Labeling Radiology Report With GPT-4 Prompt Engineering: Comparative Study of in-Context Prompting

Songsoo Kim, Donghyun Kim, Hyunjoo Shin, Jaewoong Kim, Dukyong Yoon

Submitted to: Journal of Medical Internet Research
on: March 15, 2024

Table of Contents

Labeling Radiology Report With GPT-4 Prompt Engineering: Comparative Study of in-Context Prompting

Songsoo Kim[1*] MD; Donghyun Kim[2*] MD; Hyunjoo Shin[3, 4*] MD, PhD; Jaewoong Kim[1]; Dukyong Yoon[1, 4, 5] MD, PhD

[1]Department of Biomedical Systems Informatics Yonsei University College of Medicine Seoul KR
[2]Department of Radiology Daegu Gyeongbuk Regional Office of Military Manpower Administration Daegu KR
[3]Department of Radiology, Research Institute of Radiological Science and Center for Clinical Imaging Data Science, Yongin Severance Hospital Yonsei University College of Medicine Yongin KR
[4]Center for Digital Health, Yongin Severance Hospital Yonsei University College of Medicine Yongin KR
[5]Institute for Innovation in Digital Healthcare Severance Hospital Seoul KR
[*]these authors contributed equally

Corresponding Author:
Dukyong Yoon MD, PhD
Department of Biomedical Systems Informatics
Yonsei University College of Medicine
50-1 Yonsei-ro, Seodaemun-gu
Seoul
KR

Abstract

Background: Large language models, such as Generative Pre-trained Transformer-4 (GPT-4), utilize a method known as in-context learning, which enhances the model's responses by understanding the context provided within the input text.

Objective: This study aims to assess the labeling efficacy of Generative Pre-trained Transformer-4 in radiology reports and to validate the performance enhancement through in-context learning.

Methods: In this retrospective study, radiology reports were obtained utilizing the Medical Information Mart for Intensive Care III (MIMIC-III) database, and the reports were manually labeled by two radiologists for performance evaluation. Two experimental prompts were defined for comparison: the “Basic prompt,” which included sections for “Task” and “Output,” and the “In-context prompt,” which added a “Context” section for additional information. Labeling experiments were conducted on head CT reports for multi-label classification of ten predefined labels (mass, hemorrhage, infarct, vascular, white matter, volume loss, hydrocephalus, pneumocephalus, foreign body, and fracture) - Experiment 1. Labeling abdomen CT reports for multi-label classification of actionable findings based on four different sections (gastrointestinal, genitourinary, musculoskeletal, and vascular) - Experiment 2. Precision, recall, F1-scores, and accuracy were compared between the two prompting scenarios.

Results: In Experiment 1, for most labels, In-context prompts demonstrated a notable improvement in F1 scores (up to 0.658) and accuracy (up to 0.155), except for hemorrhage and pneumocephalus labels. Statistically significant differences were observed in four labels (vascular, hydrocephalus, mass, foreign body). For Experiment 2, the In-context prompt significantly enhanced F1 scores (by up to 0.306) and accuracy (by up to 0.107) across all labels, compared to Basic prompts.

Conclusions: Our study demonstrated that Generative Pre-trained Transformer-4 with prompt engineering has commendable performance in various labeling tasks in real-world radiology reports. It offers a flexible, researcher-tailored approach to labeling tasks using in-context learning.

(JMIR Preprints 15/03/2024:58331)
DOI: https://doi.org/10.2196/preprints.58331

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.
2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?
✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Labeling Radiology Report With Generative Pre-trained Transformer-4 Prompt Engineering: Comparative Study of in-Context Prompting

Abstract

Background: Large language models, such as Generative Pre-trained Transformer-4 (GPT-4), utilize a method known as in-context learning, which enhances the model's responses by understanding the context provided within the input text.

Objective: This study aims to assess the labeling efficacy of Generative Pre-trained Transformer-4 in radiology reports and to validate the performance enhancement through in-context learning.

Methods: In this retrospective study, radiology reports were obtained utilizing the Medical Information Mart for Intensive Care III (MIMIC-III) database, and the reports were manually labeled by two radiologists for performance evaluation. Two experimental prompts were defined for comparison: the “Basic prompt,” which included sections for “Task” and “Output,” and the “In-context prompt,” which added a “Context” section for additional information. Labeling experiments were conducted on head CT reports for multi-label classification of ten predefined labels (mass, hemorrhage, infarct, vascular, white matter, volume loss, hydrocephalus, pneumocephalus, foreign body, and fracture) - Experiment 1. Labeling abdomen CT reports for multi-label classification of actionable findings based on four different sections (gastrointestinal, genitourinary, musculoskeletal, and vascular) - Experiment 2. Precision, recall, F1-scores, and accuracy were compared between the two prompting scenarios.

Results: In Experiment 1, for most labels, In-context prompts demonstrated a notable improvement in F1 scores (up to 0.658) and accuracy (up to 0.155), except for hemorrhage and pneumocephalus labels. Statistically significant differences were observed in four labels (vascular, hydrocephalus, mass, foreign body). For Experiment 2, the In-context prompt significantly enhanced F1 scores (by up to 0.306) and accuracy (by up to 0.107) across all labels, compared to Basic prompts.

Conclusions: Our study demonstrated that Generative Pre-trained Transformer-4 with prompt engineering has commendable performance in various labeling tasks in real-world radiology reports. It offers a flexible, researcher-tailored approach to labeling tasks using in-context learning.

Keywords: Radiology Report; Generative Pretrained Transformer 4; Prompt Engineering; Large Language Model; Natural Language Processing

Introduction

Artificial intelligence (AI) is being increasingly used to label free-text radiology reports for assigning their clinical relevance. These radiology report labels have wide applications, such as improving patient management, assisting in treatment planning, and enhancing the training of AI models for clinical prediction [1–4]. Therefore, these labels must be generated accurately and efficiently.

Traditional natural language processing (NLP) deep learning models such as bidirectional encoder representations from transformers (BERT), often require extensive training data for a single task, and have limited usage, leading to most radiology report data being primarily used for creating labeling models. Therefore, foundational models that can handle multiple tasks without requiring training data for each task are required [5].

Large language models (LLMs), such as Generative Pre-trained Transformer-4 (GPT-4), originally designed for general tasks, have proven effective in specialized fields such as medicine, particularly in radiology [6–8]. LLMs not only excel at medical writing, structuring radiology reports, impressions generation, advanced reasoning, and data mining [9–14], but also have demonstrated commendable performance in chest X-ray report labeling, even without domain-specific training [15, 16].

Traditional deep learning models and LLMs differ in that the latter models require tasks to be defined through natural language text, known as "Prompt." This differs from traditional approaches that typically have fixed input and output contexts. Consequently, the LLM performance can vary in the same task depending on the prompt used. "Prompt engineering," which focuses on refining these prompts to optimize LLM performance without domain-specific training, has gained attention in this regard [17, 18]. This emphasizes that the full potential of the model cannot be achieved without appropriately crafted prompts.

In the field of radiology, the importance of crafting appropriate prompts has been reported. While studies by Mukherjee et al. and Fink et al. demonstrate that different prompts can enhance task performance in radiology, the specific differences between the prompts and the exact prompting strategies used remain unclear [9, 16]. Nguyen reported the application of specific prompt engineering in decision-making tasks based on the American College of Radiology (ACR) appropriate criteria using ChatGPT and Google's Bard [19]. However, to our knowledge, no study has applied a specific prompt engineering methodology using GPT-4 in actual radiology reports for evaluating the performance of labeling tasks.

Based on this background, we aim to investigate whether GPT-4 can perform labeling tasks on complex real-world radiology reports without training data, through appropriate prompt design. Additionally, we examine whether further performance enhancements can be achieved through prompt engineering. Specifically, we utilize in-context learning, where the model learns from specific contexts provided in the prompt (Textbox 1). First, we evaluate GPT-4's multi-label classification performance on 10 predefined labels in head CT reports. Second, we evaluate GPT-4's multi-label classification performance on actionable findings developed by the ACR workgroup in abdomen CT reports [20].

Textbox 1. General example of in-context learning by GPT-4.

A. Prompt without in-context learning scenario

Human:
Label the following.
pulmonary tuberculosis:

GPT-4:
Pulmonary tuberculosis is an infectious disease primarily affecting the lungs, caused by the bacterium Mycobacterium tuberculosis. It is characterized by symptoms such as a persistent cough, often with sputum and blood, chest pain, weakness, weight loss, fever, and night sweats. Tuberculosis can spread from person to person through airborne particles expelled when the infected individual coughs or sneezes.

If you're looking to label this condition in various contexts, here are some possible labels:

Medical Condition: Pulmonary Tuberculosis
Etiological Agent: Mycobacterium tuberculosis
Type of Disease: Infectious / Communicable Disease
System Affected: Respiratory System
ICD-10 Code: A15.0 **-** A15.9 (Tuberculosis of the lung, confirmed by sputum microscopy with or without culture)
Treatment Approach: Antibiotic Regimen (e.g., Isoniazid, Rifampicin, Pyrazinamide, and Ethambutol)
Preventive Measure: Bacillus Calmette-Guérin (BCG) Vaccination
Diagnostic Test: Sputum Test, Chest X-ray, CT Scan, Tuberculin Skin Test, Interferon-Gamma Release Assays (IGRAs)
Public Health Concern: Yes, due to its contagious nature and potential for outbreak Outcome if Untreated: Potentially fatal, can lead to extensive lung damage and spread to other organs

Labeling in this context implies providing a classification or identifying the characteristics of pulmonary tuberculosis in various dimensions, such as its medical classification, pathogen responsible, its nature and impact on health, and the approach for management and prevention.

B. Prompt with in-context learning scenario

Human:
Label the following.
lung cancer: neoplasm
idiopathic pulmonary fibrosis: interstitial lung disease
asthma: inflammatory disease

pulmonary tuberculosis:

GPT-4:
pulmonary tuberculosis: infectious disease

Methods

This research utilized de-identified, publicly available datasets and did not involve direct data collection from human subjects, exempting it from Institutional Review Board (IRB) approval requirements.

Report extraction, inclusion, and exclusion criteria

Figure 1 demonstrates the study flow. Reports were derived from the Medical Information Mart for Intensive Care III (MIMIC-III) database, which is publicly accessible. The reports were retrospectively collected, and data from 53,150 patients admitted to the ICU at Beth Israel Deaconess Medical Center in Boston, USA were de-identified. The reports encompassed over two million free-text clinical notes, including diverse radiology reports from various imaging modalities [21]. This ICU dataset is sufficiently anonymized and contains diverse and complex findings, making it ideal for assessing the model's labeling capacity.

For the multi-label classification of predefined labels, we employed head CT scans, which are less explored in research than chest X-rays, despite containing various disease labels. Abdomen CT was used for labeling actionable findings owing to its clinical significance and capacity to cover a broad range of actionable findings. Within the MIMIC-III data, "Description" corresponds to various types of radiologic assessments, encompassing apparently relevant descriptions while excluding those associated with the procedure (Figure 2). The inclusion criteria comprised the following: (i) Relevant descriptions of head CT reports and abdomen CT reports not associated with the procedure. (ii) Random sampling of 200 head CT reports and 400 abdomen CT reports from the relevant descriptions. Furthermore, there were no exclusion criteria after sampling the reports.

Radiology report labeling

Two board-certified radiologists, with experience over four and five years in radiologic practice, respectively, performed the manual labeling of the radiology reports. Details on label selection and annotation guidelines are available in the Multimedia Appendix 1. If a radiology report contained findings described in the annotation instructions, the corresponding label was annotated as positive. Following the manual labeling, the level of agreement was evaluated, and in instances of discrepancy, a consensus was achieved. These labels served as the ground truth for model performance evaluation. Given the extensive length of the radiology reports, a pre-annotation process that employed regular expressions within the Python environment was introduced to improve the efficiency and accuracy of annotations. This process involved identifying keywords and phrases outlined in the annotation instructions. Subsequently, human annotators reviewed the entire document to confirm the accuracy of the pre-annotations generated by the regular expressions (Figure 3).

1) Experiment 1

Experiment 1 introduced a total of ten labels, defined as follows: mass, vascular, volume loss, infarct, white matter, hydrocephalus, foreign body, hemorrhage, fracture, pneumocephalus. Inspired by Wood's study, which utilized seven specialized categories of abnormality and five general abnormal categories in MRI report labeling, labels such as mass, vascular, volume loss, infarct, and hydrocephalus were chosen. The annotation rules for each label were clearly defined, referring to the annotation rules from the cited study [22]. Furthermore, incorporating insights from Lorga's study, labels frequently identified in noncontrast head CT scans, such as hemorrhage, fracture, and pneumocephalus, were also included [23].

1) Experiment 2

The definition of "actionable finding" in this study is based on the categories outlined by the ACR actionable reporting work group [20]. Category 1 is defined by “critical or urgent findings that require communication within minutes,” such as closed loop intestinal obstruction. Category 2 is defined by “clinically significant findings that require communication within hours,” such as intra-abdominal infections such as appendicitis and cholecystitis. Category 3 findings are incidental or unexpected but do not require immediate treatment or other action, such as liver cirrhosis.

To assess GPT-4's ability to identify urgent findings, this study merges Categories 1 and 2, defining "actionable findings" as those necessitating communication within hours. The categorization for this research incorporates findings from the gastrointestinal (GI), genitourinary (GU), musculoskeletal (MSK), and vascular sections within the abdomen and pelvis.

Findings from the lung base, such as pulmonary thromboembolism detected on abdominal CT, were excluded from this study. The study also excludes the “General” actionable section, which involves subjective decisions such as “determining that the interpreting radiologist requires immediate physician notification.” Additionally, findings that are unobservable or unevaluable via Abdomen CT, such as fetal Doppler ultrasound findings, and those outside the abdomen, such as coronary artery occlusion, were also excluded. Findings previously deemed actionable but showing no significant interval change were not classified as actionable in this study. In contrast, findings indicative of progression were considered actionable. When no comparison with previous findings was mentioned, observations were considered novel for this study and annotated accordingly.

Prompt engineering

1) Background

Prompt engineering emerges as a field of study that effectively enhances the performance of Large Language Models (LLMs) without the necessity for fine-tuning by carefully crafting the input prompt. Numerous prompt engineering techniques have been introduced, notably few-shot prompting, which enhances performance through examples; chain-of-thought prompting, which increases accuracy through stepwise reasoning; and other advanced methodologies that apply these concepts [24–26]. While various prompt engineering methodologies differ in specifics, they share the commonality of providing context to the model and guiding it through multiple processes before generating an output, thereby inducing more accurate responses.

This in-context learning is the ability of LLM to learn and apply new information or instructions based on the context provided within a given input, without requiring explicit retraining or updates to the model. This approach easily allows the model to tailor its responses based on the biases and nuances embedded in the user-provided context, potentially reinforcing the user's perspective. Consequently, this in-context learning encompasses a spectrum of learning modalities through various contexts, including one-shot learning and few-shot learning, which involve providing one or a few examples, as well as learning through instructions. It can therefore be considered an umbrella term that captures these diverse learning contexts [27, 28].

Furthermore, by structuring prompts using a template, we can systematically define and integrate the various elements of a prompt, enhancing the model's performance. This approach enhances model consistency by defining tasks, promotes in-context learning by providing context, and reduces verbosity in responses by specifying output formats [24]. The CO-STAR framework, which employs structured prompting techniques, exemplifies this approach's effectiveness. By integrating various elements and aligning the model's objectives with the context, CO-STAR played a crucial role in winning the GPT-4 prompting competition organized by the Government Technology Agency of Singapore (GovTech), highlighting the significance of structuring prompting elements [29].

2) Prompt engineering strategy used in this study

Initially, the "Basic prompt" was designed to specifically instruct the model on the Initially, the "Basic prompt" was designed to specifically instruct the model on the task and limit the output's diversity by only including the concrete "task" and "output format" section. This approach aimed to encourage the model to perform the task with greater consistency by clearly defining what was expected in the output. By restricting the output format to JavaScript Object Notation (JSON) labels, it prevented the generation of invalid outputs and unnecessary verbose sentences. Additionally, using JSON format for outputs facilitated parsing multiple labels in the output environment and streamlined the post-processing steps. From the perspective of the labeling task, no domain-specific context was provided, allowing the model to freely respond based on its knowledge.

In the "In-context prompt" to test the in-context learning capabilities of LLMs, we provided the annotation instructions used by human annotators in the labeling task as "context." Since the ground truth labels for evaluating the model's performance were derived from human annotators, it was reasonable to use the annotation instructions as the most appropriate context, containing examples and explanations for each label. Summarized annotation instructions were incorporated into the prompt's context aimed to verify the effectiveness of in-context learning. This strategy effectively guided the model towards the desired labeling method while keeping the prompt relatively short compared to providing entire examples of the report.

In the GPT Application Programming Interface (API), prompts are categorized into User Prompts (entered by human users) and System Prompts. System Prompts, often not visible to users on the ChatGPT website, define the model's role (e.g., "You are a helpful assistant"). In this study, we assigned predefined prompts to the System Prompt and allocated only the radiology reports as User Prompts. This approach was adopted to maintain concise prompts and code. By limiting the predefined prompt without assigning additional roles, we aimed to ensure the model's focus remained solely on the desired task.

Lastly, the increase in prompt length leads to a rise in the total number of tokens and, consequently, an increase in the pricing for using GPT-4 model. This was quantitatively assessed by calculating the token count of each prompt using the GPT-4 tokenizer [30]. The two predefined prompts used in the actual experiment are provided in Table 1

Table 1. Prompts used in the experiments.

	Basic prompt	**In-context prompt**
Experiment 1		
Prompt	**Task** - Categorize this report under the following labels: normal, mass, hemorrhage, infarct, vascular, white matter, volume loss, hydrocephalus, pneumocephalus, foreign body, and fracture - Usually, there is one most suitable label, but if multiple labels are deemed appropriate, several of them may be assigned (except "normal"). **Output(JSON)** ("Label": ["label 1," "label 2"])	**Task** - Review the entire Head CT Report and categorize this report under the following labels: normal, mass, hemorrhage, infarct, vascular, white matter, volume loss, hydrocephalus, pneumocephalus, foreign body, and fracture - Usually, there is one most suitable label, but if multiple labels are deemed appropriate, several of them may be assigned (except "normal"). **Context** - "Normal" : absence of other predefined labels and cannot coexist with other labels. - "Mass" : neoplasm, abscess, cyst, and other similar findings. - "Hemorrhage" : epidural hematoma, subdural hematoma, subarachnoid hemorrhage, intraparenchymal hemorrhage, and other

		similar findings. - "Infarct" : acute infarct, subacute infarct, chronic infarct and other similar findings. - "Vascular" : aneurysm, vascular steno-occlusive lesion, vascular malformation, arteriovenous fistula, and other similar findings. - “White matter” : findings describing white matter inflammation, small vessel disease, and other similar findings. - "volume loss" : diffuse brain atrophy, encephalomalacia, post-operative tissue changes, chronic infarction with volume loss, and other similar findings. - "Hydrocephalus" : acute/chronic stable hydrocephalus, ventricular enlargement, normal pressure hydrocephalus, and other similar findings. - “Pneumocephalus” : any findings suggestive of pneumocephalus on CT. - "foreign body" : shunt, clips, coils, and other materials related to surgery or procedure. - "Fracture" : any displaced/non-displaced bony fracture on skull and upper cervical vertebra **Output(JSON)** ("Label": ["label 1", "label 2"])
Token count	92	386
Experiment 2		
Prompt	**Task** - Review the entire abdomen CT Report and classify the reports into actionable and non-actionable categories. Actionable findings are defined as findings that are urgent and need to be communicated within hours. - Actionable findings should be further categorized into GI, GU, MSK, and Vascular sections (refrain from evaluating other sections). Note that there can be multiple sections of actionable findings in a single report. - Actionable findings without significant interval changes compared with previous studies are considered non-actionable. Only findings with substantial progression are defined as actionable. **Output format(JSON)** Either (“Actionable”: ["section 1", "section 2"]) or (“Non-actionable”: “NA”)	**Task** - Review the entire abdomen CT Report and classify the reports into actionable and non-actionable categories. Actionable findings are defined as findings that are urgent and need to be communicated within hours. - Actionable findings should be further categorized into GI, GU, MSK, and Vascular sections (refrain from evaluating other sections). Note that there can be multiple sections of actionable findings in a single report. - Actionable findings without significant interval changes compared with previous studies are considered non-actionable. Only findings with substantial progression are defined as actionable. **Context** Actionable findings are as below: - GI - Unexplained pneumoperitoneum - Intestinal obstruction (including closed loop intestinal obstruction - Intestinal ischemia and/or portal/mesenteric venous gas - Pseudoaneurysm or active hemorrhage (post-trauma, GI bleed, other) - Intra-abdominal organ injury (liver, spleen, pancreas, other). - Abscess, any location - Intra-abdominal infection, likely surgical or interventional candidate (appendicitis, cholecystitis, diverticulitis, abscess, other) - Large volume ascites - Pneumatosis in the bowel wall, no other signs of ischemia - GU - Torsion of testicular and ovarian. - High likelihood of ectopic pregnancy. - High-grade injuries to kidney, ureter, or bladder post-trauma. - Complications in post-operative kidney. - Obstructions in the urinary tract. - Pyonephrosis or renal abscess. - Placental abnormality - MSK - Nonspinal fractures or dislocations. - Septic arthritis. - Necrotizing fasciitis. - Bone lesions with fracture risk. - Large hematomas with potential structural compression. - Changes in fracture alignment or infection risk. - Complications with surgical hardware. - Vascular - Ruptured or leaking arterial aneurysms. - Arterial dissections or intramural hematomas. - Significant arterial stenosis or occlusion with acute symptoms. - Post-vascular access arterial pseudoaneurysms. - Abdominal aortic aneurysms exceeding 5 cm, if stable.

		- Deep venous thrombosis. **Output format(JSON)** Either (“Actionable”: ["section 1", "section 2"]) or (“Non-actionable”: “NA”)
Token count	145	534

.

3) Prompt Testing

In prompt engineering, prompts are generally refined through continuous experimentation and evaluation to determine the optimal one (9). This often involves using a small dataset separate from the test dataset (24). However, in our experiment, we tested prompts directly on the test dataset, avoiding separate prompt refining experimentation. We selected this method owing to concerns that using extra information for prompt engineering might lead to prompts overly tailored to the test dataset, which compromises generalizability. To assess the model's zero-shot performance without training or validation data, we created contexts aligned with the annotation instructions from the human annotator before the main experiment.

Detailed GPT-4 setting

The study was conducted using the GPT-4 Application Programming Interface (API) in a Python development environment, where each query was executed within a new session. To enhance the consistency of the outputs through prompt engineering, the “temperature” parameter, which governs the diversity of the model's responses, was set to zero. The zero setting compels the model to produce the most probable answer to the same query, thereby reducing the likelihood of format errors in the output and potentially increasing the reproducibility of the study. Default values were maintained for other parameters. Token counts for each prompt were determined using OpenAI’s GPT-4 tokenizer [22]. All GPT-4 inferences and responses were collected on January 19, 2024.

Experiment 1

Labels from ten predefined ones (mass, hemorrhage, infarct, vascular, white matter, volume loss, hydrocephalus, pneumocephalus, foreign body, and fracture) were assigned to a provided MIMIC-III head CT report based on the identified report findings. Multiple labels could be assigned if the report suggested multiple findings. GPT-4's labeling accuracy was assessed using two different prompts for the same report.

Experiment 2

The ACR actionable reporting work group has classified “actionable findings” in radiology reports into two categories based on the urgency of communication required and organized them according to each anatomical section. Category 1 encompasses “critical findings that require communication within minutes” (e.g., closed loop intestinal obstruction), while Category 2 includes findings that are not as urgent as Category 1 but “require communication within hours” (e.g., appendicitis or cholecystitis). To assess GPT-4's ability to distinguish clinically significant findings, in this study, Category 1 and Category 2 were deemed “actionable findings.” The task involved categorizing abdomen CT reports as either actionable or non-actionable, with the actionable findings as the ones related to gastrointestinal (GI), genitourinary (GU), musculoskeletal (MSK), or Vascular sections.

GPT-4's labeling accuracy was assessed using two different prompts for the same report.

Statistical analysis

The accuracy of the entire report was assessed from the degree of agreement between two readers in manual labeling. GPT-4's performance was evaluated in the two experiments against true labels for computing precision, recall, F1-score, and accuracy. The difference in performance metrics between the two prompting methods was determined by subtracting the Basic prompt outcomes from the In-context prompt outcomes. To evaluate the significance of metric differences between the two prompts, we performed 1000 bootstrap iterations to calculate the 95% confidence interval of the samples. Statistical significance was determined when the confidence interval excluded zero. All statistical analyses and graphical representations were conducted in Python (version 3.11.4) utilizing Pandas (version 2.1.1), SciPy (version 1.6.3), Matplotlib (version 3.4.2), and Seaborn (version 0.11.1).

Results

Baseline characteristics

The MIMIC-III radiology reports that were included are described in Figure 2. Baseline characteristics across included datasets are described in Table 2. The reports comprised 200 head CT reports, with a median word count of 279.5 (IQR, 215.5-349.75) and a sentence count of 15.5 (IQR, 11.0-19.0). A total of 174 patients (93 male) were included with median age of 62.0 years (IQR, 48.0-74.0). Further, 400 abdomen CT reports were included, with a higher median word count (570.5, IQR, 452.25-676.0) and sentence count (34.0, IQR, 25.75-41.0). Herein, 311 patients (176 males) were included with a median age of 62.0 years (IQR, 49.0-74.0).

Table 2. Baseline characteristics of included MIMIC-III datasets

	Head CT reports	**Abdomen CT reports**
Report count	200	400
Word count	279.5 (215.5-349.75)	570.5 (452.25- 676.0)
Sentence count	15.5 (11.0-19.0)	34.0 (25.75-41.0)
Patient count	174	311
Age	62.0 (48.0-74.0)	62.0 (49.0-74.0)
Sex (male)	93	176

a. Median(Q1-Q3)

Experiment 1

The label distribution of Experiment 1 was as follows: Vascular (n=131), Hemorrhage (n=114), Infarct (n=54), Foreign body (n=44), Volume loss (n=30), White matter (n=27), Hydrocephalus (n=18), Fracture (n=17), Mass (n=16), Pneumocephalus (n=7, Figure 4A). Among the reports, 7 did not have labels. For the labeled reports, there was an average of 2.37 labels per report (458 labels across 193 reports). Excellent agreement was obtained in the manual labeling of the radiology reports between the two readers (accuracy: 0.91).

Using a Basic prompt, GPT-4 demonstrated commendable performance across most labels, with F1 scores ranging from 0.784 to 1.000, except for the labels “mass” and “foreign body” (Figure 5A).

Labeling of “mass” often resulted in false positives, particularly in reports describing a “mass effect,” which led to a low precision of 0.48. The model struggled with the identification of surgical materials as “foreign body,” resulting in false negatives and a low recall of 0.159.

Notably, F1 scores for all categories except hemorrhage were enhanced in the In-context prompt. Here, the F1 score for labeling “foreign body” jumped from 0.275 to 0.933, and for labeling “mass” increased from 0.585 to 0.800. Significant statistical improvements were also observed in labels “vascular” and “hydrocephalus.” In terms of accuracy, both prompts showed strong performance overall (Basic prompt: 0.815 to 1.000, In-context prompt: 0.955 to 1.000, Figure 5B), with accuracy increasing for all labels except for “hemorrhage.” Statistically significant improvements were observed for “vascular,” “hydrocephalus,” “foreign body,” and “mass.”

Experiment 2

The label distribution of Experiment 2 was as follows: GI (n=81), GU (n=19), MSK (n=20), and Vascular (n=25, Figure 4B). There were 129 actionable and 271 non-actionable reports. An average of 1.12 labels per report were noted for those that were actionable (145 labels across 129 reports). A satisfactory agreement was found in the manual labeling of radiology reports between the two readers (accuracy: 0.81). Discrepancies occurred in cases involving non-conclusive imaging results (e.g. unclear cause of pneumoperitoneum, or infections), situations requiring subjective judgment without clear cutoffs (e.g. large volume ascites), or subjective risk assessment based solely on report descriptions (e.g. risk for pathologic fracture).

The performance of the Basic prompt was relatively low, with F1 scores ranging from 0.585 to 0.622, primarily due to a high number of false positives, which resulted in low precision. Contrastingly, the In-context prompt demonstrated a significant performance boost, with F1 scores increasing from 0.17 to 0.306 across labels, which was statistically significant (Figure 6A). Regarding accuracy, both prompts exhibited robust performance (Basic prompt: 0.763 to 0.938, In-context prompt: 0.870 to 0.988, Figure 6B), with the increase in accuracy observed across all labels being statistically significant. Detailed performance of the two experiments can be found in Table 3, Multimedia Appendix 2, and Multimedia Appendix 3.

Table 3. Performance metrics of both Basic prompt and In-context prompt in each experiment

	Precision				Recall				F1-score				Accuracy			
	Basic Prompt	In-context Prompt	Difference	95% CI	Basic Prompt	In-context Prompt	Difference	95% CI	Basic Prompt	In-context Prompt	Difference	95% CI	Basic Prompt	In-context Prompt	Difference	95% CI
Head CT																
volume loss	0.952	0.889	-0.063	(-0.174, 0.018)	0.667	0.800	0.133	(-0.057, 0.323)	0.784	0.842	0.058	(-0.078, 0.190)	0.945	0.955	0.01	(-0.020, 0.040)
infarct	0.964	1.000	0.036	(0.000, 0.089)	1.000	1.000	0	(0.000, 0.000)	0.982	1.000	0.018	(0.000, 0.048)	0.990	1.000	0.01	(0.000, 0.025)
fracture	0.944	1.000	0.056	(0.000, 0.188)	1.000	1.000	0	(0.000, 0.000)	0.971	1.000	0.029	(0.000, 0.105)	0.995	1.000	0.005	(0.000, 0.015)
vascular	0.852	0.977	0.125	(0.073, 0.183)	0.969	0.969	0	(-0.037, 0.038)	0.907	0.973	0.066	(0.031, 0.101)	0.870	0.965	0.095	(0.045, 0.145)
hydrocephalus	1.000	1.000	0	(0.000, 0.000)	0.722	0.944	0.222	(0.050, 0.421)	0.839	0.971	0.132	(0.030, 0.292)	0.975	0.995	0.02	(0.005, 0.040)
hemorrhage	0.983	0.966	-0.017	(-0.042, 0.000)	1.000	1.000	0	(0.000, 0.000)	0.991	0.983	-0.008	(-0.021, 0.000)	0.990	0.980	-0.01	(-0.025, 0.000)
mass	0.480	0.737	0.257	(0.093, 0.436)	0.750	0.875	0.125	(0.000, 0.320)	0.585	0.800	0.215	(0.081, 0.373)	0.915	0.965	0.05	(0.020, 0.085)

white matter	0.844	0.900	0.056	(-0.045, 0.179)	1.000	1.000	0	(0.000, 0.000)	0.915	0.947	0.032	(-0.033, 0.102)	0.975	0.985	0.01	(-0.010, 0.030)
pneumocephalus	1.000	1.000	0	(0.000, 0.000)	1.000	1.000	0	(0.000, 0.000)	1.000	1.000	0	(0.000, 0.000)	1.000	1.000	0	(0.000, 0.000)
foreign body	1.000	0.913	-0.087	(-0.143, 0.000)	0.159	0.955	0.796	(0.681, 0.915)	0.275	0.933	0.658	(0.519, 0.838)	0.815	0.970	0.155	(0.110, 0.225)
Abdomen CT																
GI	0.453	0.611	0.158	(0.099, 0.220)	0.827	0.987	0.16	(0.082, 0.244)	0.585	0.755	0.17	(0.112, 0.231)	0.763	0.87	0.107	(0.068, 0.148)
GU	0.383	0.633	0.25	(0.140, 0.402)	0.947	1.000	0.053	(0.000, 0.160)	0.545	0.776	0.231	(0.133, 0.348)	0.925	0.973	0.048	(0.025, 0.070)
MSK	0.419	0.690	0.271	(0.098, 0.450)	0.650	1.000	0.35	(0.136, 0.579)	0.510	0.816	0.306	(0.144, 0.481)	0.938	0.978	0.04	(0.015, 0.068)
Vascular	0.470	0.833	0.363	(0.236, 0.505)	0.920	1.000	0.08	(0.000, 0.200)	0.622	0.910	0.288	(0.182, 0.417)	0.930	0.988	0.058	(0.035, 0.083)

a. 95% confidence interval (CI) (2.5%, 97.5%)

Discussion

GPT-4 demonstrated high accuracy in labeling tasks across two experiments without using any training data. Notably, providing relevant context in the labeling task resulted in performance improvements across most evaluation metrics. In Experiment 1, the In-context prompts, compared to Basic prompts, showed an increase in F1 scores and accuracy for all labels except hemorrhage and pneumocephalus, with statistically significant differences observed in four of them (vascular, hydrocephalus, mass, and foreign body). Experiment 2 further revealed that In-context prompts significantly enhanced F1 scores and accuracy across all labels when compared to Basic prompts.

Previous labeling efforts have used various NLP techniques, from rule-based methods to domain-specific BERT models. Rule-based approaches, such as keyword or pattern searches (e.g., regular expressions), do not require training data and provide clear interpretability of results. However, defining an exhaustive set of rules is impractical, and even when rules are established, they can lead to a high number of false positives. For instance, phrases such as "No signs or evidence of infarct" could be incorrectly labeled as indicating an infarct in contexts where it does not exist. Additionally, despite efforts to create comprehensive patterns, these methods face challenges owing to the variability of sentence structures, medical abbreviations, and typos [32–34]. Conversely, domain-specific transformer models were reported to be adaptable across a range of report types [35–39]. However, to develop these models, often substantial amounts of training data are required. Moreover, once trained for a specific task, the models lose their flexibility for other tasks, which restricts their versatility. Additionally, many BERT-based models, despite good performance, lack external validation [40–43]. Consequently, these models fall short of serving as a universal tool for labeling across various radiology reports.

LLMs overcome traditional method limitations through adaptability across tasks, thus highlighting their versatility. However, they encounter challenges such as hallucination and overly varied outputs, underscoring the need for prompt engineering to use their full potential. In this study, we leveraged prompt engineering in two experiments related to labeling radiology reports for completely harnessing the potential of LLMs. Utilizing Basic prompts, GPT-4 consistently produced task-aligned outputs with a uniform format across 1,200 inferences. Furthermore, the application of additional In-context prompting significantly improved performance in domain-specific tasks.

Experiment 1 showed high performance across both prompts, likely owing to the objective nature of disease definitions, which may be included in the GPT-4 training corpus. However, the performance was lower for the “foreign body” category under the Basic prompt. When “foreign body” was provided labeling instructions as “Shunts, Clips, Coils, Other materials related to surgery or procedure,” the model successfully inferred and labeled items that were not explicitly mentioned, such as “ventriculostomy tube,” “ventriculostomy catheter,” and even “NG tube.” Moreover, in the case of “Mass,” instances such as “mass effect” were incorrectly labeled as “Mass” in the Basic prompt scenario. By simply refining the definition of mass through In-context prompting, these false positive cases were reduced. In Experiment 2, the Basic prompt showed a high false positive rate for all labels due to the model's subjective interpretation of “actionable finding,” which resulted in a generally conservative response (except for "Large volume of ascites," leading to a false negative). However, after providing a clear definition, the performance across all metrics noticeably improved, suggesting that even within the same model, the interpretation of the same task— identifying "findings require communication within hours," —can vary significantly depending on the prompt.

This suggests that tasks with a higher degree of subjectivity can derive increased benefit from the provision of contextual information in the prompts. Although in situations where GPT-4 already possesses the requisite knowledge and the answers are certain, such context might not add significant value, it becomes a highly effective method for enhancing model performance in scenarios that require specific domain knowledge. This allows for a flexible enhancement of the model's capabilities, particularly in subjective tasks. This approach not only enhances the model's capabilities but also facilitates a researcher-tailored labeling strategy, effectively unlocking the model's potential.

Furthermore, In-context prompting allows for concise prompts and efficiently furnishes domain-specific information, thereby enhancing performance. Experiment 2 shows that the additional cost of using an In-context prompt is only $0.015 per report, which is cost-effective considering the performance gain [44]. Furthermore, structured prompts, defined output formats, and a zero-temperature setting help maintain the model's consistency. These elements affect the model's consistency and accuracy, and therefore, need to be addressed to ensure the reproducibility of the study.

This study has a few limitations, which can be addressed by future work in this field. First, the performance may vary owing to the lexical complexity in reports; further, the experiments were only performed on one dataset. Meanwhile, the MIMIC database being an ICU database, contains many severe cases and complex reports. This can present a challenging situation for the LLM, thereby making our positive results even more significant. However, future studies should validate the efficacy of the models across various institutions, languages, and modalities in radiology reports. Second, we exclusively tested the GPT-4. A significant concern with cloud-based LLMs such as GPT-4 is data security, which is paramount for medical data. Although the data used in this study were sufficiently anonymized within the report, the ideal scenario would involve the use of local LLMs that can operate entirely offline [15, 45]. However, local LLMs might show less advanced reasoning capabilities than GPT-4 and would require substantial computational resources from institutions. Future studies should validate models that are both effective in medical tasks and not overly demanding on computing power. Models such as Llama2, Vicuna, and Flan-T5, previously used on medical tasks, can serve as suitable candidates [16, 34, 46].

In conclusion, our study demonstrates that Generative Pre-trained Transformer-4, with prompt engineering, exhibits commendable performance in various labeling tasks in real-world radiology reports. It offers a flexible, researcher-tailored approach to labeling tasks using in-context learning. The potential utility of these large language models in labeling reports can be beneficial for patient management, research use, and AI model training, even with limited training data.

Acknowledgements

- MD-PhD/Medical Scientist Training Program through the Korea Health Industry Development Institute, funded by the Ministry of Health & Welfare, Republic of Korea.
- Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number : HI22C0452).

Conflicts of Interest

None declared.

Abbreviations

ACR: American College of Radiology
AI: Artificial Intelligence
API: Application Programming Interface
BERT: Bidirectional Encoder Representations from Transformers
CT: Computed Tomography
GI: Gastrointestinal
GPT-4: Generative Pre-trained Transformer 4
GU: Genitourinary
ICU: Intensive Care Unit
LLM: Large Language Model
MIMIC-III: Medical Information Mart for Intensive Care III
MSK: Musculoskeletal
NLP: Natural Language Processing
JSON: JavaScript Object Notation

Multimedia Appendix 1: List of labels used in Experiments 1 and 2 and the descriptions of annotation instructions provided to annotators.

Multimedia Appendix 2: Confusion matrix classified by GPT-4 in Experiment 1

Multimedia Appendix 3: Confusion matrix classified by GPT-4 in Experiment 2

References

1. Banerjee I, Davis MA, Vey BL, Mazaheri S, Khan F, Zavaletta V, Gerard R, Gichoya JW, Patel B. Natural Language Processing Model for Identifying Critical Findings-A Multi-Institutional Study. J Digit Imaging 2023 Feb;36(1):105–113. PMID:36344632

2. Nakamura Y, Hanaoka S, Nomura Y, Nakao T, Miki S, Watadani T, Yoshikawa T, Hayashi N, Abe O. Automatic detection of actionable radiology reports using bidirectional encoder representations from transformers. BMC Med Inform Decis Mak 2021 Sep 11;21(1):262. PMID:34511100

3. Lau W, Payne TH, Uzuner O, Yetisgen M. Extraction and Analysis of Clinically Important Follow-up Recommendations in a Large Radiology Dataset. AMIA Jt Summits Transl Sci Proc 2020 May 30;2020:335–344. PMID:32477653

4. Zech JR. Using BERT Models to Label Radiology Reports. Radiology: Artificial Intelligence Radiological Society of North America; 2022 Jul 1;4(4):e220124.

5. Haupt CE, Marks M. AI-Generated Medical Advice—GPT and Beyond. JAMA 2023 Mar 27; doi: 10.1001/jama.2023.5321

6. Ueda D, Walston SL, Matsumoto T, Deguchi R, Tatekawa H, Miki Y. Evaluating GPT-4-based ChatGPT's clinical potential on the NEJM quiz. bioRxiv. 2023. doi: 10.1101/2023.05.04.23289493

7. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198. PMID:36812645

8. Jung K-H. Uncover This Tech Term: Foundation Model. Korean J Radiol 2023 Oct;24(10):1038–1041. PMID:37793672

9. Fink MA, Bischoff A, Fink CA, Moll M, Kroschke J, Dulz L, Heußel CP, Kauczor H-U, Weber TF. Potential of ChatGPT and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer. Radiology 2023 Sep;308(3):e231362. PMID:37724963

10. Adams LC, Truhn D, Busch F, Kader A, Niehues SM, Makowski MR, Bressem KK. Leveraging GPT-4 for Post Hoc Transformation of Free-text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study. Radiology 2023 May;307(4):e230725. PMID:37014240

11. Sun Z, Ong H, Kennedy P, Tang L, Chen S, Elias J, Lucas E, Shih G, Peng Y. Evaluating GPT-4 on Impressions Generation in Radiology Reports. Radiology Radiological Society of North America; 2023 Jun 1;307(5):e231259.

12. Rau A, Rau S, Zoeller D, Fink A, Tran H, Wilpert C, Nattenmueller J, Neubauer J, Bamberg F, Reisert M, Russe MF. A Context-based Chatbot Surpasses Trained Radiologists and Generic ChatGPT in Following the ACR Appropriateness Guidelines. Radiology 2023 Jul;308(1):e230970. PMID:37489981

13. Hwang SI, Lim JS, Lee RW, Matsui Y, Iguchi T, Hiraki T, Ahn H. Is ChatGPT a "Fire of Prometheus" for Non-Native English-Speaking Researchers in Academic Writing? Korean J Radiol 2023 Oct;24(10):952–959. PMID:37793668

14. Koga S. The Integration of Large Language Models Such as ChatGPT in Scientific Writing: Harnessing Potential and Addressing Pitfalls. Korean J Radiol 2023 Sep 1;24(9):924–925.

15. Cai W. Feasibility and Prospect of Privacy-preserving Large Language Models in Radiology. Radiology.

2023. p. e232335. PMID:37815443

16. Mukherjee P, Hou B, Lanfredi RB, Summers RM. Feasibility of Using the Privacy-preserving Large Language Model Vicuna for Labeling Radiology Reports. Radiology 2023 Oct;309(1):e231147. PMID:37815442

17. Zhou Y, Muresanu AI, Han Z, Paster K, Pitis S, Chan H, Ba J. Large Language Models Are Human-Level Prompt Engineers. arXiv [csLG]. 2022. Available from: http://arxiv.org/abs/2211.01910

18. Strobelt H, Webson A, Sanh V, Hoover B, Beyer J, Pfister H, Rush AM. Interactive and Visual Prompt Engineering for Ad-hoc Task Adaptation with Large Language Models. IEEE Trans Vis Comput Graph 2023 Jan;29(1):1146–1156. PMID:36191099

19. Nguyen D, Swanson D, Newbury A, Kim YH. Evaluation of ChatGPT and Google Bard Using Prompt Engineering in Cancer Screening Algorithms. Acad Radiol 2023 Dec 15; PMID:38103973

20. Larson PA, Berland LL, Griffith B, Kahn CE Jr, Liebscher LA. Actionable findings and the role of IT support: report of the ACR Actionable Reporting Work Group. J Am Coll Radiol 2014 Jun;11(6):552–558. PMID:24485759

21. Johnson AEW, Pollard TJ, Shen L, Lehman L-WH, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035. PMID:27219127

22. Wood DA, Kafiabadi S, Al Busaidi A, Guilhem EL, Lynch J, Townend MK, Montvila A, Kiik M, Siddiqui J, Gadapa N, Benger MD, Mazumder A, Barker G, Ourselin S, Cole JH, Booth TC. Deep learning to automate the labelling of head MRI datasets for computer vision applications. Eur Radiol 2022 Jan;32(1):725–736. PMID:34286375

23. Iorga M, Drakopoulos M, Naidech AM, Katsaggelos AK, Parrish TB, Hill VB. Labeling Noncontrast Head CT Reports for Common Findings Using Natural Language Processing. AJNR Am J Neuroradiol 2022 May;43(5):721–726. PMID:35483905

24. Ekin S. Prompt engineering for ChatGPT: A quick guide to techniques, tips, and best practices. 2023. doi: 10.36227/techrxiv.22683919.v2

25. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi E, Le Q, Zhou D. Chain-of-thought prompting elicits reasoning in large language models. Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, editors. arXiv [csCL]. 2022. p. 24824–24837. Available from: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

26. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. Language Models are Few-Shot Learners. arXiv [csCL]. 2020. Available from: http://arxiv.org/abs/2005.14165

27. Si C, Friedman D, Joshi N, Feng S, Chen D, He H. Measuring inductive biases of in-context learning with underspecified demonstrations. arXiv; 2023; doi: 10.48550/ARXIV.2305.13299

28. Website. doi: 10.48550/arXiv.2303.03846

29. Teo S. How I Won Singapore's GPT-4 Prompt Engineering Competition. Towards Data Science. 2023. Available from: https://towardsdatascience.com/how-i-won-singapores-gpt-4-prompt-engineering-

competition-34c195a93d41 [accessed Jan 23, 2024]

30. Website. Available from: https://platform.openai.com/tokenizer

31. OpenAI Platform. Available from: https://platform.openai.com/tokenizer [accessed Dec 7, 2023]

32. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-Ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEEE; 2017. p. 2097–2106.

33. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghgoo B, Ball R, Shpanskaya K, Seekins J, Mong DA, Halabi SS, Sandberg JK, Jones R, Larson DB, Langlotz CP, Patel BN, Lungren MP, Ng AY. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. AAAI 2019 Jul 17;33(01):590–597.

34. Alsentzer E, Rasmussen MJ, Fontoura R, Cull AL, Beaulieu-Jones B, Gray KJ, Bates DW, Kovacheva VP. Zero-shot interpretable phenotyping of postpartum hemorrhage using large language models. NPJ Digit Med 2023 Nov 30;6(1):212. PMID:38036723

35. Yan A, McAuley J, Lu X, Du J, Chang EY, Gentili A, Hsu C-N. RadBERT: Adapting Transformer-based Language Models to Radiology. Radiol Artif Intell 2022 Jul;4(4):e210258. PMID:35923376

36. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Trans Comput Healthcare New York, NY, USA: Association for Computing Machinery; 2021 Oct 15;3(1):1–23.

37. Zaman S, Petri C, Vimalesvaran K, Howard J, Bharath A, Francis D, Peters N, Cole GD, Linton N. Automatic Diagnosis Labeling of Cardiovascular MRI by Using Semisupervised Natural Language Processing of Text Reports. Radiol Artif Intell 2022 Jan;4(1):e210085. PMID:35146435

38. Kuling G, Curpen B, Martel AL. BI-RADS BERT and Using Section Segmentation to Understand Radiology Reports. J Imaging Sci Technol 2022 May 9;8(5). PMID:35621895

39. Tejani AS, Ng YS, Xi Y, Fielding JR, Browning TG, Rayan JC. Performance of Multiple Pretrained BERT Models to Automate and Accelerate Data Annotation for Large Datasets. Radiol Artif Intell 2022 Jul;4(4):e220007. PMID:35923377

40. Elbatarny L, Do RKG, Gangai N, Ahmed F, Chhabra S, Simpson AL. Applying Natural Language Processing to Single-Report Prediction of Metastatic Disease Response Using the OR-RADS Lexicon. Cancers 2023 Oct 10;15(20). PMID:37894276

41. Jupin-Delevaux É, Djahnine A, Talbot F, Richard A, Gouttard S, Mansuy A, Douek P, Si-Mohamed S, Boussel L. BERT-based natural language processing analysis of French CT reports: Application to the measurement of the positivity rate for pulmonary embolism. Research in Diagnostic and Interventional Imaging 2023 Jun 1;6:100027.

42. Weng K-H, Liu C-F, Chen C-J. Deep Learning Approach for Negation and Speculation Detection for Automated Important Finding Flagging and Extraction in Radiology Report: Internal Validation and Technique Comparison Study. JMIR Med Inform 2023 Apr 25;11:e46348. PMID:37097731

43. López-Úbeda P, Martín-Noguerol T, Luna A. Automatic classification and prioritisation of actionable BI-RADS categories using natural language processing models. Clin Radiol 2023 Sep 27; PMID:37838546

44. Pricing. Available from: https://openai.com/pricing [accessed Dec 7, 2023]

45. Kim S, Lee C-K, Kim S-S. Large Language Models: A Guide for Radiologists. Korean J Radiol 2024 Feb

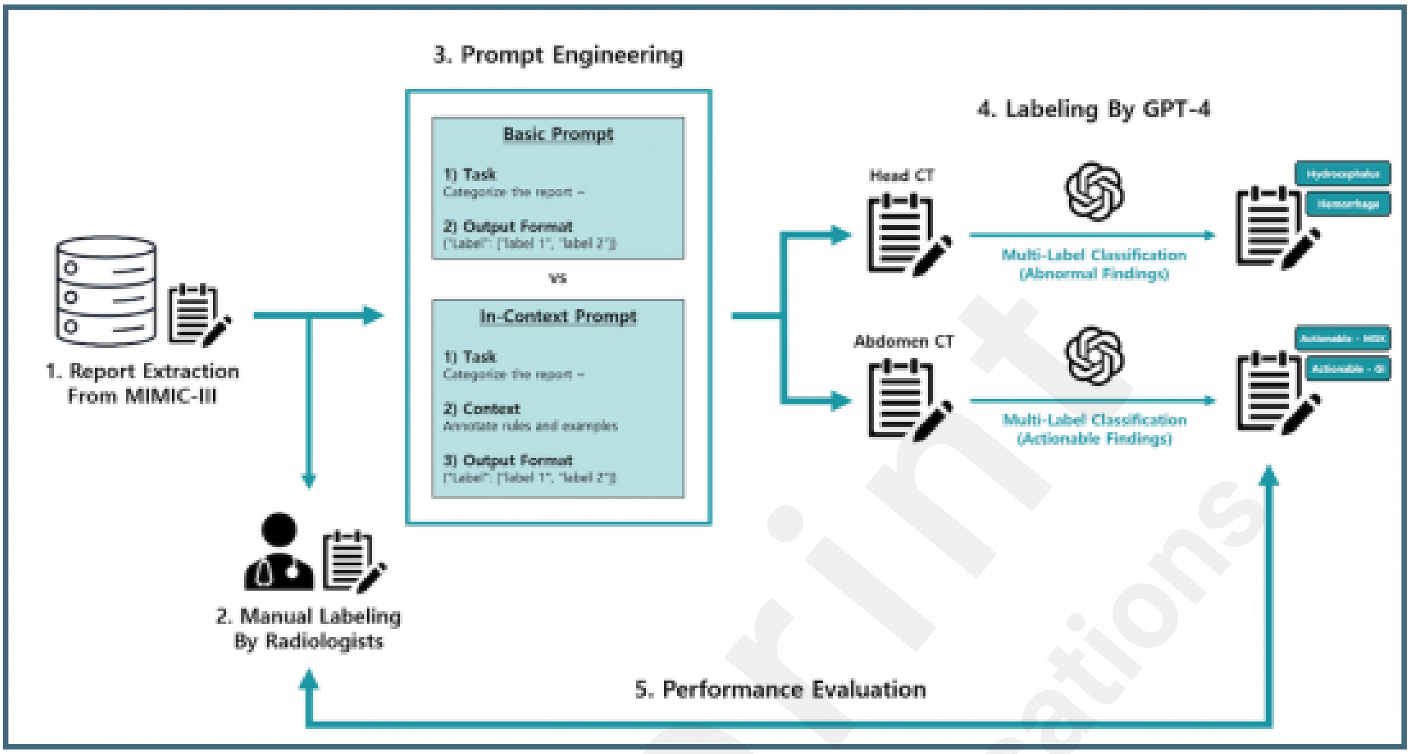
1;25(2):126–133.

46. Schmidt RA, Seah JCY, Cao K, Lim L, Lim W, Yeung J. Generative Large Language Models for Detection of Speech Recognition Errors in Radiology Reports. Radiology: Artificial Intelligence Radiological Society of North America; 2024 Jan 24; doi: 10.1148/ryai.230205

Supplementary Files

Figures

Overall flow of the study.



Inclusion and exclusion criteria of the MIMIC-III radiology reports.

Example of Head CT report extracted from MIMIC-III database and its labeling.

HISTORY: 76-year-old woman with Parkinson's disease with "large posterior circulation stroke, at OSH"; assess for bleed, thrombi, or dissection.

TECHNIQUE: Routine [**Hospital1 11**] study including contiguous 5-mm axial MDCT sections from the skull base to the vertex prior to contrast administration, with helical 1.25-mm axial sections from the level of the aortic arch through the vertex during dynamic intravenous administration of 80 mL Optiray-320. Sagittal, coronal, and axial 10-mm sections, as well as rotational 3D volume-rendered reconstructions of both the cervical and intracranial vessels, and rotational curved multiplanar reformations of the cervical vessels were reviewed on the workstation.

FINDINGS: The study is compared with the NECT of the head ([**Hospital 79244**] Hospital) obtained some nine hours earlier.

There has been no overall short-interval change in the appearance of the large, virtually complete left posterior cerebral arterial territorial infarction with extensive cytotoxic edema throughout this region and involvement of the lateral portion of the ipsilateral thalamus, likely splenium of corpus callosum and posteromedial temporal lobe. There are scattered curvilinear internal relatively hyperattenuating foci, also not significantly changed, which may represent petechial hemorrhage or, less likely, "islands" of spared brain. There is a vaguely triangular low-attenuation focus within the right hemipons, not clearly present earlier and difficult to confirm on the post-contrast images, which may be artifactual or represent additional relatively acute infarction. There is no evidence of involvement of additional vascular territories.

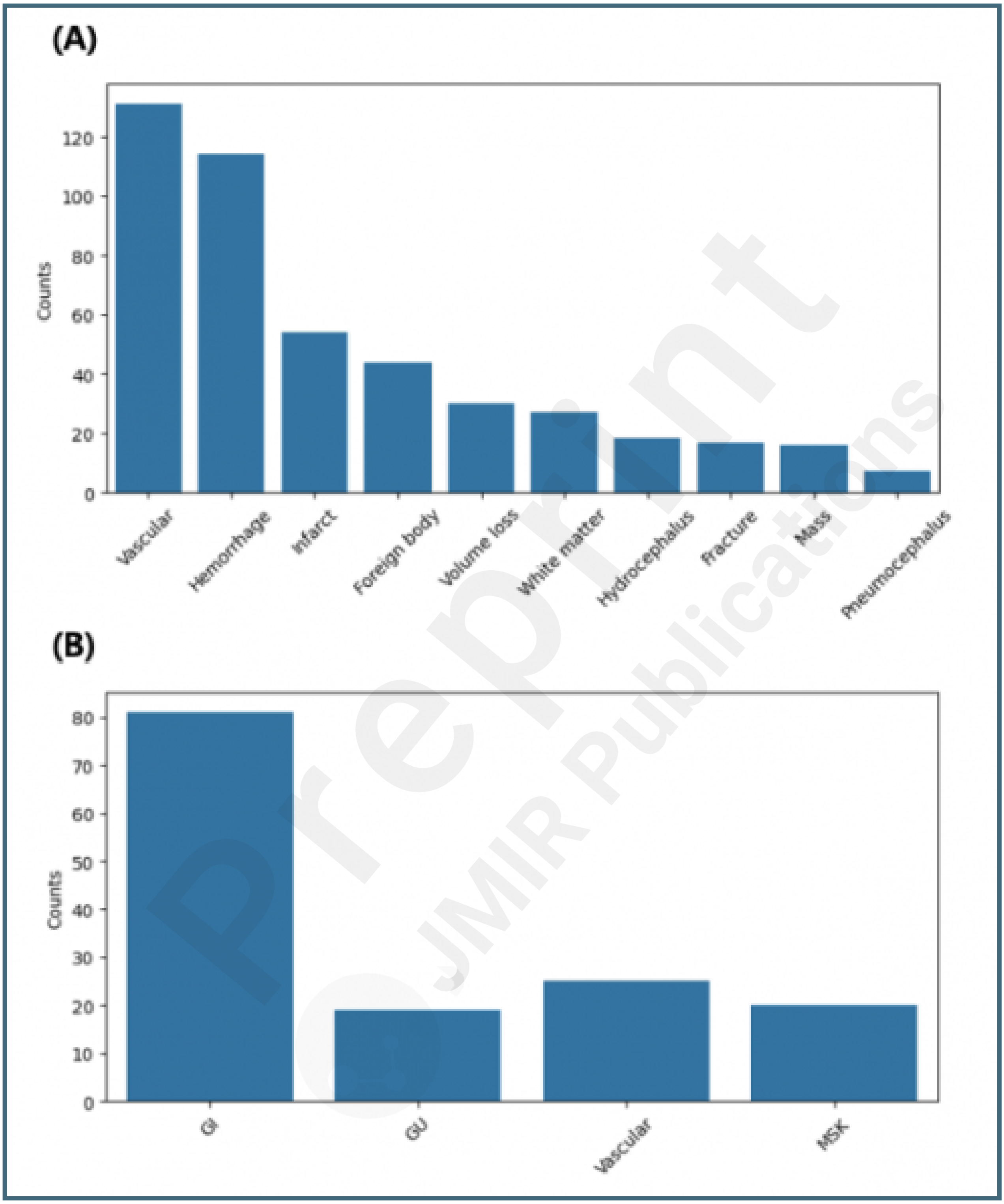
While there is atherosclerotic mural calcification involving the superior aspect of the aortic arch, as well as the left subclavian arteries, there is little atherosclerotic disease involving the common and internal carotid arteries throughout their course, to the level of the carotid termini. These vessels demonstrate normal caliber, with the left ICA measuring 6 mm at its proximal portion, just distal to the bifurcation and 5 mm at the skull base, and the right internal carotid artery measuring 7 mm proximally, just distal to the bifurcation and 5 mm, more distally, at the level of the skull base, with, therefore, no flow-limiting stenosis. The vertebral arteries are roughly co-dominant and demonstrate normal caliber, contour, and contrast enhancement throughout their course, with no flow-limiting stenosis or evidence of dissection. There is a normal appearance to the vertebrobasilar confluence, and normal contrast opacification and caliber of the principal vessels of the circle of [**Location (un) **], without significant mural irregularity or flow-limiting stenosis. Specifically, there is a normal appearance to the left posterior cerebral artery from its basilar artery origin throughout its more distal portion, which can be followed to the periphery of the infarcted vascular territory.

IMPRESSION:
1. No significant further interval extension of the large, virtually complete left PCA arterial territorial infarction since the [**Hospital 79244**] Hospital study obtained some nine hours earlier. This infarct involves the ipsilateral thalamus, medial temporal lobe and, likely, [**Last Name (un) 16610**] portions of the splenium of the corpus callosum.
2. Internal round and linear relatively hyperattenuating foci, in this context, suspicious for "petechial" hemorrhagic conversion.
3. Vaguely triangular low-attenuation focus within the right hemipons, not clearly present earlier and difficult to confirm on the post-contrast images, which may be artifactual or represent additional relatively acute infarction.
4. Unremarkable appearance to the circle of [**Location (un) **] without significant mural irregularity or flow-limiting stenosis; specifically, the left PCA is normal in caliber and opacification throughout its course through the infarcted territory, and may be recanalized.
5. Normal appearance to the common and internal carotid and vertebral arteries, bilaterally, with no significant mural irregularity or flow-limiting stenosis.

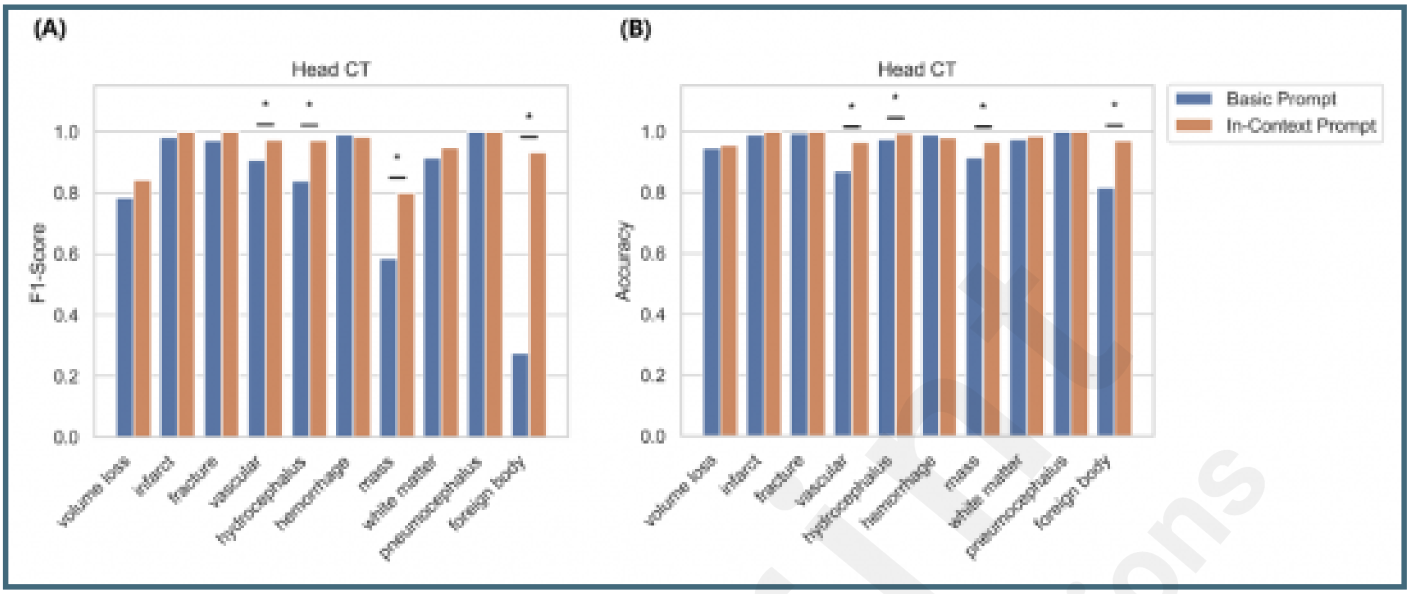
COMMENT: A preliminary interpretation of "Final read pending recons: Infarct in PCA territory, no ICH seen, COW apparently patent with left PCA intact" was discussed with the Neurology service by Dr. [**First Name (STitle) 596**] at the time of the study.

"Label": '**infarct**', '**hemorrhage**', '**vascular**'

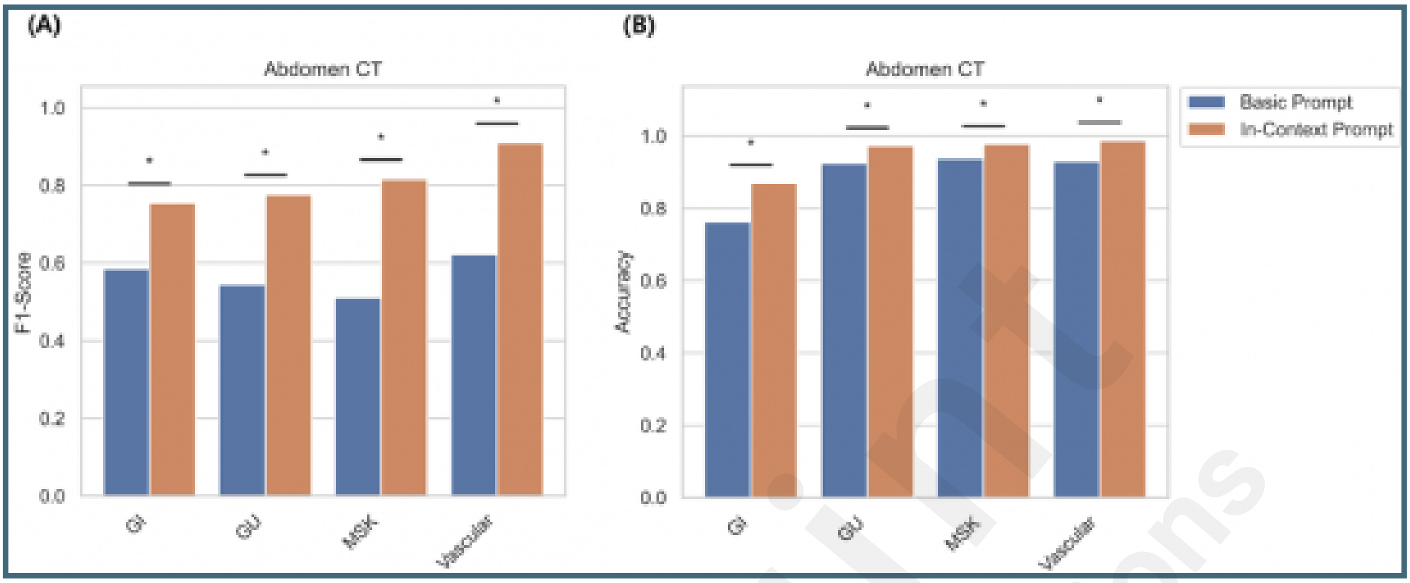
(A) Total number of labels labeled in Experiment 1. (B) Total number of labels labeled in Experiment 2.



Bar plot indicates the F1 scores and accuracy for each label in Experiment 1, as measured by two different prompts by GPT-4. Asterisk (*) indicates a statistically significant difference.



Bar plot indicates the F1 scores and accuracy for each label in Experiment 2, as measured by two different prompts by GPT-4. Asterisk (*) indicates a statistically significant difference.



Multimedia Appendixes

List of labels used in Experiments 1 and 2 and the descriptions of annotation instructions provided to annotators.
URL: http://asset.jmir.pub/assets/a37aa036b48c1671cca7378e67f3680a.pdf

Confusion matrix classified by GPT-4 in Experiment 1.
URL: http://asset.jmir.pub/assets/a17657d5069f1ab3d8aeb06bcb2a4171.png

Confusion matrix classified by GPT-4 in Experiment 2.
URL: http://asset.jmir.pub/assets/f47c91aebc0cd8f69e25f98eaa70b5cd.png