Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01522-4

Weakly supervised language models for automated extraction of critical findings from radiology reports

Check for updates

Avisha Das[1], Ish A. Talati[2], Juan Manuel Zambrano Chaves[3], Daniel Rubin[2,3] & Imon Banerjee[1,4,5] ✉

Critical findings in radiology reports are life threatening conditions that need to be communicated promptly to physicians for timely management of patients. Although challenging, advancements in natural language processing (NLP), particularly large language models (LLMs), now enable the automated identification of key findings from verbose reports. Given the scarcity of labeled critical findings data, we implemented a two-phase, weakly supervised fine-tuning approach on 15,000 unlabeled Mayo Clinic reports. This fine-tuned model then automatically extracted critical terms on internal (Mayo Clinic, $n = 80$) and external (MIMIC-III, $n = 123$) test datasets, validated against expert annotations. Model performance was further assessed on 5000 MIMIC-IV reports using LLM-aided metrics, G-eval and Prometheus. Both manual and LLM-based evaluations showed improved task alignment with weak supervision. The pipeline and model, publicly available under an academic license, can aid in critical finding extraction for research and clinical use (https://github.com/dasavisha/CriticalFindings_Extract).

Radiology reports contain substantial information about a patient including a radiologist's observations and diagnosis from images as well as critical and incidental findings[1]. Critical findings are observations or results that require immediate communication (typically within hours) with the patient's healthcare provider, since a delay in reporting such findings could cause a serious impact on the patient's health and well-being[2,3]. Moreover, these findings must be communicated to the treating clinicians who must keep track in order to render necessary medical care in a timely manner[4].

The extraction of critical findings from radiology reports presents opportunities for quality improvement and workflow optimization. Radiology reports are lengthy and suffer from inconsistencies in structure and format, which can make retrospective analysis and quality assurance challenging[3]. Each report can vary significantly in terms of the terminology used, the order in which the observations and findings are presented, and the level of detail provided[5]. This variability complicates systematic review and analysis of report content for quality assurance, research, and process improvement purposes[4]. While acute communication of critical findings occurs through direct channels like phone or chat, automated extraction tools can support downstream applications such as quality monitoring, compliance tracking, and institutional dashboards. These challenges highlight the need for reliable automated extraction of key findings from radiology reports to support these important secondary use cases.

Previous researchers have shown that artificial intelligence (AI)-based methods are useful in efficiently mining information from radiology reports and images[6]. Lakhani et al.[7] and Heilbrun et al.[8] propose systems that focus on a small dataset of chest CT reports. They propose rule-based text-mining approaches to automatically identify radiology reports containing critical findings. Mabotuwan et al.[9] proposed an automated rule-based framework to extract critical findings from 1.2M radiology exams with an approximately 90% accuracy. This work uses a pre-defined list of ten specific critical finding terms and only reports that contain one or more of these terms are identified as critical. However, such rule-based models are limited in scope and approach, and lack generalization towards external data.

In subsequent years, studies have proposed clinical BERT-based models to identify reports that contain critical findings[4,10]. But the proposed models are limited in their scope of application, focusing only on Chest CT reports. Moreover, these studies can only classify whether a report contains any critical finding(s) or not, without extracting the specific category or type of findings.

With the recent advancements in natural language processing (NLP), large language models (LLMs) can be leveraged to automatically understand and retrieve answers to queries from related textual content[11–14]. Furthermore, with domain-specific fine-tuning, LLMs can efficiently detect critical information, showing state-of-the-art performance[15,16] on retrieval tasks like

[1]Arizona Advanced AI & Innovation (A3I) Hub, Mayo Clinic Arizona, Phoenix, AZ, USA. [2]Department of Radiology, Stanford University, Stanford, CA, USA. [3]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [4]Department of Radiology, Mayo Clinic Arizona, Phoenix, AZ, USA. [5]School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, USA. ✉e-mail: Banerjee.Imon@mayo.edu

medical question-answering, biomedical entity recognition, retrieval of related knowledge and clinical guidelines[17–19]. LLMs like GPT-4 have been used in[20–22] for identification and interpretation of findings from radiology reports, largely reducing the manual effort required to parse such documents.

Use of general domain LLMs seem promising. But pre-trained AI-based systems must be aligned to domain-specific and task-specific data for more efficient performance and increased understanding of underlying clinical objectives[6]. Alignment of general domain LLMs relies on such domain and task specific data; but there exists a dearth of annotated data when it comes to the identification of critical findings in radiology reports[4,9]. Furthermore, although many academic centers maintain databases of tagged critical exams, they often lack detailed definitions and annotations of these findings. This gap makes it impractical and challenging to collect training data for extracting critical findings directly.

We propose an end-to-end pipeline to automatically identify and categorically extract critical findings from radiology reports. The major contributions of this work are as follows:

- To address the issue of scarce labeled data for model fine-tuning[4], we implement a novel weakly supervised and task-specific instruction-based training setup to align the Mistral-based LLMs for retrieval of critical and incidental findings from radiology reports across a wide variety of modalities and sites.
- We manually curate a list of 210 critical finding-based terms that denote critical findings across different anatomical sites and reports. We expand upon an initial list by the Actionable Reporting Work Group (ACR)[23] using ontological term-based expansion to create the first comprehensive list of terms to identify critical findings from a report.
- We perform a novel evaluation, comparing the efficiency of LLM-aided evaluation to human-aided metrics to empirically analyze the usefulness and feasibility of using LLM-based oracle models like GPT-4, LLaMa, for NLP extraction tasks from radiology reports at scale. To verify the model's performance and generalizability, we evaluate the system on both internal Mayo Clinic and external publicly available MIMIC-III[24] and MIMIC-IV[25] reports.

To the best of our knowledge, this is the first proposed system that can successfully identify and extract critical findings from radiology reports, irrespective of type and category.

Table 1 | Statistics of the datasets used in the development and validation - private Mayo clinic dataset and publicly available MIMIC-III and MIMIC-IV

Attributes	Dataset		
	Mayo Clinic	MIMIC-III	MIMIC-IV
Training dataset size	15,000	–	–
Test dataset size	80	123	5000
Avg. report length	476.3	53.5	199.3
Modality	Frequency (% reports)		
CT-Scan	6345 (41.4)	73 (60.9)	1490 (29.8)
XR	4127 (26.9)	35 (28.4)	1296 (25.9)
MR	3278 (21.4)	9 (7.3)	579 (11.6)
Ultrasound	1507 (9.8)	2 (1.7)	599 (12.0)
Other	73 (0.48)	4 (3.2)	1036 (20.7)
Anatomy	Frequency (% reports)		
Chest	1905 (12.4)	45 (36.6)	1834 (36.7)
Head	4770 (31.1)	30 (24.4)	1278 (25.6)
Neck	2835 (18.5)	15 (1.6)	427 (8.5)
Abdomen	5820 (39.9)	28 (31.7)	634 (12.7)
Other	–	4 (3.2)	1024 (20.5)
Result statistics	Frequency (% reports)		
Test reports with critical findings	15 of 80 (18.7)	50 of 123 (40.6)	–
Top 5 findings	Small bowel obstruction, Pulmonary embolism, Pleural effusion, Occlusion of Artery, Lesion in kidney	Small bowel obstruction, Ischemic bowel, Subarachnoid hemorrhage, Parenchymal hemorrhage, Subdural Hematoma	–

Results

Dataset

We collected a private dataset of radiology reports of various modalities (MR, CT, Radiograph, US) and anatomy (chest, abdomen, head, extremities) from four Mayo Clinic sites (Arizona, Rochester, Florida and Mid-West). For finetuning the model, we randomly selected 15,000 reports documented between 2015 and 2021 at Mayo Clinic and created a held-out dataset of randomly selected 80 reports for our internal model evaluation (see Table 1). To measure the models' generalizability, we created an expert-annotated smaller dataset of 123 reports of varying modalities randomly selected from the publicly available MIMIC-III corpus[24]. The inter-annotator agreement for both manually annotated datasets was assessed using Cohen's Kappa score (κ). For the MIMIC-III dataset, the Kappa score was $\kappa = 0.738$, while for the Mayo Clinic dataset, it was $\kappa = 0.798$. Additionally, to evaluate the model performance on a large-scale dataset, we randomly selected 5000 radiology reports from the publicly available MIMIC-IV[25] (see Table 1).

Quantitative performance

We evaluate the proposed pipeline performance using two *smaller manually annotated test datasets* - (a) internal hold-out test set of randomly selected 80 radiology reports from Mayo Clinic, and (b) external test set of 123 radiology reports from MIMIC-III, and a *large-scale test dataset* of 5000 randomly selected radiology reports from MIMIC-IV. We compare the performance of the weakly finetuned models (WFT) with the baseline pre-trained versions of the model (PT). We provide more details on the task template and prompts for zero-shot and few-shot-based weak label generation, as well as the examples used in few-shot prompting in the Supplementary Sections A and B respectively. Figure 1 demonstrates the model performance for two human-based metrics (ROUGE-2, BLEU) and two LLM-based metrics (G-Eval, Prometheus). Model performance on the large-scale dataset is evaluated using LLM-based metrics, as shown in Fig. 2. The performance of the models are also evaluated using the classification-based metrics like F1-score, precision and recall, calculated on the Mayo Clinic and MIMIC-III datasets as shown in Table 2. For qualitative evaluation and further error analysis, we present examples of the extracted critical finding terms in Table 3. A larger set of human-based and LLM-based metrics report a detailed overview of the model performance on the small-scale human annotated internal and external datasets and large-scale MIMIC dataset respectively in the Supplementary Tables 1 and 2 (Supplementary Section 1).

The quantitative results on the *small-scale* internal and external validation datasets (Fig. 1) show that with weakly supervised fine-tuning, the Mistral and BioMistral models perform consistently better than the pre-trained baseline. Moreover, the fine-tuned Mistral model performance is better than the BioMistral LLM. The models that were weakly fine-tuned with labels generated using few-shot prompting in Phase I, consistently performed better on the quantitative metrics. On the internal Mayo test set, the weakly fine-tuned Mistral model achieves a 48% Rouge-2 score, performing better than the pre-trained Mistral and BioMistral models. Similarly, the fine-tuned BioMistral model comes a close second, achieving 41%

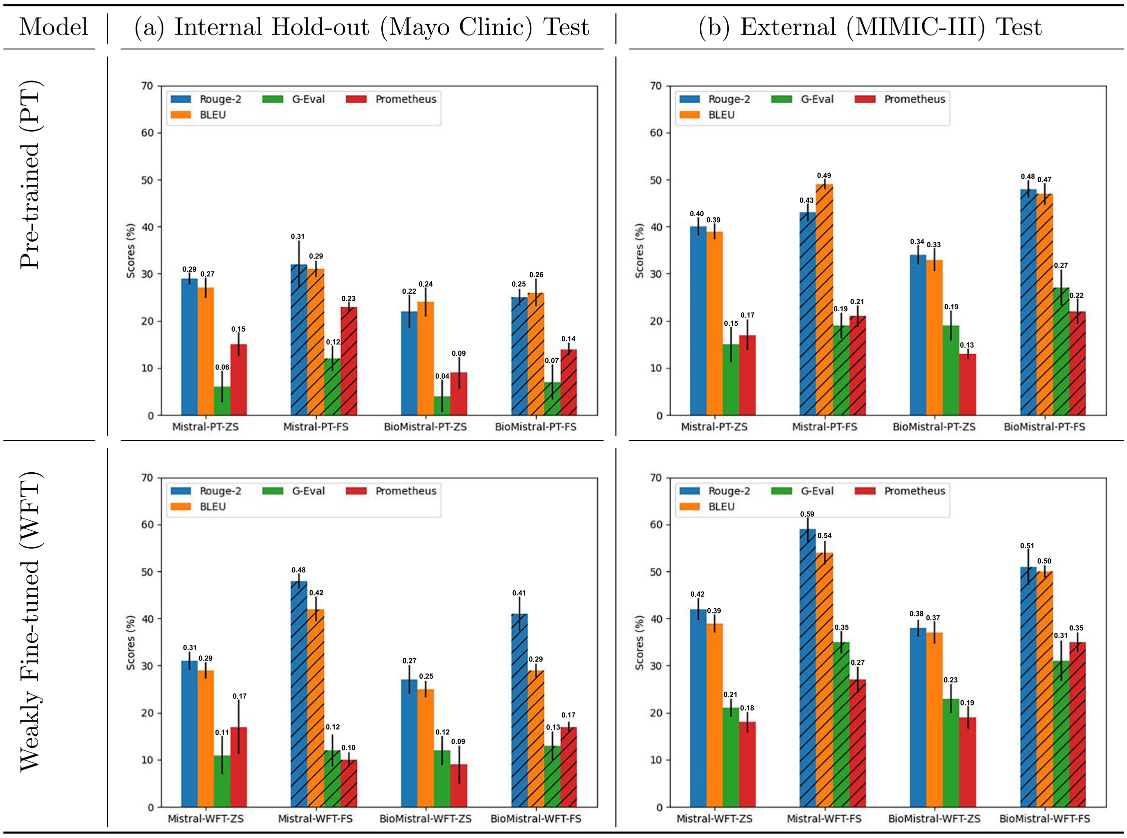
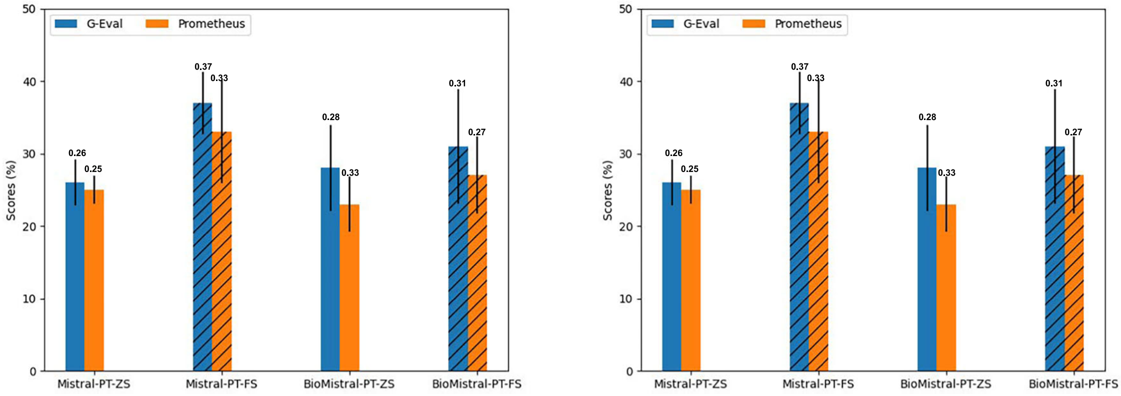


Fig. 1 | Model performance on human-annotated test datasets. a This column shows the performance on the Internal hold-out (Mayo Clinic) test set; **b** This column shows the performance on the External (MIMIC-III) test set. Each row corresponds to the pre-trained (PT) and weakly fine-tuned (WFT) models. For each model, we plot the histogram bars for Rouge-1 (blue), BLEU (yellow), G-Eval (green), and Prometheus (red). For each model, we consider two prompting techniques: Zero-shot (ZS) and Few-shot (FS) for weak label generation. The score for each metric (normalized between 0 and 1) is added to the top of its corresponding bar plot. Error bars denote standard deviations. The scores for models trained using weak labels generated by FS-based prompting are shown with ▨.



(a) Pre-trained (PT) models

(b) Weakly Fine-tuned (WFT) models

Fig. 2 | Model performance on large-scale test dataset from MIMIC-IV. LLM-aided large-scale evaluation on **a** Pre-trained models and **b** Weakly Fine-tuned models. For each model, we plot the histogram bars for G-Eval (blue) and Prometheus (yellow). For each model, we consider two prompting techniques: Zero-shot (ZS) and Few-shot (FS) for weak label generation. The score for each metric (normalized between 0 and 1) is added to the top of its corresponding bar plot. Error bars denote standard deviations. The scores for models trained using weak labels generated by FS-based prompting are shown with ▨.

Rouge-2 score on the internal test data. On the external MIMIC-III test reports we observe a similar trend, with the weakly fine-tuned Mistral and BioMistral models performing better than their pre-trained unsupervised baselines with Rouge-2 scores of 59% and 51% respectively.

This observation is also supported by the F1-score results of weakly fine-tuned BioMistral (BioMistral-WFT) model, especially when the fine-tuning was done with labels generated using few-shot prompting. The weakly fine-tuned model consistently delivers the best performance across both the internal hold-out dataset from Mayo Clinic and the external MIMIC-III dataset and achieves the highest F1-scores, reaching 0.57 on both datasets, suggesting that combining weak supervision with domain-specific pre-training can

Table 2 | Results on human-annotated small-scale test datasets - Internal hold-out (Mayo Clinic) and External (MIMIC-III)

Dataset	Models	Prompting setup	Metrics		
			Precision	Recall	F1-score
Internal hold-out test	Mistral-PT	Zero-Shot	0.38	0.33	0.35
		Few-Shot	0.41	0.56	0.47
	Mistral-WFT	Zero-Shot	0.57	0.42	0.48
		Few-Shot	0.63	0.41	0.49
	BioMistral-PT	Zero-Shot	0.47	0.23	0.31
		Few-Shot	0.53	0.31	0.39
	BioMistral-WFT	Zero-Shot	0.63	0.51	0.56
		Few-Shot	0.68	0.49	0.57
External test	Mistral-PT	Zero-Shot	0.41	0.37	0.39
		Few-Shot	0.53	0.41	0.46
	Mistral-WFT	Zero-Shot	0.42	0.47	0.44
		Few-Shot	0.45	0.51	0.48
	BioMistral-PT	Zero-Shot	0.38	0.29	0.33
		Few-Shot	0.41	0.37	0.39
	BioMistral-WFT	Zero-Shot	0.57	0.45	0.50
		Few-Shot	0.65	0.51	0.57

PT Pre-trained, *WFT* Weakly Fine-tuned.

significantly improve performance in clinical language processing tasks.

We also compared the LLM-aided metrics against the human annotator. Figures 1 and 2 show that though these metrics are not ideal for scoring overall LLM performance, the LLM-aided metrics consistently evaluated human annotation with higher score (referred as 'Human' in Supplementary Table D1) than neural model extraction, an observation consistent with human-based evaluation metrics. Figure 2 (and Supplementary Table 2) demonstrates the results of the models for the LLM-aided metrics on the *large-scale* dataset from MIMIC-IV. While the models perform moderately on the LLM-based scoring metrics, the scores are similar to the overlap-based measures on the human-annotated small-scale test dataset. This shows that while not completely perfect, with proper domain-knowledge on critical findings, these automated LLM-based scoring algorithms can be a promising alternative to human-aided metrics for large-scale validation task. Within the scope of both human-based (Rouge-2, BLEU, etc.) and LLM-based evaluation (G-eval, Prometheus), we observe that the weakly supervised models (WFT) perform better than the pre-trained LLMs (PT) on both internal and external validation datasets. We also observe that the general domain Mistral LLMs perform better than the biomedical domain BioMistral models.

Error analysis - To understand and visualize the performance and quality of the extraction by the weakly supervised models, we examine correctly predicted and falsely predicted examples as shown in Table 3. The examples show that the model completely and/or partly extracts critical terms from the reports where they are clearly demarcated or indicated (Table 3-Examples 1 and 2). These indications could be presence of phrases that indicate urgency or the need to communicate to physicians for further follow-ups, like usage of *'concerning'*, *'new'*. However, where the models clearly fail, are the reports with sentences that indicate the chronic critical findings along with the mention of the critical term, e.g. 'No significant interval change..' (Table 3-Example 3). The instruction-tuned LLMs are unsupervised in nature, and therefore have no prior exposure to complicated chronic instances. A chronic finding may not always be critical, it depends on the textual context of the finding - so the model must be trained to carefully discern between new, known and expected critical findings before raising a false alarm. Moreover, Table 3-Example 4 presents a report might only contain a single sentence that reports the absence of the critical finding(s), i.e. a non-critical report. In such an instance, an extractive LLM cannot process the negative sentences and extracts the critical findings terms, thus marking the radiology report as a false positive.

Discussion

We proposed an end-to-end LLM-based pipeline for extracting critical findings from a wide-variety of radiology reports, in terms of modalities and anatomical regions. Given the complexity of the critical finding extraction task and wide-range of potential categorization, the extraction is formulated as a weakly supervised problem without the requirement for manually labeled data for model training. We leverage the task-following generative properties of the instruction-tuned Mistral-class of models and follow a systematic prompt engineering process to create task-specific and comprehensive instructions for the best generative and extractive performance. We selected the Mistral class of models for this extraction task since Mistral models are optimized for inference speed and task performance. Additionally in Supplementary Section 2, we present the quantitative and qualitative results of LLaMa-13B[13] models (pre-trained and weakly fine-tuned) on the human-annotated test datasets in Supplementary Tables 3 and 4. The performance of LLaMa-13B model is similar to the Mistral-7B model performance. We evaluated both generic Mistral-7B and domain-specific BioMistral-7B weakly supervised fine-tuned models with zero-shot and few-shot and compared the performance against baseline pre-trained models. We also perform an in-depth analysis of model performance using varying strategies, as shown in examples (A) through (D) in Supplementary Section 3.

We observe that the weakly supervised models (WFT) perform better than the pre-trained LLMs (PT) on both internal and external validation datasets. However, we do observe a difference in performance between the general domain Mistral and the biomedical BioMistral models, with the BioMistral models scoring consistently lower on the evaluation metrics. BioMistral model is trained only on biomedical literature[26] which is different from the textual content and structure of radiology reports. Therefore, BioMistral model has limited knowledge about radiology reporting as well as the terms that denote critical or incidental findings are limited in the model vocabulary, thus limiting model performance. On the contrary, Mistral models have been trained on web-scraped textual content and may contain partial knowledge about the radiology reports available on the web[14]. Using few-shot prompting technique for label generation helps guide and align these instruction-based Mistral models using task-specific examples, leading to comparatively better extractive performance[27].

We see that the models comparatively have higher performance scores on the external MIMIC-III dataset, than the internal Mayo Clinic dataset which is primarily because of the nature of the dataset - the average report length of Mayo Clinic dataset is 476.3 words, which is much higher than the length of MIMIC-III reports (53.5). Greater textual content is essentially not always beneficial for aligning an instruction-tuned model, and may cause the model to hallucinate or generate misinformation[12]. We have performed a stratified analysis of the models' performance across varying report lengths, as shown in Supplementary Section 4. The quantitative results supporting this observation are shown in Supplementary Tables 5 (external MIMIC-III test data) and 6 (internal Mayo test data). Furthermore, the qualitative examples showing the model performance for short-length, medium-length, and long-length documents also shown in the Supplementary Tables 7, 8, and 9 respectively. Moreover, the reports from the Mayo Clinic dataset, tend to have a higher frequency of *negative* sentences or text that indicate the absence of critical findings, as shown in Table 3-Example 3. Therefore, the model picks up these sentences as false positive signals, leading to erroneous predictions.

Compared to the current literature[4,7–10] which is primarily focused on developing supervised machine learning or rule-based model for critical finding extraction with narrower scope (e.g. single modality or specific

Table 3 | Examples of extracted critical terms from de-identified MIMIC-III radiology reports

Example 1. A positive example from a Chest CT report showing critical findings in Chest and Musculoskeletal regions.	
Radiology report	1. Multifocal bilateral pneumonia with right lung cavitary lesions, right calcified granulomas, and right pleural plaques are very concerning for reactivation tuberculosis with a component of right upper lobe necrotizing pneumonia. 2. Enlarged pulmonary artery suggesting underlying pulmonary hypertension. 3. Hard and soft plaque throughout the aorta with narrowing of the origin of the celiac artery. Finding # 1 was discussed with Dr. First Name8 (NamePattern2) Last Name (NamePattern1) 92986 by phone at 3 : 40 p. m. on 2192-9 -11 immediately after discovery and attending review.
Critical findings in the report	
Groundtruth	'reactivation tuberculosis', 'necrotizing pneumonia'
Mistral-PT-FS	'necrotizing pneumonia'
Mistral-WFT-FS	'reactivation tuberculosis'
BioMistral-PT-FS	NONE
BioMistral-WFT-FS	'reactivation tuberculosis', 'necrotizing pneumonia'
Example 2. A positive example from a Head CT report showing critical findings in Head and Neck regions.	
Radiology report	1) New right parietal intraparenchymal hemorrhage with extensive edema. 2) Stable cerebellar hemorrhage, increased infratentorial edema. 3) Increase in subarachnoid hemorrhage. Findings were discussed with clinical team.
Critical findings in the report	
Groundtruth	'intraparenchymal hemorrhage', 'subarachnoid hemorrhage'
Mistral-PT-FS	'subarachnoid hemorrhage'
Mistral-WFT-FS	'subarachnoid hemorrhage', 'parenchymal hemorrhage'
BioMistral-PT-FS	'parenchymal hemorrhage'
BioMistral-WFT-FS	'parenchymal hemorrhage'
Example 3. A false negative example from a report showing critical findings in Head and Neck regions.	
Radiology Report	1) No significant interval change in the appearance of the comminuted C2 fracture, with components bilaterally involving the body/pedicle, or a hangman's type fracture, and displaced fracture fragment of the anterior inferior aspect of the C2 body, an extensive teardrop-type fracture. Persisting mild rotary subluxation without significant encroachment upon the cervical cord. Fracture lines remain apparent without significant interval callus information. 2) Healing first and second right rib and sternal fractures.
Critical findings in the report	
Groundtruth	'teardrop-type fracture', 'hangman's type fracture'
Mistral-PT-FS	NONE
Mistral-WFT-FS	NONE
BioMistral-PT-FS	NONE
BioMistral-WFT-FS	NONE
Example 4. A false positive example from a report showing critical findings in Gastrointestinal regions.	
Radiology Report	No evidence for intestinal obstruction or free intraperitoneal gas.
Critical findings in the report	
Groundtruth	NONE
Mistral-PT-FS	'bowel obstruction'
Mistral-WFT-FS	'intestinal obstruction', 'free air'
BioMistral-PT-FS	'bowel obstruction', 'hematoma'
BioMistral-WFT-FS	'intestinal obstruction'

PT Pretrained, *WFT* Weakly Fine-tuned, *FS* Few-Shot weak label generation.

anatomical region), we propose a weakly supervised pipeline for automated extraction of critical findings irrespective of the report type and category of the findings. In Supplementary Section 5, we perform a subgroup analysis of the models' performance under varying anatomical sites to demonstrate difference in their critical term extraction capabilities. The quantitative results and qualitative examples are included in Supplementary Tables 10 and 11 respectively. The proposed pipeline was weakly-supervised with LLM generated labels and able to extract a wide variety of critical findings from both internal and external test sets without manual curation of labeled data. Due to the wide and unknown/rare category for potential critical findings, we can only present string overlapping metrics for performance which evaluates exact string matching, often without evaluating semantic context. Even with weakly supervised training, the proposed framework was able to achieve moderate overlap with human annotation. Such automated extraction of critical findings from radiology reports may reduce risk of human error regarding flagging and allows for timely communication of serious conditions such as tumors, fractures, or internal bleeding, leading to prompt and appropriate treatment[4,9]. Furthermore, our system holds promise for retrospective data analysis, ensuring that significant health issues are not overlooked in radiology reports. An additional future direction involves adapting the system to facilitate timely communication of critical findings to referring physicians. This could lead to more effective follow-up based on current guidelines, better management of conditions, and ultimately improved patient outcomes.

The study has several limitations. We view this work primarily as a proof-of-concept demonstrating that weakly supervised approaches can improve performance in scenarios where labeled data is scarce. We are aware that despite showing an improved performance, the approach when compared to existing technology, still remains mediocre and cannot be adopted as an immediately deployable clinical solution. The performance metrics, while showing clear improvement over baselines, indicate that additional refinement is needed before these models would be suitable for

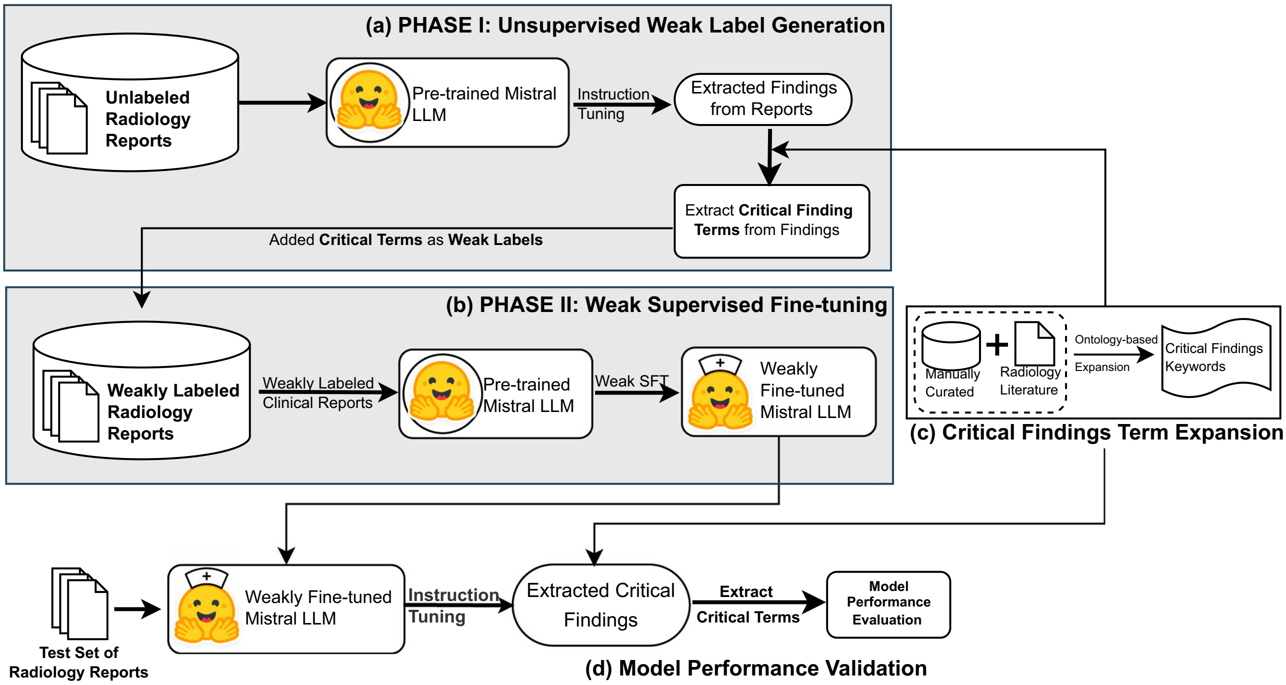

(a) PHASE I: Unsupervised Weak Label Generation
Unlabeled Radiology Reports
Pre-trained Mistral LLM
Instruction Tuning
Extracted Findings from Reports
Extract Critical Finding Terms from Findings
Added Critical Terms as Weak Labels
(b) PHASE II: Weak Supervised Fine-tuning
Weakly Labeled Radiology Reports
Weakly Labeled Clinical Reports
Pre-trained Mistral LLM
Weak SFT
Weakly Fine-tuned Mistral LLM
Manually Curated
Radiology Literature
Ontology-based Expansion
Critical Findings Keywords
(c) Critical Findings Term Expansion
Test Set of Radiology Reports
Weakly Fine-tuned Mistral LLM
Instruction Tuning
Extracted Critical Findings
Extract Critical Terms
Model Performance Evaluation
(d) Model Performance Validation

Fig. 3 | Overall weakly supervised training pipeline for critical findings extraction and performance validation. The complete pipeline is composed of the following modules: **a** Phase I : This module generates the weak labels using unsupervised instruction-tuned Mistral class of LLMs; **b** Phase II: This module demonstrates the weakly supervised fine-tuning of the Mistral LLMs using the weak labels generated in Phase I; **c** This is module shows the expansion of critical findings terms and keywords. This expanded list is used for the key-based critical finding extraction; **d** This final pipeline demonstrates the model performance validation using internal and external datasets.

routine clinical use. Even after additional curation of critical finding term list from prior publications, ACR lists and ontology-based expansion, there may exist rare critical findings that are not captured by our curated list and thus lead to missed critical findings. In the absence of critical terms, the weakly supervised LLMs do not retrieve any findings, thus flagging an otherwise critical radiology report as non-threatening. Another reason for moderate performance in the term matching metrics is not considering partial term matching, e.g. ‘spinal injury’ vs. ‘injury of spine’. We also plan on extending our model to add the ability to extract critical findings with an increased granularity, where the model can be trained to identify new, known or expected and uncertain or unexpected findings. Finally, we plan to include a vision language model to directly analyze images to identify critical findings with currently trained LLM model as language encoding backbone.

The results of our study demonstrate that the proposed weakly supervised pipeline not only performs better than baseline pre-trained models but also has the potential to improve patient outcomes. By automating the extraction of critical findings from radiology reports, the pipeline reduces the risk of human oversight, ensuring that critical findings such as intracranial haemorrhage, spinal injury, pulmonary embolism, etc. are identified and communicated promptly. This can improve long-term treatment outcomes, and ultimately enhance overall patient care. This underscores the practical utility of our approach in clinical settings, where timely and accurate information can make a significant difference in patient outcomes.

In summary, we highlight the importance of promptly communicating critical findings from radiology reports to physicians for timely patient management. We leverage instruction-tuned Mistral-based large language models (LLMs) to identify and extract these critical findings from radiology reports. To address the lack of labeled datasets due to the rarity of such events, we propose a two-phase weakly supervised fine-tuning approach on Mistral models using unlabeled radiology reports. The weakly fine-tuned model was tested on internal and external datasets and evaluated with both human annotated-based and LLM-based metrics, showing that weakly supervised fine-tuning improves model performance significantly.

Methods

Given the complexity of critical finding extraction and wide-range of potential categorization, we formulated this as a weakly-supervised model learning problem where we leverage weak labels generated by LLM. We leverage the task-following generative properties of the instruction-tuned Mistral-class of models[14,26]. We also follow a systematic prompt engineering process to create task-specific and comprehensive instructions for the best generative and extractive performance[15,16]. Figure 3 demonstrates the overall pipeline of the proposed two-phase pipeline - (a) Phase I: The unsupervised instruction-tuned weak label generation; and (b) Phase II: Weakly supervised training framework for automated critical finding extraction. We further validate pipeline performance on small-scale manually annotated held-out internal and external datasets, along with a large-scale external dataset. The internal dataset was extracted from radiology reports from four Mayo Clinic sites (Arizona, Rochester, Florida, and Mid-West) after ethical review by the Mayo Clinic Institutional Review Board (IRB 21-009503) The external datasets came from MIMIC-III (small-scale annotated test) and MIMIC-IV (large-scale test). In the following sections, we provide a detailed description of our framework and evaluation methods.

Weakly supervised training pipeline

To tackle the challenge regarding lack of annotated database for critical findings, we propose a two-phase weakly supervised training approach to align generic Large Language Models (LLMs) for extracting critical findings efficiently from radiology reports which are documented in a domain-specific clinical language.

In Phase I (Generating Weak Labels) of the pipeline, we leverage the instruction-tuned pre-trained Mistral-based large language models (LLMs), Mistral-7B and BioMistral-7B[14,26] to extract the critical findings in the input reports in an unsupervised manner. We rely on the Mistral class of models[14] since Mistral model architecture is an efficient version of Meta’s LLaMa model[13] that leverages grouped-query attention to significantly accelerate model inference, thus reducing computational complexity while keeping model performance consistent. For understanding the effect of clinical

Table 4 | Prompt instructions for extraction of critical findings

Instruction type	Prompt text
Definition of CRITICAL and INCIDENTAL Findings	CRITICAL findings are life threatening imaging findings that need to be communicated immediately.
	INCIDENTAL findings are non-life-threatening findings, but significant enough that they need to be communicated within a short period of time.
Instruction of Retrieval Task	Based on these definitions of the Critical and Incidental findings, find the CRITICAL findings and INCIDENTAL findings mentioned in the report.

Table 5 | Examples of critical finding term expansions

Original term	Expanded term list
Spine injury	SI; Spinal Injury; Traumatic injury of spinal cord and vertebral column; Traumatic injury of spine; Traumatic injury of vertebral column
Foreign body	FB; Exogenous material; Foreign material
Pulmonary embolism	PE
Pneumoperitoneum	PP; Peritoneal cavity free air
Cord compression	SCC; Spinal cord compression; Compression of spinal cord

domain-specific training, we leverage two pre-trained versions of the Mistral model in parallel - (a) Mistral-7B-Instruct[28], a general domain model with 7B parameter size; and (b) BioMistral-7B[29], a version of the Mistral model trained on 1.47M biomedical literature documents (about 3B tokens) from PubMed Central[26]. We use these instruction-tuned LLMs as an initial step to retrieve the critical findings from radiology reports as weak labels.

Using task-based prompt engineering, the instruction-tuned LLM can be aligned and guided to follow well-defined rules while generating or retrieving textual content[12]. Quality and usability of the model output is largely dependent on the nature of specific input prompt used to align the model during inference[15]. The prompt definitions and task instruction are shown in Table 4. For the task of extracting critical findings, the prompts are engineered to instruct the model to identify and retrieve the findings that appear as critical from the given reports based on linguistic style of reporting, order of documentation as well as definition of the findings[30]. In Phase I, the experimental setup is unsupervised and involves prompt-specific generation. We consider two prompting techniques - *zero-shot* (no task-specific examples) and *few-shot* (five task-specific examples). Given the fact that extraction of critical finding is a highly complex task and needs proper understanding of the severity of the findings, each prompt consists of the definitions of ‘critical’ and ‘incidental’ findings in a report, followed by the description of the task to be performed. While zero-shot method of prompting just includes the task definitions and instructions in the query body, few-shot prompting strategy includes examples showing the task input and desired output, to specifically align for in-context task-based learning[16]. The prompt templates and the few-shot examples used in prompting are shown in Supplementary Sections 6 and 7 respectively.

Mistral-based LLMs are generative transformer models and during inference outputs are descriptive textual content. Hence, the proposed pipeline incorporates a module for the *automated extraction of key terms or phrases that represent the critical findings* in the radiology reports, using a list-based term matching algorithm. To map the extracted findings from the reports to the actual terms, we first curate a *manually verified and comprehensive list of critical findings keywords and phrases*. The list of terms was collected from different academic institutions (Stanford, UCLA, Yale, Emory), to which we appended additional terms and phrases from a list developed by American College of Radiology’s Actionable Reporting Work Group (ACR)[23]. But radiologists often use abbreviations to refer to findings, for example ‘PE’ instead of ‘Pulmonary Embolism’, or there are spelling errors due to dictation (e.g., ‘hemorrhage’ or ‘haemorrhage’). To incorporate additional keywords that share the same ontological relations with the curated manual list of critical key terms and phrases, we follow an *ontology-based expansion of critical findings keywords*. We used the ‘PyMedTermino2’ Python library[31] to parse NCBO[32] ontologies such as SNOMED_CT, ICD10, and UMLS, and include synonym-based searches. We specifically used the SNOMED_CT ontology for our approach and starting with a reference list of 102 terms, expanded the list to 210 terms. Table 5 has a list of some sample terms and their corresponding expansions. We have included the complete list of terms in the Supplementary Section 8 (Supplementary Tables 12 and 13).

In our pipeline, we check for the mentions of the terms/phrases, from the curated list, in the output text containing the extracted findings generated by the Mistral LLMs. We use the Python library “fuzzywuzzy”[33] to extract the keywords from the text using exact term-matching algorithms. We do not consider partial or relaxed term-matching to avoid incorrect flagging if part of a critical finding term appears in an otherwise non-critical radiology report. For a given unlabeled radiology report, the output of the proposed Phase I pipeline, is a set of terms and/or phrases that represent the critical findings in the radiology report and if there is no mention of any critical findings the extracted term is *null*. These are then augmented to the report as weak labels for fine-tuning the weakly supervised model during Phase II of the pipeline.

The main objective for Phase II (Weakly Supervised Fine-tuning) is to fine-tune the Mistral-based models using weakly supervised labels for an improved automated extraction of specific critical findings from radiology reports. Fine-tuning the pretrained LLM in a weakly supervised setting helps to refine the model’s understanding and enhance its ability to discern critical information in the required format[12,34].

We first combine the ‘weak labels’ generated in Phase I with the unlabeled reports. For each report that has one or more key terms representing critical findings, the label is represented by those extracted keywords as a list. For a non-critical report, the label is an empty list. During training, we further append the task-specific instruction to the report text and its weak label list to emulate the instruction-tuning setup used to train the Mistral class of models. The main goal is to train an instruction-tuned generative LLM that can automatically identify and retrieve critical findings terms from radiology reports.

We use the standard fine-tuning configurations along with LoRA-specific hyperparameters[35] for faster and computationally efficient training of the Mistral and BioMistral models. The LoRA parameter values used were - attention dimension = 16, scaling rate $\alpha = 4$, dropout rate = 0.1. We also used weighted Adam optimizer with a constant learning rate scheduler, with an initial learning rate of 0.0002 and weight decay of 0.001. We ran the fine-tuning on four NVIDIA RTX A5000 GPUs for 20 epochs for a total of 3 hours.

Evaluation strategy

We describe the setup for the manual and automated metrics used to evaluate the performance of the models during both unsupervised Phase I and weakly supervised Phase II. To measure overall performance of the models and validate the quality and correctness of the extracted critical findings, we evaluate on both internal (held-out reports from Mayo Clinic) and external (reports from publicly available MIMIC-III) datasets. We evaluate the performance of two types of Mistral-based models - Mistral-7B[14] and BioMistral-7B[26]. The instruction-tuned versions of these models (referred as *PT*) are considered the baseline models, for comparison with the weakly fine-tuned LLMs (referred as *WFT*). Two separate prompting methods (zero-shot and few-shot) were used to generate the weak labels (Phase I). We evaluate the models using two strategies - Human-based (small-scale) and automated LLM-based (large-scale).

Human-based evaluation. compares the critical findings automatically extracted by the LLMs with the manually extracted critical findings mentions in the radiology reports. This is the preferred method of evaluating generative models and consists of assessing the quality and correctness of the generated content by comparing with human annotated labels as the ground truth. Given the large variety of potential findings, standard accuracy-based metrics are not suitable for this evaluation. Therefore, we report the models' performance score using the lexical-similarity metrics like BLEU[36], ROUGE[37] (-1 and -2), and METEOR[38]. These metrics quantify the overlap between the human-generated and LLM-generated mentions of critical findings. Additionally, we also look at the semantic similarity between the model generated labels and the manual labels using RadBERTScore. This metric uses BERTScore-based[39] evaluation algorithm but replaces the vanilla BERT model with RadBERT[40], a pre-trained RoBERTa-based model specifically fine-tuned on 4M radiology reports[41]. However, manual evaluation can be time-consuming and costly, and can only be performed on a smaller scale due to required effort.

LLM-based evaluation. leverage the causal reasoning and extensive knowledge of pre-trained LLMs, like GPT-4[11] and LLaMa[13], to automatically rate quality of the extracted textual content, while achieving results comparable to a human annotator[17,42]. The main usefulness of LLM-based metrics lies in evaluating the performance of the proposed models on large-scale datasets of radiology reports, where manual annotation is challenging and time-consuming. So in addition to evaluating the smaller human-annotated internal (Mayo Clinic) and external (MIMIC-III) datasets with LLM-based scorers, we additionally use these metrics to evaluate the extracted critical findings from a large-scale dataset of 5000 random radiology reports from MIMIC-IV.

We report the model performance using two LLM-aided evaluation metrics - G-Eval[42] and Prometheus[43]. G-Eval[44] uses task-specific chain-of-thought prompting[45] to instruct OpenAI's GPT-4[11] model to evaluate the text. Prometheus[46] also follows an approach similar to G-Eval, but instead of GPT-4, uses a pre-trained LLaMa-13B model as the scoring algorithm. These LLM-based scoring algorithms are instruction-tuned and require an evaluation metric definition and task-based prompt to rate model performance. We use the *correctness* metric to score the quality of the generated output from the models. The score is a continuous value between 0 to 1, with a higher value denoting an output closer to the ground-truth or gold-standard annotations, indicating better model performance. Based on the nature of the test dataset, we have two separate definitions of the correctness evaluation task.

For the *human-annotated smaller test datasets*, we have the human or ground truth labels. We provide the following task definition to the scoring LLM - 'Given the human label, determine if the model output is correct. Consider partial matches.' We use the LLM-based scorers to compare the human annotations with the model-extracted critical findings; we do not provide additional information like the report metadata, impressions, findings, etc. However, since we do not have human annotated labels for the *large-scale dataset*, we slightly change the definition and objective of the evaluation task. Since MIMIC-IV is a publicly available dataset, we provide the report text and the extracted critical findings by the models. We also provide the definition of critical finding, along with the correctness definition in the evaluation task prompt. The evaluation task definition given to the scoring algorithm is - 'Given the report and the definition of critical findings, determine if the model output is correct. Consider partial matches.' The examples of the prompts used in LLM-aided automated evaluation have been included in the Supplementary Section 9. In the absence of human labels, our main objective here is to leverage the prior knowledge of these LLM-based scoring algorithms for rating the quality of automated large-scale dataset annotation using weakly supervised LLMs.

Data availability

The Mayo Clinic data supporting the findings of this study are private and not available publicly. Corresponding author can be reached for data usage agreement for approving the access for the private data. The MIMIC-III and MIMIC IV data sets are publicly available at https://physionet.org/content/mimiciii/1.4/ and https://physionet.org/content/mimiciv/2.2/ respectively.

Code availability

Our code and trained models are publicly shared under open-source academic license at https://github.com/dasavisha/CriticalFindings_Extract. In our implementation, we used Python 3.8.10, and the following open-source libraries: torch = 1.13.1, tqdm = 4.66.1, pandas = 2.7.1, numpy = 1.21.0, transformers = 4.35.2, huggingface-hub = 0.19.4, accelerate = 0.25.0, and fuzzywuzzy = 0.18.0.

Received: 10 September 2024; Accepted: 17 February 2025;
Published online: 08 May 2025

References

1. Cronin, P. & Rawson, J. V. Review of research reporting guidelines for radiology researchers. *Acad. Radiol.* **23**, 537–558 (2016).
2. Berlin, L. Communicating findings of radiologic examinations: whither goest the radiologist's duty? *Am. J. Roentgenol.* **178**, 809–815 (2002).
3. Sistrom, C. L. & Langlotz, C. P. A framework for improving radiology reporting. *J. Am. Coll. Radiol.* **2**, 159–167 (2005).
4. Banerjee, I. et al. Natural language processing model for identifying critical findings-a multi-institutional study. *J. Digital Imaging* **36**, 105–113 (2023).
5. Clinger, N. J., Hunter, T. B. & Hillman, B. J. Radiology reporting: attitudes of referring physicians. *Radiology* **169**, 825–826 (1988).
6. Van Leeuwen, K. G., de Rooij, M., Schalekamp, S., van Ginneken, B. & Rutten, M. J. How does artificial intelligence in radiology improve efficiency and health outcomes? *Pediatric Radiol.* **52**, 2087–2093 (2022).
7. Lakhani, P., Kim, W. & Langlotz, C. P. Automated detection of critical results in radiology reports. *J. Digital Imaging* **25**, 30–36 (2012).
8. Heilbrun, M. E., Chapman, B. E., Narasimhan, E., Patel, N. & Mowery, D. Feasibility of natural language processing–assisted auditing of critical findings in chest radiology. *J. Am. Coll. Radiol.* **16**, 1299–1304 (2019).
9. Mabotuwana, T., Hall, C. S. & Cross, N. Framework for extracting critical findings in radiology reports. *J. Digital Imaging* **33**, 988–995 (2020).
10. Jiang, Z. et al. Learning to summarize chinese radiology findings with a pre-trained encoder. *IEEE Trans. Biomed. Eng.* **70**, 3277–3287 (2023).
11. Achiam, J. et al. GPT-4 technical report. https://arxiv.org/abs/2303.08774 (2023).
12. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. neural Inf. Process. Syst.* **35**, 27730–27744 (2022).
13. Touvron, H. et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
14. Jiang, A. Q. et al. Mistral 7b. Preprint at https://arxiv.org/abs/2302.13971 (2023).
15. Wei, J. et al. Finetuned language models are zero-shot learners. Preprint at https://arxiv.org/abs/2109.01652 (2021).
16. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
17. Liu, L. et al. A survey on medical large language models: Technology, application, trustworthiness, and future directions. Preprint at https://arxiv.org/abs/2406.03712 (2024).

18. Wang, B. et al. Pre-trained language models in biomedical domain: A systematic survey. *ACM Comput. Surv.* **56**, 1–52 (2023).
19. Wang, J., Yang, Z., Yao, Z. & Yu, H. JMLR: Joint medical llm and retrieval training for enhancing reasoning and professional question answering capability. Preprint at https://arxiv.org/abs/2402.17887 (2024).
20. Woo, K.-m. C. et al. Evaluation of GPT-4 ability to identify and generate patient instructions for actionable incidental radiology findings. *J. Am. Med. Informatics Assoc.* **31**, 1983–1993 (2024).
21. Bhayana, R. et al. Use of GPT-4 with single-shot learning to identify incidental findings in radiology reports. *Am. J. Roentgenol.* **222**, e2330651 (2024).
22. Kim, S. H. et al. Boosting LLM-assisted diagnosis: 10-minute llm tutorial elevates radiology residents' performance in brain mri interpretation. *medRxiv*, https://www.medrxiv.org/content/10.1101/2024.07.03.24309779v1.full.pdf (2024).
23. Larson, P. A., Berland, L. L., Griffith, B., Kahn Jr, C. E. & Liebscher, L. A. Actionable findings and the role of it support: report of the acr actionable reporting work group. *J. Am. Coll. Radiol.* **11**, 552–558 (2014).
24. Johnson, A. E. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 1–9 (2016).
25. Johnson, A. E. et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **10**, 1 (2023).
26. Labrak, Y. et al. Biomistral: A collection of open-source pretrained large language models for medical domains. Preprint at https://arxiv.org/abs/2402.10373 (2024).
27. Zambrano Chaves, J. et al. Rales: a benchmark for radiology language evaluations. *Adv. Neural Inf. Process. Syst.* **36**, 74429–74454 (2023).
28. Mistral-7B huggingface model card. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2 (2023). Accessed: 2024-09-30.
29. BioMistral-7B huggingface model card. https://huggingface.co/BioMistral/BioMistral-7B (2024). Accessed: 2024-09-30.
30. Gramopadhye, O. et al. Few shot chain-of-thought driven reasoning to prompt llms for open ended medical question answering. Findings of the Association for Computational Linguistics: EMNLP 2024 (2024).
31. PyMedTermino toolkit. https://owlready2.readthedocs.io/en/latest/pymedtermino2.html (2024). Accessed: 2024-09-30.
32. NCBO ontologies. https://bioportal.bioontology.org/ (2011). Accessed: 2024-09-30.
33. FuzzyWuzzy python library. https://pypi.org/project/fuzzywuzzy/ (2020). Accessed: 2024-09-30.
34. Wei, J. et al. Emergent abilities of large language models. *Transact. Mach. Learn. Res.* https://doi.org/10.48550/arXiv.2206.07682 (2022).
35. Hu, E. J. et al. Lora: Low-rank adaptation of large language models. In *Proceedings of The Tenth International Conference on Learning Representations* (ICLR 2022).
36. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* 311–318 (ACM, 2002).
37. Lin, C.-Y. *Rouge: A package for automatic evaluation of summaries*. Text summarization branches out 74–81 (Association for Computational Linguistics, 2004).
38. Banerjee, S. & Lavie, A. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In: *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* 65–72 (ACL, 2005).
39. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. Bertscore: Evaluating text generation with bert. *Proceedings of International Conference on Learning Representations* (2019).
40. RadBERT model. https://github.com/zzxslp/RadBERT (2024). Accessed: 2024-09-30.
41. Yan, A. et al. RadBERT: Adapting transformer-based language models to radiology. *Radiology* **4**, e210258 (2022).
42. Liu, Y. et al. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (2023)
43. Kim, S. et al. Prometheus: Inducing fine-grained evaluation capability in language models. In: *The Twelfth International Conference on Learning Representations* (2023), OpenReview.net.
44. G-Eval model. https://github.com/nlpyang/geval (2023). Accessed: 2024-09-30.
45. Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **35**, 24824–24837 (2022).
46. Prometheus model. https://github.com/prometheus-eval/prometheus (2023). Accessed: 2024-09-30.

Acknowledgements

The work is supported in part by a grant from Phillips Healthcare.

Author contributions

A.D., I.T., D.R., and I.B. analyzed and interpreted the data used to train and evaluate the automated critical findings extraction pipeline. A.D. developed the pipeline, performed the empirical evaluation of the pipeline, and was a major contributor in writing the manuscript. I.T., D.R. and J.M.Z.C. provided the definitions and few-shot examples used in prompting setup. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01522-4.

Correspondence and requests for materials should be addressed to Imon Banerjee.