Optimizing Clinical Data Availability: Extracting Pulmonary Embolism Diagnoses from Radiology Impressions with GPT-4o

Mohammed Mahyoub[1,2*], Kacie Dougherty[1], Ajit Shukla[1*]

[1]Advanced Analytics and Solutions, Virtua Health, Marlton, New Jersey 08053

[2]School of Systems Science and Industrial Engineering, Binghamton University, Binghamton, New York, 13902

*Correspondence: mmahyoub@virtua.org; ashukla@virtua.org

Abstract

Background: Pulmonary embolism (PE) is a life-threatening condition that requires timely diagnosis to reduce mortality. Radiology reports, particularly the Impression sections, play a critical role in diagnosing PE. However, manually extracting this information from large volumes of reports is challenging. This study aims to develop an advanced natural language processing (NLP) system using GPT-4o to automatically extract PE diagnoses from radiology report impressions, enhancing clinical workflows and decision-making.

Materials and Methods: We developed two text classification models: a fine-tuned Clinical Longformer (as a baseline model) and GPT-4o. Models were trained using 1,000 radiology report impressions and validated on 200 samples, with a post-deployment evaluation conducted using 500 operational records. The primary dataset was sourced from an electronic medical record relational database, and key metrics such as sensitivity, specificity, and F1 score were used to evaluate model performance.

Results: GPT-4o achieved superior performance with 100% sensitivity, specificity, and F1 score, outperforming the Clinical Longformer. Post-deployment, GPT-4o continued to perform flawlessly, identifying all positive PE cases without false positives or false negatives. The model successfully streamlined the clinical workflow, reducing the burden of manual review and enhancing diagnostic accuracy.

Keywords: pulmonary embolism, natural language processing, GPT-4o, Clinical Longformer, text classification, radiology reports

1. Introduction

Pulmonary embolism (PE) is a serious medical condition where a blood clot blocks one of the pulmonary arteries in the lungs, typically originating from a vein in the lower limbs [1], [2], [3]. This blockage can significantly impede blood flow, leading to reduced oxygen levels in the blood and potential lung tissue damage. PE is critical because it can cause sudden, life-threatening complications such as cardiac dysfunction and other acute admissions [4], [5]. Prompt diagnosis and treatment are crucial to improve outcomes and reduce the risk of mortality [6], [7].

Clinical imaging techniques commonly used for diagnosing pulmonary embolisms include pulmonary computed tomography angiography (CTA), combined CT venography and pulmonary angiography (CVPA), and multi-detector CT angiography (MDCTA) [8], [9], [10]. The analysis and outcomes of these modalities are recorded in radiology reports which describe the presence or absence of emboli, their location, size, and impact on pulmonary circulation. Radiology reports are structured documents that capture the conditions observed from radiology images [11], [12]. Typically, the most important parts of these reports are the Findings and Impression sections [13].

The Impression section provides a clinically precise summary of the patient's status, typically summarizing the key findings and diagnoses from the Findings section [14]. Therefore, the diagnosis of PE is highly likely to be mentioned in the Impression section. Early documentation of PE and its extraction in the (electronic medical record) EMR system, and consequently in clinical workflows, is crucial for improving patient outcomes. In this study, we aim to develop an advanced transformer-based text classification model to extract PE diagnoses from the Impression section of radiology reports, expediting structured data availability and enhancing quality of care through evidence-based practices.

Natural Language Processing (NLP) techniques have been increasingly utilized in the field of radiology, particularly in extracting critical information from radiology reports such as diagnoses [15]. Studies have shown that NLP, combined with machine learning and deep learning algorithms, can effectively extract relevant information from radiology reports [16], [17], [18]. These techniques enable the automatic identification and extraction of critical findings such as pleural effusion, pulmonary infiltrate, and pneumonia, aiding in the classification of reports consistent with bacterial pneumonia [19]. Furthermore, NLP algorithms have been developed to detect specific findings like acute pulmonary embolism in radiology reports, showcasing the potential of NLP in enhancing diagnostic processes [20], [21].

The application of NLP in radiology reports extends to various medical conditions, including pulmonary embolism. Studies have demonstrated the effectiveness of NLP in structuring the content of radiology reports, thereby increasing their value and aiding in the classification of pulmonary oncology according to the TNM classification system, a standard for staging cancer

[22]. Additionally, NLP has been used to identify ureteric stones in radiology reports and to build cohorts for epidemiological studies, showcasing the versatility of NLP in medical research [23].

Recent studies have demonstrated the effectiveness of Clinical-Longformer in various clinical NLP tasks. For instance, it has been utilized to identify incarceration status from medical records, showcasing good sensitivity and specificity compared to traditional keyword-based methods [24]. Additionally, Clinical-Longformer has been successfully applied in the classification of clinical notes for automated ICD coding, where it outperformed other models in accuracy [25], [26]. This capability to accurately interpret and classify clinical text is crucial for improving healthcare delivery and ensuring proper coding for reimbursement purposes.

On the other hand, advanced versions of the GPT family like GPT-4 and GPT-4o, generative language model, has been recognized for their versatility in clinical applications, particularly in generating and summarizing clinical information [27], [28]. Its multimodal capabilities allow it to process not only text but also images and audio, enhancing its utility in diverse clinical settings [29]. GPT-4 has been employed in clinical trial matching, where it automates eligibility screening, thus streamlining the recruitment process for clinical studies [30].

This study aims to develop an advanced NLP system to automatically extract pulmonary embolism diagnoses from radiology report impressions. The key contributions of this study are:

- Enhance and accelerate clinical data availability to improve the quality of care through evidence-based approaches.
- Develop an advanced NLP system tailored for clinical text, which examines two technologies: Clinical Longformer and GPT-4o.
- Deploy the developed system as a cloud-based web application, addressing a gap often found in Clinical AI research.

- Evaluate the model both before and after deployment.

2. Methodology

In this section, we provide a comprehensive overview of the study's methodology. Subsequently, we explore the text classification approach, followed by a detailed description of the dataset utilized. We then describe the models applied in this research. Additionally, we discuss the deployment pipeline of the selected model. Finally, we outline the evaluation metrics employed to assess the model's performance.

2.1. Overview

The primary objective of this study is to develop and deploy an AI solution capable of extracting Pulmonary Embolism (PE) diagnoses from radiology report impressions. After defining this goal, the research proceeds through four distinct phases. In the first step, we determine the appropriate data sources, match the data fields to the clinical database, and extract them for further analysis. This is followed by preprocessing and transforming the data to make it suitable for model development. The second step involves creating and testing different models, then choosing the one with the best results to proceed. The selected model is then implemented during the third step. In the final step, we track the model's performance in real-world conditions and evaluate how it affects operational outcomes.

2.2. Radiology impressions text classification

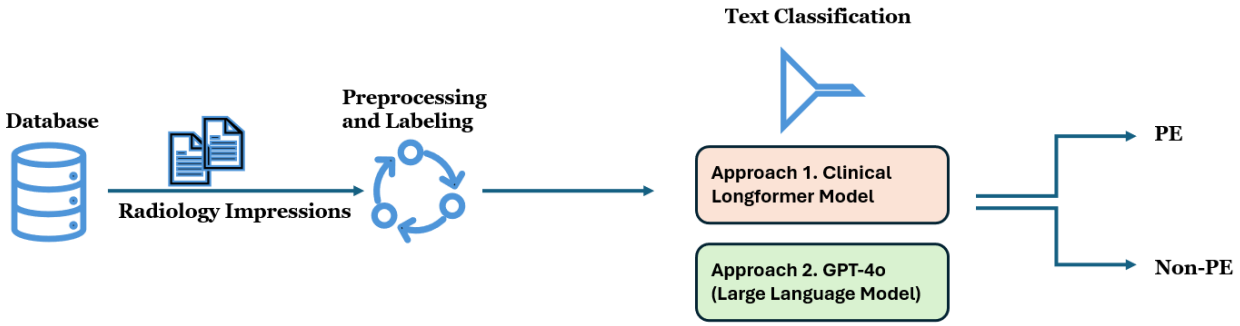


Figure 1. Radiology impressions text classification.

Text classification is a fundamental task in natural language processing (NLP) that involves categorizing text into predefined labels based on its content. This task is widely used in applications such as sentiment analysis, spam detection, and medical report classification. Text classification models typically preprocess the data by tokenizing the text and transforming it into numerical representations suitable for machine learning. Various approaches, such as rule-based methods, machine learning algorithms, or deep learning models, are then employed to make predictions based on patterns in the text.

Figure 1 illustrates the process of classifying radiology report impressions to identify PE cases, which is adopted in this study. The workflow begins with the extraction of radiology impressions from a clinical database. The impressions are preprocessed by consolidating line-wise text, removing unnecessary spaces, and applying labels to prepare the data for analysis. Following this, two different text classification models are employed: Approach 1 utilizes a Clinical Longformer Model, while Approach 2 involves GPT-4o, a large language model. Both models

classify the impressions into two categories: PE and Non-PE. The goal is to determine whether a diagnosis of PE is present in each radiology report impression.

2.3. Data

The data used in this study was sourced from the electronic medical record relational database, with the primary data element being the impressions of radiology reports. These impressions, which contain key diagnostic information, were consolidated from line-wise data and cleaned to remove extraneous spaces. This process ensured that the data was formatted appropriately for analysis and modeling.

The training dataset consists of 1,000 samples, which were randomly selected from radiology reports generated between January 1, 2024, and June 30, 2024. For the validation dataset, 200 samples were randomly drawn from radiology reports collected in July 2024. Additionally, a separate testing dataset, consisting of 500 observations, was sampled randomly from operational data received between August 1, 2024, and August 31, 2024. The characteristics of the training and validation datasets are outlined in Table 1. The testing dataset characteristics will be discussed in the following section.

As shown in Table 1, the training dataset contains 1,000 observations, with an average of 43.64 words (or 32.73 tokens, where a token is approximately three-fourths of a word) per report impression. The training data includes 235 occurrences of pulmonary embolism term, with 36 positive cases for pulmonary embolism and 964 negative cases. The validation dataset, consisting of 200 observations, has a slightly lower average word count per report impression, at 40.18 words. There are 46 occurrences of pulmonary embolism term in the validation dataset. Also, there are 8 positive cases and 192 negative cases.

Table 1. Training and Validation Data Characteristics.

Metric	**Training Dataset**	**Validation Dataset**
Number of observations	1,000	200
Average number of words (a token is approximately $\frac{3}{4}$ of a word)	43.64	40.18
Average number of pulmonary emboli/embolism occurrences	235	46
Average number of positive cases for pulmonary embolism	36	8
Average number of negative cases for pulmonary embolism	964	192

2.4. Fine-tuned Clinical Longformer classifier

The Clinical Longformer is a specialized transformer model designed to handle long clinical documents, overcoming the typical limitations of standard transformer models such as BERT, which can process sequences up to 512 tokens [31]. Clinical Longformer incorporates a sparse attention mechanism that allows it to efficiently process sequences up to 4,096 tokens, making it ideal for handling lengthy clinical narratives. Pre-trained on large clinical datasets, it is particularly effective in capturing long-term dependencies in medical text. In this study, the Clinical Longformer is fine-tuned to classify radiology impressions for identifying pulmonary embolism, leveraging its ability to process comprehensive radiology reports impressions without truncating important contextual information.

We fine-tuned the Clinical Longformer model on a GPU server with 48 GB of memory. The fine-tuning parameters were as follows:

- Batch size: 4

- Gradient accumulation steps: 8
- Learning rate: 2e-5
- Number of epochs: 5
- Optimizer: AdamW
- Learning rate scheduler: Linear

2.5. GPT-4o classifier

The methodology for utilizing GPT-4o in the text classification of radiology impressions, specifically for PE diagnosis, is based on a combination of chain-of-thought (COT) reasoning and few-shot learning techniques. As outlined in Figure 2(a), the process begins by initializing an empty list to store the generated labels. GPT-4o is then prompted using a COT and few-shot learning template, where relevant examples of radiology impressions with their corresponding labels (PE or Non-PE) are presented to the model. The temperature parameter is set to zero to minimize randomness in the model's predictions. For each radiology impression in the dataset, the system inserts the impression into the prompt, calls the GPT-4o API, and receives a response that indicates whether a PE diagnosis is present. The resulting labels are appended to the list for further analysis and validation.

As shown in Figure 2(b), the prompt includes a persona where GPT-4o is defined as a clinical AI assistant proficient in radiology, capable of interpreting complex medical language. The prompt further provides detailed steps, starting with studying example impression-label pairs, followed by reading through the target impression to extract potential diagnoses. The model is tasked with determining whether PE is indicated in the impression and returns the output as a structured JSON object. This methodology leverages GPT-4o's advanced language comprehension

capabilities to classify radiology reports efficiently, using both clinical reasoning and context learned from the few-shot examples.

2.6. Deployment pipeline

The deployment pipeline for the pulmonary embolism (PE) classification model, illustrated in Figure 3(a), integrates a combination of on-premises and Azure cloud services to create a streamlined and scalable system. The process begins with data being sourced from an on-premises SQL server, which stores radiology report impressions. These impressions are transferred to an Azure SQL database, where they are stored and prepared for further analysis. This architecture utilizes a direct interaction between Azure SQL, an Azure Web App, and the Azure OpenAI service. The Azure OpenAI service, hosting the GPT-4o model, is invoked by the Azure Web App to perform text classification on the radiology impressions and return pulmonary embolism classification results. These results are then stored back in the Azure SQL database. The web app fetches the results from Azure SQL and displays them for end users.

As shown in Figure 3(b), the web app was built using Python Flask for the backend, along with HTML, CSS, and JavaScript for the frontend. The interface allows users to query the system by submitting a patient's medical record number to retrieve the corresponding PE classification result. Users can also refresh the data or download the results for further analysis. The table on the right displays relevant patient information, including patient IDs, encounter IDs, admission times, and the PE classification results. This interface serves as a convenient tool for healthcare professionals to quickly identify patients with a PE diagnosis, improving clinical decision-making and patient outcomes by providing prompt, automated insights.

Algorithm Chain of Thought (COT) and Few-Shot Learning Pulmonary Embolism Classification

Input: List of impressions

Output: List of labels (Yes if PE presents and No otherwise)

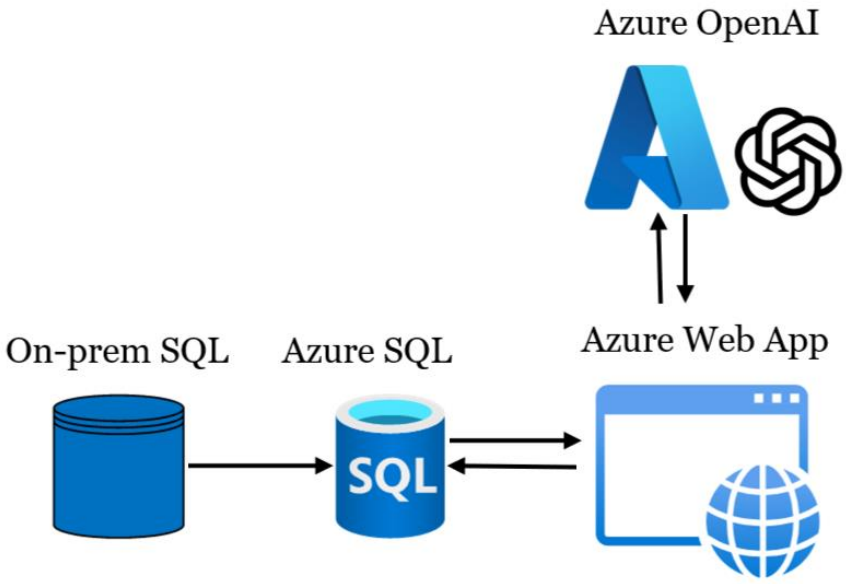
1: Initialize an empty list for extracted labels.
   labels = [ ]
2: Initialize COT and few-shot learning prompt template.
3: Set the temperature parameter to 0.0.
4: **for** each item in the list of impressions **do**
5: Insert the current impressions item in the prompt.
6: Send the formatted prompt and temperature parameter to the GPT-4o model (API call)
7: Parse the response.
8: Append the generated label: *labels.append(response["answer"]*.
9: **end for**
10: Return the list of generated labels.

(a)

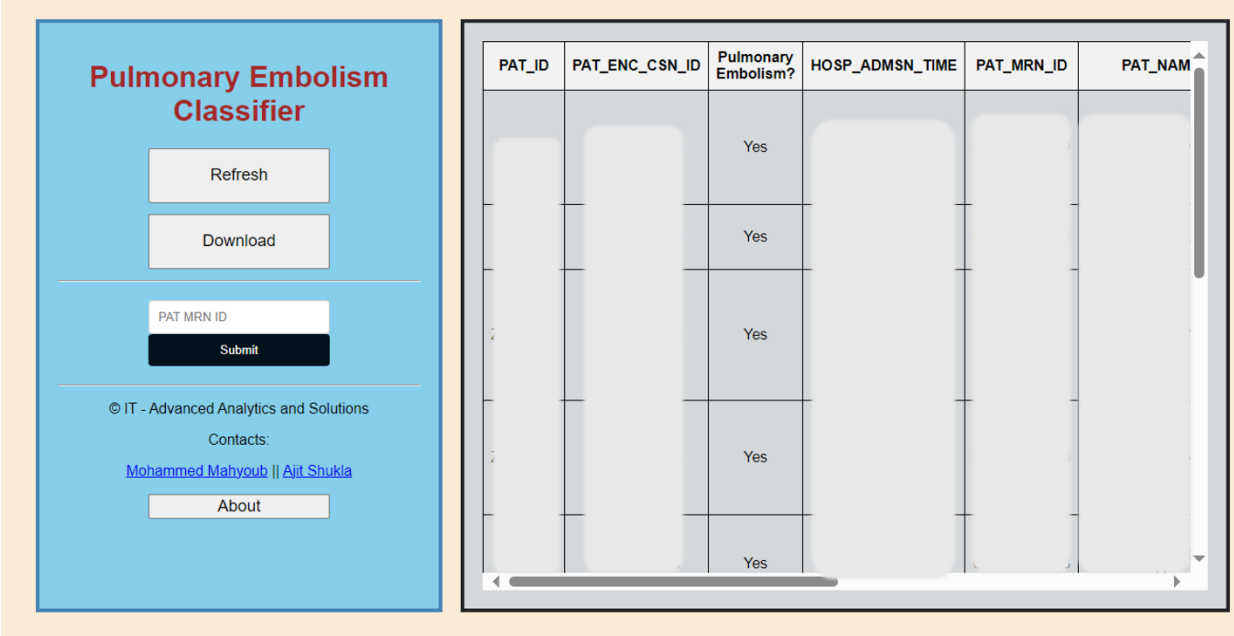
```
prompt_template = '''
# Persona
You are a clinical AI assistant who is expert in radiology. You are capable of annotating clinical text.

# Steps:
- Study the impression-label examples.
    - Examples
        Example 1:
            Impression: No acute inflammatory process is present within the thorax.
            Label: No
        Example 2:
            Impression: 1. No evidence for noncalcified pulmonary nodule or thoracic  adenopathy.
            2.    Evidence of prior granulomatous disease. 3. Stable 1.2 x 1.3 cm left adrenal adenoma.
            LUNG-RADS CATEGORY:  1, Negative (risk of malignancy < 1%). MANAGEMENT:
            If the patient continues to meet the criteria for lung cancer screening,
            recommend annual screening LDCT in 12 months.
            Label: No
        Example 3:
            Impression: 1.  Multiple right-sided pulmonary emboli .
            No findings to suggest right ventricular strain. Other findings as above.
            Label:Yes
- Read through the clinical impression of the radiology report.
- Extract possible diagnoses.
- Determine if the impressions has pulmonary embolism or not.
- Return the answer in JSON object with an 'answer' key that labels the following impression with:
    * Yes (if the impression has a pulmonary embolism diagnosis)
    * No (if the impression does not indicate pulmonary embolism diagnosis)

Impression: {impression}
'''
```

(b)

Figure 2. GPT-4o for radiology impressions classification. Extraction of pulmonary embolism diagnosis.



(a)



(b)

Figure 3. Deployment pipeline and consumption Web App. Only positive cases are displayed.

2.7. Evaluation metrics

To evaluate the performance of the pulmonary embolism (PE) classification model, we employed several commonly used metrics:

- *Confusion Matrix*: A table that summarizes the model's predictions by showing the number of true positives (correctly predicted PE cases), true negatives (correctly predicted non-PE cases), false positives (non-PE cases incorrectly classified as PE), and false negatives (PE cases incorrectly classified as non-PE). This matrix provides a detailed view of model performance.
- *Sensitivity (Recall)*: The proportion of actual PE cases that the model correctly identified. It measures the model's ability to detect positive cases (PE) and is defined as the ratio of true positives to the sum of true positives and false negatives.

  Specificity: The proportion of actual non-PE cases that the model correctly identified. It reflects the model's ability to avoid false positives, calculated as the ratio of true negatives to the sum of true negatives and false positives.
- *Precision*: The proportion of predicted PE cases that were correctly identified. It measures the accuracy of the model's positive predictions and is calculated as the ratio of true positives to the sum of true positives and false positives.
- *F1 Score*: A harmonic mean of precision and recall, which provides a balanced measure of the model's performance, especially in cases of imbalanced data. It is particularly useful for evaluating the trade-off between precision and recall in the context of PE classification.

3. Results and Discussion

This section presents the research findings. First, we evaluate the Clinical Longformer and GPT-4o models using the validation dataset during the development phase. Next, we assess the performance of the deployed model (GPT-4o) post-deployment. Lastly, we discuss the benefits and clinical implications of the pulmonary embolism classifier.

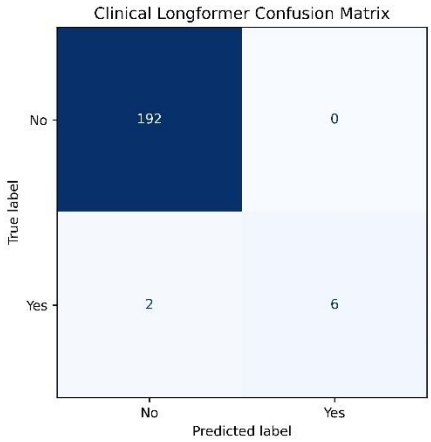
3.1. Models evaluation

Figure 4 illustrates the evaluation of two models, the fine-tuned Clinical Longformer and GPT-4o, on the validation dataset for the task of pulmonary embolism extraction from radiology report impressions. Figure 4(a) and Figure 4(b) present the confusion matrices for each model, highlighting their classification performance. The Clinical Longformer (a) misclassified two positive cases as negative (false negatives), achieving a sensitivity of 75%. Meanwhile, GPT-4o (b) perfectly classified all cases, achieving a sensitivity of 100%. Both models demonstrated flawless classification of negative cases, with a specificity of 100%. These confusion matrices suggest that GPT-4o excels in capturing all positive instances of pulmonary embolism.

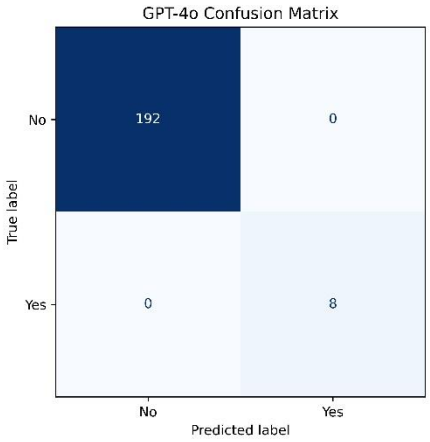
In Figure 4(c), the performance metrics across both models are compared. The Clinical Longformer achieves an F1 score of 0.86, reflecting a balance between precision and recall, with its lower sensitivity (0.75) slightly lowering its overall performance. In contrast, GPT-4o's F1 score is a perfect 1.0, indicating superior performance in identifying pulmonary embolism cases. The 100% accuracy and perfect metrics across sensitivity, specificity, and F1 score suggest that GPT-4o is more reliable in the critical task of detecting pulmonary embolisms from unstructured radiology reports – impressions section, minimizing the risk of missing positive cases and ensuring more comprehensive clinical decision support in a real-world setting.

Table 2. Post-Deployment Testing Data Characteristics. A random sample of 500 records was taken from the operational dataset in August 2024 for post-deployment evaluation.

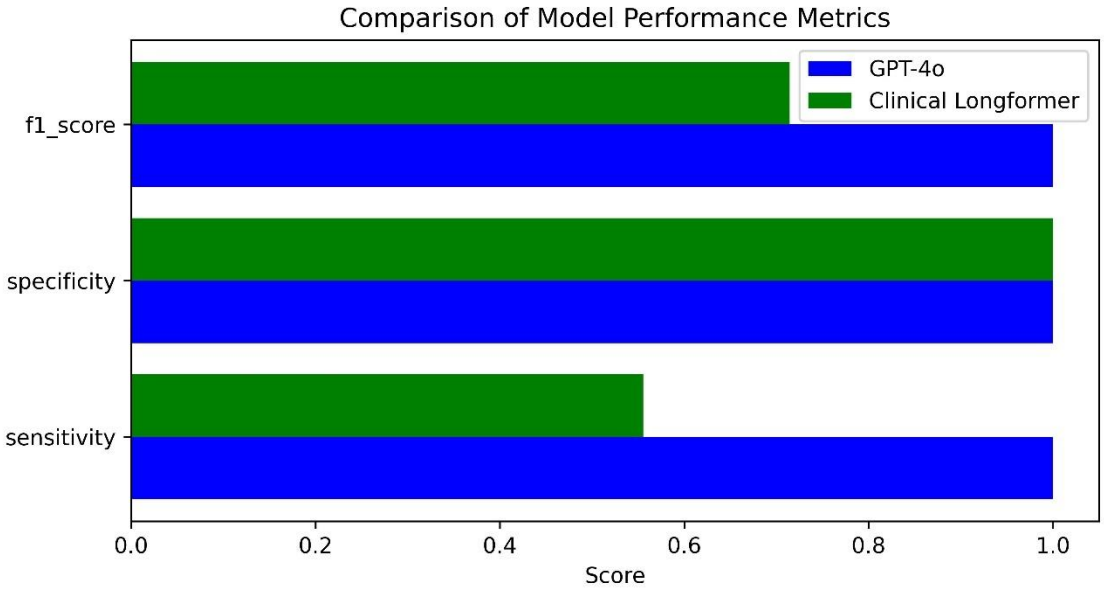
Metric	**Value**
Number of observations	500
Average number of words (a token is approximately $\frac{3}{4}$ of a word)	43.76
Average number of pulmonary emboli/embolism occurrences	106
Average number of positive cases for pulmonary embolism	18
Average number of negative cases for pulmonary embolism	482



(a)



(b)



(c)

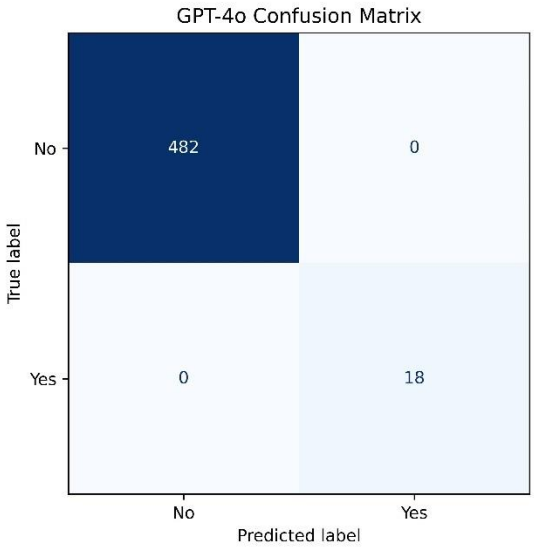
Figure 4. Evaluation of Models on the Validation Dataset (Pre-Deployment). (a) Confusion matrix for the fine-tuned Clinical Longformer. (b) Confusion matrix of GPT-4o. (c) Comparison of validation metrics across both models.

3.2. Post-deployment performance

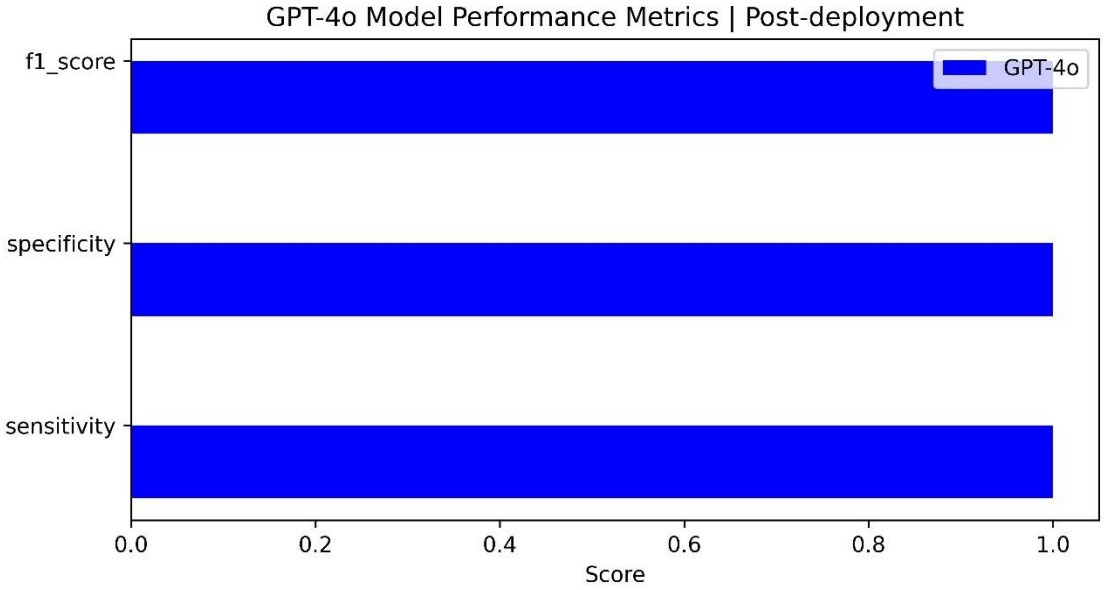
Based on GPT-4o's exceptional performance during the validation phase, where it achieved perfect metrics across all categories, it was selected for deployment in the operational setting. The post-deployment evaluation of the GPT-4o model was conducted using a randomly selected dataset of 500 records from the operational data collected in August 2024. Table 2 provides a summary of the dataset characteristics, including 18 positive cases of pulmonary embolism and 482 negative cases. The dataset also contained 106 mentions of pulmonary embolism-related terms, reflecting the richness and complexity of the radiology reports used for evaluation.

Figure 5 presents the post-deployment performance of GPT-4o. The confusion matrix in subfigure (a) shows that the model correctly classified all cases, with no false positives or false negatives. The model achieved perfect sensitivity (1.0), indicating that it successfully identified all 18 positive cases of pulmonary embolism. Likewise, the model's specificity was also 1.0, as it correctly classified all 482 negative cases.

The overall performance metrics of GPT-4o, as shown in subfigure (b), demonstrate the robustness of the model in the real-world setting. With an F1 score of 1.0, the model maintains an optimal balance between precision and recall, providing confidence that it can accurately detect pulmonary embolism cases in practice. These results indicate that GPT-4o continues to perform exceptionally well post-deployment, offering reliable support in the classification of pulmonary embolism from radiology reports (impressions), a critical task in the clinical setting.



(a)



(b)

Figure 5. Post-Deployment Evaluation of GPT-4o. (a) Confusion matrix of GPT-4o on the post-deployment dataset. (b) Performance metrics of GPT-4o, including F1 score, sensitivity, and specificity.

3.3. Operational and clinical implications

The post-deployment results of GPT-4o highlight several important operational and clinical implications. First, the model's ability to maintain high sensitivity and specificity in a real-world setting ensures that it can reliably detect pulmonary embolism cases without missing any true positives or generating false alarms. This is critical in a clinical environment where missed cases of pulmonary embolism can lead to severe consequences for patient outcomes, while false positives can result in unnecessary follow-up tests or treatments.

From an operational standpoint, the model's flawless performance on both the validation and post-deployment datasets minimizes the need for manual review, streamlining the workflow for radiologists and other healthcare professionals. By accurately classifying cases, GPT-4o reduces the cognitive and time burden on clinicians, allowing them to focus their attention on more complex cases or other clinical tasks. Additionally, the model's high precision ensures that healthcare resources are used more efficiently, reducing unnecessary interventions and improving overall patient care.

Clinically, the deployment of GPT-4o enhances the decision support available to clinicians, providing a reliable tool for the timely detection of pulmonary embolism from radiology reports. This early identification can lead to faster diagnosis and treatment, improving patient outcomes and potentially reducing mortality. Furthermore, the consistent performance of GPT-4o in identifying pulmonary embolism across different datasets demonstrates its robustness and adaptability, making it a valuable asset in diverse clinical settings.

4. Conclusion

In conclusion, this study presents an efficient approach to automating the extraction of pulmonary embolism (PE) diagnoses from radiology report impressions using advanced natural language processing (NLP) models. By comparing the performance of a fine-tuned Clinical Longformer and GPT-4o, we demonstrated that GPT-4o outperforms in terms of sensitivity, specificity, and overall accuracy, both pre- and post-deployment. The deployment of GPT-4o within a clinical setting offers significant operational and clinical advantages, including the reduction of manual review, enhanced clinical decision support, and the timely detection of PE cases. This model's robustness in handling real-world clinical data suggests it can play a crucial role in improving patient outcomes by providing more accurate and faster diagnostic insights. Future work may explore expanding this approach to other medical conditions and further refining the integration of NLP-based models into clinical workflows to continue improving the quality and efficiency of healthcare delivery.

Author Contributions

Mohammed Mahyoub: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. *Kacie Dougherty*: Writing – review & editing, Formal analysis, Validation. *Ajit Shukla*: Project administration, Resources, Writing – review & editing, Validation.

Data Availability Statement

The data analyzed in this study is subject to the following licenses/ restrictions: data in the present study are not available due to agreements made with the IRB of Virtua Health.

Ethics Statement

The studies involving humans were approved by Virtua Health Institutional Review Board FWA00002656. The studies were conducted in accordance with the local legislation and institutional requirements. The ethics committee/institutional review board waived the requirement of written informed consent for participation from the participants or the participants' legal guardians/next of kin because the research involved no more than minimal risk to subjects, could not practically be carried out without the waiver, and the waiver will not adversely affect the rights and welfare of the subjects. This requirement of consent was waived on the condition that, when appropriate, the subjects will be provided with additional pertinent information about participation.

References

[1] A. H. Tanra, A. T. Lopa, T. Esa, and D. E. Rauf, "Diagnostic Value of Platelet Indices in Patients with Pulmonary Embolism," *Indones. J. Clin. Pathol. Med. Lab.*, vol. 27, no. 1, pp. 22–26, 2020.

[2] W. Deng and W. Gao, "Cathepsin causal association with pulmonary embolism: a Mendelian randomization analysis," 2024, Accessed: Aug. 14, 2024. [Online]. Available: https://www.researchsquare.com/article/rs-4191858/latest

[3] M. D. Lyhne, J. A. Kline, J. E. Nielsen-Kudsk, and A. Andersen, "Pulmonary vasodilation in acute pulmonary embolism – a systematic review," *Pulm. Circ.*, vol. 10, no. 1, pp. 1–16, Jan. 2020, doi: 10.1177/2045894019899775.

[4] S.-L. Zhang *et al.*, "Case Report: Resuscitation of patient with tumor-induced acute pulmonary embolism by venoarterial extracorporeal membrane oxygenation," *Front. Cardiovasc. Med.*, vol. 11, p. 1322387, 2024.

[5] G. Grusova, L. Lambert, J. Zeman, A. Lambertova, and J. Benes, "The additional value of esophageal wall evaluation and secondary findings in emergency patients undergoing CT pulmonary angiography," *Iran. J. Radiol.*, vol. 15, no. 1, 2018, Accessed: Aug. 14, 2024. [Online]. Available: https://brieflands.com/articles/iranjradiol-63466.html

[6] C. Becattini, M. C. Vedovati, and G. Agnelli, "Diagnosis and prognosis of acute pulmonary embolism: focus on serum troponins," *Expert Rev. Mol. Diagn.*, vol. 8, no. 3, pp. 339–349, May 2008, doi: 10.1586/14737159.8.3.339.

[7] J. Simpson and A. López-Candales, "Elevated Brain Natriuretic Peptide and Troponin I in a Woman with Generalized Weakness and Chest Pain," *Echocardiography*, vol. 22, no. 3, pp. 267–271, Mar. 2005, doi: 10.1111/j.0742-2822.2005.03192.x.

[8] Y. Zhou, H. Shi, Y. Wang, A. R. Kumar, B. Chi, and P. Han, "Assessment of correlation between CT angiographic clot load score, pulmonary perfusion defect score and global right ventricular function with dual-source CT for acute pulmonary embolism," *Br. J. Radiol.*, vol. 85, no. 1015, pp. 972–979, 2012.

[9] B. Lapergue *et al.*, "Diagnostic yield of venous thrombosis and pulmonary embolism by combined CT venography and pulmonary angiography in patients with cryptogenic stroke and patent foramen ovale," *Eur. Neurol.*, vol. 74, no. 1–2, pp. 69–72, 2015.

[10] H. Yuan, Y. Shao, Z. Liu, and H. Wang, "An improved faster R-CNN for pulmonary embolism detection from CTPA images," *IEEE Access*, vol. 9, pp. 105382–105392, 2021.

[11] J. D. Segrelles, R. Medina, I. Blanquer, and L. Martí-Bonmatí, "Increasing the Efficiency on Producing Radiology Reports for Breast Cancer Diagnosis by Means of Structured Reports: A Comparative Study," *Methods Inf. Med.*, vol. 56, no. 03, pp. 248–260, 2017, doi: 10.3414/ME16-01-0091.

[12] J. M. Nobel, E. M. Kok, and S. G. F. Robben, "Redefining the structure of structured reporting in radiology," *Insights Imaging*, vol. 11, no. 1, p. 10, Dec. 2020, doi: 10.1186/s13244-019-0831-6.

[13] M. P. Hartung, I. C. Bickle, F. Gaillard, and J. P. Kanne, "How to Create a Great Radiology Report," *RadioGraphics*, vol. 40, no. 6, pp. 1658–1670, Oct. 2020, doi: 10.1148/rg.2020200020.

[14] J. R. Wilcox, "The written radiology report.," *Appl. Radiol.*, vol. 35, no. 7, 2006, Accessed: Aug. 14, 2024. [Online]. Available: https://cdn.agilitycms.com/applied-radiology/PDFs/Issues/AR_07-06_Wilcox.pdf

[15] A. Casey *et al.*, "A systematic review of natural language processing applied to radiology reports," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, p. 179, Dec. 2021, doi: 10.1186/s12911-021-01533-7.

[16] X. Fei, P. Chen, L. Wei, Y. Huang, Y. Xin, and J. Li, "Quality management of pulmonary nodule radiology reports based on natural language processing," *Bioengineering*, vol. 9, no. 6, p. 244, 2022.

[17] A.-D. Pham *et al.*, "Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings," *BMC Bioinformatics*, vol. 15, no. 1, p. 266, Dec. 2014, doi: 10.1186/1471-2105-15-266.

[18] S. Yu *et al.*, "Classification of CT pulmonary angiography reports by presence, chronicity, and location of pulmonary embolism with natural language processing," *J. Biomed. Inform.*, vol. 52, pp. 386–393, 2014.

[19] S. Meystre, R. Gouripeddi, J. Tieder, J. Simmons, R. Srivastava, and S. Shah, "Enhancing comparative effectiveness research with automated pediatric pneumonia detection in a multi-institutional clinical repository: a PHIS+ pilot study," *J. Med. Internet Res.*, vol. 19, no. 5, p. e162, 2017.

[20] T. Cai *et al.*, "Natural Language Processing Technologies in Radiology Research and Clinical Applications," *RadioGraphics*, vol. 36, no. 1, pp. 176–191, Jan. 2016, doi: 10.1148/rg.2016150080.

[21] P. Lakhani, W. Kim, and C. P. Langlotz, "Automated Detection of Critical Results in Radiology Reports," *J. Digit. Imaging*, vol. 25, no. 1, pp. 30–36, Feb. 2012, doi: 10.1007/s10278-011-9426-6.

[22] S. Puts, M. Nobel, C. Zegers, I. Bermejo, S. Robben, and A. Dekker, "How natural language processing can aid with pulmonary oncology tumor node metastasis staging from free-text radiology reports: algorithm development and validation," *JMIR Form. Res.*, vol. 7, p. e38125, 2023.

[23] A. Y. Li and N. Elliot, "Natural language processing to identify ureteric stones in radiology reports," *J. Med. Imaging Radiat. Oncol.*, vol. 63, no. 3, pp. 307–310, Jun. 2019, doi: 10.1111/1754-9485.12861.

[24] T. Huang *et al.*, "Identifying incarceration status in the electronic health record using large language models in emergency department settings," *J. Clin. Transl. Sci.*, vol. 8, no. 1, p. e53, 2024.

[25] M. A. Ayden, M. E. Yuksel, and S. E. Y. Erdem, "A two-stream deep model for automated ICD-9 code prediction in an intensive care unit," *Heliyon*, vol. 10, no. 4, 2024, Accessed: Oct. 11, 2024. [Online]. Available: https://www.cell.com/heliyon/fulltext/S2405-8440(24)01991-1

[26] D. Kim, H. Yoo, and S. Kim, "An Automatic ICD Coding Network Using Partition-Based Label Attention," Nov. 15, 2022, *arXiv*: arXiv:2211.08429. Accessed: Oct. 11, 2024. [Online]. Available: http://arxiv.org/abs/2211.08429

[27] Y. Miyazaki *et al.*, "Performance and Errors of ChatGPT-4o on the Japanese Medical Licensing Examination: Solving All Questions Including Images with Over 90% Accuracy," *JMIR Med Educ*, 2024, Accessed: Oct. 11, 2024. [Online]. Available: https://s3.ca-central-1.amazonaws.com/assets.jmir.org/assets/preprints/preprint-63129-submitted.pdf

[28] V. M. Builoff *et al.*, "Evaluating AI Proficiency in Nuclear Cardiology: Large Language Models take on the Board Preparation Exam," *medRxiv*, pp. 2024–07, 2024.

[29] L. Lian, "Comparative Study of GPT-4.0, ERNIE Bot 4.0, and GPT-4o in the 2023 Chinese Medical Licensing Examination," 2024, Accessed: Oct. 11, 2024. [Online]. Available: https://www.researchsquare.com/article/rs-4639770/latest

[30] J. Beattie *et al.*, "Utilizing Large Language Models for Enhanced Clinical Trial Matching: A Study on Automation in Patient Screening," *Cureus*, vol. 16, no. 5, 2024, Accessed: Oct. 11, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11162699/

[31] Y. Li, R. M. Wehbe, F. S. Ahmad, H. Wang, and Y. Luo, "Clinical-Longformer and Clinical-BigBird: Transformers for long clinical sequences," Apr. 15, 2022, *arXiv*: arXiv:2201.11838. Accessed: Oct. 10, 2024. [Online]. Available: http://arxiv.org/abs/2201.11838