Journal of the Brazilian Computer Society, 2025, 31:1, doi: 10.5753/jbcs.2025.5767

RagPharma: A RAG-Based Chatbot for Medicine Leaflets with a Dual-Dataset Evaluation Framework

Letícia C. Navarro [**Federal University of Espírito Santo** | ***lnavarro@lcad.inf.ufes.br***]
Filipe Mutz [**Federal University of Espírito Santo** | ***filipe.mutz@ufes.br***]
Thiago M. Paixão [**Federal Institute of Espírito Santo** | ***thiago.paixao@ifes.edu.br***]
Guilherme G. Zanetti [**Federal University of Espírito Santo** | ***guilherme.zanetti@lcad.inf.ufes.br***]
Claudine Badue [**Federal University of Espírito Santo** | ***claudine@lcad.inf.ufes.br***]
Alberto F. De Souza [**Federal University of Espírito Santo** | ***alberto@lcad.inf.ufes.br***]
Thiago Oliveira-Santos [**Federal University of Espírito Santo** | ***todsantos@lcad.inf.ufes.br***]

Laboratório de Computação Aplicada, Universidade Federal do Espírito Santo, Avenida Fernando Ferrari, 514, Goiabeiras, ES, 29075-910, Brazil.

Received: 31 March 2025 • **Accepted:** 31 July 2025 • **Published:** 23 October 2025

Abstract Despite being essential sources of information, Brazilian medicine package leaflets remain underutilized due to their complexity and lack of user-friendly tools for information retrieval. Currently, there are no chat-based systems in Portuguese designed to assist patients in accessing and understanding leaflet content. To address this gap, we present **RagPharma**, a novel Retrieval-Augmented Generation (RAG) system that integrates professional medicine leaflets into a chat interface to answer patient queries. During RagPharma's development, we observed that evaluation performance was significantly higher when using questions derived from the same dataset used to build the system. This led to the identification of a critical evaluation *bias*, often overlooked in RAG applications. In response, we propose a novel **dual-dataset evaluation framework**, which separates the knowledge base and the evaluation source in distinct, but related, datasets. Experimental results confirmed the presence of bias when using overlapping datasets and demonstrated the reliability of our dual-dataset methodology. Under this new evaluation scheme, RagPharma achieved 81% accuracy using the Mistral 7B model—representing a 60% improvement over standalone LLMs. These findings validate both the effectiveness of RagPharma and the importance of unbiased evaluation strategies in domain-specific RAG systems.

Keywords: Retrieval-Augmented Generation, Large Language Models, Healthcare Chatbots, Perplexity Accuracy, Medicine Package Leaflets

1 Introduction

Medicine package leaflets (or simply *leaflets*) are essential documents that provide critical information about drugs, including composition, usage instructions, contraindications, and potential side effects [Dos Santos *et al.*, 2019]. Despite their importance, many patients overlook these documents, either due to their complexity or difficulty in accessing specific information.

Several approaches have been proposed for medical chatbots, from agents for Italian medicine leaflets [Minutolo *et al.*, 2022] and Chinese QA systems [Tian *et al.*, 2019], to classification-based models [Ahmed *et al.*, 2024] and RAG-based solutions using official medical documents [Lunardi *et al.*, 2024; Torres *et al.*, 2024]. Still, there are no dedicated strategies for healthcare communication in Brazilian Portuguese.

In particular, there are no chat-based systems designed to support users in querying medicine leaflets in this language. Existing platforms[1] allow users to view leaflets, but they lack intelligent mechanisms for answering targeted questions. To address this gap, we propose **RagPharma**, a Retrieval-Augmented Generation (RAG) system that integrates brazilian medicine package leaflets into a conversational interface, enabling more accessible and accurate information retrieval for patients.

RagPharma presented a high performance when evaluated with questions derived from the same source used to feed the RAG system. However, follow-up experiments showed that accuracy drops significantly when the evaluation questions come from a dataset different from the knowledge base, even if they have a high intersection of subjects. This discrepancy indicated a potential *bias* caused by the high similarity between questions used for evaluation and as training or retrieval source.

Therefore, the second - and most important - contribution of this work is a thorough experimental evaluation to confirm the supposed bias and the proposal of a novel evaluation methodology for RAG-based systems that mitigates this bias. In the dual-dataset evaluation framework, a dataset serves as external knowledge base for the RAG system, while evaluation questions are generated from a second independent dataset.

As illustrated in Figure 1, the traditional RAG approach—referred to here as RAG-Single Dataset—uses the same dataset both to retrieve relevant information and to generate the responses being evaluated. In contrast, our proposed

[1] sara.com.br, consultaremedios.com.br

method, RAG-Dual Dataset, employs two separate datasets with equivalent information expressed differently: one for the retrieval process and another for evaluation, thereby minimizing potential overlap bias.

When applying the framework to evaluate RagPharma, we benefit from ANVISA's[2] regulations that enforce the creation of independet leaflets for healthcare professionals and for the general population (patients). Professional leaflets are used as the external knowledge base for RagPharma, while patient leaflets serve as source for evaluation. Multiple-choice questions (MCQs) generated from patient leaflets are used to measure the system's ability to correctly answer queries using only the information available in professional leaflets. This separation ensures that the knowledge used to construct the system is distinct from the knowledge used to evaluate it, thereby reducing bias and improving the reliability of the assessment. Moreover, the use of MCQs allows for a more interpretable and accurate evaluation compared to traditional NLP metrics (such as BLEU or ROUGE), which often focus on surface-level similarities rather than the correctness of the conveyed information.

Experimental results confirmed the presence of bias when using a single dataset as knowledge base and evaluation. Furthermore, the proposed dual-dataset evaluation showed improved robustness and interpretability. RagPharma achieved 81% accuracy in answering MCQs using the Mistral 7B model, representing a 60% improvement over standalone LLMs.

The contributions of this paper are summarized as follows:

- Development of **RagPharma**, a novel RAG-based chat application for Brazilian medicine leaflets;
- Empirical demonstration of *bias* in RAG evaluation when training and test data overlap;
- Proposal of a new evaluation framework leveraging a **dual-dataset** approach for more reliable and unbiased assessment of RAG systems;
- A multiple-choice question-answering dataset in Portuguese, based on medicine leaflet content.[3]

The remainder of this paper is organized as follows. Section 2 reviews the related work, including recent evaluation frameworks for RAG systems, medical chatbot applications, and multiple-choice question generation techniques. Section 3 details the proposed system, RagPharma, including its architecture, embedding strategy, and retrieval mechanisms. Section 4 describes the experimental methodology, including dataset preparation, evaluation metrics, and the dual-dataset approach. Section 5 reports and discusses the results, highlighting the performance differences between biased and unbiased evaluation strategies. Finally, Section 6 concludes the paper and outlines directions for future work.

2 Related Work

Medical chatbots have shown promising results, offering natural language support for both patients and professions. Section 2.1 discusses some of these applications and highlight limitations of current approaches. In a complementary direction, the growing adoption of RAG systems has heightened the need for effective evaluation methods, especially in sensitive domains such as healthcare. As pointed out in Section 2.2, previous works explored key aspects such as the quality of the generated responses, the relevance of retrieved context, and the faithfulness of the provided information. Nonetheless, an important aspect, the generalization ability to questions from the same domain, but different from those used as knowledge base, received little attention. Several evaluation frameworks use multiple-choice questions as interpretable strategies for assessing a model's ability to understand and reproduce factual knowledge. However, the automatic generation of such evaluation questions still challenging and Section 2.3 presents recent works that focus on this issue.

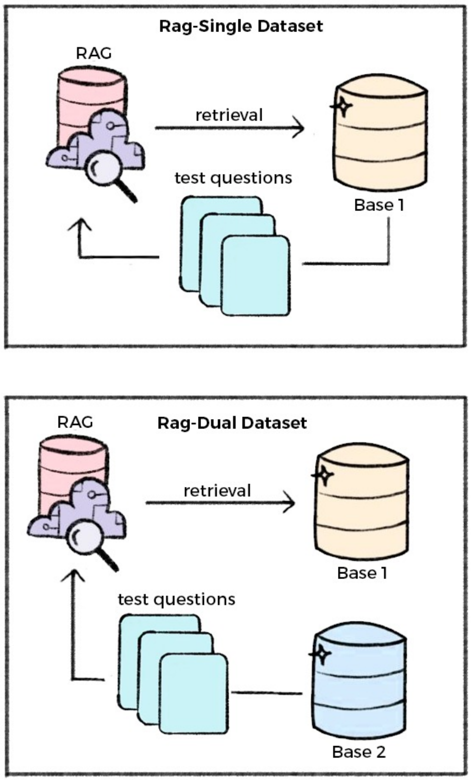

Figure 1. Comparison between *Rag-Single* and *Rag-Dual* setups. While *Rag-Single* uses the same dataset for retrieval and evaluation, *Rag-Dual* employs distinct datasets, mitigating overlap bias.

2.1 Medical Chatbot Applications

A wide range of conversational systems has been proposed in recent years to support healthcare communication, leveraging both traditional machine learning and, more recently, large language models (LLMs). These systems vary considerably in their design, scope, and use of external knowledge sources. For instance, Minutolo *et al.* [2022] proposed a conversational agent that allows users to query italian leaflets in natural language. The system retrieves information from a structured knowledge base built from official documents and was shown to be effective in improving health literacy in a user study. Tian *et al.* [2019] introduced ChiMed, a medical question-answering corpus built from a Chinese forum where

[2] ANVISA is the Brazilian National Health Surveillance Agency

[3] Can be accessed here: `https://github.com/i2ca/ragpharma/`

licensed physicians respond to patient queries. Although the study focuses on corpus construction and analysis, it highlights that the "adopted answer′′ flag in the dataset is not a reliable indicator of answer quality. Predictive analyses using models such as CNN, LSTM, and DRMM (Deep Relevance Matching Model) revealed that answer adoption may be influenced by factors unrelated to relevance or correctness. Along similar lines, Kim *et al.* [2024] proposed the **MEDIC** chatbot, aimed at identifying drug interactions in cancer patients. The system processes images of medication packaging, applies OCR to extract drug names, and then checks for interactions by querying specialized databases such as the Drug Authorization List. The evaluation included metrics such as accuracy, character error rate, and word error rate, along with an overall performance analysis of the chatbot.

Other works adopt distinct strategies for classification and response generation. Ahmed *et al.* [2024] introduced the Robotic Medical Support Chatbot (RMSCB), which provides basic clinical guidance using machine learning techniques. The system relies on two probabilistic models: the Pre-Fixed Class Label Model (PFCLM), which categorizes user queries into predefined disease classes, and the Pre-Fixed Answer and Question Model (PFAQM), which generates responses containing first-aid information and general recommendations. Although it does not use LLMs, the system represents a complementary approach based on supervised learning.

Finally, Lunardi *et al.* [2024] developed a RAG-based conversational agent designed to provide information about PILs, leveraging the LangChain framework for document retrieval. The system was evaluated by comparing the generated responses to reference texts using embeddings and cosine similarity, and by testing different strategies for text segmentation and prompting.

2.2 RAG Evaluation Frameworks

The evaluation of RAG systems has received increasing attention in the literature, given the complexity of assessing the quality of responses produced by such models. The **RAGAS** framework [Es *et al.*, 2024] proposes an automated, reference-free approach based on LLMs to assess three core dimensions: *faithfulness*, *answer relevance*, and *context relevance*. **ARES** [Saad-Falcon *et al.*, 2023] builds upon this by using synthetic query–passage–answer triplets to fine-tune lightweight models (LLM judges), and refines the evaluation through prediction-powered inference (PPI), which enables the estimation of scores with confidence intervals using a small set of human-labeled examples.

Torres *et al.* [2024] also proposed an evaluation framework for RAG systems in the context of pharmaceutical leaflets. Their approach used topic segmentation via BERTTopic and over one thousand drug leaflets as a data source. The quality of the responses was evaluated using standard NLP metrics such as BLEU, ROUGE, and BLEURT. While valuable, this method still relies on surface-level text comparisons, which may fail to capture deeper semantic nuances.

Liu *et al.* [2023] introduced the **RECALL** benchmark to evaluate the robustness of LLMs when exposed to counterfactual information. Their results show that models often favor misleading content, highlighting vulnerabilities in current evaluation approaches. Separately, Xiong *et al.* [2024] proposed **MIRAGE**, a large-scale benchmark of 7,663 medical multiple-choice questions for assessing RAG configurations. They also developed **MEDRAG**, a toolkit that boosts performance by optimizing retriever, corpus, and snippet selection

2.3 Generation and Evaluation of Multiple-Choice Questions

The use of MCQs as an evaluation strategy has gained attention, especially in educational contexts and assessing LLM-based systems. Mucciaccia *et al.* [2025] proposed a method for automatically generating MCQs from textual contexts using LLMs combined with prompt engineering. Their work also includes an automatic validation process for the generated questions, analyzing aspects such as format, grammar, and relevance. To minimize bias, different LLMs were used for the generation and validation stages, with GPT-4o serving as the final evaluator.

Hybrid approaches have also been explored. Maheen *et al.* [2022] introduced a system that generates MCQs by identifying informative sentences from a given context. The system uses BERT to generate embeddings, followed by K-Means clustering to select the most relevant sentences based on a scoring mechanism. Correct answers are extracted using named entity recognition, while distractor (incorrect alternatives) are generated from external sources such as WordNet, Wiktionary, and Google Searches.

Cheung *et al.* [2023] compared MCQs generated by ChatGPT with those written by human experts. Using two medical textbooks, both created 50 questions, which were evaluated by experts on five criteria. The results showed that ChatGPT's questions were nearly as good as human-generated ones, except in domain relevance, while ChatGPT generated them in 20 minutes compared to over 3 hours for humans.

These studies demonstrate the potential and challenges of automatic MCQ generation, particularly in specialized domains such as medicine. In the present work, MCQs were used to enable interpretable and objective evaluation of system performance.

3 RagPharma

RagPharma is an advanced question-answering system designed to facilitate access to information present in medicine package leaflets. It combines natural language processing and information retrieval technologies to provide reliable and relevant answers. An overview of the system is presented in Figure 2. A set of leaflets, usually in PDF format, is used as a knowledge base for the system. These documents are first cleaned and structured using an LLM. The professional leaflets were used as the knowledge base for the RAG, as they are more comprehensive than patient leaflets. After the data cleaning, an embedding model transforms the texts into vector representations, which are further stored in a dataset. Once the system is deployed, whenever it receives a question about a medication, it maps the question into an embedding and searches in the dataset for the leaflet with the most similar embedding. Finally, the question and the leaflet content are

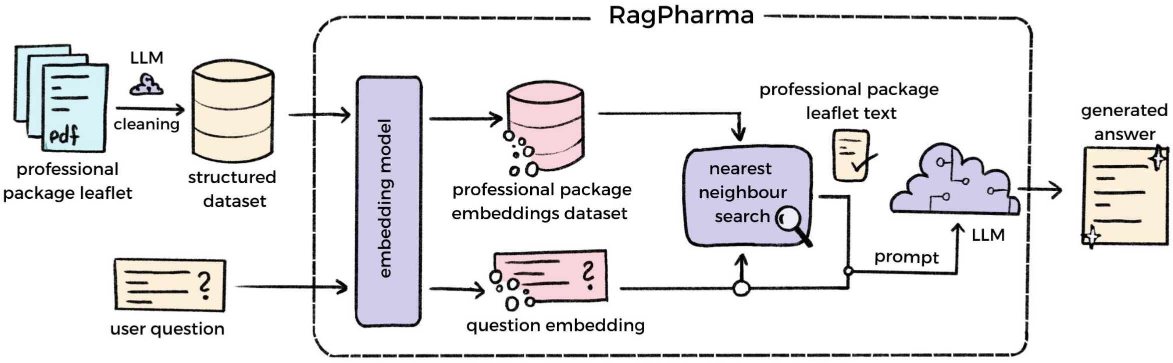

Figure 2. Overview of RagPharma. The user's question is converted into an embedding representation and, compared to pre-generated embeddings from medicine package leaflets. The most similar leaflet is used as context for the LLM to generate an accurate response.

combined into a single prompt, which is then passed to the LLM to generate a response.

3.1 Embedding Dataset Creation

First, textual information is extracted from PDF with an automatic tool. Since this process is subject to noise, a data-cleaning step is performed with the aid of an LLM. As a byproduct, the information from the leaflet is also structured according to their sections. The professional leaflet contains on average 10 sections of medication information that include: Composition, Indications, Efficacy Results, among others.

Although leaflets are official technical-scientific healthcare documents with a standardized format, this format is not always respected in the PDF files. To deal with this variation, we employed the LLaMA 3 8B model [AI@Meta, 2024] as an information extraction tool, allowing for the separation of each section present in the leaflets. LLaMA was selected for its strong performance in accurately identifying section topics during preliminary tests, as well as for its smaller environmental footprint compared to larger models. The instructions also specify not to return any line breaks and to correct any spelling errors if they exist. Due to the 8192-token context limit of LLaMA 3 8B, some PDFs had to be split for extraction, while others that were of context larger than 8k were discarded in this approach. This was essential for extracting only the relevant parts of the leaflets and producing a consistent dataset.

The text used for generating the embeddings comprises of all sections of the leaflet, excluding the introduction, references, and laboratory information. This approach focuses on the essential content of the medicinal leaflet and prevents chunk segmentation from leaving contextless and uninformative parts, such as excerpts containing only side effects without mentioning the medication's name. As a result, each chunk corresponds to a full medical leaflet. By including all relevant sections within each entry, we ensure that each segment contains sufficient information to answer specific questions related to that leaflet.

The selection of the embedding model was guided by detailed experiments, as presented in Section 4.2.

3.2 RAG System

RAG systems consist of a dataset, an embedding model that processes both the dataset and the user query, and a language model that answers the query using the most relevant context based on a similarity function. In our system, the dot product was used as the similarity function, given its central role in computing relevance between embeddings. We chose not to use a vector store due to the small size of the dataset, relying instead on CSV files. However, the framework is fully compatible with vector stores if needed. The implemented code is available in the Github repository referenced in the Section *Availability of Data and Materials*. .

When a user asks a question, the system converts it into embeddings using the embedding model. Then, by comparing the embeddings generated for each leaflet with the question embedding, the system identifies and returns the most relevant leaflet to answer the question. In this work, only one leaflet is retrieved, meaning that only the closest embedding is considered. Retrieving two or more leaflets could confuses the model, leading to inaccurate answers.

4 Experimental Methodology

An LLM receives the question and the most relevant leaflet and generates the final answer for the user. The language model receives a prompt structured as follows: an instruction to answer the question based on this context, followed by the retrieved leaflet is marked as context, and finally, the user's question. Having the context allows it to create more concise answers on the topic without hallucinating. Three LLMs were investigated for this task: Mistral 7B v.02 [Jiang *et al.*, 2023], LLaMA 3 8B, and Phi3 [Abdin *et al.*, 2024] from Microsoft. All models have comparable size, between 7 and 8 billion parameters, and are popular for delivering good results despite being smaller than larger models available such as GPT-3.5 and Gemini. Similar to the choice of LLM model for text cleaning, these models also offer a lower environmental impact due to their smaller size. All selected models are multilingual, with support for Portuguese, which is essential since the system requires responses to be generated in that

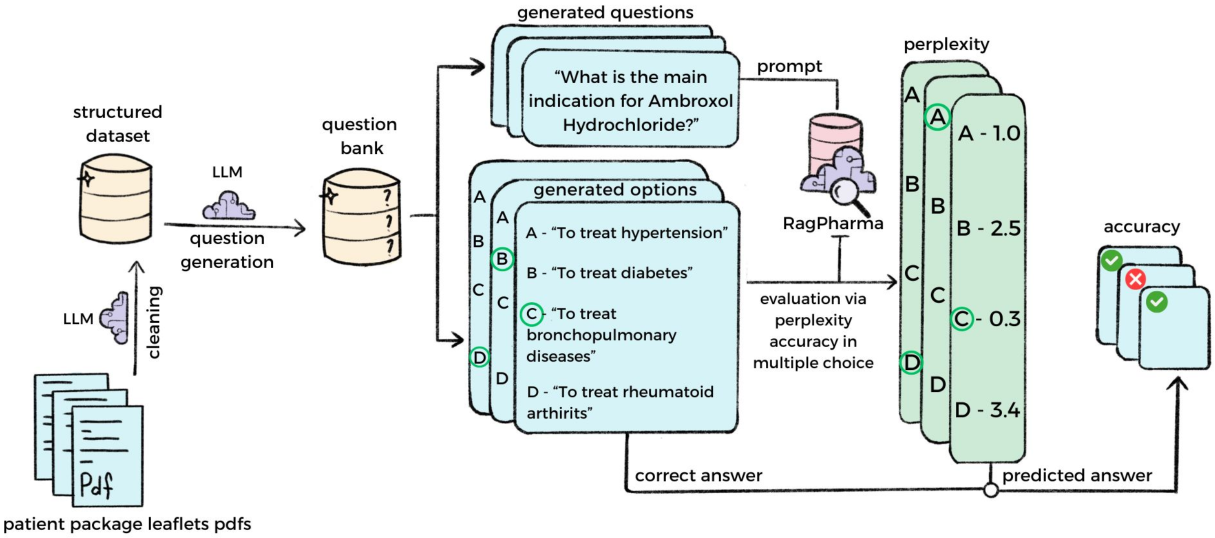
generated questions
"What is the main indication for Ambroxol Hydrochloride?"
prompt
perplexity
structured dataset
question bank
LLM
question generation
LLM
cleaning
generated options
A - "To treat hypertension"
B - "To treat diabetes"
C - "To treat bronchopulmonary diseases"
D - "To treat rheumatoid arthirits"
A
B
C
D
RagPharma
evaluation via perplexity accuracy in multiple choice
A - 1.0
B - 2.5
C - 0.3
D - 3.4
accuracy
correct answer
predicted answer
Pdf
patient package leaflets pdfs

Figure 3. Evaluation process illustrated, involving dataset structuring, question generation, and selection based on perplexity, a metric assessing language model predictive performance.

language. The models were used in their 'instruct' version[4], which besides being trained with web documents, mathematical formulas, and code, are trained with instructions allowing for a better chat experience.

All three models have approximately 8k of context length. Because the cleaned leaflets are more concise than the original ones, this context length comfortably accommodates all sections of the leaflets.

This section outlines the experimental methodology, including a detailed description of the used dataset and the evaluation framework. To conduct a comparative analysis, experiments were carried out with the three LLMs under investigation.The experiments include the evaluation of standalone LLMs (i.e., without RAG) as a baseline. Additionally, the LLMs were assessed when paired with RagPharma using both the proposed dual-dataset approach (LLM + RagPharma-dual) and the single-dataset approach (LLM + RagPharma-single) to highlight potential bias in the evaluation. This approach enabled us to assess whether the proposed system improves the performance of the language models in the specific task of answering questions from leaflets.

It is important to note that RagPharma is specialized in answering questions about individual medications but is not designed to handle comparative questions about drugs. For example, it can answer questions like "Is headache a side effect of ibuprofen?" but not questions like "Which one has more side effects, Neosaldina or Dorflex". The system also provides the possibility of retrieving the original leaflet content, allowing users to verify the source of the generated response in cases where hallucination is suspected.

The details of the computational environment used for conducting the experiments are provided in Appendix A.1.

4.1 Dataset

The data files were obtained through web scraping on the ANVISA website[5]. Due to the high volume of data, only 368 leaflets were used. The professional and patient leaflet datasets were pre-processed as described in Section 3.1. The professional leaflets were used as source for the RAG dataset due to their greater completeness and details. The patient leaflets were used for evaluation purposes, i.e., as a reference for formulating questions, ensuring that all information presented to patients were supported and grounded in the broader and more technical guidelines of the professional leaflet.

The nature of the dataset enabled the separation in two related datasets. This is the basis for a thorough and less biased evaluation, as the data is paired in terms of information while remaining distinct in text (i.e., written in different styles). As an example of the same information presented in different ways, the Appendix A.2 shows two leaflets for the same medication: one intended for patients and other for healthcare professionals.

4.2 Embeddings Model Experiment

For the embedding model, we used Sentence Transformers and conducted experiments with three variants: a small multilingual model, the `Multilingual-E5-small` model [Wang *et al.*, 2024] with strong performance on Massive Text Embedding Benchmark MTEB, a Portuguese-only model adapted from BERTimbau, the `Legal-BERTimbau-sts-base-ma-v2` [Souza *et al.*, 2020; Fonseca *et al.*, 2016; Real *et al.*, 2020; May, 2021] and the same BERTimbau model after fine-tuning with questions and sentences related to medications.

The training dataset for the embedding model consisted of pairs combining medication-related questions and drug names drawn from the full dataset. Each pair was assigned a high similarity score when the drug in the question matched the associated medication name, and a low score when the pair referred to different drugs. A total of 3 million such tuples

[4] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct, https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2, https://huggingface.co/microsoft/Phi-3-small-8k-instruct

[5] https://www.gov.br/anvisa/pt-br/sistemas/bulario-eletronico

were generated. The questions were manually written and the medication names were automatically substituted—without the use of an LLM. While similar in structure, these questions were not copied from the LLM-generated question set used for test.

To evaluate the quality of the embedding model, we used real questions from the LLM-generated dataset, along with their corresponding medication names, to assess whether the model was capable of correctly identifying the associated drug.

4.3 Performance Assessment

What side effects may occur with prolonged administration of Mud Oral?
A) Adrenal suppression 1.027 ✓
B) Fever or chest pain 85.0499
C) Stomach pain or diarrhea 15.055
D) Headache or sore throat 39.726

What is a common side effect that may occur with the use of Verapamil Hydrochloride?
A) Stevens-Johnson syndrome 2.623
B) Renal failure **1.294** ✗
C) Headache and dizziness 2.747
D) Hyperglycemia 5.286

What are the common side effects that may occur with the use of Riluzole and relate to the nervous system?
A) Diarrhea or stomach pain 5.022
B) Anaphylactic reactions or angioedema 5.964
C) Anemia or pancreatitis 29.975
D) Headache, dizziness, and paresthesia 1.733 ✓

Figure 4. Examples of three generated multiple-choice questions. Each question includes four answer options, with perplexity scores shown in orange. The correct answers are highlighed in bold, and a checkmark or cross indicates whether the model selected the correct option. These specific questions and answers were translated into English for illustrative purposes.

Currently, there are no standard method for evaluating the answer-generation capabilities of LLMs in specific domains. Therefore, in this work, we propose an evaluation process, illustrated in Figure 3, that leverages the specific characteristics of the data in our application.

Firstly, the patient and professional datasets are structured, cleaned, and preprocessed with the assistance of an LLM. The leaflets were obtained in PDF format, which prevented their direct use. With the assistance of an LLM, the content was segmented into standardized sections (e.g., indications, dosage, side effects), and the relevant text from each was extracted, discarding visual and structural elements from the PDFs. The resulting information was organized into CSV files, ensuring a structured and consistent representation for each leaflet. More details can be seen in the repository.Secondly, each processed patient leaflet is passed through an LLM for generating multiple-choice questions regarding the content of the leaflet. For each question, there are 4 generated answers: 1 correct and 3 incorrect. These questions and answers are organized in a question bank used as ground truth to assess RagPharma's performance. RagPharma receives each question as if it was asked by a user, retrieving the most relevant leaflet to provide an answer. The core of the evaluation lies in the last step. Instead of generating a textual response, the model's answer is assumed to be the most likely option it would produce from the four choices, based on the perplexity metric [Jelinek *et al.*, 1977], conditioned on the provided context. This approach enables a more objective evaluation based on the accuracy metric computation, as both correct and incorrect answers are known beforehand. The LLM used for cleaning the leaflets and creating the questions was LLaMA 3 8b. Figure 4 illustrates examples of generated multiple-choice questions along with their answer options and the model's selection process. The rest of this section details the question generation procedure, answer selection using the perplexity metric, accuracy evaluation, and other complementary performance metrics based on text similarity.

4.3.1 Question Generation

The patient information leaflet consists of nine sections, specifically written for the patient, and includes information such as indications, contraindications, dosage, posology, side effects, and overdose instructions.

The generation process relies on LLMs, which are effective in creating multiple-choice questions from texts, similar to questions posed by humans. In this process, all sections of the patient leaflet were provided as input to ensure the model has the context necessary for generating the question. The input for the LLM consists of three elements: the section from which the question will be generated, an example question to guide the generation, and the full leaflet sections. The LLM was prompted to return a JSON-like string with the question, a correct answer, and a list of wrong answers. After creation, the questions underwent a two-step verification process conducted with the LLM to ensure their consistency in terms of context coherence and format integrity (i.e., 1 correct answer and 3 incorrect answers). In the first step, the LLM was prompted with the generated question and the related context to verify whether they are aligned. In the second step, the LLM was prompted with the generated question, the respective context, and the answers assigned as 'wrong' to verify whether they were indeed incorrect and could not be interpreted as correct by the model. An equivalent procedure was performed for the correct answers. Questions that failed verification were discarded, leaving a total of 1,084 multiple-choice questions. It is important to mention that the generated question bank was later manually reviewed to remove flawed cases, such as questions with multiple correct answers, blank correct answers (i.e., options without text), or questions unrelated to the topic. As a result, 239 questions were removed, leaving a total of 845 questions. The following question is an example of a generated question unrelated to the topic (in this case, that is not likely to be performed by patients): "How long after applying Mud Oral is it recommended that the patient develops a thin film?".

4.3.2 Perplexity-based Answer Selection

LLM-based systems are typically used to generate textual responses when prompted by a user. For the objective evaluation proposed in this work, however, the softmax LLM outputs (probabilities) are analyzed to assess how closely a candidate answer matches the model's expected output given the context, with the closest match selected as the LLM's preferred answer. This procedure involves the perplexity metric, defined as

$$\text{PPL}(X) = \exp\left\{-\frac{1}{t}\sum_{i=1}^{t}\log p_\theta(x_i|x_{<i})\right\}, \qquad (1)$$

where $X = (x_1, \ldots, x_t)$ represents a sequence of t tokens, $x_{<i}$ represents the tokens preceding x_i, and $p_\theta(x_i|x_{<i})$ denotes the probability of the i-th token, given the preceding tokens as context, for a model parametrized by θ. Perplexity measures the overall uncertainty of the model in predicting the next token, with lower values indicating a higher degree of compatibility between the model and the sequence.

In our application, $X_{gen} = (x_1, \ldots, x_g)$ represents the generation prompt tokens, which can be used for chat-like response generation. The remaining $t - g$ tokens, X_{ans}, belong to a candidate answer. To compute PPL, the LLM is fed with X_{gen} and autoregressively predicts the probabilities for the remaining $t - g$ tokens using teacher forcing. We set $p_\theta(x_i|x_{<i}) = 1$ for $i \leq g$, ensuring that the metric focuses solely on the answer.

Although the generation prompt is not explicitly used in Equation 1, it influences the probability outputs related to the candidate answer. Thus, context augmented with RAG-retrieved content is expected to affect the probability outputs in a way that minimizes perplexity.

4.3.3 Accuracy Evaluation

The quality of the responses generated by LLM systems depends on the generation prompt provided to the model. In this work, we evaluate LLMs using two formats of generation prompts. The first format, designed for standalone LLM evaluation, includes a question as context and an instruction to answer it. The second format, used for evaluating RAG-augmented LLMs, additionally incorporates retrieved content from the embeddings dataset.

For a given question, the generation prompt is constructed and appended to each candidate answer from the four possible options. The model selects the answer by feeding the LLM with the prompt and candidate answer, choosing the one that results in the lowest perplexity. If the selected answer matches the correct option, it is considered a true positive; otherwise, it is a false positive. Conversely, an incorrect option that is not selected is a true negative, while a false negative occurs when the correct option is not selected.

The accuracy is measured by iterating over the questions and computing the percentage of correct answers provided by the model, which is given by

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \qquad (2)$$

where TP, TN, FP, and FN represent the total number of true positives, true negatives, false positives, and false negatives, respectively.

It is important to note that the prompt sent to the model does not include the multiple-choice options; they are only used for answer selection via perplexity. This approach is more consistent with a system designed as a generative model for answering user questions in a chat-like manner, rather than a multiple-choice answering system (which would be more suitable for a quiz, for example).

4.4 Complementary Evaluation

A complementary analysis was conducted by examining text generation when the models were fed with the generation prompt. The generated responses were compared to the correct answers for each question using common similarity metrics from the NLP literature: BLEU Papineni *et al.* [2002], which measures n-gram precision; ROUGE Lin [2004], which assesses n-gram overlap; and BertScore Zhang *et al.* [2019], which uses embeddings to compare semantic similarity between texts. While these metrics do not strongly indicate whether the model's response is correct, they help evaluating the quality of the generated text, complementing the perplexity-based accuracy evaluation. Another analysis verifying the accuracy of RAG retrieval was conducted and reported in appx. A.2

5 Results and Discussion

In this section, we present and analyze the results of our experiment using the RagPharma system. We begin by evaluating different embedding models to understand their impact on retrieval quality. Next, we assess the accuracy of generated responses under both single- and dual-dataset settings. We then perform a complementary evaluation using widely adopted NLP metrics to further analyze metric behavior. This is followed by a qualitative analysis of representative cases to provide deeper insight into system performance. Finally, we discuss the main limitations of our current approach.

5.1 Embedding Model Evaluation

This experiment evaluates the retrieval effectiveness of different embedding models by measuring their ability to identify the correct leaflet based on the given question. As shown in Table 2, the multilingual model achieved the highest accuracy (0.936), followed by the fine-tuned BERTimbau model (0.571). Although BERTimbau (base) initially performed poorly (0.134), its substantial improvement after fine-tuning highlights the potential of Portuguese-language models when adapted to domain-specific tasks with a small amount of targeted training.

Despite the performance of the fine-tuned BERTimbau, the multilingual model was selected for subsequent experiments due to its superior results and more robust pretraining, which likely contributed to its generalization ability in complex medical queries.

Table 1. Complementary Analysis with other metrics, Bert Score (BS), BLEU, and ROUGE.

Model	BS (F1)	BS (Recall)	BS (Precision)	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Standalone Mistral 7B	0.743	0.794	0.700	0.0051	0.0958	0.0267	0.0850
Mistral 7B + RagPharma-dual	0.805	0.867	0.755	0.0676	0.2759	0.1932	0.2611
Standalone LLaMA 3 8B	0.779	0.777	0.783	0.0266	0.1786	0.0660	0.1701
LLaMA 3 8B + RagPharma-dual	0.859	0.889	0.833	0.194	0.4754	0.3697	0.4598
Standalone Phi 3 8B	0.792	0.805	0.781	0.0277	0.1913	0.0749	0.1744
Phi 3 8B + RagPharma-dual	0.839	0.882	0.803	0.124	0.4048	0.3056	0.3888

Table 2. Results of embedding models on the retrieval task, with their respective accuracy (Acc.).

Embedding Model	Acc.
Multilingual (small)	0.936
BERTimbau (base)	0.134
BERTimbau (fine-tuned)	0.571

5.2 Accuracy Evaluation

Table 3 presents the accuracy results for the standalone LLMs (baseline), LLMs + RagPharma-dual, and LLMs + RagPharma-single (biased). The utilized LLMs were: Mistral 7B, LLaMA 3 8B, and Phi 3 8B. The results highlight the effectiveness of RagPharma in enhancing the performance of LLMs in answering questions about medicine leaflets. All tested models showed significant performance improvements when integrated with RagPharma. As expected, the standalone LLMs achieved the lowest accuracy perplexity, demonstrating that they lack sufficient information to effectively answer questions about Brazilian medication leaflets. For dual-dataset evaluation, the RAG-augmented LLMs achieved an average accuracy of 0.805, significantly outperforming the standalone models. Notably, the highest accuracy was achieved by RAG-augmented LLMs under the single-dataset (biased) setup, with LLaMA reaching an accuracy perplexity of 0.949. On average, RagPharma-single outperformed RagPharma-dual by 11.64 percentage points. This performance gain indicates that the model answers more effectively when the question-generation and retrieval components share the same underlying dataset—supporting the previously discussed bias and reinforcing the hypothesis that single-dataset evaluation inflates results due to data leakage.

For dual-dataset evaluation, Mistral 7B achieved the highest overall performance with an accuracy of 0.810, representing a 59.14% increase compared to the standalone Mistral 7B. This substantial improvement highlights a significant quality enhancement. Similarly, LLaMA 3 8B and Phi 3 8B showed notable improvements across all evaluated metrics. These findings suggest that integrating RagPharma greatly enhances the models' ability to interpret and respond to questions based on medicine leaflets, indicating that the LLMs lacked sufficient knowledge about the leaflets during pre-training.

It is worth noting that although the LLaMA 3 8B model was used for both question generation and answering in one of the setups—potentially introducing additional bias—two other models were also used to answer the questions. This allowed us to analyze the obtained results under conditions less susceptible to generation-induced bias. Notably, the best performance was achieved by one of the models (Mistral) not involved in the question generation process.

Table 3. Results of the revised questions (854 questions) with the reported accuracy (Acc.).

Model	Acc.
Standalone Mistral 7B	0.512
Mistral 7B + RagPharma-dual	0.821
Mistral 7B + RagPharma-single	0.940
Standalone LLaMA 3 8B	0.415
LLaMA 3 8B + RagPharma-dual	0.800
LLaMA 3 8B + RagPharma-single	0.949
Standalone Phi 3 8B	0.449
Phi 3 8B + RagPharma-dual	0.795
Phi 3 8B + RagPharma-single	0.929

5.3 Complementary Evaluation

In addition to the objective accuracy evaluation, we assessed the models' ability to generate text (answers) similar to the correct ones. The results of this experiment are shown in Table 1. Given the observed bias condition, the analysis is restricted to the standalone LLMs and the RAG-augmented LLMs using a dual-dataset approach. A significant improvement was observed in all NLP metrics with the use of RagPharma to augment the LLMs' capabilities in answering questions. Although distance-based metrics like BLEU and ROUGE did not show as substantial gains, this can be attributed to the fact that they are based on measuring token distance. They are sensitive to synonyms and word order, which can distort quality evaluation when the content is semantically correct but differs lexically or syntactically. Therefore, a more specialized metric would be recommended for a more robust evaluation in these circumstances.

5.4 Qualitative Analysis

Table 4 illustrates user questions and the answers generated by RagPharma using Mistral 7B, the best-accuracy model. In the example, the questions and answers were presented in

English instead of Portuguese, the original language used for system operation and evaluation.

Noticeably, the model was able to generate correct answers for questions I and II, with some additional phrases typical from the LLM chat pre-training. Question III exhibits a scenario where the answer is correct but differs substantially from the provided correct answer, highlighting a limitation encountered in this study where multiple valid answers are possible.

Table 4. Sample questions displaying the correct answer alongside responses generated by RagPharma.

Question I: What is the maximum daily recommended dose for long-term treatment with Verapamil hydrochloride? **Correct Answer:** 480 mg **RagPharma Answer:** The maximum daily dose for long-term treatments with Verapamil Hydrochloride should not exceed 480 mg.
Question II: What are the symptoms of side effects that may occur with the use of Neo Mistatin cream? **Correct Answer:** irritation and sensitivity, including burning sensation and itching **RagPharma Answer:** The symptoms of known side effects associated with the use of Neo Mistatin cream include irritation and sensitivity, including a burning sensation and itching.
Question III: What is the very common reaction ($>$1/10) that can occur with the use of Oxaliplatin and is it related to hematological function? **Correct Answer:**Anemia **RagPharma Answer:** A very common ($>$1/10) hematologic reaction that can occur with the use of Oxaliplatin is leukopenia.

5.5 Limitations

The main limitation of the proposed evaluation method is the requirement for a dual dataset, where the same information must be present in different forms across multiple datasets. Although this study proposes a very powerful evaluation method, it targets specific data not commonly found in standard benchmarks, but still enabling better evaluation of real world RAG applications. Additionally, as observed in the qualitative results, the model's responses may be correct but still fail to match the ground truth, as seen in the red column of Table 4.

Another limitation is the possibility of LLM hallucination, which can lead to the generation of inaccurate information. For this reason, it is important to emphasize that the goal of RagPharma is not to replace medical professionals, but to assist users in better understanding the contents of the leaflet. In case of any doubts, a healthcare professional should always be consulted.

6 Conclusion

This work introduced **RagPharma**, a novel RAG-based chat system for Brazilian medicine leaflets, addressing the lack of Portuguese-language tools for patient-centered access to medical information. During evaluation, we identified a performance discrepancy when comparing questions derived from the same source used in the system's construction with those from an independent source, revealing a potential *bias* in traditional RAG evaluation.

To address this, we proposed a **dual-dataset evaluation framework** that leverages the distinction between *patient* and *professional* leaflets, ensuring independence between retrieval and assessment data. Results confirmed the presence of bias in single-dataset evaluations and showed that RagPharma, when evaluated under the dual-dataset setting, achieved 81% accuracy using the Mistral 7B model—outperforming standalone LLMs by 60%. These findings enable future users of the framework to better estimate the real expected performance of the system after the deployment since the questions performance by the end users will not be withdraw from the same RAG retrieval dataset. The dual-dataset evaluation gives a less overestimated result since it also this not share the same dataset in the evaluation and retrieval.

Our embedding experiments revealed that even models trained specifically in Portuguese can underperform compared to multilingual models when dealing with highly specialized tasks—such as understanding medical terms and drug names. These findings highlight the importance of domain-specific fine-tuning, even within a single language, to ensure reliable performance in specialized applications. Moreover, the strong performance of the multilingual model underscores how a high-quality pretraining phase, especially one involving diverse and extensive data, can significantly enhance a model's generalization ability in domain-specific retrieval tasks.

The results demonstrate the feasibility of the proposed framework, supported by high accuracy in multiple-choice evaluations. Additionally, the qualitative analysis confirmed that the system can deliver informative and user-friendly responses in a conversational format.

Future work will focus on improving retrieval for complex queries, such as handling questions that mention multiple medications, using multi-hop retrieval to address broader queries—for example, identifying which medications are suitable for headaches and extending the dual-dataset evaluation framework to other domains. In particular, we plan to investigate strategies for approximating dual-dataset setups in scenarios where naturally paired data sources (such as professional and patient leaflets) are not available. Furthermore, we aim to investigate question generation strategies that account for cultural and linguistic biases, with the goal of producing more realistic and user-aligned queries. This includes the use of data augmentation, paraphrasing techniques, or synthetic dataset generation to simulate evaluation under knowledge/-source separation constraints.

Declarations

Authors' Contributions

All authors contributed to the conceptualization, methodology, and formal analysis of the study. TOD was responsible for supervision and providing resources. LCN led the manuscript writing and was also in charge of investigation, data curation, and software development. FM, TMP, GGZ, and TOD participated in writing, reviewing, and editing the manuscript.

Acknowledgment

This study was financed in part by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Brazil) - Finance Code 001; Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil); and Fundação de Amparo à Pesquisa do Espírito Santo (FAPES, Brazil) - grant 2021-07KJ2.

Availability of Data and Materials

We provide all resources—including code, datasets, prompts, and scripts—in the GitHub repository: https://github.com/i2ca/ragpharma/.

Conflicts of Interest

The authors declare that they have no conflict of interest related to this work.

References

Abdin, M., Jacobs, S. A., Awan, A. A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., *et al.* (2024). Phi-3 technical report: A highly capable language model locally on your phone. DOI: 10.48550/arXiv.2404.14219.

Ahmed, S. T., Fathima, A. S., Nishabai, M., Sophia, S., *et al.* (2024). Medical chatbot assistance for primary clinical guidance using machine learning techniques. *Procedia Computer Science*, 233:279–287. DOI: 10.1016/j.procs.2024.03.217.

AI@Meta (2024). Llama 3 model card. Available at:https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Cheung, B. H. H., Lau, G. K. K., Wong, G. T. C., Lee, E. Y. P., Kulkarni, D., Seow, C. S., Wong, R., and Co, M. T.-H. (2023). Chatgpt versus human in generating medical graduate exam multiple choice questions—a multinational prospective study (hong kong sar, singapore, ireland, and the united kingdom). *PloS one*, 18(8):e0290691. DOI: 10.1371/journal.pone.0290691.

Dos Santos, D. J. L., Feitosa, M., Sena, E., and Dalcin, F. (2019). A importância da bula para o usuário de medicamentos. *Brazilian Journal of Surgery & Clinical Research*, 27(1). Available at:https://www.mastereditora.com.br/periodico/20190607_201024.pdf.

Es, S., James, J., Anke, L. E., and Schockaert, S. (2024). Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158. DOI: 10.48550/arXiv.2309.15217.

Fonseca, E., Santos, L., Criscuolo, M., and Aluisio, S. (2016). Assin: Avaliacao de similaridade semantica e inferencia textual. In *Computational Processing of the Portuguese Language-12th International Conference, Tomar, Portugal*, pages 13–15. Available at:https://sites.google.com/view/assin2/.

Jelinek, F., Mercer, R. L., Bahl, L. R., and Baker, J. K. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63. DOI: 10.1121/1.2016299.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7b.

Kim, W. T., Shin, J., Yoo, I.-S., Lee, J.-W., Jeon, H. J., Yoo, H.-S., Kim, Y., Jo, J.-M., Hwang, S., Lee, W.-J., Park, S., and Kim, Y.-J. (2024). Medication extraction and drug interaction chatbot: Generative pretrained transformer-powered chatbot for drug-drug interaction. *Mayo Clinic Proceedings: Digital Health*, 2(4):611–619. DOI: 10.1016/j.mcpdig.2024.09.001.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81. Available at:https://aclanthology.org/W04-1013/.

Liu, Y., Huang, L., Li, S., Chen, S., Zhou, H., Meng, F., Zhou, J., and Sun, X. (2023). Recall: A benchmark for llms robustness against external counterfactual knowledge. *arXiv preprint arXiv:2311.08147*. DOI: 10.48550/arxiv.2311.08147.

Lunardi, R., Coppola, P., *et al.* (2024). Conversational-agent for patient information leaflet. In *Proceedings of the 14th Italian Information Retrieval Workshop*, pages 70–73. Available at:https://ceur-ws.org/Vol-3802/paper14.pdf.

Maheen, F., Asif, M., Ahmad, H., Ahmad, S., Alturise, F., Asiry, O., and Ghadi, Y. Y. (2022). Automatic computer science domain multiple-choice questions generation based on informative sentences. *PeerJ Computer Science*, 8:e1010. DOI: 10.7717/peerj-cs.1010.

May, P. (2021). Machine translated multilingual sts benchmark dataset. Available at:https://github.com/PhilipMay/stsb-multi-mt.

Minutolo, A., Damiano, E., De Pietro, G., Fujita, H., and Esposito, M. (2022). A conversational agent for querying italian patient information leaflets and improving health literacy. *Computers in Biology and Medicine*, 141:105004. DOI: 10.1016/j.compbiomed.2021.105004.

Mucciaccia, S. S., Meireles Paixão, T., Wall Mutz, F., Santos Badue, C., Ferreira de Souza, A., and Oliveira-Santos, T. (2025). Automatic multiple-choice question generation and evaluation systems based on llm: A study case with university resolutions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2246–2260, Abu Dhabi, UAE. Association for Computational Linguistics. Available at:`https://aclanthology.org/2025.coling-main.154/`.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. DOI: 10.3115/1073083.1073135.

Real, L., Fonseca, E., and Oliveira, H. G. (2020). The assin 2 shared task: a quick overview. In *International Conference on Computational Processing of the Portuguese Language*, pages 406–412. Springer. DOI: 10.1007/978-3-030-41505-1$_3$9.

Saad-Falcon, J., Khattab, O., Potts, C., and Zaharia, M. (2023). Ares: An automated evaluation framework for retrieval-augmented generation systems. *arXiv preprint arXiv:2311.09476*. DOI: 10.18653/v1/2024.naacl-long.20.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*. DOI: 10.1007/978-3-030-61377-8$_2$8.

Tian, Y., Ma, W., Xia, F., and Song, Y. (2019). Chimed: A chinese medical corpus for question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 250–260. DOI: 10.18653/v1/W19-5027.

Torres, J. J. G., Bîndilă, M. B., Hofstee, S., Szondy, D., Nguyen, Q.-H., Wang, S., and Englebienne, G. (2024). Automated question-answer generation for evaluating rag-based chatbots. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC-COLING 2024*, pages 204–214. Available at:`https://aclanthology.org/2024.cl4health-1.25/`.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., and Wei, F. (2024). Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*. DOI: 10.48550/arXiv.2402.05672.

Xiong, G., Jin, Q., Lu, Z., and Zhang, A. (2024). Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251. DOI: 10.18653/v1/2024.findings-acl.372.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *CoRR*, abs/1904.09675. DOI: 10.48550/arXiv.1904.09675.

A Appendix

A.1 Computational Environment

The experiments were carried out on a machine equipped with an NVIDIA GeForce RTX 4090 GPU, Intel(R) Core(TM) i7-10700 CPU @ 2.90GHz, 32GiB RAM, running Ubuntu 22.04.2 LTS. The framework was built using Python3 with the HuggingFace libraries.

A.2 Leaflet comparison

Below, we compare two sections from the leaflet of the same medication—one intended for patients and the other for healthcare professionals—highlighting how the same information is conveyed in a similar manner, yet tailored to the target audience.

PATIENT

PARA QUE ESTE MEDICAMENTO É INDICADO?
WHAT IS THIS MEDICATION INDICATED FOR?

Buscopan® Composto é indicado para o tratamento dos sintomas de cólicas intestinais, estomacais, urinárias, das vias biliares, dos órgãos sexuais femininos e menstruais.

Buscopan® Composto is indicated for the treatment of symptoms of intestinal, stomach, urinary, biliary tract, female reproductive organ, and menstrual cramps.

QUANDO NÃO DEVO USAR ESTE MEDICAMENTO?
WHEN SHOULD I NOT USE THIS MEDICATION?

Você não deve usar Buscopan® Composto se tiver alergia a analgésicos semelhantes à dipirona (como isopropilaminofenazona, propifenazona, fenazona, fenilbutazona), ao butilbrometo de escopolamina ou a algum outro componente do produto. Isto inclui, por exemplo, o desenvolvimento de agranulocitose (febre, dor de garganta ou alteração da boca e garganta, associados à ausência ou diminuição de células brancas no sangue) após o uso destas substâncias.
You should not use Buscopan® Composto if you are allergic to analgesics similar to dipyrone (such as isopropylaminophenazone, propyphenazone, phenazone, phenylbutazone), to butylscopolamine bromide, or to any other component of the product. This includes, for example, the development of agranulocytosis (fever, sore throat, or mouth and throat changes associated with a lack or reduction of white blood cells) after using these substances.

O uso também não é indicado se tiver asma induzida por analgésicos, ou se desenvolver reações anafilactoides (manifestações na pele e inchaço dos lábios, língua e garganta) ou broncoespasmo (estreitamento das vias respiratórias) após tomar analgésicos (como paracetamol, salicilatos, diclofenaco, ibuprofeno, indometacina ou naproxeno).

Use is also not recommended if you have analgesic-induced asthma or if you develop anaphylactoid reactions (such as skin manifestations and swelling of the lips, tongue, and throat) or bronchospasm (narrowing of the airways) after taking analgesics (such as paracetamol, salicylates, diclofenac, ibuprofen, indomethacin, or naproxen).

PATIENT

Você também não deve usar Buscopan® Composto se tiver comprometimento da medula óssea (ex.: após algum tratamento medicamentoso com agentes citostáticos, que inibem o crescimento ou a reprodução das células) ou comprometimento no sistema formador de elementos do sangue; deficiência genética da enzima glicose-6-fosfato-desidrogenase, tendo risco aumentado de alterações do sangue; porfiria hepática aguda intermitente (doença do metabolismo do sangue que provoca alterações na pele e sistema nervoso); glaucoma (aumento da pressão dentro do olho); aumento da próstata com dificuldade para urinar; estreitamento da passagem do conteúdo no estômago e intestinos; íleo paralítico ou obstrutivo (não funcionamento do intestino); megacólon (dilatação da parte final dos intestinos); taquicardia (batimentos cardíacos acelerados); miastenia gravis (doença que provoca fraqueza muscular), se estiver no terceiro trimestre de gravidez ou amamentando.

You should also not use Buscopan® Composto if you have bone marrow impairment (e.g., after treatment with cytostatic agents that inhibit cell growth or reproduction) or impairment in the blood cell-producing system; a genetic deficiency of the enzyme glucose-6-phosphate dehydrogenase, which increases the risk of blood disorders; acute intermittent hepatic porphyria (a blood metabolism disorder that affects the skin and nervous system); glaucoma (increased pressure inside the eye); enlarged prostate with difficulty urinating; narrowing of the passage through the stomach or intestines; paralytic or obstructive ileus (intestinal blockage or failure); megacolon (dilation of the lower part of the intestines); tachycardia (rapid heartbeat); myasthenia gravis (a condition that causes muscle weakness); if you are in the third trimester of pregnancy or are breastfeeding.

O comprimido revestido de Buscopan® Composto também é contraindicado em condições hereditárias raras de intolerância à galactose. Buscopan® Composto é contraindicado a partir dos 6 meses de gravidez. Este medicamento não deve ser utilizado por mulheres grávidas sem orientação médica. Informe imediatamente seu médico em caso de suspeita de gravidez.

The coated tablet form of Buscopan® Composto is also contraindicated in rare hereditary conditions of galactose intolerance. Buscopan® Composto is contraindicated from the sixth month of pregnancy onward. This medication should not be used by pregnant women without medical advice. Inform your doctor immediately in case of suspected pregnancy.

PROFESSIONAL

INDICAÇÕES
INDICATIONS

Buscopan Composto é indicado para o tratamento sintomático de estados espástico-dolorosos e cólicas do trato gastrintestinal, das vias biliares, do trato geniturinário e do aparelho genital feminino (dismenorreia).

Buscopan Composto is indicated for the symptomatic treatment of painful spasms and cramps of the gastrointestinal tract, biliary tract, genitourinary tract, and female reproductive system (dysmenorrhea).

CONTRAINDICAÇÕES
CONTRAINDICATIONS

Buscopan® Composto é contraindicado nos casos de: *Buscopan® Composto is contraindicated in the following cases:*

- Hipersensibilidade prévia a pirazolonas ou pirazolidinas (como dipirona, isopropilaminofenazona, propifenazona, fenazona, fenilbutazona), ao butilbrometo de escopolamina ou a qualquer outro componente do produto. *Previous hypersensitivity to pyrazolones or pyrazolidines (such as dipyrone, isopropylaminophenazone, propyphenazone, phenazone, phenylbutazone), to butylscopolamine bromide, or to any other component of the product.*
- Desenvolvimento anterior de agranulocitose após uso dessas substâncias. *Previous development of agranulocytosis after using these substances.*
- Síndrome de asma induzida por analgésico ou intolerância analgésica do tipo urticária-angioedema. *Analgesic-induced asthma syndrome or analgesic intolerance such as urticaria-angioedema.*
- Comprometimento da função da medula óssea ou doenças do sistema hematopoiético. *Impaired bone marrow function or disorders of the hematopoietic system.*
- Deficiência genética de glicose-6-fosfato-desidrogenase (risco de hemólise). *Genetic deficiency of glucose-6-phosphate dehydrogenase (risk of hemolysis).*
- Porfiria hepática aguda intermitente. *Acute intermittent hepatic porphyria.*
- Glaucoma. *Glaucoma.*
- Hipertrofia da próstata com retenção urinária. *Prostate enlargement with urinary retention.*
- Estenose mecânica do trato gastrintestinal. *Mechanical stenosis of the gastrointestinal tract.*
- Íleo paralítico ou obstrutivo. *Paralytic or obstructive ileus.*

PROFESSIONAL

- Megacólon. *Megacolon.*
- Taquicardia. *Tachycardia.*
- Miastenia gravis. *Myasthenia gravis.*
- Terceiro trimestre de gravidez. *Third trimester of pregnancy.*
- Amamentação. *Breastfeeding.*

O comprimido revestido de Buscopan® Composto também é contraindicado em condições hereditárias raras de intolerância à galactose. *The coated tablet form of Buscopan® Composto is also contraindicated in rare hereditary conditions of galactose intolerance.*