Original Paper

Extracting Pulmonary Embolism Diagnoses From Radiology Impressions Using GPT-4o: Large Language Model Evaluation Study

Mohammed Mahyoub[1,2], MS; Kacie Dougherty[1], PhD; Ajit Shukla[1], MBA

[1]Virtua Health, Marlton, NJ, United States

[2]School of Systems Science and Industrial Engineering, Binghamton University, Binghamton, NY, United States

Corresponding Author:
Mohammed Mahyoub, MS
Virtua Health
301 Lippincott Drive, 3rd Fl.
Marlton, NJ, 08053
United States
Phone: 1 8888478823
Email: mmahyoub@virtua.org

Abstract

Background: Pulmonary embolism (PE) is a critical condition requiring rapid diagnosis to reduce mortality. Extracting PE diagnoses from radiology reports manually is time-consuming, highlighting the need for automated solutions. Advances in natural language processing, especially transformer models like GPT-4o, offer promising tools to improve diagnostic accuracy and workflow efficiency in clinical settings.

Objective: This study aimed to develop an automatic extraction system using GPT-4o to extract PE diagnoses from radiology report impressions, enhancing clinical decision-making and workflow efficiency.

Methods: In total, 2 approaches were developed and evaluated: a fine-tuned Clinical Longformer as a baseline model and a GPT-4o-based extractor. Clinical Longformer, an encoder-only model, was chosen for its robustness in text classification tasks, particularly on smaller scales. GPT-4o, a decoder-only instruction-following LLM, was selected for its advanced language understanding capabilities. The study aimed to evaluate GPT-4o's ability to perform text classification compared to the baseline Clinical Longformer. The Clinical Longformer was trained on a dataset of 1000 radiology report impressions and validated on a separate set of 200 samples, while the GPT-4o extractor was validated using the same 200-sample set. Postdeployment performance was further assessed on an additional 200 operational records to evaluate model efficacy in a real-world setting.

Results: GPT-4o outperformed the Clinical Longformer in 2 of the metrics, achieving a sensitivity of 1.0 (95% CI 1.0-1.0; Wilcoxon test, P<.001) and an F_1-score of 0.975 (95% CI 0.9495-0.9947; Wilcoxon test, P<.001) across the validation dataset. Postdeployment evaluations also showed strong performance of the deployed GPT-4o model with a sensitivity of 1.0 (95% CI 1.0-1.0), a specificity of 0.94 (95% CI 0.8913-0.9804), and an F_1-score of 0.97 (95% CI 0.9479-0.9908). This high level of accuracy supports a reduction in manual review, streamlining clinical workflows and improving diagnostic precision.

Conclusions: The GPT-4o model provides an effective solution for the automatic extraction of PE diagnoses from radiology reports, offering a reliable tool that aids timely and accurate clinical decision-making. This approach has the potential to significantly improve patient outcomes by expediting diagnosis and treatment pathways for critical conditions like PE.

(JMIR Med Inform 2025;13:e67706) doi: 10.2196/67706

KEYWORDS

pulmonary embolism; large language models; LLMs; natural language processing; GPT-4o; Clinical Longformer; text classification; radiology reports

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 [1-3]. 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, combined computed tomography venography and pulmonary angiography, and multidetector computed tomography angiography [8-10]. 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 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 the quality of care through evidence-based practices.

Natural language processing (NLP) techniques have been increasingly used 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 [16-18]. 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 tumor, nodes, and metastasis classification system, a standard for staging cancer [22]. In addition, 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 used to identify incarceration status from medical records, showcasing good sensitivity and specificity compared to traditional keyword-based methods [24]. In addition, Clinical Longformer has been successfully applied in the classification of clinical notes for automated *ICD* (*International Classification of Diseases*) coding, where it outperformed other models in accuracy [25,26]. This capability to accurately interpret and classify clinical text is crucial for improving health care 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 models, have been recognized for their versatility in clinical applications, particularly in generating, summarizing, classifying, and extracting clinical information [27-33]. Its multimodal capabilities allow it to process not only text but also images and audio, enhancing its utility in diverse clinical settings [34]. GPT-4 has been used in clinical trial matching, where it automates eligibility screening, thus streamlining the recruitment process for clinical studies [35].

This study aims to develop an large language model (LLM)–based extraction system to automatically extract pulmonary embolism diagnoses from radiology report impressions. The key contributions of this study are mentioned in Textbox 1.

Textbox 1. Key contributions.

- Enhance and accelerate clinical data availability to improve the quality of care through evidence-based approaches.
- Develop a reliable diagnoses extraction system, which examines 2 technologies: Clinical Longformer and GPT-4o.
- Deploy the developed system as a cloud-based web application, addressing a gap often found in clinical artificial intelligence research.
- Evaluate the model both before and after deployment.

Methods

Overview

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 used. We then describe the models applied in this research. In addition, we discuss the deployment pipeline of the selected model. Finally, we outline the evaluation metrics used to assess the model's performance.

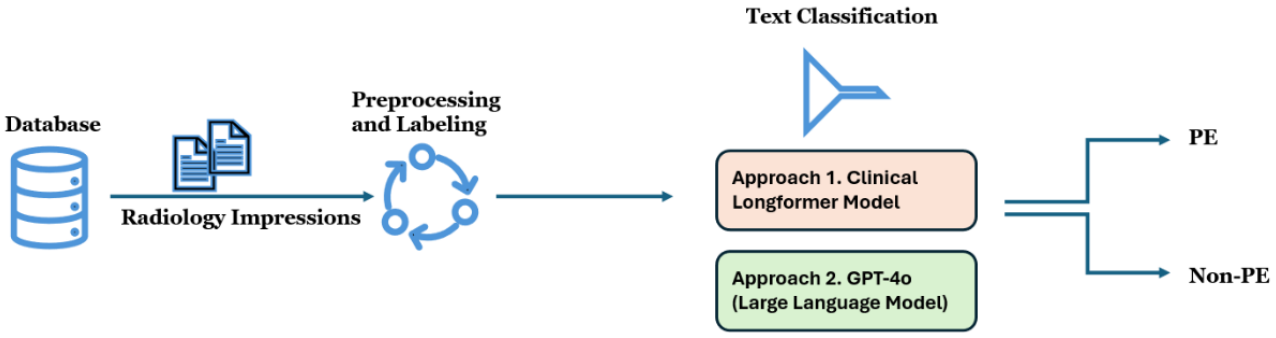
The primary objective of this study is to develop and deploy an artificial intelligence (AI) solution capable of extracting PE diagnoses from radiology report impressions. After defining this goal, the research proceeds through 4 distinct phases. In the first step, for collecting training and validation datasets, we determined the appropriate data tables in the Virtua Health clinical database and extracted radiology impressions data. This is followed by preprocessing and transforming the data to make it suitable for model development. The second step involves creating and testing 2 models (Clinical Longformer and GPT-4o), then choosing the one with the best results to proceed. The selected model is then implemented during the third step. In the final step, the performance of the model was tracked in real-world conditions and evaluated on how it affected operational outcomes.

Radiology Impressions Text Classification

Text classification is a fundamental task in 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. An NLP approach is then used 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, 2 different text classification models are used: Approach 1 uses a Clinical Longformer Model, while Approach 2 involves GPT-4o, a LLM. Both models classify the impressions into 2 categories: PE and non-PE. The goal is to determine whether a diagnosis of PE is present in each radiology report impression.

Figure 1. Radiology impressions text classification. PE: pulmonary embolism.



Data

The data used in this study was sourced from the electronic medical record relational database of Virtua Health, New Jersey, 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 were formatted appropriately for analysis and modeling.

The training dataset consists of 1000 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. In addition, a separate testing dataset, consisting of 200 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.

Table 1. Training and validation data characteristics.

Metric	Training dataset	Validation dataset
Number of observations	1000	200
Average number of words (a token is approximately three-fourths of a word)	43.64	40.18
Number of pulmonary embolism-positive cases	36	100
Number of pulmonary embolism-negative cases	964	100

As shown in Table 1, the training dataset contains 1000 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 (1:1 ratio of classes), has a slightly lower average word count per report impression, at 40.18 words.

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

Bidirectional Encoder Representations from Transformers (BERT), which can process sequences up to 512 tokens [36]. Clinical Longformer incorporates a sparse attention mechanism that allows it to efficiently process sequences up to 4096 tokens, making it ideal for handling lengthy clinical narratives. Pretrained 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 report impressions without truncating important contextual information.

We fine-tuned (full fine-tuning) the Clinical Longformer model on a graphics processing unit server with 48 GB of memory. The fine-tuning hyperparameters were mentioned in Textbox 2.

Textbox 2. Clinical Longformer fine-tuning hyperparameters.

- Batch size: 4
- Gradient accumulation steps: 8
- Learning rate: 2e-5
- Number of epochs: 5
- Optimizer: AdamW
- Learning rate scheduler: Linear

GPT-4o Classifier

The methodology for using 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 Textbox 3, 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, 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.

Textbox 3. Extraction of pulmonary embolism diagnosis using GPT-4o-based chain of thought (COT) and few-shot learning algorithm.

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 chain of thought 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.

Figure 2. Extraction of pulmonary embolism diagnosis prompt template.

```
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}
'''
```

Deployment Pipeline

The deployment pipeline for the PE classification model, illustrated in Figure 3, 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 uses 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 4, 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 PE classification results. This interface serves as a convenient tool for health care professionals to quickly identify patients with a PE diagnosis, improving clinical decision-making and patient outcomes by providing prompt, automated insights. The Web App has been operationalized at Virtua Health, New Jersey.

Figure 3. Deployment pipeline.

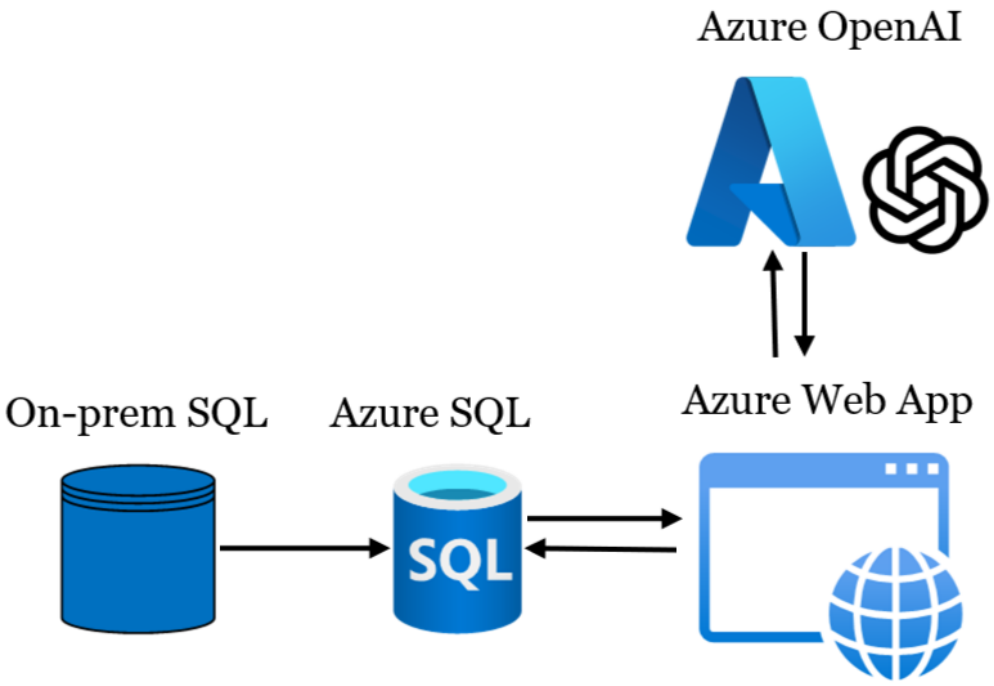
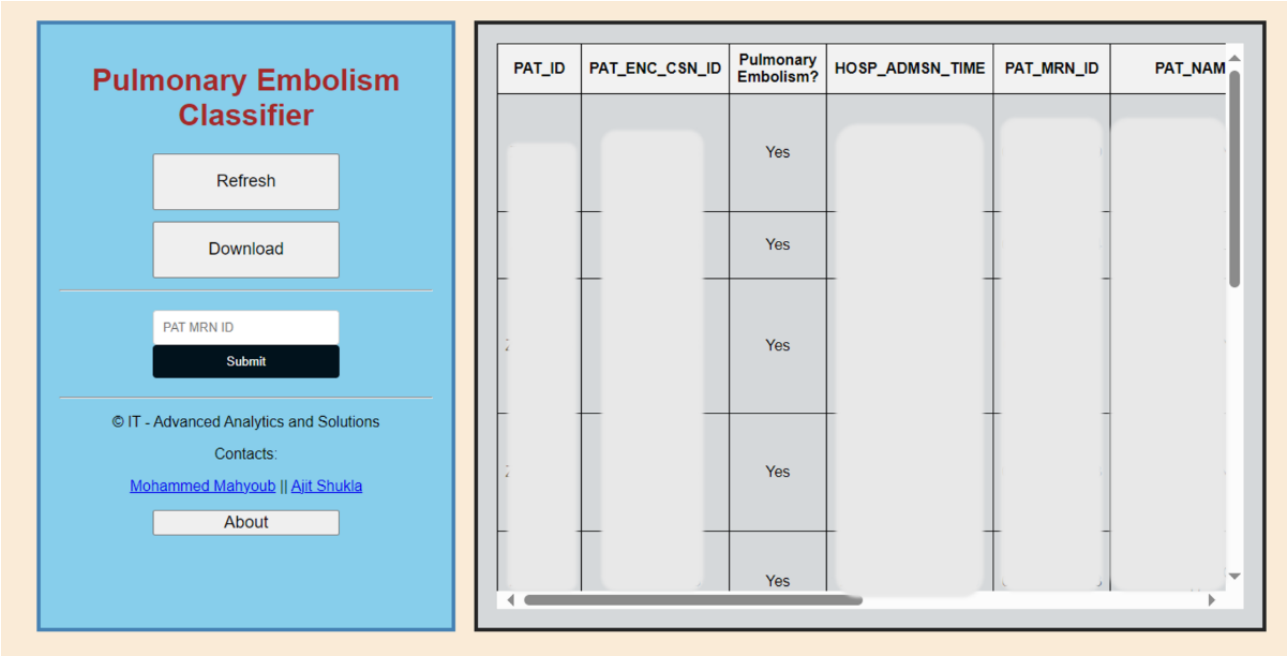


Figure 4. Web App. Only positive cases are displayed. This Web App has been operationalized at Virtua Health, New Jersey.



Evaluation Metrics and Statistical Testing

To evaluate the performance of the PE classification model, we used several commonly used metrics (Textbox 4).

Textbox 4. Classification evaluation metrics.

- 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.
- F_1-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.

To rigorously evaluate and compare model performance, we used nonparametric bootstrap sampling with 1000 iterations on both the validation and postdeployment datasets. In each bootstrap iteration, samples were drawn with replacement from the dataset, and standard evaluation metrics—sensitivity, specificity, and F_1-score—were calculated. This process yielded 1000 metric estimates per model (GPT-4o and Clinical Longformer), providing an empirical distribution for each metric.

From these distributions, the mean and 95% CI were derived using percentile-based estimation. This approach enables robust quantification of performance uncertainty without assuming normality.

To statistically compare the models, we conducted paired, 2-sided Wilcoxon signed rank tests on the bootstrapped metric distributions. This nonparametric test assesses whether the differences in paired metric values across bootstraps are statistically significant. A significance level of α=.01 was used to determine the threshold for significance.

The same bootstrapping procedure was applied to the postdeployment dataset to calculate mean metric values and corresponding 95% CI. However, statistical significance testing was conducted only on the validation dataset to ensure controlled model comparisons under consistent evaluation conditions. In the postdeployment setting, only the GPT-4o model was in use.

Ethical Considerations

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 or institutional review board waived the requirement of written informed consent for participation from the participants or the participants' legal guardians or 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 for consent was waived on the condition that, when appropriate, the subjects will be provided with additional pertinent information about participation.

Results

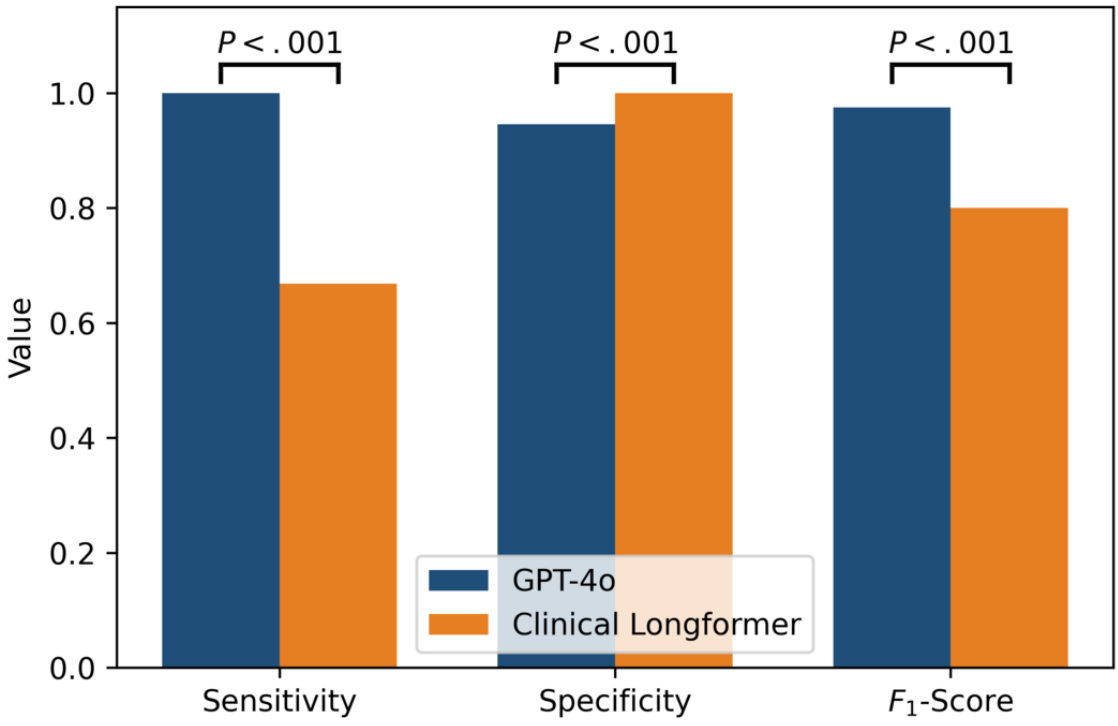
Overview

This section presents the research findings. First, we evaluate the Clinical Longformer and GPT-4o models using the validation dataset during the development phase. Then, we assess the performance of the deployed GPT-4o model post deployment.

Models Evaluation

Figure 5 presents a comparative analysis of evaluation metrics, including sensitivity, specificity, and F_1-score, between GPT-4o and Clinical Longformer. The results demonstrate statistically significant differences across all 3 metrics (P<.001), as determined by the Wilcoxon signed-rank test.

Figure 5. Evaluation metrics comparison of Clinical Longformer (baseline model) and GPT-4o.



GPT-4o exhibits superior sensitivity, achieving a perfect sensitivity of 1.0 (95% CI 1.0-1.0), in contrast to Clinical Longformer, which attains a sensitivity of 0.6684 (95% CI 0.5644-0.7604). This indicates that GPT-4o captures all positive cases correctly, whereas Clinical Longformer has a notable false negative rate. For specificity, Clinical Longformer achieves a perfect score of 1.0 (95% CI 1.0-1.0), slightly surpassing GPT-4o, which attains a specificity of 0.9456 (95% CI 0.9022-0.989). Although both models perform well in distinguishing negative cases, GPT-4o exhibits a marginally higher false positive rate. Regarding the F_1-score, GPT-4o significantly outperforms Clinical Longformer, with an F_1-score of 0.975 (95% CI 0.9495-0.9947) compared to 0.8002 (95% CI 0.7215-0.8639) for Clinical Longformer. This improvement reflects GPT-4o's higher balance between precision and recall, leading to superior overall classification performance.

The statistically significant differences across all 3 metrics underscore GPT-4o's robust generalization and enhanced performance in comparison to Clinical Longformer, suggesting its potential for improved clinical applications.

Postdeployment Performance

Based on GPT-4o's excellent performance during the validation phase, where it achieved superior metrics, it was selected for deployment in the operational setting. The postdeployment evaluation of the GPT-4o model was conducted using a dataset of 200 records, which was selected using stratified sampling from the operational data. Table 2 provides a summary of the dataset characteristics, including 100 positive cases of pulmonary embolism and 100 negative cases.

Table 2. Postdeployment testing data characteristics. A stratified random sample of 200 records was taken from the operational dataset for postdeployment evaluation.

Metric	Values
Number of observations	200
Average number of words (a token is approximately three-fourths of a word)	43.76
Number of pulmonary embolism-positive cases	100
Number of pulmonary embolism-negative cases	100

In postdeployment, sensitivity remains at a perfect 1.0 (95% CI 1.0-1.0), demonstrating the model's ability to correctly identify all positive cases without false negatives. Specificity, however, shows a slight decrease compared to predeployment values, with a point estimate of approximately 0.94 (95% CI 0.8913-0.9804), suggesting a modest increase in false positives. The F_1-score remains high at approximately 0.97 (95% CI 0.9479-0.9908), indicating a strong balance between precision and recall. While GPT-4o maintains robust predictive performance postdeployment, with high sensitivity and a stable F_1-score, the slight decline in specificity highlights a potential area for further refinement.

Discussion

Principal Findings

This study underscores the efficacy of GPT-4o in automating the extraction of PE diagnoses from radiology report impressions. GPT-4o demonstrated excellent performance across sensitivity, specificity, and F_1-scores during both validation and postdeployment evaluations. These findings suggest that GPT-4o's advanced language understanding capabilities allow it to capture subtle contextual and semantic nuances in radiology impressions, which are often critical for accurate diagnosis.

This success is attributed to GPT-4o's ability to leverage large-scale training on diverse datasets, which enhances its generalizability and adaptability. In addition, its deployment as a cloud-based tool offers scalability, making it accessible to health care systems of varying sizes. These attributes make GPT-4o a transformative tool in clinical AI, setting a benchmark for future applications in radiology and other diagnostic domains. This success has also been recognized in various studies, as GPT-4o has demonstrated promise in addressing critical tasks across different areas of radiology [37-41].

Error Analysis

As indicated in the Results section, GPT-4o scored 1.0 (95% CI 1.0-1.0) on sensitivity. However, the specificity was 0.94 (95% CI 0.8913-0.9804), hinting at a false positive rate of around 6%. Here, we look at 3 of these false positives to uncover any commonality.

The first sample, labeled as "Yes" while the ground truth (human labeler) says otherwise, has the following impressions text: "Mild right upper lobe infiltrate. Heterogeneous density of the pulmonary artery, which could represent pulmonary embolus." This is a borderline case, where the text indicates a likelihood of PE without firm confirmation. The model was able to bring attention to this case for further evaluation. Also, ground truth labeling may suggest the need for further validation.

The second sample's impressions text states "(1) Limited examination as mentioned above, particularly further evaluation of the pulmonary arteries in the right lung. Questionable subsegmental pulmonary embolism may be present in the right middle lobe. Follow-up would be helpful as clinically indicated. (2) Large right pleural effusion. Adjacent lung consolidation suggests atelectasis. (3) Mild ground glass infiltrates in the right upper lobe and left lower lobe. This may represent mild pneumonitis. (4) Small left pleural effusion. (5) Incidental findings as above." The model indicated the presence of PE. If we look at the text, we can see that the radiologist indicated a questionable PE. This case is similar to the previous one, in

which the model labels likelihoods and questionable diagnoses as indicative of PE, which could be a useful way to bring clinicians' attention to the case for further evaluation.

The third sample impressions text is "(1) Small linear filling defects involving the right lower lobe, left upper lobe, and left lower lobe, likely sequela of chronic pulmonary embolism. No convincing findings of acute pulmonary embolism. No imaging findings of right heart strain. (2) Interval enlargement of multiple pulmonary nodules measuring up to approximately 12 mm. Further correlation with PET/CT versus follow-up computed tomography in 3 months is recommended. (3) Areas of mild subpleural atelectasis or scar with mild interstitial thickening that may be related to chronic change, although mild pulmonary edema or infection would be difficult to entirely exclude." The text indicates that "No convincing findings of acute pulmonary embolism." This resulted in a negative ground truth label. However, the model interpreted the "likely sequela of chronic pulmonary embolism" as an indicator of PE. The latter suggests that filling defects were a result of a past medical condition. Thus, this might be a definite false positive.

In summary, analyzing some of the false positives indicates the model's ability to bring attention to borderline cases that might hint at the presence of PE, allowing for further clinical validation. However, the model sometimes results in definite false positives when the radiology text ambiguously discusses PE in the context of past conditions.

Operational and Clinical Implications

The postdeployment results of GPT-4o show several important operational and clinical implications. First, the model's sensitivity and specificity in a real-world setting indicate that it can help distinguish between subjects who have pulmonary embolisms and those who do not, without many false positives or negatives. This is important in a clinical setting where missing cases of pulmonary embolism have implications for patient outcomes, and false positives can lead to unnecessary further investigation or treatment. At Virtua Health, the model successfully identified over 700 positive PE cases in 2024.

From an operational perspective, the model's excellent accuracy on both the validation and postdeployment testing datasets means that the model can be operated in an autonomous manner which would help reduce the workload of radiologists and other health care professionals. As a result, it can classify cases correctly, thus taking some of the pressure and time off clinicians and their ability to concentrate on other, more challenging or time-consuming cases. Furthermore, the model's high precision means that health care resources are used more effectively, with fewer unnecessary interventions and therefore better patient care.

Clinically, the use of GPT-4o increases the decision support that clinicians can receive by offering a correct answer to the question regarding the presence of pulmonary embolism based on the information presented in the radiology reports. This early recognition can result in early diagnosis and management and therefore better patient results and possibly reduced mortality. Moreover, the stability of the performance of GPT-4o in identifying pulmonary embolisms in various datasets suggests robustness and generality, which could be useful in various clinical settings.

Ethical Issues in the Use of AI Models in Health Care

The employment of AI models in health care induces multifaceted ethical issues that need to be deliberated on. The first and foremost concern is the privacy of the patients and the security of the information because medical information is private. This includes strong network security, policy compliance with laws like HIPAA (Health Insurance Portability and Accountability Act) and General Data Protection Regulation, and data exposure minimization during training and use of the model. It is crucial to ensure that AI systems work within these legal and ethical parameters to ensure the credibility of their use.

Another important factor is compliance with regulations because the application of AI in health care has to be compatible with the standards that are set on a regional and international level. This includes making sure that the process of training, validation, and integration into the clinical workflow of the model is well documented. These processes are easily understandable and can be explained to health care providers, patients, and regulators, thus establishing the ethical and legal sustainability of technology.

Another important issue is bias and fairness as AI models trained on small and biased datasets are likely to emit biased output. For instance, if the training data is biased then the predictions of the model may be unfair to some patients and thus lead to unequal care. This is particularly a concern in health care where discrimination can have severe impacts on minority populations, stressing the need to ensure that datasets are inclusive, and model evaluation focuses on fairness.

Finally, it is important that in using AI in the various fields of life, transparency and accountability are considered to create confidence in the use of AI. The results of the model must be easy to understand and explain to the clinicians and patients, hence the need for explainable AI. Moreover, it is crucial to subject the models to routine assessments and performance tracking to provide evidence of accountability and trustworthiness in the long run.

To address these ethical issues, there is a need for ongoing assessment, open reporting, and interaction between AI creators, health care deliverers, and regulators. Thus, the main aspects that can be improved are the aspects that when improved will help achieve the best possible results with the use of AI in health care while at the same time adhering to ethical principles and protecting the rights of patients.

Bias Evaluation

LLMs are generally pretrained on large corpora of texts from the internet. As a result, inherent biases in human language can influence model outputs. Recent LLMs, such as GPT-4o, undergo posttraining alignment to reduce biased responses. This serves as the first line of defense against bias. Second, in our case, the LLM is instructed to return a single label (yes or no), which may further reduce the potential for hallucinations and bias. LLMs typically require more tokens to reveal their

underlying tendencies, which can include bias from the pretraining stage. Finally, we evaluated the model's performance using the F_1-score across different age groups and gender profiles.

Table 3 summarizes the F_1-score across different sex and age groups to assess the consistency of the model's performance across demographic subpopulations. The model demonstrates comparable F_1-scores for both female (0.970, 95% CI 0.930-1.000) and male (0.972, 95% CI 0.937-1.000) groups, indicating no apparent sex-based disparity in classification performance.

Table 3. F_1-score across sex and age groups.

Variable and group	Values, n (%)	F_1-score (95% CI)
Sex		
Female	98 (51)	0.970 (0.930-1.000)
Male	102 (49)	0.972 (0.937-1.000)
Age		
Under 30 years	10 (5)	1 (1-1)
30-50 years	25 (12.5)	0.963 (0.867-1.000)
50-70 years	93 (46.5)	0.982 (0.951-1.000)
70-90 years	62 (31)	0.966 (0.906-1.000)
90 years and older	10 (5)	0.823 (CI 0.500-1.000)

Across age groups, the F_1-scores remain generally high, with minor variations. The highest score is observed in individuals under 30 years old (F_1-score=1.000), although this group comprises only 5% of the sample, suggesting potentially limited generalizability due to the small sample size. The model performs well for the majority age groups—50-70 years (F_1-score=0.982, 95% CI 0.951-1.000) and 70-90 years (F_1-score=0.966, 95% CI 0.906-1.000)—which collectively represent over 75% of the dataset. However, performance decreases slightly for individuals aged 90 and above (F_1-score=0.823, 95% CI 0.500-1.000), likely due to a smaller sample size and increased clinical complexity in this cohort. Overall, the model demonstrates robust and consistent performance across most demographic groups, with slight variation observed at the extremes of age distribution.

Explainability

GPT-4o is an LLM built on transformer architecture, a type of neural network that uses an attention mechanism to learn semantics. Due to its complexity, the model is inherently a black box. However, since the model generates output by continuing the input text based on user instructions, the output can be interpreted as the most probable completion (especially when the temperature parameter is set to 0.0).

In our case, the instructions are to analyze radiology impressions and return a "Yes or No," indicating the presence or absence of PE, respectively. In real-world decision-making, users are informed that the LLM generates responses based on the most likely continuation of the input text according to the instructions provided. The generated label is returned to users along with the corresponding impressions for validation and confirmation.

Exploring explainable artificial intelligence methods is highly encouraged to better understand how input text influences LLM output, improving predictability and transparency. Techniques such as Shapley Additive Explanations (SHAP) and attention mechanism analysis can support this effort. In our case, implementation was challenging due to the use of a large proprietary LLM (ie, GPT-4o), for which we do not have access to the model architecture needed for such analyses. Therefore, leveraging open-source LLMs in the future could enable the application of these techniques.

The GPT-4o-based PE classifier is delivered through a web application, requiring user involvement to confirm the model's results. In summary, our application adopts a human-in-the-loop approach. While this supports practical use, the broader explainability of LLM output remains an ongoing area of research aimed at better understanding the underlying mechanisms of next-token prediction.

Limitations

As the study shows promising results, there are important limitations, and their implications cannot be overlooked. A major limitation is the size of the dataset used in this study, which is 1000 training samples, 200 validation samples, and 200 postdeployment testing samples. Another limitation is the data imbalance in the training dataset (for the Clinical Longformer). With such a restricted number of examples, the model may not be able to generalize its learning to the full extent of the clinical scenarios that are likely to be encountered in the real world. This could lead to it being less accurate in real-world use, where report structures can differ significantly, and larger datasets are usually used to increase model confidence and stability.

Another limitation has to do with the model's generalization. The training data was collected from a single health care organization, Virtua Health, which means that the language and cultural features that are unlikely to be encountered in reports from other hospitals or countries may not be well-represented in the model. As a result, this may restrict the applicability of

the model to other settings unless it is retrained (or prompt-engineered) with local data.

The study also found that the Clinical Longformer, which was used as a baseline, could not grasp some of the semantic details, resulting in the misclassification of 2 positive PE cases. This suggests that future work should build on this study to enhance the architecture of transformer models specific to clinical tasks, especially those that entail processing medical text with sophistication.

Some issues may occur when implementing the model in the real world even though GPT-4o performed excellently in controlled experiments. Real-world data often contains errors, omissions, or other noise that the model may not be able to handle properly. In addition, there are potential threats to consistency in reporting practices that may lead to inconsistencies in the information captured by the model.

Future Work

To overcome these limitations, future work should focus on the growth of the dataset, which should be expanded to include a bigger and more complex set of samples that would encompass different types of radiology reports, patients, and institutions. This will increase the size of the dataset and the variety of the samples, which in turn will increase the model's coverage and stability. Data imbalance in the training dataset of the Clinical Longformer should be tackled by using data augmentation techniques to test the impact of validation performance. Furthermore, the model must be tested in different health care contexts to determine its effectiveness, which means that external validation is critical. External validation can be achieved through collaborations with other health systems or testing on publicly available datasets. To mitigate real-world operationalization challenges, future implementation should include routine performance audits to monitor model accuracy over time, periodic retraining with updated real-world data to maintain robustness, and regular human validation checks to ensure data quality and reflect evolving documentation practices. These strategies can help maintain reliability despite variability in data quality and reporting standards. Overall, these measures will assist in ensuring that the model is not only efficient but also easily implementable in various health care environments.

Conclusions

In conclusion, this study features an effective LLM-based approach to the automation of the extraction of PE diagnoses from radiology report impressions. We then compared the performance of a fine-tuned Clinical Longformer and GPT-4o and found that GPT-4o outperformed in terms of sensitivity, specificity, and overall accuracy both before and post deployment. The integration of GPT-4o into clinical practice has several operational and clinical benefits, including decreasing the need for manual review, improving clinical decision support, and detecting cases of PE more quickly. Furthermore, the model's capability to lighten the load of manual review to a great extent is a key contributor to improving the workflow of the diagnostic process, thereby allowing clinicians to channel their efforts toward more challenging problems. These qualities make it suitable for application in other clinical settings with the potential to go beyond PE diagnosis to other medical conditions that require comprehensive and accurate analysis of radiology reports. Not only does its integration enhance efficiency but also the quality of clinical decision-making processes. Future work may involve expanding this approach to other medical conditions and improving the integration of NLP-based models into clinical workflows to keep on enhancing the quality of health care delivery.

Data Availability

The data analyzed in this study are subject to the following licenses or restrictions: data in this study are not available due to agreements made with the institutional review board of Virtua Health.

Authors' Contributions

MM contributed to conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, writing original drafts, review, and editing. KD managed writing review and editing, formal analysis, and validation. AS handled project administration, resources, writing review and editing, and validation.

Conflicts of Interest

None declared.

References

1. Tanra AH, AT L, T E, DE R. Diagnostic value of platelet indices in patients with pulmonary embolism. Indonesian J. Clin. Pathol. Med. Lab. 2020;27(1):22-26. [doi: 10.24293/ijcpml.v27i1.1625]
2. Deng W, Gao W. Cathepsin Causal Association with Pulmonary Embolism: A Mendelian Randomization Analysis. 2024. URL: https://www.researchsquare.com/article/rs-4191858/latest [accessed 2024-08-14]
3. Lyhne MD, Kline JA, Nielsen-Kudsk JE, Andersen A. Pulmonary vasodilation in acute pulmonary embolism - a systematic review. Pulm Circ. 2020;10(1):2045894019899775. [FREE Full text] [doi: 10.1177/2045894019899775] [Medline: 32180938]
4. Zhang SL, Zhang QF, Li G, Guo M, Qi X, Xing XH, et al. Case Report: resuscitation of patient with tumor-induced acute pulmonary embolism by venoarterial extracorporeal membrane oxygenation. Front Cardiovasc Med. 2024;11:1322387. [FREE Full text] [doi: 10.3389/fcvm.2024.1322387] [Medline: 38426120]

5. Grusova G, Lambert L, Zeman J, Lambertova A, Benes J. The additional value of esophageal wall evaluation and secondary findings in emergency patients undergoing CT pulmonary angiography. Iran J Radiol Brieflands. 2018;15(1):e63466. [FREE Full text] [doi: 10.5812/iranjradiol.63466]
6. Becattini C, Vedovati MC, Agnelli G. Diagnosis and prognosis of acute pulmonary embolism: focus on serum troponins. Expert Rev Mol Diagn. 2008;8(3):339-349. [doi: 10.1586/14737159.8.3.339] [Medline: 18598112]
7. Simpson J, López-Candales A. Elevated brain natriuretic peptide and troponin I in a woman with generalized weakness and chest pain. Echocardiography. 2005;22(3):267-271. [doi: 10.1111/j.0742-2822.2005.03192.x] [Medline: 15725164]
8. Zhou Y, Shi H, Wang Y, Kumar AR, Chi B, Han P. 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. 2012;85(1015):972-979. [FREE Full text] [doi: 10.1259/bjr/40850443] [Medline: 21976633]
9. Lapergue B, Decroix JP, Evrard S, Wang A, Bendetowicz D, Offroy MA, 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. 2015;74(1-2):69-72. [doi: 10.1159/000437261] [Medline: 26228469]
10. Yuan H, Shao Y, Liu Z, Wang H. An improved faster R-CNN for pulmonary embolism detection from CTPA images. IEEE Access. 2021;9:105382-105392. [doi: 10.1109/access.2021.3099479]
11. Segrelles JD, Medina R, Blanquer I, Martí-Bonmatí L. Increasing the efficiency on producing radiology reports for breast cancer diagnosis by means of structured reports. A comparative study. Methods Inf Med. 2017;56(3):248-260. [doi: 10.3414/ME16-01-0091] [Medline: 28220929]
12. Nobel JM, Kok EM, Robben SGF. Redefining the structure of structured reporting in radiology. Insights Imaging. 2020;11(1):10. [FREE Full text] [doi: 10.1186/s13244-019-0831-6] [Medline: 32020396]
13. Hartung MP, Bickle IC, Gaillard F, Kanne JP. How to create a great radiology report. Radiographics. 2020;40(6):1658-1670. [doi: 10.1148/rg.2020200020] [Medline: 33001790]
14. Wilcox JR. The written radiology report. AR. 2006;35(7):33-37. [FREE Full text] [doi: 10.37549/AR1440]
15. Casey A, Davidson E, Poon M, Dong H, Duma D, Grivas A, et al. A systematic review of natural language processing applied to radiology reports. BMC Med Inform Decis Mak. 2021;21(1):179. [FREE Full text] [doi: 10.1186/s12911-021-01533-7] [Medline: 34082729]
16. Fei X, Chen P, Wei L, Huang Y, Xin Y, Li J. Quality management of pulmonary nodule radiology reports based on natural language processing. Bioengineering (Basel). 2022;9(6):244. [FREE Full text] [doi: 10.3390/bioengineering9060244] [Medline: 35735487]
17. Pham AD, Névéol A, Lavergne T, Yasunaga D, Clément O, Meyer G, et al. Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. BMC Bioinformatics. 2014;15(1):266. [FREE Full text] [doi: 10.1186/1471-2105-15-266] [Medline: 25099227]
18. Yu S, Kumamaru KK, George E, Dunne RM, Bedayat A, Neykov M, et al. Classification of CT pulmonary angiography reports by presence, chronicity, and location of pulmonary embolism with natural language processing. J Biomed Inform. 2014;52:386-393. [FREE Full text] [doi: 10.1016/j.jbi.2014.08.001] [Medline: 25117751]
19. Meystre S, Gouripeddi R, Tieder J, Simmons J, Srivastava R, Shah S. Enhancing comparative effectiveness research with automated pediatric pneumonia detection in a multi-institutional clinical repository: a PHIS+ pilot study. J Med Internet Res JMIR Publication. 2017;19(5):e162. [doi: 10.2196/jmir.6887]
20. Cai T, Giannopoulos AA, Yu S, Kelil T, Ripley B, Kumamaru KK, et al. Natural language processing technologies in radiology research and clinical applications. Radiographics. 2016;36(1):176-191. [FREE Full text] [doi: 10.1148/rg.2016150080] [Medline: 26761536]
21. Lakhani P, Kim W, Langlotz CP. Automated detection of critical results in radiology reports. J Digit Imaging. 2012;25(1):30-36. [FREE Full text] [doi: 10.1007/s10278-011-9426-6] [Medline: 22038514]
22. Puts S, Nobel M, Zegers C, Bermejo I, Robben S, Dekker A. 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. 2023;7:e38125. [FREE Full text] [doi: 10.2196/38125] [Medline: 36947118]
23. Li AY, Elliot N. Natural language processing to identify ureteric stones in radiology reports. J Med Imaging Radiat Oncol. 2019;63(3):307-310. [doi: 10.1111/1754-9485.12861] [Medline: 30720244]
24. Huang T, Socrates V, Gilson A, Safranek C, Chi L, Wang EA, et al. Identifying incarceration status in the electronic health record using large language models in emergency department settings. J Clin Transl Sci. 2024;8(1):e53. [FREE Full text] [doi: 10.1017/cts.2024.496] [Medline: 38544748]
25. Ayden M, Yuksel M, Erdem S. A two-stream deep model for automated ICD-9 code prediction in an intensive care unit. Heliyon Elsevier. 2024;10(4):e25960. [FREE Full text]
26. Kim D, Yoo H, Kim S. An Automatic ICD Coding Network Using Partition-Based Label Attentio. 2022. URL: http://arxiv.org/abs/2211.08429 [accessed 2024-10-11]
27. Miyazaki Y, Hata M, Omori H, Hirashima A, Nakagawa Y, Et? M, 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. URL: https://s3.ca-central-1.amazonaws.com/assets.jmir.org/assets/preprints/preprint-63129-submitted.pdf [accessed 2024-10-11]

28. Builoff V, Shanbhag A, Miller RJ, Dey D, Liang JX, Flood K, et al. Evaluating AI proficiency in nuclear cardiology: large language models take on the board preparation exam. J Nucl Cardiol. 2024:102089. [doi: 10.1016/j.nuclcard.2024.102089] [Medline: 39617127]
29. Wals Zurita AJ, Miras Del Rio H, Ugarte Ruiz de Aguirre N, Nebrera Navarro C, Rubio Jimenez M, Muñoz Carmona D, et al. The transformative potential of large language models in mining electronic health records data: content analysis. JMIR Med Inform. 2025;13:e58457. [FREE Full text] [doi: 10.2196/58457] [Medline: 39746191]
30. Cardamone NC, Olfson M, Schmutte T, Ungar L, Liu T, Cullen SW, et al. Classifying unstructured text in electronic health records for mental health prediction models: large language model evaluation study. JMIR Med Inform. 2025;13:e65454. [FREE Full text] [doi: 10.2196/65454] [Medline: 39864953]
31. Yoon D, Han C, Kim DW, Kim S, Bae S, Ryu JA, et al. Redefining health care data interoperability: empirical exploration of large language models in information exchange. J Med Internet Res. 2024;26:e56614. [FREE Full text] [doi: 10.2196/56614] [Medline: 38819879]
32. Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K. Large language models for mental health applications: systematic review. JMIR Ment Health. 2024;11:e57400. [FREE Full text] [doi: 10.2196/57400] [Medline: 39423368]
33. Shin D, Kim H, Lee S, Cho Y, Jung W. Using large language models to detect sepression from user-generated diary text data as a novel approach in digital mental health screening: instrument validation study. J Med Internet Res. 2024;26:e54617. [FREE Full text] [doi: 10.2196/54617] [Medline: 39292502]
34. Lian L. Comparative Study of GPT-4.0, ERNIE Bot 4.0, and GPT-4o in the 2023 Chinese Medical Licensing Examination. URL: https://www.researchsquare.com/article/rs-4639770/latest [accessed 2024-10-11]
35. Beattie J, Neufeld S, Yang D, Chukwuma C, Gul A, Desai N, et al. Utilizing large language models for enhanced clinical trial matching: a study on automation in patient screening. Cureus. 2024;16(5):e60044. [FREE Full text] [doi: 10.7759/cureus.60044] [Medline: 38854210]
36. Li Y, Wehbe RM, Ahmad FS, Wang H, Luo Y. Clinical-Longformer and Clinical-BigBird: Transformers for long clinical sequences. URL: http://arxiv.org/abs/2201.11838 [accessed 2024-10-10]
37. Kanzawa J, Kurokawa R, Kaiume M, Nakamura Y, Kurokawa M, Sonoda Y, et al. Evaluating the role of GPT-4 and GPT-4o in the detectability of chest radiography reports requiring further assessment. Cureus. 2024;16(12):e75532. [doi: 10.7759/cureus.75532] [Medline: 39803046]
38. Ferreira TL, Oliveira MC, de AVT. Comparative study of large language models for lung-RADS classification in portuguese CT reports. IEEE; 2024. Presented at: IEEE 24th International Conference on Bioinformatics and Bioengineering (BIBE); 2024 November 27-29:1-8; Kragujevac, Serbia. URL: https://ieeexplore.ieee.org/abstract/document/10820460/ [doi: 10.1109/bibe63649.2024.10820460]
39. Beşler MS, Oleaga L, Junquero V, Merino C. Evaluating GPT-4o's performance in the official European board of radiology exam: a comprehensive assessment. Acad Radiol. 2024;31(11):4365-4371. [doi: 10.1016/j.acra.2024.09.005] [Medline: 39294055]
40. Li D, Gupta K, Bhaduri M, Sathiadoss P, Bhatnagar S, Chong J. Comparative diagnostic accuracy of GPT-4o and LLaMA 3-70b: proprietary vs. open-source large language models in radiology. Clin Imaging. 2025;118:110382. [doi: 10.1016/j.clinimag.2024.110382] [Medline: 39740646]
41. Busch F, Prucker P, Komenda A, Ziegelmayer S, Makowski MR, Bressem KK, et al. Multilingual feasibility of GPT-4o for automated Voice-to-Text CT and MRI report transcription. Eur J Radiol. 2025;182:111827. [FREE Full text] [doi: 10.1016/j.ejrad.2024.111827] [Medline: 39566177]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations from Transformers
COT: chain-of-thought
HIPAA: Health Insurance Portability and Accountability Act
ICD: International Classification of Diseases
LLM: large language model
NLP: natural language processing
PE: pulmonary embolism
SHAP: Shapley Additive Explanations

Edited by A Castonguay; submitted 18.10.24; peer-reviewed by D Chrimes, J Huang; comments to author 14.01.25; revised version received 30.01.25; accepted 13.03.25; published 09.04.25

Please cite as:
Mahyoub M, Dougherty K, Shukla A
Extracting Pulmonary Embolism Diagnoses From Radiology Impressions Using GPT-4o: Large Language Model Evaluation Study
JMIR Med Inform 2025;13:e67706
URL: https://medinform.jmir.org/2025/1/e67706
doi: 10.2196/67706
PMID: