>
- Hadi, M. U., Al-tashi, Q., Qureshi, R., Shah, A., Muneer, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Hassan, S. Z., Shoman, M., Wu, J., Mirjalili, S., & Shah, M. (2024). Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects. *TechRxiv*, 1–47. <https://doi.org/10.36227/techrxiv.23589741.v2>
- Jain, N., Saifullah, K., Wen, Y., Kirchenbauer, J., Shu, M., Saha, A., Goldblum, M., Geiping, J., & Goldstein, T. (2023). *Bring Your Own Data! Self-Supervised Evaluation for Large Language Models*. <http://arxiv.org/abs/2306.13651>
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. Le, Lavril, T., Wang, T., Lacroix, T., & Sayed, W. El. (2023). *Mistral 7B: Vol. 7b. ? 179*. <http://arxiv.org/abs/2310.06825>
- Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). *Challenges and Applications of Large Language Models*. <http://arxiv.org/abs/2307.10169>
- Kartini, K., Jayani, N. I. E., Octaviyanti, N. D., Krisnawan, A. H., & Avanti, C. (2019). Standardization of Some Indonesian Medicinal Plants Used in "Scientific Jamu." *IOP Conference Series: Earth and Environmental Science*, 391(1), 012042. <https://doi.org/10.1088/1755-1315/391/1/012042>



- Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T.-Y. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6). <https://doi.org/10.1093/bib/bbac409>
- OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., ... Zoph, B. (2023). *GPT-4 Technical Report*. 4, 1–100. <http://arxiv.org/abs/2303.08774>
- Putri, L. S. E., Dasumiati, D., Kristiyanto, K., Mardiansyah, M., Malik, C., Leuvinadrie, L. P., & Mulyono, E. A. (1970). Ethnobotanical study of herbal medicine in Ranggawulung Urban Forest, Subang District, West Java, Indonesia. *Biodiversitas Journal of Biological Diversity*, 17(1), 172–176. <https://doi.org/10.13057/biodiv/d170125>
- Radeva, I., Popchev, I., Doukovska, L., & Dimitrova, M. (2024). Web Application for Retrieval-Augmented Generation: Implementation and Testing. *Electronics*, 13(7), 1361. <https://doi.org/10.3390/electronics13071361>
- Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3, 121–154. <https://doi.org/10.1016/j.iotcps.2023.04.003>
- Ren, X., Zhou, P., Meng, X., Huang, X., Wang, Y., Wang, W., Li, P., Zhang, X., Podolskiy, A., Arshinov, G., Bout, A., Piontkovskaya, I., Wei, J., Jiang, X., Su, T., Liu, Q., & Yao, J. (2023). *PanGu- $\Sigma$ : Towards Trillion Parameter Language Model with Sparse Heterogeneous Computing*. <http://arxiv.org/abs/2303.10845>
- Sholikhah, E. N. (2016). Indonesian medicinal plants as sources of secondary metabolites for pharmaceutical industry. *Journal of the Medical Sciences (Berkala Ilmu Kedokteran)*, 48(04), 226–239. <https://doi.org/10.19106/JMedSci004804201606>
- Sianipar, E. A. (2021). The Potential of Indonesian Traditional Herbal Medicine as Immunomodulatory Agents: A Review. *International Journal of Pharmaceutical Sciences and Research*, 12(10), 5229–5237. [https://doi.org/10.13040/IJPSR.0975-8232.12\(10\).5229-37](https://doi.org/10.13040/IJPSR.0975-8232.12(10).5229-37)
- Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., ... Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972), 172–180. <https://doi.org/10.1038/s41586-023-06291-2>
- Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachgar, S., Mansfield, P., Prakash, S., Green, B., Dominowska, E., Arcas, B. A. y., ... Natarajan, V. (2023). *Towards Expert-Level Medical Question Answering with Large Language Models*. <http://arxiv.org/abs/2305.09617>
- Sumarni, W., Sudarmin, S., & Sumarti, S. S. (2019). The scientification of jamu: a study of Indonesian's traditional medicine. *Journal of Physics: Conference Series*, 1321(3), 032057. <https://doi.org/10.1088/1742-6596/1321/3/032057>
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models*. <http://arxiv.org/abs/2302.13971>
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., ... Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. <http://arxiv.org/abs/2307.09288>
- Wang, T., Yu, P., Tan, X. E., O'Brien, S., Pasunuru, R., Dwivedi-Yu, J., Golovneva, O., Zettlemoyer, L., Fazel-Zarandi, M., & Celikyilmaz, A. (2023). *Shepherd: A Critic for Language Model Generation*. <https://arxiv.org/abs/2308.04592v1>
- Zareie, A., Sahebkar, A., Khorvash, F., Bagherniya, M., Hasanzadeh, A., & Askari, G. (2020). Effect of cinnamon on migraine attacks and inflammatory markers: A randomized double-blind placebo-controlled trial. *Phytotherapy Research*, 34(11), 2945–2952. <https://doi.org/10.1002/ptr.6721>





- Zhang, K., Zhou, R., Adhikarla, E., Yan, Z., Liu, Y., Yu, J., Liu, Z., Chen, X., Davison, B. D., Ren, H., Huang, J., Chen, C., Zhou, Y., Fu, S., Liu, W., Liu, T., Li, X., Chen, Y., He, L., ... Sun, L. (2024). BiomedGPT: A Generalist Vision-Language Foundation Model for Diverse Biomedical Tasks. *Nature Medicine*. <https://doi.org/10.1038/s41591-024-03185-2>
- Zhang, T., Huang, Z., Wang, Y., Wen, C., Peng, Y., & Ye, Y. (2022). Information Extraction from the Text Data on Traditional Chinese Medicine: A Review on Tasks, Challenges, and Methods from 2010 to 2021. *Evidence-Based Complementary and Alternative Medicine*, 2022, 1–19. <https://doi.org/10.1155/2022/1679589>
- Zhu, X., Li, J., Liu, Y., Ma, C., & Wang, W. (2023). *A Survey on Model Compression for Large Language Models*. <https://arxiv.org/abs/2308.07633v4>

