Comparative Study of Large Language Models for Lung-RADS Classification in Portuguese CT Reports

1st Tarcísio Lima Ferreira
Computing Institute, (UFAL)
Federal University of Alagoas (UFAL)
Maceió, Brazil
tlf@ic.ufal.br

2nd Marcelo Costa Oliveira
Computing Institute, (UFAL)
Federal University of Alagoas (UFAL)
Maceió, Brazil
oliveiramc@ic.ufal.br

3rd Thales Miranda de Almeida Vieira
Computing Institute, (UFAL)
Federal University of Alagoas (UFAL)
Maceió, Brazil
thales@ic.ufal.br

2024 IEEE 24th International Conference on Bioinformatics and Bioengineering (BIBE) | 979-8-3315-1862-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIBE63649.2024.10820460

***Abstract*—Lung cancer has the highest mortality rate among all cancer types for both males and females. It is estimated that lung cancer accounts for 21% of cancer deaths in each gender. This alarming statistic highlights the significant impact of lung cancer on overall cancer mortality, underscoring the urgent need for effective prevention, early detection, and treatment strategies to combat this disease. Lung cancer screening (LCS) is a process that involves carefully selecting high-risk individuals, primarily current or former heavy smokers. It includes annual low-dose computed tomography scans and meticulous interpretation of results followed by appropriate follow-up care. Adherence to LCS follow-up is essential for maximizing the life-saving benefits of this preventive measure. Multiple professional societies, such as the American College of Radiology (ACR) and Fleischner Society, have published guidelines for managing patients with pulmonary nodules. Lung CT Screening Reporting & Data System is a quality assurance tool designed to standardize the reporting of lung cancer screening CT scans and provide consistent management recommendations. In this context, this work aims to evaluate whether large language models (LLM) could accurately identify and extract lung nodules' characteristics from unstructured chest CT reports in the Portuguese, based on the Lung-RADS classification system. This work assessed the effectiveness of three LLMs: Gemini, GPT-4-o, Llama-3 70B, and a BERT model BioBERTpt. Our findings indicate that LLMs, especially GPT-4-o, have significant potential in automating the extraction of lung nodule characteristics for Lung-RADS classification, which could aid radiologists in their work. Notably, GPT-4-o with few-shot learning using Prompt 4 emerged as the best model, achieving an F1-score of 0.89. Our results highlight the potential of LLMs to assist radiologists in accurately classifying lung nodules according to the Lung-RADS criteria, streamlining the diagnostic process.**

***Index Terms*—Information extraction, Named entity recognition, Natural language processing, Large language model, Chest CT report, Lung cancer, Lung-RADS**

I. Introduction

Lung cancer has the highest mortality rate among all cancer types for both males and females. In 2023, an estimated 238,340 people (117,550 men and 120,790 women) were diagnosed with lung cancer, and 127,070 people died from the disease. This alarming statistic highlights the significant impact of lung cancer on overall cancer mortality, underscoring the urgent need for effective prevention, early detection, and treatment strategies to combat this disease [25].

Lung cancer screening (LCS) is a process that involves carefully selecting high-risk individuals, primarily current or former heavy smokers. It includes annual low-dose computed tomography (LDCT) scans and meticulous interpretation of results followed by appropriate follow-up care [3]. LCS is crucial for early detection and improved outcomes. A study of the National Lung Screening Trial (NLST) revealed that individuals who underwent annual LDCT scans experienced a significant 20% reduction in lung cancer mortality [23].

Adherence to LCS follow-up is essential for maximizing the life-saving benefits of this preventive measure. However, a recent American College of Radiology (ACR) analysis revealed a low follow-up rate in the American population of just 22% in 12 months [9]. Multiple professional societies, such as the ACR and Fleischner Society, have published guidelines for managing patients with pulmonary nodules.

So, the use of guidelines in screening programs holds significant importance because it aims to minimize the need for excessive follow-up exams, offering enhanced guidance to radiologists, clinicians, and patients [10]. These guidelines provide healthcare professionals with evidence-based recommendations for the appropriate diagnostic and treatment approaches for patients with suspected or confirmed nodules [12].

Guidelines recommended by LCS programs utilize a structured reporting system to ensure clarity and consistency in reported information. This system should include details on the characteristics of lung nodules, such as the quantity, location, and size, guideline-based recommendations for the surveillance of small lung nodules, and descriptions of other potentially actionable findings [21].

Lung CT Screening Reporting & Data System (Lung-RADS®) is a quality assurance tool designed to standardize the reporting of lung cancer screening CT scans and provide consistent management recommendations [24]. The characteristics of the lung nodules determine each Lung-RADS index; the greater the risk of malignancy, the higher the Lung-RADS

index. The type and interval of the examination change depending on this index. The follow-up examination for a nodule with a lower index is an LDCT in 12 months while for a nodule with a higher index, a PET CT or biopsy is recommended [24]. Thus, extracting structured Lung-RADS scores from free-text radiology reports can significantly enhance the efficiency of LCS programs.

Determining follow-up examinations for an individual LCS CT based on the Lung-RADS score is straightforward. However, extracting the information in a structured way to classify the lung nodule according to the Lung-RADS criteria is not trivial. This clinical information is often stored in an unstructured form as free text. Converting this data into a structured format can be time-consuming and may only effectively capture some aspects of the information [19].

Natural Language Processing (NLP) can assist radiologists in real time by suggesting the appropriate Lung-RADS category and identifying reports that lack sufficient data to assign the correct Lung-RADS classification [2] [16] [17]. NLP can help reduce errors and minimize false-positives and negatives [6]. Advances in NLP provide a promising avenue for automated extraction of Lung-RADS malignancy index data from the unstructured content of radiology reports.

Current state-of-art works in NLP have used architectures based on transformers [26]. The Bidirectional Encoder Representations from Transformers (BERT) have demonstrated exceptional performance across various tasks, from natural language understanding tasks and text classification to Question answering and text generation [4]. Building upon the BERT foundation, domain-specific models like BioBERT (pre-trained on extensive biomedical corpora), ClinicalBERT (pre-trained on approximately 2 million MIMIC-III clinical notes) and PubMedBERT (pre-trained in PubMed literature) have been developed [20]. These models were applied to various clinical NLP tasks, showing that models pre-trained in clinical domains outperform base BERT models and BERT models pre-trained in other areas.

Recent advancements in transformer-based Large Language Models (LLMs), pre-trained on massive text corpora, have dramatically expanded natural language processing capabilities. These models, characterized by their neural network architecture and vast parameter counts, have performed remarkably on various language tasks. Prominent examples of LLMs include PaLM, LLaMA, and GPT-4 [22]. In the realm of clinical text, LLMs have shown promise in tasks such as information extraction, Question answering, summarization, and even aiding in clinical decision-making. By leveraging their ability to process and comprehend complex medical language, LLMs offer the potential to enhance efficiency, accuracy, and patient care within the healthcare industry [1] .

Previous research in information extraction has primarily focused on Chinese and English idioms, limiting the understanding of this phenomenon in other languages [18] [15]. Existing studies often compare variations within a single model architecture, such as BERT or its derivatives, or exclusively examine the effectiveness of different Large Language Models (LLMs). This research, however, introduces a different approach by directly comparing BERT-based models with LLMs in the context of Portuguese idiom, offering a more comprehensive analysis of their capabilities.

In this context, this work aims to evaluate whether large language models (LLM) could accurately identify and extract lung nodules' characteristics from unstructured chest CT reports in the Portuguese, based on the Lung-RADS classification system, as well as compare the effectivness of these LLMs to determine which model performs best. This work assessed the effectiveness of three LLMs: Gemini, GPT-4-o, Llama-3 70B, and a BERT model BioBERTpt. Our findings indicate that LLMs, especially GPT-4-o, have significant potential in automating the extraction of lung nodule characteristics for Lung-RADS classification, which could aid radiologists in their work.

II. Materials and Methods

A. Dataset

Our dataset has 963 chest CT reports in the Portuguese idiom collected from January 01, 2022, to April 03, 2023, from University Hospital of Alagoas. Upon signing the Consent Form, the reports of patients undergoing chest CT for any indication were obtained. It is important to highlight that all patient data has been anonymized.

For the annotation scheme, we adopted the Inside-Outside-Beginning (IOB) format. This schema makes it possible to identify correctly and delimit the entities in a text, facilitating the processing and analysis of specific information.

- The I-prefix indicates that a tag is inside an entity;
- The O-prefix indicates that the token does not belong to any entity;
- The B-prefix indicates that the tag is at the beginning of an entity that follows another chunk of this entity without O tags between the two chunks;

The 963 texts from the CT reports were divided into specific proportions: 90% for training, and 10% for testing in the Named Entity Recognition (NER) task. In Information Extraction (IE) task, we utilized 100 CT reports to perform the Question-answering procedure.

Based on the Lung-RADS guidelines [24], we defined eight Questions related to pulmonary nodules for the IE task. A thoracic radiologist with 15+ years of experience manually labeled the named entities (NEs) and the answers to the Questions utilized for IE. The Questions and their corresponding statistics are presented in Table I.

B. Models

We fine-tuned the BERT model using BioBERTpt [7] to serve as a baseline for the traditional supervised learning approach. The model weights were initialized using the Transformers library, available at HuggingFace [14], and we used the PyTorch (version 2.0.1) implementation of the model. We fine-tuned the model and used the grid search with the following hyperparameter values: Batch size {4, 8, 16}, Epochs = {5, 10}, and Learning Rate = {2e-3, 1e-3}. The

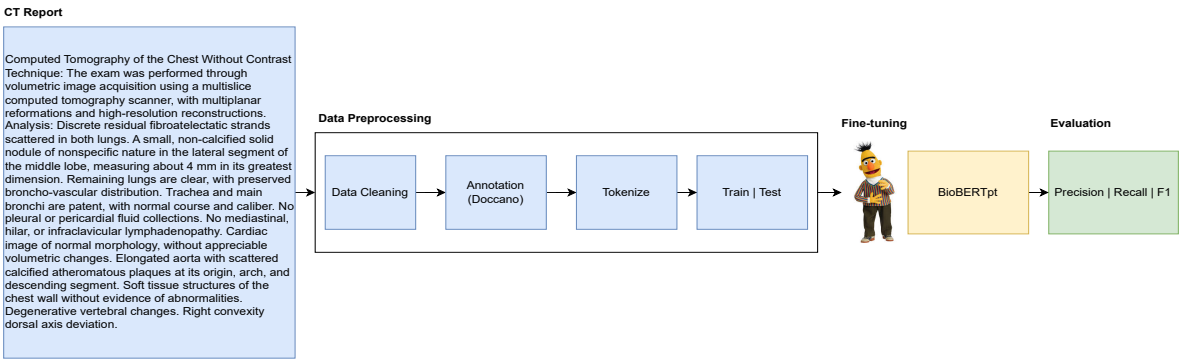

Fig. 1. Methodology scheme applied in NER task.

Stochastic Gradient Descent (SGD) = 0.01 was the adopted optimizer, with a momentum = 0.9, and the GPU utilized in the fine-tuning task was an RTX3060 12GB.

For the LLMs models, we used the specific versions: GPT-4o (gpt-4o-2024-05-13) [13], Gemini 1-5 Flash (gemini-1.5-flash) [8], and the Llama-3 70B [5]. In the context of LLMs, the temperature parameter regulates the uncertainty or randomness in the generation process. This parameter typically ranges from 0 (completely deterministic) to 1 or higher (resulting in increasingly random and diverse outputs) [15]. For GPT-4 and Llama-3 models, the temperature was set to 0 to minimize randomness in response generation. Using a lower temperature, we limited the model's tendency to take creative leaps, ensuring more predictable and consistent outputs. This is important in IE tasks, where the accuracy and reliability of the information extracted are crucial.

Requests for the three LLM models were made via API. We used the OpenAI API for GPT-4o, the Together AI API for Llama-3 70B, and the Google AI API for the Gemini Flash 1.5 model. During the tests conducted in this study, the cost of GPT-4o was US$5.00 per 1 million tokens for input and US$15.00 per 1 million tokens for output. For Llama-3 70B, the cost was approximately US$0.88 per 1 million tokens for both input and output.

C. Data Preprocessing

We performed data cleaning on all reports. First, special characters were removed from the reports. After that, each report was uploaded to the Doccano annotation tool [11]. Six NEs in the Portuguese idiom were used to label the text, and these NEs corresponded to the characteristics of the pulmonary nodules. The NEs used were Atenuação (Attenuation), Calcificação (Calcification), Bordas (Edges), Achado (Finding), Localização (Localization) and Tamanho (Size). These characteristics were chosen based on Lung-RADS guidelines [24].

As a result, a JSON file containing the labeling information for all reports was generated. Next, we split every report and its labeling information into sentences and tokenize them with the BERT tokenizer.

Finally, each sequence of integers representing a report and its labeling information was padded to a fixed size. This was necessary because models like BERT require a specific input sequence length. The maximum size of the lists containing the converted reports was 497 tokens.

The value of the max token length was set to 512. As a result, any texts with fewer than 512 tokens were padded with zeros at the end of the list to match the maximum length. This specific token quantity was employed because it aligns with the token limit of the BioClinicalBertpt model. Padding tokens were designated with a distinctive tag: '–PADDING–'.

D. Prompt engineering

Based on the work of [18], the Prompts used for IE were designed, as shown in Figure 2. The Prompt templates consist of three parts: (1) Original CT report; (2) IE instructions and an unfilled Question table; (3) Additional requirements for the IE task. In this work, the LLMs were instructed to respond with "No" as the default answer for Questions that do not have corresponding information in the given CT report.

We provide annotated reports with completed tables to enhance LLMs' task comprehension and result accuracy. The K-Nearest Neighbors (KNN) algorithm was used to identify the reports most similar to the test reports. From our database of 963 reports, only those that contained answers to all the Questions were selected. After the filtering process, the dataset was narrowed down to 300 reports. One hundred reports were used for testing, while the remaining 200 were utilized as examples for few-shot learning.

The testing framework uses 1, 2, and 6 examples in the few-shot learning approach. Two of the six examples are the most similar to the test report, while the other four are examples where the LLMs faced difficulty extracting information. This methodology results in six different Prompt combinations.

- Prompt 1 with one example;
- Prompt 1 with two examples;
- Prompt 1 with six examples;
- Prompt 2 with one example;
- Prompt 2 with two examples;
- Prompt 2 with six examples;

TABLE I
PULMONARY NODULES QUESTIONS AND THE STATISTICS OF THE ANNOTATED ANSWERS.

No.	Question	Answer type	Answer statistic
1	Report ID	Numerical	-
2	Is the nodule solid	Boolean	10 (Positive)
3	Is the nodule soft tissue, semisolid or subsolid	Boolean	7 (Positive)
4	Is the nodule ground glass	Boolean	2 (Positive)
5	Is the nodule spiculated or irregular	Boolean	4 (Positive)
6	Is the nodule calcified	Boolean	61 (Positive)
7	Nodule location	Categorical	21 (RUL) 10 (RML) 28 (RLL) 18 (LUL) 17 (LLL) 7 (Others)
8	Nodule size	Numerical	5.41 mm $\pm$ 3,27

RUL - Right upper lobe RML - Right middle lobe RLL - Right lower lobe LUL - Left upper lobe LLL - Left lower lobe

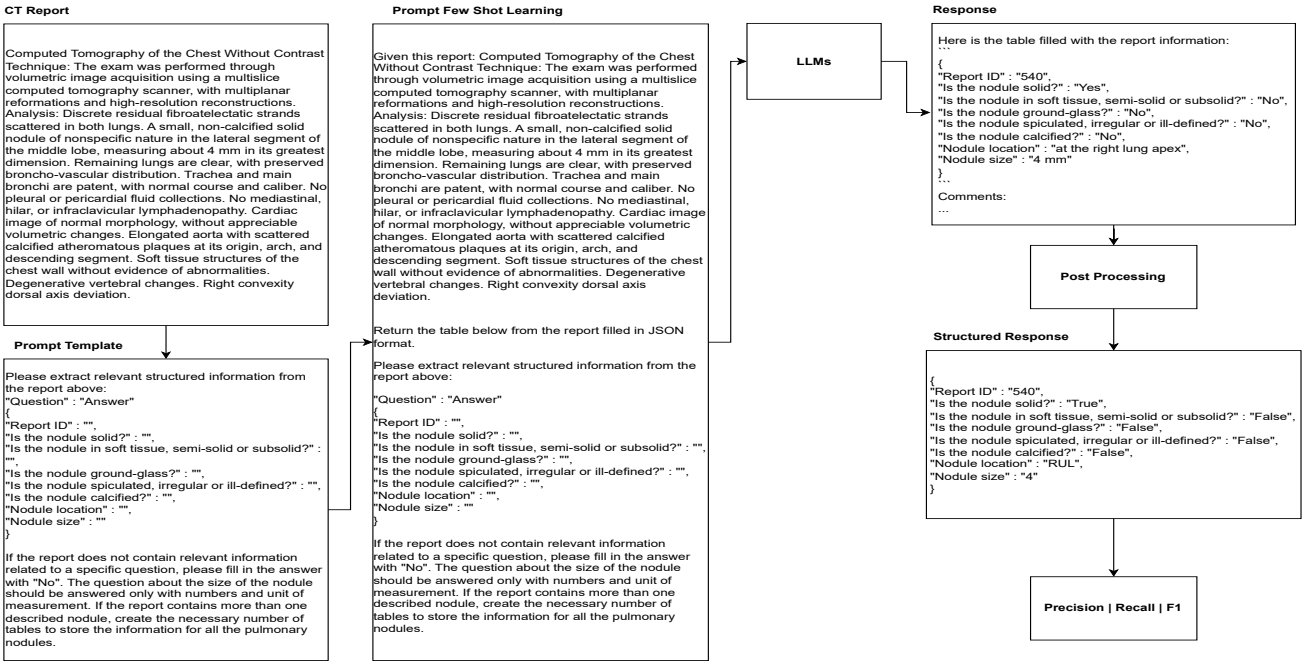

Fig. 2. Methodology scheme applied in IE task.

This setup allows us to evaluate the LLMs' effectiveness under various conditions and explains their ability to handle similar and challenging examples.

E. Questions answering using LLMs

The CT reports are combined with Prompt models to generate answers to Questions. This combined Prompt is submitted via API to the LLMs, and their responses are obtained. A new request is made for each CT report, preventing previous requests from influencing the IE results. Additionally, the LLMs' responses are requested in JSON format to facilitate post-processing of the results. The responses from these language models do not always consist solely of the completed Question table. Consequently, any additional text is disregarded, as it is irrelevant to the analysis. The focus is exclusively on extracting the content in the form of the Question table.

F. Post-processing for structured information extraction

The LLMs were instructed to extract only the answers from the provided table in the Prompts. However, the responses were not always structured. To address this issue, post-processing was applied to convert unstructured responses into a structured format. This post-processing involved removing all text except the table containing the Questions and answers. Besides, when the LLM did not provide an answer in the table, leaving the space blank, or responded with 'Not informed,' the answer was considered as 'No.' For Questions 2 to 6, the answers 'Yes' and 'No' were converted into Boolean values.

Regular expressions were used to identify the number and unit of measurement to extract the nodule size, which was then converted to millimeters. For the location Question, keywords such as 'right,' 'left,' 'middle,' 'upper,' and 'lower' were employed to categorize the answer into six formats: 'right upper lobe,' 'right middle lobe,' 'left lower lobe,' 'left upper lobe,' 'left lower lobe,' and 'others.' The answer 'others' was used when the specific location of the pulmonary nodule was unclear.

G. Evaluation

To evaluate the effectiveness of LLMs for the IE task and BioBERTpt for the NER task, we utilized precision (P), recall (R), and F1-score (F1) metrics. For the nodule size Question, the following strategy was applied to calculate these metrics:

- If the extracted value matched the gold standard value and both were not empty, it was recorded as a true positive.
- If both the extracted and gold standard results were empty, it was considered a true negative.
- If the extracted result was empty while the gold standard result was not, it was classified as a false negative.
- If the extracted result was not empty, but the gold standard result was empty, it was classified as a false positive.
- If the extracted and gold standard results were not empty but differed, it was also considered a false positive.

III. Experimental Results and Discussion

Tables II, III, and IV, present a detailed analysis of BioBERTpt and four LLMs: Gemini, GPT-4-o, Llama-3 70B, in terms of their effectiveness at extracting key characteristics of pulmonary nodules from Portuguese chest CT reports. Table V compares the effectiveness of BioBERTpt specifically with the other three LLMs.

TABLE II
Effectiveness BioBERTpt for each Entity.

Entity	P	R	F1
Finding	0.90	0.91	0.91
Attenuation	0.88	0.78	0.82
Edges	0.67	0.43	0.52
Calcification	0.98	0.98	0.98
Localization	0.89	0.88	0.88
Size	0.92	0.83	0.87

A. Zero-shot effectiveness with different Prompts

The effectiveness of Gemini 1.5 Flash, GPT-4-o, and Llama-3 70B in zero-shot learning with Prompt 1 (P1) and Prompt 2 (P2) are shown in Table III.

As expected, using a more elaborate Prompt with prior medical knowledge (third part of the Prompt) increased the F1-score for most of the seven Questions across the three LLMs evaluated. The most significant effectiveness improvements were observed in GPT-4-o and Llama-3 70B for Question 3, where the F1-score for GPT-4-o increased from 0.17 to 1.00, and for Llama-3 70B, it increased from 0.10 to 1.00. However, we also noted a decrease in the F1-score for some Questions. In Gemini 1.5 Flash, the F1-score for Question 2 decreased by 0.08, while for the GPT-4-o model, the F1-score decreased by 0.20 in Question 1 and 0.04 in Question 2.

Our results support the findings of [18], demonstrating that more elaborate Prompts incorporating prior medical knowledge can significantly enhance model effectiveness in zero-shot learning tasks. We utilized a Prompt similar to theirs but extended our analysis to include the GPT model and the effectiveness of Llama-3 70B and Gemini 1.5 Flash. Even though we used a different dataset from the previous study our results were consistent with the original study's findings.

The improvement in F1-scores, particularly for Question 3 in the GPT-4o and Llama-3 70B models, indicates that these models benefit from additional and contextual information. However, the observed decreases in effectiveness, such as in Gemini 1.5 Flash and GPT-4o, suggest that the complexity or structure of the Prompts may sometimes confuse the models or overwhelm their processing capabilities.

B. Few-shot effectiveness with different Prompts

The effectiveness of Gemini 1.5 Flash, GPT-4-o, and Llama-3 70B in few-shot learning with different Prompts are shown in Table IV.

For Gemini 1.5 Flash, the F1-score of most Questions improved as the Prompts included more instructions to extract information and included more input examples. Specifically, the F1-score for Question 1 increased by 0.26 from Prompt 1 to Prompt 6, while Question 2 F1-score remained unchanged. Question 3 also saw a 0.26 increase in the F1-score, but Question 4 experienced a slight decrease of 0.08. The F1-score for Question 5 improved by 0.20, and for both Questions 6 and 7, there was a modest increase of 0.01 in the F1-score.

The same trend was not observed for GPT-4-o. The most significant increase in F1-score occurred for Question 3, where the F1-score increased from 0.24 in Prompt 1 to 1.00 in Prompt 6. However, in Question 1, there was a decrease of 0.23 in the F1-score from Prompt 1 to Prompt 6. Question 2 had an increase of 0.09 in the F1-score, while Question 4 improved by 0.17. Question 5 experienced a 0.24 increase, and Question 7 had a slight gain of 0.05 in the F1-score. The F1-score of Question 6 remained unchanged.

Interestingly, when analyzing the results with Prompt 4, GPT-4-o emerged as the most effective model among those evaluated. The most significant increase was again in Question 3, where the F1-score rose from 0.24 in Prompt 1 to 1.00 in Prompt 4. However, in Question 1, there was a more pronounced decrease of 0.27 in the F1-score from Prompt 1 to Prompt 4. Question 4 improved by 0.15, while Question 5 had an increase of 0.06, and Question 6 showed a slight gain of 0.01 in the F1-score. Both Questions 2 and 7 remained unchanged in F1-score.

Llama-3 70B also was benefited from the improved Prompts, with an increase in the F1-score for all Questions between Prompts 1 and 6. Specifically, the F1-score for Question 1 increased by 0.23, and Question 3 saw a substantial

TABLE III
ZERO SHOT LLMS EFFECTIVENESS PROMPT 1 AND PROMPT 2

No.	Question	Gemini 1.5 Flash - P1			GPT-4-o - P1			Llama-3 - 70 B - P1		
		P	R	F1	P	R	F1	P	R	F1
1	Is the nodule solid	0.45	1.00	0.63	0.77	1.00	0.87	0.36	1.00	0.53
2	Is the nodule soft tissue, semisolid or subsolid	0.64	1.00	0.78	0.64	1.00	0.78	0.54	1.00	0.70
3	Is the nodule ground glass	0.22	1.00	0.36	0.10	1.00	0.17	0.05	1.00	0.10
4	Is the nodule spiculated or irregular	0.60	0.75	0.67	0.57	1.00	0.73	0.31	1.00	0.47
5	Is the nodule calcified	0.95	0.57	0.71	0.95	0.57	0.71	0.93	0.61	0.73
6	Nodule location	0.91	1.00	0.95	0.89	1.00	0.94	0.90	1.00	0.95
7	Nodule size	0.88	1.00	0.94	0.82	1.00	0.90	0.85	1.00	0.92
		Gemini 1.5 Flash - P2			GPT-4-o - P2			Llama-3 - 70 B - P2		
		P	R	F1	P	R	F1	P	R	F1
1	Is the nodule solid	0.50	1.00	0.67	0.50	1.00	0.67	0.91	1.00	0.95
2	Is the nodule soft tissue, semisolid or subsolid	0.54	1.00	0.70	0.58	1.00	0.74	0.64	1.00	0.78
3	Is the nodule ground glass	0.40	1,00	0.57	1.00	1.00	1.00	1.00	1.00	1.00
4	Is the nodule spiculated or irregular	0.57	1.00	0.73	0.75	0.75	0.75	0.36	1.00	0.53
5	Is the nodule calcified	0.95	0.95	0.95	0.95	0.57	0.71	0.95	0.98	0.97
6	Nodule location	0.91	1.00	0.95	0.92	1.00	0.96	0.95	1.00	0.97
7	Nodule size	0.91	1.00	0.95	0.89	1.00	0.94	0.90	1.00	0.95

gain of 0.87. Question 2 showed a smaller improvement of 0.09, while Question 4 rose by 0.17. The F1-score for Question 5 increased by 0.24, Question 6 saw a modest gain of 0.02, and Question 7 experienced a slight rise of 0.05.

The general improvement in F1-scores in Gemini 1.5 Flash, GPT-4-o, and Llama-3 70B with more intricate Prompts in a few-shot scenario reflects these models' ability to learn from additional examples and detailed instructions. Our results were similar to those reported in Yan Hu et al. [15], despite utilizing a different dataset, varying the input Prompts, and adopting a distinct strategy for comparing the effectiveness of the supervised model with the LLMs. In their work, an input Prompt for the LLMs was used, where an HTML text file was generated with NEs marked using an HTML span tag to highlight these entities. In contrast, we adopted an information extraction strategy based on a Question table. Additionally, while their study employed GPT-3.5, GPT-4, and BioClinicalBERT, our work explored the effectiveness of different LLMs, such as GPT-4o, Llama-3 70B, and Gemini 1.5 Flash. This consistency in results further underscores the robustness of these models when provided with carefully crafted Prompts and an appropriate context.

C. Effectiveness comparison LLMs x BioBERTpt

We compared the effectiveness of LLMs using zero-shot and few-shot learning in the task of IE to the effectiveness of BioBERTpt in the task of NER.

Among the models that received fine-tuning, the BioBERTpt model that achieved the best results utilized the following hyperparameters: a batch size of 8, 10 epochs, and a learning rate of 1e-3. The effectiveness of this model is shown in the Table II.

To compare the effectiveness of LLMs with BioBERTpt, we carried out some experiments using the following evaluation metrics: The metrics for NE ATTENUATION were compared with the average metrics of Questions 1-3. The metrics for NE EDGES were compared to the metrics of Question 4. The metrics for NE CALCIFICATION were compared to the metrics of Question 5. The metrics for NE LOCALIZATION were compared to the metrics of Question 6. Finally, the metrics for NE SIZE were compared to the metrics of Question 7.

Based on Table V, it is evident that LLMs outperform BioBERTpt in extracting information. BioBERTpt only excelled in identifying whether the lung nodule is calcified. Llama-3 70B demonstrated the best effectiveness among all the models evaluated, achieving the highest F1-scores in Questions 1, 2, 3, 6, and 7.

The comparison between LLMs and BioBERTpt reveals that, although BioBERTpt outperforms in specific tasks, such as calcification identification, LLMs significantly surpass BioBERTpt in other information extraction areas. This suggests that while BioBERTpt is effective in highly specialized tasks, LLMs offer flexibility and overall effectiveness across a broader range of tasks, with GPT-4-o particularly excelling.

Therefore, GPT-4-o with few-shot learning using Prompt 4 is the best model for extracting lung nodule characteristics from Portuguese chest CT reports to assist the radiologist in calculating the Lung-RADS index. However, the effectiveness of this model diminishes for reports with more intricate descriptions of findings.

Despite their superior effectiveness, most LLMs are paid services, with few offering free access, and they also raise concerns regarding the handling of sensitive data. Although the promising results, our study has some limitations.

First, we limited our evaluation to a specific set of LLMs: GPT-4-o, Llama-3 70B, and Gemini 1.5 Flash. In future work, we plan to include other models, such as Llama-3.1, Gemini 1.5 Pro, and GPT-4.

Second, we only analyzed reports that contained all the information from the Question table. If all available reports were included, the effectiveness we observed might not be assured, as LLMs could struggle to extract the desired information depending on the format and content of those additional reports.

Third, we employed a basic few-shot learning approach. In

TABLE IV
FEW SHOT LLMS EFFECTIVENESS PROMPT 1 - PROMPT 6

No.	Question	Gemini 1.5 Flash - P1			GPT-4-o - P1			Llama-3 - 70 B - P1		
		P	R	F1	P	R	F1	P	R	F1
1	Is the nodule solid	0.53	1.00	0.69	1.00	1.00	1.00	0.63	1.00	0.77
2	Is the nodule soft tissue, semisolid or subsolid	0.54	1.00	0.70	0.64	1.00	0.78	0.44	1.00	0.61
3	Is the nodule ground glass	0.13	1.00	0.24	0.13	1.00	0.24	0.07	1.00	0.13
4	Is the nodule spiculated or irregular	0.75	0.75	0.75	0.50	0.75	0.60	0.33	1.00	0.50
5	Is the nodule calcified	0.95	0.62	0.75	0.96	0.87	0.91	0.92	0.59	0.72
6	Nodule location	0.91	1.00	0.95	0.93	1.00	0.96	0.93	1.00	0.96
7	Nodule size	0.89	1.00	0.94	0.90	1.00	0.95	0.87	1.00	0.93
		Gemini 1.5 Flash - P2			GPT-4-o - P2			Llama-3 - 70 B - P2		
		P	R	F1	P	R	F1	P	R	F1
1	Is the nodule solid	0.67	1.00	0.80	0.83	1.00	0.91	1.00	1.00	1.00
2	Is the nodule soft tissue, semisolid or subsolid	0.64	1.00	0.78	0.64	1.00	0.78	0.54	1.00	0.70
3	Is the nodule ground glass	0.14	1.00	0.25	0.14	1.00	0.25	0.10	1.00	0.18
4	Is the nodule spiculated or irregular	0.60	0.75	0.67	0.50	0.75	0.60	0.33	1.00	0.50
5	Is the nodule calcified	0.95	0.62	0.75	0.97	0.93	0.95	0.93	0.67	0.78
6	Nodule location	0.90	1.00	0.95	0.94	1.00	0.97	0.93	1.00	0.96
7	Nodule size	0.86	1.00	0.92	0.90	1.00	0.95	0.88	1.00	0.94
		Gemini 1.5 Flash - P3			GPT-4-o - P3			Llama-3 - 70 B - P3		
		P	R	F1	P	R	F1	P	R	F1
1	Is the nodule solid	0.67	1.00	0.80	0.59	1.00	0.74	0.77	1.00	0.87
2	Is the nodule soft tissue, semisolid or subsolid	0.54	1.00	0.70	0.58	1.00	0.74	0.47	1.00	0.64
3	Is the nodule ground glass	0.33	1.00	0.50	1.00	1.00	1.00	0.67	1.00	0.80
4	Is the nodule spiculated or irregular	0.60	0.75	0.67	0.75	0.75	0.75	0.44	1.00	0.62
5	Is the nodule calcified	0.97	0.92	0.94	0.97	0.93	0.95	0.95	0.98	0.97
6	Nodule location	0.91	1.00	0.95	0.94	1.00	0.97	0.95	1.00	0.97
7	Nodule size	0.90	1.00	0.95	0.92	1.00	0.96	0.92	1.00	0.96
		Gemini 1.5 Flash - P4			GPT-4-o - P4			Llama-3 - 70 B - P4		
		P	R	F1	P	R	F1	P	R	F1
1	Is the nodule solid	0.71	1.00	0.83	0.71	1.00	0.83	0.82	0.90	0.86
2	Is the nodule soft tissue, semisolid or subsolid	0.54	1.00	0.70	0.64	1.00	0.78	0.50	1.00	0.67
3	Is the nodule ground glass	0.33	1.00	0.50	1.00	1.00	1.00	0.67	1.00	0.80
4	Is the nodule spiculated or irregular	0.60	0.75	0.67	0.75	0.75	0.75	0.43	0.75	0.55
5	Is the nodule calcified	0.97	0.95	0.96	0.97	0.97	0.97	0.95	0.98	0.97
6	Nodule location	0,89	1,00	0.94	0.95	1.00	0.97	0.96	1.00	0.98
7	Nodule size	0.88	1.00	0.94	0.91	1.00	0.95	0.90	1.00	0.95
		Gemini 1.5 Flash - P5			GPT-4-o - P5			Llama-3 - 70 B - P5		
		P	R	F1	P	R	F1	P	R	F1
1	Is the nodule solid	0.83	1.00	0.91	0.83	1.00	0.91	1.00	1.00	1.00
2	Is the nodule soft tissue, semisolid or subsolid	0.50	1.00	0.67	0.54	1.00	0.70	0.54	1.00	0.70
3	Is the nodule ground glass	0.33	1.00	0.50	1.00	1.00	1.00	0.67	1.00	0.80
4	Is the nodule spiculated or irregular	0.60	0.75	0.67	0.60	0.75	0.67	0.60	0.75	0.67
5	Is the nodule calcified	0.97	0.92	0.94	0.97	0.97	0.97	0.97	0.95	0.96
6	Nodule location	0.92	1.00	0.96	0.94	1.00	0.97	0.97	1.00	0.98
7	Nodule size	0.90	1.00	0.95	0.92	1.00	0.96	0.91	1.00	0.95
		Gemini 1.5 Flash - P6			GPT-4-o - P6			Llama-3 - 70 B - P6		
		P	R	F1	P	R	F1	P	R	F1
1	Is the nodule solid	0.91	1.00	0.95	0.63	1.00	0.77	1.00	1.00	1.00
2	Is the nodule soft tissue, semisolid or subsolid	0.54	1.00	0.70	0.58	1.00	0.74	0.58	1.00	0.74
3	Is the nodule ground glass	0.33	1.00	0.50	1.00	1.00	1.00	0.67	1.00	0.80
4	Is the nodule spiculated or irregular	0.60	0.75	0.67	0.60	0.75	0.67	0.50	0.75	0.60
5	Is the nodule calcified	0.98	0.92	0.95	0.97	0.95	0.96	0.97	0.97	0.97
6	Nodule location	0.92	1.00	0.96	0.93	1.00	0.96	0.97	1.00	0.98
7	Nodule size	0.90	1.00	0.95	0.90	1.00	0.95	0.91	1.00	0.95

TABLE V
EFFECTIVENESS BIOBERTPT X LLM MODELS

					Gemini 1.5 Flash - P6			GPT-4-o - P4			Llama-3 70B - P6		
Entity	P	R	F1	Question	P	R	F1	P	R	F1	P	R	F1
Attenuation	0.88	0.78	0.82	1-3	0.59	1.00	0.72	0.78	1.00	**0.87**	0.75	1.00	0.85
Edges	0.67	0.43	0.52	4	0.60	0.75	0.67	0.75	0.75	**0.75**	0.50	0.75	0.60
Calcification	0.98	0.98	**0.98**	5	0.98	0.92	0.95	0.97	0.97	0.97	0.97	0.97	0.97
Localization	0.89	0.88	0.88	6	0.92	1.00	0.96	0.95	1.00	0.97	0.97	1.00	**0.98**
Size	0.92	0.83	0.87	7	0.90	1.00	**0.95**	0.91	1.00	**0.95**	0.91	1.00	**0.95**

future work, we plan to explore more advanced techniques, such as the chain-of-thought method [27], to assess whether

they yield improved effectiveness.

The results underscore the potential of LLMs, particularly GPT-4-o, as valuable tools in assisting radiologists with the extraction of lung nodule characteristics for accurate Lung-RADS classification. By automating the extraction process, these models can significantly reduce the time and effort required for radiologists to interpret CT reports, allowing them to focus more on patient care. Furthermore, the increased accuracy and consistency in Lung-RADS scoring facilitated by these models can lead to more precise follow-up recommendations. Patients benefit from timely and accurate diagnoses, which are critical for early detection and effective treatment of lung cancer. This advancement in NLP thus holds promise for enhancing the efficiency and effectiveness of lung cancer screening programs, contributing to better healthcare delivery.

IV. Conclusion

In this study, we utilized three LLMs and a BERT model to extract structured information from unstructured radiology reports. We found that the effectiveness of LLMs improved when using Prompts that included prior medical knowledge and reports, whereas these models previously had difficulty extracting information. The best effective model was GPT-4-o using Prompt 4, surpassing even the BioBERTpt model fine-tuned with a batch size of 8, 10 epochs, and a learning rate of 1e-3, which served as the baseline model. GPT-4-o with few-shot learning using Prompt 4 emerged as the best model, achieving an F1-score of 0.89, outperforming Gemini 1.5 Flash with few-shot learning using Prompt 6 (F1-score of 0.81), Llama-3 70B with few-shot learning using Prompt 6 (F1-score of 0.86), and BioBERTpt model (F1-score of 0.83).

Our results highlight the potential of LLMs to assist radiologists in accurately classifying lung nodules according to the Lung-RADS criteria, streamlining the diagnostic process. By automating and enhancing the accuracy of information extraction, these models are expected to reduce the workload for radiologists and improve the consistency of follow-up recommendations. Finally, we hope this will benefit patients by enabling more timely and precise detection and management of lung cancer.

This work was supported by the following research agencies: Ministry of Health, CNPq, SESAU-AL and Alagoas Research Foundation (FAPEAL).

References

[1] Suhana Bedi, Sneha S. Jain, and Nigam H. Shah. Evaluating the clinical benefits of llms. *Nature Medicine*, Jul 2024.

[2] Sebastian E Beyer, Brady J McKee, Shawn M Regis, Andrea B McKee, Sebastian Flacke, Gilan El Saadawi, and Christoph Wald. Automatic Lung-RADS™ classification with a natural language processing system. *J Thorac Dis*, 9(9):3114–3122, September 2017.

[3] Mark E Deffebach and Linda Humphrey. Lung cancer screening. *Surg Clin North Am*, 95(5):967–978, October 2015.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[5] Abhimanyu Dubey et al. The llama 3 herd of models, 2024.

[6] Dexter P. Mendoza et al. Lung-rads category 3 and 4 nodules on lung cancer screening in clinical practice. *American Journal of Roentgenology*, 219(1):55–65, 2022. PMID: 35080453.

[7] Elisa Terumi Rubel Schneider et al. BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online, November 2020. Association for Computational Linguistics.

[8] Gemini Team et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024.

[9] Gerard A. Silvestri et al. Outcomes from more than 1 million people screened for lung cancer with low-dose ct imaging. *CHEST*, 164(1):241–251, 2023.

[10] Heber MacMahon et al. Guidelines for management of incidental pulmonary nodules detected on ct images: From the fleischner society 2017. *Radiology*, 284(1):228–243, 2017. PMID: 28240562.

[11] Hiroki Nakayama et al. doccano: Text annotation tool for human, 2018. Software available from https://github.com/doccano/doccano.

[12] Michael K. Gould et al. Evaluation of individuals with pulmonary nodules: When is it lung cancer?: Diagnosis and management of lung cancer, 3rd ed: American college of chest physicians evidence-based clinical practice guidelines. *Chest*, 143(5, Supplement):e93S–e120S, 2013.

[13] OpenAI et al. Gpt-4 technical report, 2024.

[14] Thomas Wolf et al. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[15] Yan Hu et al. Improving large language models for clinical named entity recognition via prompt engineering, 2024.

[16] Tarcísio Lima Ferreira, Marcelo Costa Oliveira, and Thales Miranda De Almeida Vieira. Lung-rads + ai: A tool for quantifying the risk of lung cancer in computed tomography reports. In *2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 292–297, 2023.

[17] Amir Gandomi, Eusha Hasan, Jesse Chusid, Subroto Paul, Matthew Inra, Alex Makhnevich, Suhail Raoof, Gerard Silvestri, Brett C. Bade, and Stuart L. Cohen. Evaluating the accuracy of lung-rads score extraction from radiology reports: Manual entry versus natural language processing. *International Journal of Medical Informatics*, 191:105580, 2024.

[18] Danqing Hu, Bing Liu, Xiaofeng Zhu, Xudong Lu, and Nan Wu. Zero-shot information extraction from radiological reports using chatgpt. *International Journal of Medical Informatics*, 183:105321, 2024.

[19] Kory Kreimeyer, Matthew Foster, Abhishek Pandey, Nina Arya, Gwendolyn Halford, Sandra F Jones, Richard Forshee, Mark Walderhaug, and Taxiarchis Botsis. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *Journal of Biomedical Informatics*, 73:14–29, 2017.

[20] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019.

[21] Peter J. Mazzone, Gerard A. Silvestri, Lesley H. Souter, Tanner J. Caverly, Jeffrey P. Kanne, Hormuzd A. Katki, Renda Soylemez Wiener, and Frank C. Detterbeck. Screening for lung cancer: Chest guideline and expert panel report. *Chest*, 160(5):e427–e494, 2021.

[22] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey, 2024.

[23] National Lung Screening Trial Research Team. Lung cancer incidence and mortality with extended follow-up in the national lung screening trial. *J Thorac Oncol*, 14(10):1732–1742, June 2019.

[24] American College of Radiology. Lung-rads® v2022. https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Lung-Rads. Accessed: 2023-05-01.

[25] Rebecca L Siegel, Kimberly D Miller, Nikita Sandeep Wagle, and Ahmedin Jemal. Cancer statistics, 2023. *CA Cancer J Clin*, 73(1):17–48, January 2023.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[27] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.