Original Article

Healthc Inform Res. 2025 July;31(3):295-309.
https://doi.org/10.4258/hir.2025.31.3.295
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

In-Context Learning with Large Language Models: A Simple and Effective Approach to Improve Radiology Report Labeling

Songsoo Kim[1,*], Donghyun Kim[2,*], Jaewoong Kim[1], Jalim Koo[3], Jinsik Yoon[4], Dukyong Yoon[1,5,6]

[1]Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Korea
[2]Department of Radiology, Central Draft Physical Examination Office of Military Manpower Administration, Daegu, Korea
[3]Department of Radiology, Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul, Korea
[4]Department of Integrative Medicine, Yonsei University College of Medicine, Seoul, Korea
[5]Center for Digital Health, Yongin Severance Hospital, Yonsei University College of Medicine, Yongin, Korea
[6]Institute for Innovation in Digital Healthcare, Severance Hospital, Seoul, Korea

Objectives: This study assessed the effectiveness of in-context learning using Generative Pre-trained Transformer-4 (GPT-4) for labeling radiology reports. **Methods:** In this retrospective study, radiology reports were obtained from the Medical Information Mart for Intensive Care III database. Two structured prompts—the "basic prompt" and the "in-context prompt"—were compared. An optimization experiment was conducted to assess consistency and the occurrence of output format errors. The primary labeling experiments were performed on 200 unseen head computed tomography (CT) reports for multi-label classification of predefined labels (Experiment 1) and on 400 unseen abdominal CT reports for multi-label classification of actionable findings (Experiment 2). **Results:** The inter-reader accuracies in Experiments 1 and 2 were 0.93 and 0.84, respectively. For multi-label classification of head CT reports (Experiment 1), the in-context prompt led to notable increases in F1-scores for the "foreign body" and "mass" labels (gains of 0.66 and 0.22, respectively). However, improvements for other labels were modest. In multi-label classification of abdominal CT reports (Experiment 2), in-context prompts produced substantial improvements in F1-scores across all labels compared to basic prompts. Providing context equipped the model with domain-specific knowledge and helped align its existing knowledge, thereby improving performance. **Conclusions:** In-context learning with GPT-4 consistently improved performance in labeling radiology reports. This approach is particularly effective for subjective labeling tasks and allows the model to align its criteria with those of human annotators for objective labeling. This practical strategy offers a simple, adaptable, and researcher-oriented method that can be applied to diverse labeling tasks.

Keywords: Radiology, Natural Language Processing, Medical Informatics, Artificial Intelligence, Computer-Assisted Diagnosis

Submitted: July 29, 2024, **Revised:** July 10, 2025, **Accepted:** July 22, 2025

Corresponding Author
Dukyong Yoon
Department of Biomedical Systems Informatics, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea.
Tel: +82-31-5189-8450, E-mail: dukyong.yoon@yonsei.ac.kr (https://orcid.org/0000-0003-1635-8376)
*These authors contributed equally to this work.

I. Introduction

Radiology report labeling significantly enhances the utility and value of reports, enabling their use beyond communication with physicians. Extracting clinical information as labels from reports provides a valuable resource for research [1,2]. These labels typically serve as ground truth for training artificial intelligence (AI) models [3], supporting the development of numerous predictive algorithms. By extracting detailed clinical nuances, researchers can define tailored cohorts for their studies [4,5]. This approach achieves greater precision than extraction using structured data from electronic medical records. Furthermore, rapid identification of findings in reports facilitates timely alerts for urgent issues [6,7] and enables follow-up recommendations [8], thereby improving patient care coordination.

Despite the growing need for accurate and efficient radiology report labeling, the process remains highly specialized and challenging, extending beyond simple text classification. One major challenge is the variability in how radiologists describe medical findings, even when referring to the same observation. As a result, an accurate understanding of both the clinical context and the imaging findings is essential for appropriate labeling. Another challenge is the broad spectrum of labeling tasks, ranging from straightforward disease categorization via keyword extraction to complex, high-level interpretive tasks requiring substantial medical expertise. Thus, a deep understanding of the labeling topic and the clinical context is crucial for proper categorization of radiology reports.

Recent advances in large language models (LLMs) have demonstrated significant potential for radiology report labeling, often surpassing traditional methods. Unlike rule-based approaches that rely on language- and institution-specific dictionaries [9], or earlier deep learning models, such as bidirectional encoder representations from transformers, that require domain-specific fine-tuning [3], modern LLMs excel at interpreting the nuanced context present in radiology reports. This strength has been especially valuable for tasks requiring a comprehensive understanding, such as report structuring, impression generation, and error detection [10-13].

In studies employing LLMs for radiology, prompt engineering—the process of carefully crafting prompts to enhance performance—is critical [14]. Several studies have implemented prompt engineering using a small subset of data prior to the main experiment and observed increases in F1-score or accuracy as a result [1,2,7,15]. However, specific details regarding prompt modifications and their direct impact on outcomes are often lacking, with most studies merely noting that prompts underwent iterative adjustments. Providing context knowledge has been shown to improve disease labeling performance in chest X-rays [16]. This technique, known as in-context learning (ICL), allows LLMs to acquire specific knowledge without changing internal parameters [17]. Unlike chain-of-thought prompting, which focuses on enhancing reasoning [18], or few-shot learning, which depends on careful selection of examples [19], ICL is particularly well-suited for subjective labeling tasks where identical findings may receive different labels depending on clinical context [20]. Nonetheless, existing studies rarely provide detailed explanations of how clinical context is chosen or systematically evaluate its quantitative and qualitative effects on model performance across different labeling tasks and difficulty levels.

In this study, we assessed the performance benefits of ICL on two distinct labeling tasks: labeling diseases on head computed tomography (CT) and urgent findings on abdominal CT. We further present both quantitative and qualitative analyses of how ICL influences the performance of these labeling tasks.

II. Methods

This study utilized de-identified, publicly available datasets and did not involve direct data collection from human subjects, exempting it from Institutional Review Board approval requirements. Figure 1 illustrates the study flow.

1. Data Curation

1) Report extraction, inclusion, and exclusion criteria

Radiology reports were sourced from the Medical Information Mart for Intensive Care III (MIMIC-III), one of the most extensively validated open-source databases [21]. MIMIC-III contains over 2 million anonymized free-text clinical notes, including a wide range of radiology reports, from 53,150 intensive care unit (ICU) patients at Beth Israel Deaconess Medical Center. The dataset is thoroughly anonymized, encompasses diverse findings, and has undergone rigorous quality control, making it suitable for evaluating GPT-4.

We selected head CT reports for multi-label classification of disease labels and abdominal CT reports for labeling actionable findings, due to their clinical importance and broad representation. In MIMIC-III, the "Description" column includes the relevant radiological assessments. We randomly sampled 220 head CT and 420 abdominal CT reports from these descriptions, excluding those related to procedures.

https://doi.org/10.4258/hir.2025.31.3.295

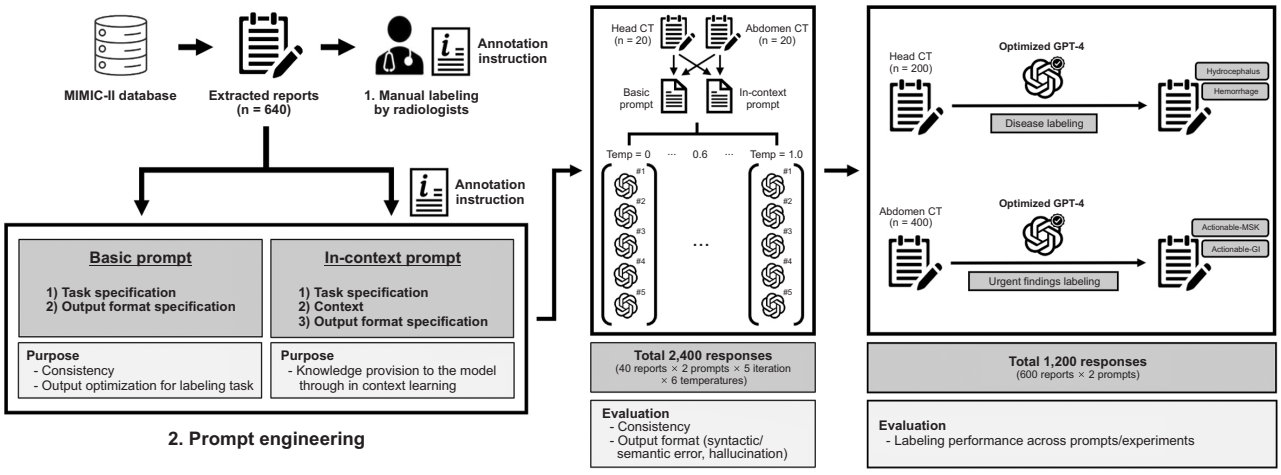

Figure 1. Overall flow of the study. The radiology reports obtained from the MIMIC-III database were manually labeled by radiologists. Prompt engineering involved developing "basic" and "in-context" prompts to provide context for in-context learning that effectively summarizes the instructions used by the human annotators. A parameter optimization experiment was conducted across multiple temperature settings to evaluate consistency and output format. The main labeling experiment utilized optimized GPT-4 models for disease and urgent findings labeling on head CT (n = 200) and abdominal CT (n = 400) reports, with a focus on evaluating the labeling performance of two different prompts. CT: computed tomography.

2) Manually labeling reports

Two board-certified radiologists (with 4 and 6 years of experience) manually labeled the radiology reports. For multi-label classification of head CT reports (Experiment 1), they assigned 10 labels (mass, hemorrhage, infarct, vascular, white matter, volume loss, hydrocephalus, pneumocephalus, foreign body, and fracture) to each report based on identified findings, allowing multiple labels per report. For multi-label classification of abdominal CT reports (Experiment 2), they labeled urgent findings and corresponding anatomical sections (gastrointestinal, genitourinary, musculoskeletal, and vascular) according to the American College of Radiology actionable reporting work group's classification [22]. Category 1 and 2 actionable findings—those requiring action within hours—were defined as "actionable findings within hours" to assess GPT-4's ability to identify clinically significant findings. Discrepancies between annotators were resolved by consensus. The complete rationale and references for the specific labels used in Experiments 1 and 2 are provided in Supplement A, Section I. Method *(*Experiments 1 and 2). A detailed description of the manual labeling workflow—including the pre-annotation strategy and the consensus procedure for resolving inter-reader discrepancies—is available in Supplement A.

2. Prompt Engineering

We designed two structured prompts: a "basic prompt," which included "task" and "output" sections, and an "in-context prompt," which added a "context" section (see Figure 1, Tables 1 and 2). The "task" section provided stepwise instructions, while the "output format" specification was intended to minimize verbosity, prevent hallucinations, and ensure output consistency. For post-processing, we used JavaScript Object Notation (JSON), a computer-friendly format.

In the "in-context prompt," the "context" section supplied additional contextual information—in this case, the annotation instructions used by human annotators for labeling. These instructions were carefully composed based on the hypothesis that ICL could enhance the model's labeling accuracy [17]. Our approach emphasizes brevity and clarity, unlike methods that present full reports and correct labels as examples (few-shot prompting) or employ stepwise reasoning (chain-of-thought prompting), which tend to produce unnecessarily long inputs or outputs.

A detailed rationale for our prompt design, along with an in-depth explanation of why this strategy is optimal, is provided in Supplement A.

3. Parameter Optimization Experiment

Although previous studies have explored the diverse capabilities of LLMs, their robustness and consistency remain insufficiently characterized. The flexibility of LLMs enables versatility across many tasks, but it can also lead to inconsistency and bias [23,24]. For example, GPT-4 has exhibited limited robustness and repeatability on radiology board-style examinations, highlighting the need for optimization [24].

To evaluate model consistency and determine the optimal

Table 1. Prompts used in the Experiment 1

Basic prompt	In-context prompt
Task - Categorize this report under the following labels: mass, hemorrhage, infarct, vascular, white matter, volume loss, hydrocephalus, pneumocephalus, foreign body, and fracture - If multiple labels are deemed appropriate, several of them may be assigned (except "normal"). Output (JSON) {"Label": ["label 1", "label 2"]} or {"Label": "none"}	Task - Review the entire Head CT Report and categorize this report under the following labels: mass, hemorrhage, infarct, vascular, white matter, volume loss, hydrocephalus, pneumocephalus, foreign body, and fracture - If multiple labels are deemed appropriate, several of them may be assigned (except "none"). **Context** { **"Mass": [** **"neoplasm",** **"abscess",** **"cyst",** **"other similar findings"** **],** **"Hemorrhage": [** **"epidural hematoma",** **"subdural hematoma",** **"subarachnoid hemorrhage",** **"intraparenchymal hemorrhage",** **"other similar findings"** **],** **"Infarct": [** **"acute infarct",** **"subacute infarct",** **"chronic infarct",** **"other similar findings"** **],** **"Vascular": [** **"aneurysm",** **"vascular steno-occlusive lesion",** **"vascular malformation",** **"arteriovenous fistula",** **"other similar findings"** **],** **"White matter": [** **"findings describing white matter inflammation",** **"small vessel disease",** **"other similar findings"** **],** **"Volume loss": [** **"diffuse brain atrophy",** **"encephalomalacia",** **"post-operative tissue changes",** **"chronic infarction with volume loss",**

Continued on the next page.

 https://doi.org/10.4258/hir.2025.31.3.295

Table 1. Continued

Basic prompt	In-context prompt
	"other similar findings" **],** **"Hydrocephalus": [** **"acute/chronic stable hydrocephalus",** **"ventricular enlargement",** **"normal pressure hydrocephalus",** **"other similar findings"** **],** **"Pneumocephalus": [** **"any findings suggestive of pneumocephalus on CT"** **],** **"Foreign body": [** **"shunt",** **"clips",** **"coils",** **"other materials related to surgery or procedure"** **],** **"Fracture": [** **"any displaced/non-displaced bony fracture on skull",** **"upper cervical vertebra"** **]** **}** Output (JSON) {"Label": ["label 1", "label 2"]} or {"Label": "none"}

"temperature" parameter—which controls the diversity of responses—we conducted an experiment using 40 reports (20 head CT and 20 abdominal CT). Two prompts per report were evaluated, and the temperature was varied from 0 to 1.0 in increments of 0.2, with 5 iterations per setting, yielding 2,400 responses. "Inconsistency" was defined as the proportion of responses deviating from the most frequent answer across the five iterations. "Output format error" was the sum of syntactic errors (JSON inaccuracies or minor structural issues), semantic errors (correct but imprecise labels), and undefined label errors (labels not present in the predefined set). Experiments were conducted on our institution's private Azure OpenAI platform, using the OpenAI application program interface (API) in a Python environment, with each query executed in a new session and in accordance with the PhysioNet Credentialed Data Use Agreement.

4. Main Labeling Experiment

For the main GPT-4 labeling experiment, we used 200 previously unseen head CT reports and 400 previously unseen abdominal CT reports, applying two prompts per task. GPT-4's performance was compared to the reference labels using precision, recall, F1-score, and accuracy. A qualitative review was also conducted to evaluate the effects of ICL on GPT-4's labeling.

5. Statistical Analysis

Inter-reader agreement for manual labeling was measured as accuracy on a per-label basis. The difference in performance metrics between the two prompting methods was calculated by subtracting the outcomes of the basic prompt from those of the in-context prompt. To assess the statistical significance of metric differences between the two prompts, we performed 1,000 bootstrap iterations to obtain the 95%

Table 2. Prompts used in the Experiment 2

Basic prompt	In-context prompt
Task - Review the entire abdomen CT Report and classify the reports into actionable and non-actionable categories. Actionable findings are those that need to be urgently communicated within hours. - Actionable findings should be further categorized into GI, GU, MSK, and Vascular sections (refrain from evaluating other sections). Note: a single report may contain multiple sections of actionable findings. - Actionable findings without significant interval changes compared with those of 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 those that need to be urgently communicated within hours. - Actionable findings should be further categorized into GI, GU, MSK, and Vascular sections (refrain from evaluating other sections). Note: a single report may contain multiple sections of actionable findings. - Actionable findings without significant interval changes compared with those of 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”: [**

Continued on the next page.

 https://doi.org/10.4258/hir.2025.31.3.295

Table 2. Continued

Basic prompt	In-context prompt
	"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) {"Actionable": ["section 1", "section 2"]} or {"Non-actionable": "NA"}

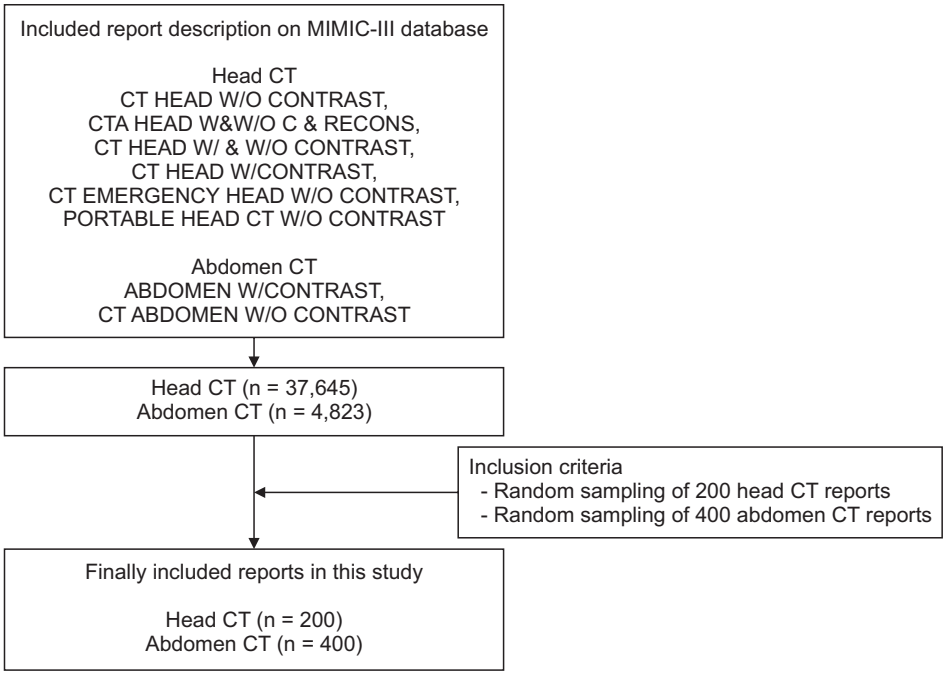

Figure 2. Inclusion and exclusion criteria of the MIMIC-III radiology reports. CT: computed tomography.

confidence intervals. Statistical significance was determined when the confidence interval did not include zero. All statistical analyses and data visualizations were performed in Python (version 3.11.4) using Pandas (version 2.1.1), SciPy (version 1.6.3), Matplotlib (version 3.4.2), and Seaborn (version 0.11.1).

III. Results

1. Baseline Characteristics

The characteristics of the MIMIC-III radiology reports are shown in Figure 2, with baseline characteristics summarized in Table 3. The dataset comprised 200 head CT scans, with a median word count of 279.5 and a median sentence count of 15.5. These head CT scans represented 174 patients (93 male), with a median age of 62.0 years. For abdominal CT reports, 400 reports were included, with a higher median word count (570.5) and sentence count (34.0). This cohort included 311 patients (176 males), also with a median age of 62.0 years. Examples of MIMIC-III reports are provided in Supplementary Figure S1.

2. Parameter Optimization Experiment

Inconsistency rates declined from 4%–12% at a temperature setting of 1.0 to 0%–3% at temperature 0 (Figure 3A). The output format error likewise decreased from 4%–8% at temperature 1.0 to 0% at temperature 0 (Figure 3B). Among the errors, syntactic errors were most frequent (27 cases), followed by semantic errors (17 cases), such as the use of "infarction" instead of "infarct." Only one undefined label error occurred, producing the label "actionable or non-actionable." No hallucinations (i.e., generation of entirely novel labels) were observed. Based on these results, temperature 0 was considered optimal, as it yielded the lowest inconsistency and output format error rates.

3. Multi-label Classification for Head CT (Experiment 1)

Excellent agreement was observed between the manual labels by the two readers (accuracy = 0.93). Across the labeled reports, an average of 2.29 labels per report was assigned (458 labels across 193 reports). The label distribution was: 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), and pneumocephalus (n = 7) (Figure 4A). Seven reports were not assigned any labels.

Using the basic prompt, GPT-4 demonstrated strong performance across most labels, with F1-scores ranging from 0.784 to 1.000, except for "mass" and "foreign body" (Figure 5A). Notably, the F1 scores for "foreign body" and "mass" increased from 0.275 to 0.933 (ΔF1 = 0.658; 95% confidence interval [CI], 0.519–0.838) and from 0.585 to 0.800 (ΔF1 = 0.215; 95% CI, 0.081–0.373), respectively. The score for "hydrocephalus" also improved, rising from 0.839 to 0.971 (ΔF1 = 0.132; 95% CI, 0.030–0.292). While performance improved for most labels, the extent of improvement was generally modest.

Labeling of "mass" often resulted in false positives. With the basic prompt, especially in cases describing a "mass effect," the model struggled to identify surgical materials as "foreign body," resulting in frequent false negatives (recall = 0.159). When "foreign body" labeling instructions explicitly included "shunts, clips, coils, or other materials related to surgery or procedure," the model successfully inferred not only these items but also others, such as "ventriculostomy tube," "ventriculostomy catheter," and "NG tube," even when not explicitly listed in the instructions.

Table 3. Baseline characteristics of the included MIMIC-III datasets

	CT reports	
	Head	Abdomen
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 (yr)	62.0 (48.0–74.0)	62.0 (49.0–74.0)
Sex, male	93	176

Values are presented as number or median (interquartile range).
CT: computed tomography.

4. Multi-label Classification for Abdominal CT (Experiment 2)

Manual labeling of abdominal CT reports demonstrated moderate agreement between the two readers (accuracy = 0.84). Among actionable reports, an average of 1.12 labels per report was assigned (145 labels across 129 reports). The distribution of actionable labels was as follows: n = 81 for

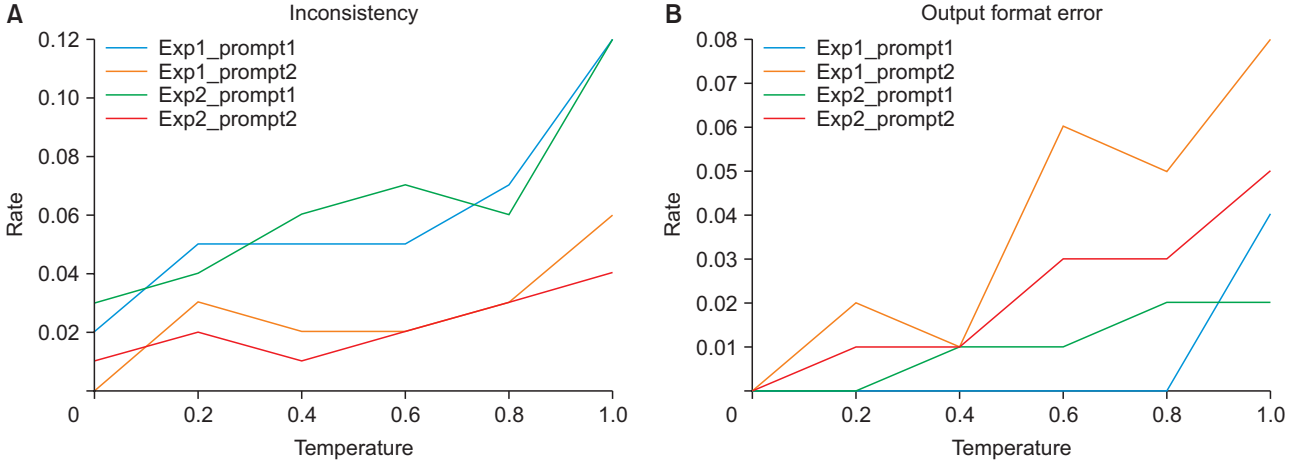

Figure 3. Optimization experiment. (A) Rates of inconsistency by temperature for different prompts in two experiments. (B) Rates of output format errors by temperature for different prompts in both experiments.

https://doi.org/10.4258/hir.2025.31.3.295

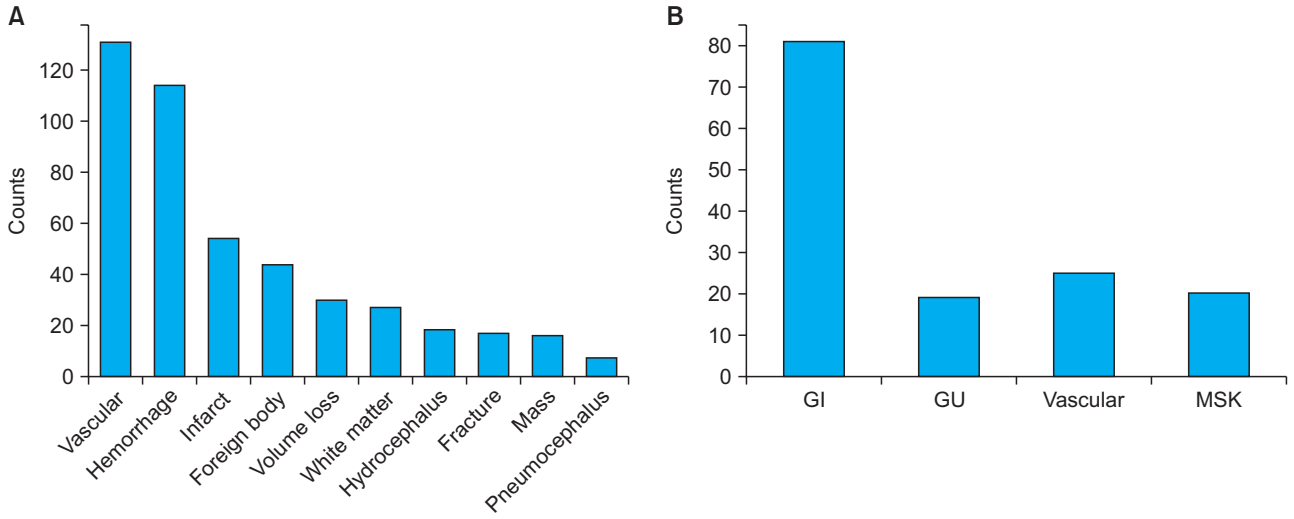

Figure 4. Total numbers of labels in (A) Experiment 1 - Multi-label classification for head CT, and (B) Experiment 2 - Multi-label classification for abdominal CT. CT: computed tomography, GI: gastrointestinal, GU: genitourinary, MSK: musculoskeletal.

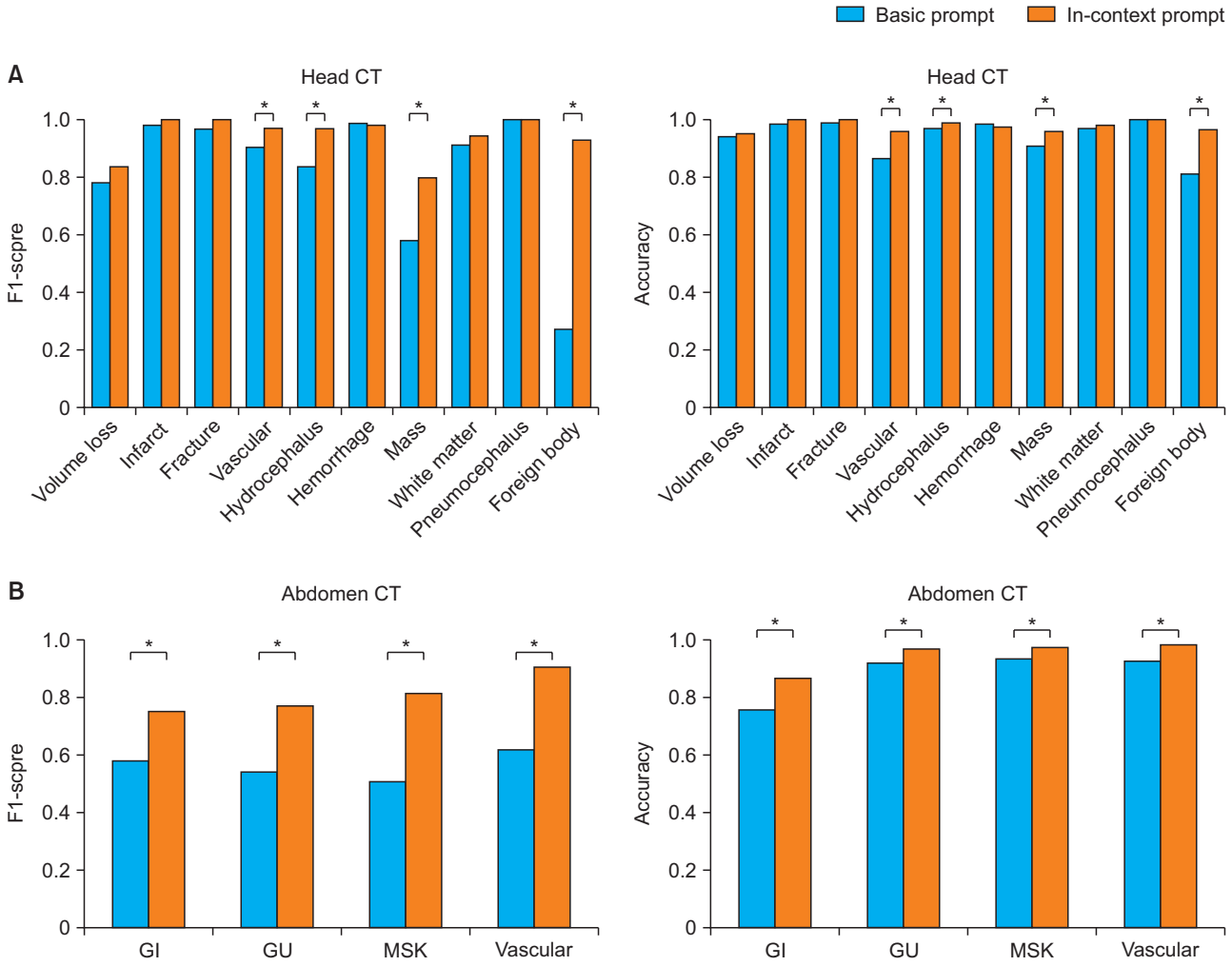

Figure 5. (A) Bar plot indicates the F1-scores and accuracy for each label in Experiment 1, as measured by two different prompts for GPT-4. (B) Bar plot indicates the F1-scores and accuracy for each label in Experiment 2, as measured by two different prompts for GPT-4. An asterisk (*) indicates a statistically significant difference. CT: computed tomography, GI: gastrointestinal, GU: genitourinary, MSK: musculoskeletal.

gastrointestinal (GI), n = 19 for genitourinary (GU), n = 20 for musculoskeletal (MSK), and n = 25 for vascular (Figure 4B). In total, 129 actionable and 271 non-actionable reports were identified. Discrepancies occurred primarily in cases with inconclusive imaging findings (e.g., unclear cause of pneumoperitoneum or infections), situations requiring subjective judgment without clear cutoffs (e.g., large-volume ascites), or risk assessments based solely on report text (e.g., risk for pathologic fracture).

The basic prompt yielded relatively low performance, with F1-scores ranging from 0.585 to 0.622, primarily due to frequent false positives. In contrast, the in-context prompt led to significant performance gains, with F1-scores across labels increasing from 0.17 to 0.306: GI ΔF1 = 0.170 (95% CI, 0.112–0.231), GU ΔF1 = 0.231 (95% CI, 0.133–0.348), MSK ΔF1 = 0.306 (95% CI, 0.144–0.481), and vascular ΔF1 = 0.288 (95% CI, 0.182–0.417), all statistically significant (Figure 5B). Most of the performance improvement was attributed to a decrease in false positives across label categories. Additionally, GPT-4 demonstrated the ability to identify several clinical contexts not explicitly stated in the instructions.

Detailed results for both experiments are presented in Table 4, Figure 5, and Supplementary Figures S2 and Figure S3. Representative cases and their descriptions, based on a review of model responses where ICL improved labeling after context was provided, are summarized in Table 5. Analysis of failed cases appears in Supplementary Table S4.

IV. Discussion

We observed significant performance gains in GPT-4 via ICL across two radiology report labeling experiments. The performance improvement was particularly marked in Experiment 2, where baseline performance and inter-reader agreement were lower. The model not only utilized concepts explicitly presented in the instructions, but also inferred additional, related concepts, further enhancing its labeling effectiveness. Moreover, our results confirm that low-temperature settings for GPT-4 resulted in highly consistent model behavior and strong adherence to the specified output format.

Previous labeling studies have employed a range of natural language processing approaches, from rule-based methods to deep learning models. Rule-based techniques, such as keyword or pattern searches (e.g., regular expressions), are interpretable but can generate false positives and often struggle with variations in sentence structure, medical abbreviations, and typographical errors [5,10]. Traditional deep learning models have demonstrated adaptability to various report types [25–27], but they require substantial training data and, once trained, may lack flexibility for other tasks. Furthermore, many of these models have not undergone external validation [28–30], limiting their generalizability as universal tools for labeling diverse radiology reports.

LLMs address many limitations of traditional approaches through their versatility. However, they function fundamentally differently from existing models, and their unique challenges must be addressed for optimal use. Their probabilistic outputs can result in variability in both output format and clinical judgment for identical queries—a phenomenon unfamiliar to traditional models and clinicians alike. Thus, before evaluating model accuracy, it is important to assess the extent of this variability for a given task. Our optimization experiments demonstrated that, by specifying an appropriate output format and utilizing low-temperature settings, LLMs can perform consistently for the same task.

By conducting two labeling experiments, we showed that ICL effectively improves labeling performance while supporting that the performance improvement is attributed to different mechanisms in each case. In multi-label classification for head CT (Experiment 1), the significant increase in labels such as “foreign body” and “mass” may be due to the alignment of the model with our desired output rather than an increase in domain knowledge. Considering the extensive training corpus of GPT-4, the definitions of labels such as “foreign body” or “mass” may not be absent. This is supported by the fact that the performance of the basic prompt was reasonably good for most labels. The substantial performance improvement in certain labels suggested that the model initially interpreted these labels differently from human annotators. The ability of the model to understand the intent conveyed by the instructions by observing a few cases and applying this understanding further contributed to performance improvement. Thus, providing annotation rules can enhance the potential of the model by reminding it of the correct label intent, including when the primary goal is not to supply domain-specific knowledge.

In multi-label classification for abdominal CT (Experiment 2), we observed a general increase in performance, probably because of an increase in task-specific domain knowledge rather than model alignment. The definition of an actionable finding can still be subjective at a specific level [7,22], and the American College of Radiology guideline used in this study is not a public document. Therefore, it may not have been included in the corpus or may have been learned in combination with various other information. In these

 https://doi.org/10.4258/hir.2025.31.3.295

Table 4. Performance metrics of both the "basic prompt" and "in-context prompt" in each experiment

	Precision					Recall					F1-score					Accuracy				
	BP	ICP	Δ	95% CI		BP	ICP	Δ	95% CI		BP	ICP	Δ	95% CI		BP	ICP	Δ	95% CI	
				Lower	Upper				Lower	Upper				Lower	Upper				Lower	Upper
Head CT																				
Volume loss	0.952	0.889	−0.063	−0.174	0.018	0.667	0.8	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	0.036	0.000	0.089	1	1	0	0.000	0.000	0.982	1	0.018	0.000	0.048	0.99	1	0.01	0.000	0.025
Fracture	0.944	1	0.056	0.000	0.188	1	1	0	0.000	0.000	0.971	1	0.029	0.000	0.105	0.995	1	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	0.87	0.965	0.095	0.045	0.145
Hydrocephalus	1	1	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	1	0	0.000	0.000	0.991	0.983	−0.008	−0.021	0.000	0.99	0.98	−0.01	−0.025	0.000
Mass	0.48	0.737	0.257	0.093	0.436	0.75	0.875	0.125	0.000	0.320	0.585	0.8	0.215	0.081	0.373	0.915	0.965	0.05	0.020	0.085
White matter	0.844	0.9	0.056	−0.045	0.179	1	1	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	1	0	0.000	0.000	1	1	0	0.000	0.000	1	1	0	0.000	0.000	1	1	0	0.000	0.000
Foreign body	1	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	0.97	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	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	0.69	0.271	0.098	0.450	0.65	1	0.35	0.136	0.579	0.51	0.816	0.306	0.144	0.481	0.938	0.978	0.04	0.015	0.068
Vascular	0.47	0.833	0.363	0.236	0.505	0.92	1	0.08	0.000	0.200	0.622	0.91	0.288	0.182	0.417	0.93	0.988	0.058	0.035	0.083

CT, computed tomography, GI: gastrointestinal, GU: genitourinary, MSK: musculoskeletal, BP: basic prompt, ICP: In-context prompt, Δ: difference between BP and ICP, CI: confidence interval.

Table 5. Example of how labeling performance improved by understanding and applying the given context in the "in-context prompt" scenario

	Report example	Explanation	Basic prompt	In-context prompt
Experiment 1	Findings …(truncated)… The coil in the anterior communicating artery is visualized. This examination does not reveal any evidence of residual lumen of the aneurysm; however, because of artifact from the coil, a tiny residual lumen cannot be completely excluded. …(truncated)…	The term "Foreign body" in the "Basic prompt" was presumably interpreted in a different sense in the absence of labeling. However, after the examples of the label were provided, it accurately labeled the report's "coil" by understanding the context of the label as surgical material or device.	Vascular	Vascular, Foreign body
	Findings: A predominantly hyperdense right thalamic hemorrhage is present measuring up to 2.8 cm in greatest dimension. This hemorrhage demonstrates a peripheral rim of hypodensity likely representing edema. The hemorrhage exerts a mass effect upon the right lateral ventricle without significant right to left midline shift. …(truncated)…	'Mass effect' was incorrectly labeled as 'mass' in the "Basic prompt." After the examples of the label 'mass' as neoplastic lesions were provided, it did not label the "mass-effect of hemorrhage" as a "Mass."	Hemorrhage, Mass	Hemorrhage
	Findings: …(truncated)… The degree of intraventricular hemorrhage is unchanged from the exam of [**2123-1-16**]. The ventricles are more dilated than those of the previous exam. The ventricular catheter is in an unchanged position. …(truncated)…	Although the term "hydrocephalus" is not directly mentioned in the report, "In-context prompt" indicated a situation of secondary hydrocephalus based on "ventricle dilatation," "intraventricular hemorrhage," and "catheter placement" and provided the appropriate label. Additionally, the "catheter" was a procedure-related material and provided the appropriate "foreign body" label.	(none)	Foreign body, Hydrocephalus
Experiment 2	Findings: ...(truncated)... Significant progression in osseous metastases involving the lumbar spine, entire pelvis, and both proximal femurs. No pathologic sacral or femoral fracture; however, large metastatic involvement in the superior aspect of the sacrum is at high risk for pathologic fracture. Moreover, abutment and encasement of the exiting left S1 and S2 nerve roots is observed. …(truncated)…	After "Bone lesions with fracture risk" was provided in the "In-context prompt," the report was appropriately labeled based on the determination of whether contents such as "Significant osseous metastases" and "Pathologic fracture" matched the given description.	(none)	MSK
	Findings: ...(truncated)... Left hydronephrosis and delayed contrast excretion, with blood in the left renal pelvis. This appears to be secondary to a filling defect, which may represent a blood clot or a mass, in the left ureter. ...(truncated)...	Although not directly mentioned, the report contained information such as "hydronephrosis" and "filling defect," interpreted them as "obstruction in the urinary tract" provided in the "In-context prompt" and then provided the appropriate label.	(none)	GU

MSK: musculoskeletal, GU: genitourinary.

subjectivity tasks, the ICL of GPT-4 effectively provides task-specific knowledge within the prompt and enables studies to induce positive bias in the desired direction depending on the context provided.

Both experiments demonstrated that supplying conceptual information—without overly specific examples or verbatim sentences—enables the model to flexibly apply similar clinical concepts in radiology reports. This flexibility is a major advantage when using LLMs. While such human-written instructions require domain expertise, they are considerably more feasible and effective than attempting to include every real-world example within the prompt.

The strengths and implications of this study are threefold. First, we showed that providing annotation instructions consistent with those used by human annotators offered meaningful context to the model and improved performance. This approach was efficient, enabling the model to generalize beyond directly provided information, and subjective tasks especially benefited from contextual prompts. Second, we proposed a simple, reusable, and efficient ICL framework for radiology report labeling, utilizing existing prompt engineering techniques. The prompt components can be flexibly adapted, suggesting that targeted guidance for subjective issues can support research-specific labeling needs. Third, we quantitatively assessed model consistency and confirmed the importance of prompt and parameter optimization in radiology report labeling tasks.

Nonetheless, this study has several limitations that future research can address. First, all experiments were conducted using the MIMIC database. Although this ICU dataset included severe and complex cases—posing a meaningful challenge for LLMs and reinforcing the significance of our findings—the generalizability of our results warrants careful consideration. The MIMIC-CXR reports reflect a specific institutional format and clinical environment, and any potential biases in our manually labeled validation dataset could further limit generalizability. Future studies should validate these findings across different institutions, clinical settings, languages, and radiology modalities, especially examining ICL's performance with other imaging modalities (e.g., MRI, CT, ultrasound) that feature unique reporting complexities and disease spectra. Second, we evaluated only GPT-4, as larger LLMs with many parameters are thought to benefit most from ICL, and GPT-4, as the largest available model, was thus selected for this evaluation [30]. Subsequent work should focus on validating open-source models that are effective yet less computationally demanding [5,12,15].

In conclusion, our study demonstrates that GPT-4 with ICL significantly and consistently enhances performance on radiology report labeling tasks. This practical approach offers a simple, flexible, and researcher-adaptable method that can be broadly applied to diverse labeling scenarios. By leveraging ICL, the utility of radiology reports can be further extended to support research, artificial intelligence model development, and improved patient care coordination.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

The research was supported by the MD-Phd/Medical Scientist Training Program through the Korea Health Industry Development Institute, funded by the Ministry of Health & Welfare; and the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare (No. HI22C0452).

Datasets analyzed during the study are available in the MIMIC-III database, which is a publicly and freely available database developed by the MIT Lab for Computational Physiology. Researchers seeking to use this database must complete the required training and obtain approval from an ethics review board.

ORCID

Songsoo Kim (https://orcid.org/0000-0002-6908-4324)
Donghyun Kim (https://orcid.org/0000-0002-9353-775X)
Jaewoong Kim (https://orcid.org/0000-0002-5706-0181)
Jalim Koo (https://orcid.org/0000-0001-6387-2277)
Jinsik Yoon (https://orcid.org/0009-0001-8351-3797)
Dukyong Yoon (https://orcid.org/0000-0003-1635-8376)

Supplementary Materials

Supplementary materials can be found via https://doi.org/10.4258/hir.2025.31.3.295.

References

1. Fink MA, Bischoff A, Fink CA, Moll M, Kroschke J, Dulz L, et al. Potential of ChatGPT and GPT-4 for data mining of Free-Text CT reports on lung cancer. Radiol-

ogy 2023;308(3):e231362. https://doi.org/10.1148/radiol.231362
2. Gu K, Lee JH, Shin J, Hwang JA, Min JH, Jeong WK, et al. Using GPT-4 for LI-RADS feature extraction and categorization with multilingual free-text reports. Liver Int 2024;44(7):1578-87. https://doi.org/10.1111/liv.15891
3. Zech JR. Using BERT models to label radiology reports. Radiol Artif Intell 2022;4(4):e220124. https://doi.org/10.1148/ryai.220124
4. 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. https://doi.org/10.1186/s12911-021-01533-7
5. Alsentzer E, Rasmussen MJ, Fontoura R, Cull AL, Beaulieu-Jones B, Gray KJ, et al. Zero-shot interpretable phenotyping of postpartum hemorrhage using large language models. NPJ Digit Med 2023;6(1):212. https://doi.org/10.1038/s41746-023-00957-x
6. Banerjee I, Davis MA, Vey BL, Mazaheri S, Khan F, Zavaletta V, et al. Natural language processing model for identifying critical findings-a multi-institutional study. J Digit Imaging 2023;36(1):105-13. https://doi.org/10.1007/s10278-022-00712-w.
7. Woo KC, Simon GW, Akindutire O, Aphinyanaphongs Y, Austrian JS, Kim JG, et al. Evaluation of GPT-4 ability to identify and generate patient instructions for actionable incidental radiology findings. J Am Med Inform Assoc 2024;31(9):1983-93. https://doi.org/10.1093/jamia/ocae117
8. 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;2020:335-44.
9. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. Proc AAAI Conf Artif Intell 2019;33(1):590-7. https://doi.org/10.1609/aaai.v33i01.3301590
10. Adams LC, Truhn D, Busch F, Kader A, Niehues SM, Makowski MR, et al. Leveraging GPT-4 for Post Hoc Transformation of Free-text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study. Radiology. 2023;307(4):e230725. https://doi.org/10.1148/radiol.230725
11. Sun Z, Ong H, Kennedy P, Tang L, Chen S, Elias J, et al. Evaluating GPT4 on Impressions Generation in Radiology Reports. Radiology. 2023;307(5):e231259. https://doi.org/10.1148/radiol.231259
12. Mukherjee P, Hou B, Lanfredi RB, Summers RM. Feasibility of using the privacy-preserving large language model vicuna for labeling radiology reports. Radiology 2023;309(1):e231147. https://doi.org/10.1148/radiol.231147
13. Kim S, Kim D, Shin HJ, Lee SH, Kang Y, Jeong S, et al. Large-scale validation of the feasibility of GPT-4 as a proofreading tool for head CT reports. Radiology 2025;314(1):e240701. https://doi.org/10.1148/radiol.240701
14. Nguyen D, Swanson D, Newbury A, Kim YH. Evaluation of ChatGPT and Google Bard using prompt engineering in cancer screening algorithms. Acad Radiol 2024;31(5):1799-804. https://doi.org/10.1016/j.acra.2023.11.002
15. Schmidt RA, Seah JC, Cao K, Lim L, Lim W, Yeung J. Generative large language models for detection of speech recognition errors in radiology reports. Radiol Artif Intell 2024;6(2):e230205. https://doi.org/10.1148/ryai.230205
16. Savage CH, Park H, Kwak K, Smith AD, Rothenberg SA, Parekh VS, et al. General-purpose large language models versus a domain-specific natural language processing tool for label extraction from chest radiograph reports. AJR Am J Roentgenol 2024;222(4):e2330573. https://doi.org/10.2214/AJR.23.30573
17. Dong Q, Li L, Dai D, Zheng C, Ma J, Li R, et al. A survey on in-context learning [Internet]. Ithaca (NY): arXiv.org; 2024 [cited at 2025 Jul 1]. Available from: https://arxiv.org/abs/2301.00234.
18. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst 2022;35:24824-37.
19. Liu J, Shen D, Zhang Y, Dolan B, Carin L, Chen W. What makes good in-context examples for GPT-3? [Internet]. Ithaca (NY): arXiv.org; 2021 [cited at 2025 Jul 1]. Available from: https://arxiv.org/abs/2101.06804.
20. Rouzrokh P, Khosravi B, Faghani S, Moassefi M, Vera Garcia DV, Singh Y, et al. Mitigating Bias in Radiology Machine Learning: 1. Data Handling. Radiol Artif Intell 2022;4(5):e210290. https://doi.org/10.1148/ryai.210290
21. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035. https://doi.org/10.1038/sdata.2016.35
22. Larson PA, Berland LL, Griffith B, Kahn CE Jr, Liebscher LA. Actionable findings and the role of IT sup-

port: report of the ACR Actionable Reporting Work Group. J Am Coll Radiol 2014;11(6):552-8. https://doi.org/10.1016/j.jacr.2013.12.016
23. Stureborg R, Alikaniotis D, Suhara Y. Large language models are inconsistent and biased evaluators [Internet]. Ithaca (NY): arXiv.org; 2024 [cited at 2025 Jul 1]. Available from: https://arxiv.org/abs/2405.01724.
24. Krishna S, Bhambra N, Bleakney R, Bhayana R. Evaluation of reliability, repeatability, robustness, and confidence of GPT-3.5 and GPT-4 on a radiology board-style examination. Radiology 2024;311(2):e232715. https://doi.org/10.1148/radiol.232715
25. Yan A, McAuley J, Lu X, Du J, Chang EY, Gentili A, et al. RadBERT: adapting transformer-based language models to radiology. Radiol Artif Intell 2022;4(4):e210258. https://doi.org/10.1148/ryai.210258
26. Zaman S, Petri C, Vimalesvaran K, Howard J, Bharath A, Francis D, et al. Automatic diagnosis labeling of cardiovascular MRI by using semisupervised natural language processing of text reports. Radiol Artif Intell 2021;4(1):e210085. https://doi.org/10.1148/ryai.210085
27. 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;4(4):e220007. https://doi.org/10.1148/ryai.220007
28. Weng KH, Liu CF, Chen CJ. 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;11:e46348. https://doi.org/10.2196/46348
29. Lopez-Ubeda P, Martin-Noguerol T, Luna A. Automatic classification and prioritisation of actionable BI-RADS categories using natural language processing models. Clin Radiol 2024;79(1):e1-e7. https://doi.org/10.1016/j.crad.2023.09.009
30. Wei J, Wei J, Tay Y, Tran D, Webson A, Lu Y, et al. Larger language models do in-context learning differently [Internet]. Ithaca (NY): arXiv.org; 2023 [cited at 2025 Jul 1]. Available from: https://arxiv.org/abs/2303.03846.