Journal of Imaging Informatics in Medicine
https://doi.org/10.1007/s10278-025-01505-7

Role of Model Size and Prompting Strategies in Extracting Labels from Free-Text Radiology Reports with Open-Source Large Language Models

Bardia Khosravi[1,2] · Theo Dapamede[3] · Frank Li[3] · Zvipo Chisango[4] · Anirudh Bikmal[4] · Sara Garg[4] · Babajide Owosela[4] · Amirali Khosravi[2] · Mohammadreza Chavoshi[3] · Hari M. Trivedi[3] · Cody C. Wyles[2] · Saptarshi Purkayastha[5] · Bradley J. Erickson[1] · Judy W. Gichoya[3]

Received: 8 January 2025 / Revised: 25 March 2025 / Accepted: 13 April 2025

Abstract

Extracting accurate labels from radiology reports is essential for training medical image analysis models. Large language models (LLMs) show promise for automating this process. The purpose of this study is to evaluate how model size and prompting strategies affect label extraction accuracy and downstream performance in open-source LLMs. Three open-source LLMs (Llama-3, Phi-3 mini, and Zephyr-beta) were used to extract labels from 227,827 MIMIC-CXR radiology reports. Performance was evaluated against human annotations on 2000 MIMIC-CXR reports, and through training image classifiers for pneumothorax and rib fracture detection tested on the CANDID-PTX dataset ($n = 19{,}237$). LLM-based labeling outperformed the CheXpert labeler, with the best LLM achieving 95% sensitivity for fracture detection versus CheXpert's 51%. Larger models showed better sensitivity, while chain-of-thought prompting had variable effects. Image classifiers showed resilience to labeling noise when tested externally. The choice of test set labeling schema significantly affected reported performance—a classifier trained on Llama-3 with chain-of-thought labels achieved AUCs of 0.96 and 0.84 for pneumothorax and fracture detection respectively when evaluated against human annotations, compared to 0.91 and 0.73 when evaluated on CheXpert labels. Open-source LLMs effectively extract labels from radiology reports at scale. While larger pre-trained models generally perform better, the choice of model size and prompting strategy should be task specific. Careful consideration of evaluation methods is critical for interpreting classifier performance.

Keywords Large language models · Natural language processing · Noisy labeling · Label extraction · Open-source

Introduction

Deep learning (DL) models have revolutionized various domains, including medical image analysis, by leveraging large-scale datasets to learn complex patterns and make accurate predictions [1–4]. However, the success of these models, specifically data-hungry supervised algorithms, heavily relies on the availability of quality labeled data [5]. Obtaining high-quality labels for medical imaging datasets is a challenging task, as it requires expert knowledge and is time-consuming and expensive. Manual annotation by medical experts is often necessary to ensure the accuracy and reliability of the labels, but this process becomes impractical when dealing with large-scale datasets containing hundreds of thousands of images.

One approach to mitigate this challenge is to leverage the information contained in the accompanying radiology

Bradley J. Erickson and Judy W. Gichoya are co-senior authors.

✉ Judy W. Gichoya
judywawira@emory.edu

1 Department of Radiology, Mayo Clinic, Rochester, MN, USA

2 Department of Orthopedic Surgery, Mayo Clinic, Rochester, MN, USA

3 Department of Radiology, Emory University, 101 Woodruff Circle, Atlanta, GA 30322, USA

4 Emory School of Medicine, Atlanta, GA, USA

5 School of Informatics and Computing, Indiana University, Indianapolis, IN, USA

Published online: 05 May 2025

reports. Several natural language processing (NLP) tools have been developed specifically for this purpose, utilizing rule-based methods to identify and extract key findings from radiology reports [6–8]. However, these rule-based NLP tools have limitations in understanding complexities of language, particularly in handling negations (e.g., "no evidence of" or "absence of") that are commonly used in radiology reports, and nuanced expressions (e.g., "cardiomegaly" and "enlarged cardiac silhouette") that convey the same meaning. This variability poses challenges for rule-based systems, which rely on predefined patterns and keywords to extract information.

Recent advancements in large language models (LLMs) provide a promising solution to address limitations of rule-based NLP tools. LLMs have opened new avenues for automating complex tasks in medical domains such as impression generation, report simplification, and question answering [9–13]. These models, trained on vast amounts of text data can understand and generate human-like text, making them useful for automating label extraction from medical reports. LLMs' deep understanding of language and context allows them to handle negations and nuanced expressions more effectively than rule-based systems, potentially leading to more accurate and efficient label extraction [14].

Previous studies have explored using LLMs for label extraction from medical reports. Early efforts primarily utilized proprietary models, which, while powerful, pose challenges related to accessibility, cost, and compliance with regulations such as the Health Insurance Portability and Accountability Act (HIPAA) [15–17]. The closed-source nature of these models restricts the extent to which they can be customized and validated in various clinical environments. Recently, there has been a shift towards the use of open-source models. Although these open-source models offer greater flexibility and transparency, their performance frequently falls short compared to proprietary models [18, 19].

As new LLMs are released at an unprecedented pace with improved capabilities, a practical question emerges—would updating extracted labels with each new model generation meaningfully improve downstream task performance, or are image classifiers sufficiently robust to label noise that such updates provide diminishing returns? Additionally, the impact of model size and prompting strategies on label extraction accuracy and downstream tasks' performance has not been thoroughly investigated.

This work addresses these limitations and investigates several key questions: (1) How effective are open-source LLMs compared to traditional rule-based methods for extracting labels from radiology reports? (2) What is the impact of model size on label extraction accuracy? (3) Does chain-of-thought prompting improve performance in medical text analysis tasks? (4) How do labeling errors propagate to downstream image classification performance? Our study leverages open-source LLMs to extract labels from a large-scale radiology dataset containing more than 220,000 reports. We compare the performance of different open-source LLMs with varying model sizes and prompting strategies. Additionally, we leverage all seven labeling schema variants (three different LLMs with and without chain-of-thought prompting, plus the traditional CheXpert labeler) to train separate pathology classifiers for detecting pneumothorax and rib fractures. Their performance is tested on an external dataset of 19,000 radiographs with radiologist image-level annotations. This approach allowed us to directly assess how different labeling methods affect downstream model performance.

Methods

Data

This retrospective study utilized two publicly available datasets: the MIMIC-CXR dataset from Beth Israel Deaconess Medical Center, MA, USA, and the CANDID-PTX dataset from Dunedin Hospital, New Zealand [20–23]. MIMIC-CXR dataset consists of 377,110 posteroanterior (PA), anteroposterior (AP), and lateral chest radiographs from 65,379 unique patients, accompanied by 227,827 radiology reports, as some series have multiple instances with different projection planes. The CANDID-PTX dataset comprises 19,237 frontal (PA and AP) chest radiographs with radiologist annotations for rib fracture and pneumothorax, provided as segmentation masks, along with their corresponding radiology reports. The dataset characteristics are summarized in Table 1.

Manual Labeling

A subset of 2000 reports from the MIMIC-CXR dataset was randomly sampled for manual annotation. The randomization was stratified by the pre-existing CheXpert labeler-generated pathology labels [6]. Four medical students were tasked with identifying two distinct pathologies, pneumothorax and rib fracture. The annotation categories included presence, absence, and uncertain labels. Absence was assigned when the presence of a pathology was not explicitly mentioned, while uncertain was used for cases where the report was equivocal in reporting the presence of a pathology.

The annotators underwent a 2-h training session with a supervising radiologist to recognize these pathologies in radiology reports. After that, each annotator annotated 1000 reports. Two individuals annotated each report, and discrepancies were resolved by the supervising radiologist

Table 1 Study population characteristics

Variable	MIMIC-CXR[a]	CANDID-PTX
Dataset statistics		
Dataset origin	MA, USA	New Zealand
Number of studies	227,827	19,237
Number of images	377,110	19,237
Number of patients	65,379	19,237
Demographic information		
Mean age (SD, yrs.)	62.6 (17.8)	60.1 (20.1)
Sex (female)	32,416 (52.4%)	8,929 (46.4%)
Pathology distribution[b]		
Pneumothorax	*Positive:* 15,819 (4.2%)	*Positive:* 3,196 (16.6%)
Fracture	*Positive:* 8,658 (2.3%)	*Positive:* 335 (1.7%)

[a]Demographic information is only available for 61,868 patients

[b]The distributions presented in the MIMIC-CXR dataset come from the CheXpert-Labeler results, while those in CANDID-PTX are based on image-level pathology annotations by radiologists

with 7 y of experience. During adjudication, if the supervising radiologist determined that a case was equivocal, it was removed from the 2000 manually labeled reports for that specific pathology. However, these uncertain cases were not excluded from the classifier training set. An online user interface was developed to facilitate the annotation process, with the code available on the study's GitHub repository (https://github.com/Emory-HITI/TextAnnotationUI).

LLM-Based Labeling

Three open-source LLMs with varying sizes were employed for automated labeling: Llama- 3.0 (8 billion parameters), Phi- 3 mini (4 billion parameters), and Zephyr-beta (7 billion parameters) [24–26]. These models have a context length of more than 8000 tokens, ensuring the model can access the entire radiology report when generating a response. The models were deployed locally on a server with an NVIDIA A100 GPU (80 GB RAM). Models were hosted through a vLLM server that batches multiple requests, enabling parallel processing of multiple reports [27]. A concurrency factor of 12 was used for the Llama and Zephyr models, while this number was set to 24 for the smaller Phi- 3 mini model.

Initial prompt engineering was performed on 500 reports not included in the manually labeled set. A temperature of 0 was used, and all generations were seeded with a fixed value to ensure reproducibility. The final prompt was applied to label the datasets using zero-shot prompting and chain-of-thought (CoT) prompting [28]. Zero-shot prompting involves providing the model with a task description and input without example input and output pairs. In contrast, CoT prompting encourages the model to explain its reasoning step-by-step before providing the final answer, which has been previously shown to improve LLM performance on reasoning tasks [28]. Please refer to Supplemental Material to see individual prompts. Despite being instructed to produce outputs in JSON format—a structured data format featuring key-value pairs—and being provided with a response template, the models did not always fulfill this request [19]. Therefore, rule-based post-processing was employed to handle instances where the models did not generate valid JSON output, with manual correction applied to any remaining failures (see "Results"). The CheXpert labeler, a rule-based NLP method regarded as the reference standard in many previous studies, was used for comparison [8, 29, 30]. The MIMIC-CXR and CANDID-PTX datasets were labeled using all 7 schemas (three LLMs with and without CoT, and the baseline CheXpert labeler). As mentioned, the CANDID-PTX dataset includes human annotations on the actual images in the form of segmentation masks. Following the recommendations of the original authors, we utilized a *U-ones* approach where all uncertain CheXpert labels are considered positive [6].

We packaged our code—RadPrompter—for composable and reproducible prompting and have released it publicly for the medical research community to build upon and improve (https://github.com/Emory-HITI/RadPrompter).

Image Classification Model Training

The frontal chest radiographs ($n = 243{,}334$) in the MIMIC-CXR dataset were split into training (80%) and tuning (20%) sets at the patient level. Based on the extracted labels, seven image classifiers were trained for pneumothorax and rib fracture detection. The CANDID-PTX dataset was used for external testing. Images were preprocessed by scaling intensities to [− 1, 1], resizing to 512 × 512 pixels (with aspect ratio preservation), and applying histogram equalization to match the MIMIC-CXR-JPG dataset. Data augmentation strategies, including horizontal and vertical flipping, Gaussian noise addition, rotation (± 25°), scaling (90–110%), translation (± 10%), mixup, and CutMix were applied during training [31, 32]. All pre-processing steps and augmentations were applied through the MONAI package (v1.2) [33].

A ConvNeXt-base model (86 million parameters), pretrained on ImageNet, was used as the base architecture [34, 35]. The models were trained using binary cross-entropy loss. Lion optimizer with a learning rate and weight decay of 0.00001 and 0.0003, respectively, and exponential moving weight averaging (EMA) with a decay factor of 0.9999 was employed to stabilize training [1, 36].

Evaluation

Model performance was evaluated using the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC) to account for class imbalance. Ninety-five (95)% confidence intervals were obtained using 1000 rounds of bootstrapping. Confusion matrices were used to assess agreement between LLM-generated labels and human annotations on the sampled 2000 MIMIC-CXR reports. Inter-annotator agreement was assessed using Gwet's AC1 metric, which is suitable for multi-class scenarios with severe class-imbalance [37]. As mentioned, reports that were deemed equivocal by the supervising radiologist were excluded from this part of the evaluation. To assess the impact of labeling schemas on downstream classifier performance, each classifier was evaluated against labels extracted from the CANDID-PTX radiology reports using the seven different labeling schemas, as well as the gold standard expert segmentations provided with the dataset, resulting in a total of eight references for comparison.

This approach allows for investigating how errors introduced by the labeling process propagate into the final image classification models. By comparing the performance of classifiers trained on different labeling schemas, we elucidate the influence of model size, prompting strategy, and label generation method on the effectiveness of pathology classification models in a scenario where only radiology reports are available. Furthermore, by labeling the test set using different schemas, we evaluated the perceived performance of models trained on the same source dataset. This approach highlights how the choice of labeling schema for the test set can impact the models' reported performance, even when trained on identical data. For instance, the same classification model might achieve different performance metrics when evaluated against different labeling schemas (e.g., an AUROC of 0.85 when evaluated against CheXpert labels versus 0.80 when evaluated against chain-of-thought LLM annotations) on the same test set. These variations arise not from changes in the model's inherent capabilities, but from differences in how each labeling method interprets radiological findings, particularly in borderline cases.

In addition to performance metrics, the processing time for each labeling schema and model on each dataset was recorded and reported to highlight the trade-offs between different models and prompting strategies regarding both performance and computational efficiency. Statistical significance was set at $p < 0.05$, and paired *t*-tests were used to compare model performance.

Results

Agreement with Human Annotations

The annotators demonstrated high agreement, with Gwet's AC1 coefficient of 0.988 (95% CI: 0.983–0.993, $p < 0.001$) for identifying pneumothorax, and 0.959 (95% CI: 0.950–0.968, $p < 0.001$) for identifying rib fracture, indicating substantial inter-annotator reliability that was not due to chance. The performance of the different models and prompting strategies in extracting pneumothorax and rib fracture labels were evaluated against human annotations and compared with a conventional rule-based tool (CheXpert Labeler) on a subset of 2000 reports from the MIMIC-CXR dataset. The performance metrics, including accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score, are presented in Table 2. Confusion matrices for pneumothorax and rib fractures are visualized in Figs. 1 and 2, respectively.

All LLM-based labeling methods outperformed the conventional CheXpert labeler in extracting labels from radiology reports. The difference in performance was notable particularly for rib fracture detection, which is a more nuanced label compared to pneumothorax that can most of the time be recognized with word filtering. For example, the sensitivity of the best-performing LLM (Llama with CoT) was 95% (39/41) for rib fractures, compared to just 51% (21/41) for the CheXpert labeler. Similarly, in the case of pneumothorax, the CheXpert labeler had the highest false positive rate of 34% (14/41) compared to other LLM-based methods.

Model size appeared to notably impact model sensitivity, with larger models generally yielding better results. This trend was evident for extracting both pneumothorax and fracture labels from MIMIC-CXR reports. The use of CoT prompting generally had detrimental effects on model performance, especially on smaller models. For pneumothorax detection, the F1 scores for the Llama, Zephyr, and Phi- 3 models without CoT were 0.93, 0.93, and 0.91, respectively. With the addition of CoT, the F1 scores were 0.93, 0.92, and 0.81, respectively. This drop, especially in the case of Phi- 3, suggests that the reasoning stream is confusing the model, resulting in decreased model performance. Examples of this phenomenon is presented in Supplementary Table E1. This mixed pattern is also present in other metrics, such as sensitivity and specificity.

Impact on Downstream Classification Performance

The impact of the different labeling schemas on the performance of downstream image classification models was

Table 2 Performance of different labeling methods for extracting pneumothorax and rib fracture labels against human annotations on a subset of 2,000 reports from the MIMIC-CXR dataset. Values presented as mean (95% CI). Abbreviations: Acc, Accuracy; Sen, Sensitivity; Spe, Specificity; NPV, Negative predictive value; PPV, Positive predictive value; CXP Labeler, CheXpert Labeler; CoT, Chain-of-Thought prompting

Model	Prompting	Acc	Sen	Spe	NPV	PPV	F1 Score
Pneumothorax							
CXP Labeler	N/A	.992 (.988-.995)	.932 (.821–1.00)	.993 (.989-.996)	.999 (.997–1.00)	.660 (.513-.804)	.770 (.652-.867)
Phi- 3	0-shot	.998 (.995-.999)	.862 (.724-.969)	1.00 (.998–1.00)	.998 (.996-.999)	.963 (.875–1.00)	.908 (.809-.980)
	CoT	.995 (.992-.998)	.691 (.518-.870)	1.00 (1.00–1.00)	.995 (.992-.998)	1.00 (1.00–1.00)	.814 (.683-.930)
Zephyr beta	0-shot	.998 (.995-.999)	.964 (.880–1.00)	.998 (.996–1.00)	.999 (.998–1.00)	.901 (.781–1.00)	.930 (.851-.986)
	CoT	.998 (.995-.999)	.967 (.889–1.00)	.998 (.996-.999)	1.00 (.998–1.00)	.879 (.757-.972)	.920 (.833-.981)
LLama- 3	0-shot	.998 (.996-.999)	1.00 (1.00–1.00)	.998 (.996-.999)	1.00 (1.00–1.00)	.879 (.763-.971)	.935 (.866-.986)
	CoT	.998 (.996-.999)	1.00 (1.00–1.00)	.998 (.996-.999)	1.00 (1.00–1.00)	.875 (.741-.972)	.933 (.851-.986)
Rib Fracture							
CXP Labeler	N/A	.977 (.970-.983)	.508 (.342-.652)	.986 (.981-.991)	.990 (.985-.994)	.435 (.294-.579)	.466 (.326-.585)
Phi- 3	0-shot	.990 (.985-.994)	.736 (.590-.864)	.995 (.992-.998)	.994 (.991-.997)	.750 (.609-.878)	.740 (.615-.840)
	CoT	.990 (.985-.994)	.660 (.513-.796)	.997 (.994-.999)	.993 (.989-.996)	.817 (.677-.943)	.727 (.597-.831)
Zephyr-beta	0-shot	.986 (.980-.991)	.953 (.868–1.00)	.987 (.981-.991)	.999 (.997–1.00)	.598 (.479-.716)	.733 (.632-.825)
	CoT	.987 (.982-.991)	.878 (.769-.970)	.989 (.984-.993)	.997 (.995-.999)	.628 (.500-.750)	.730 (.624-.825)
LLama- 3	0-shot	.983 (.977-.988)	.951 (.878–1.00)	.984 (.978-.989)	.999 (.997–1.00)	.548 (.423-.662)	.693 (.585-.788)
	CoT	.986 (.981-.991)	.952 (.878–1.00)	.987 (.982-.992)	.999 (.997–1.00)	.610 (.490-.740)	.741 (.641-.833)

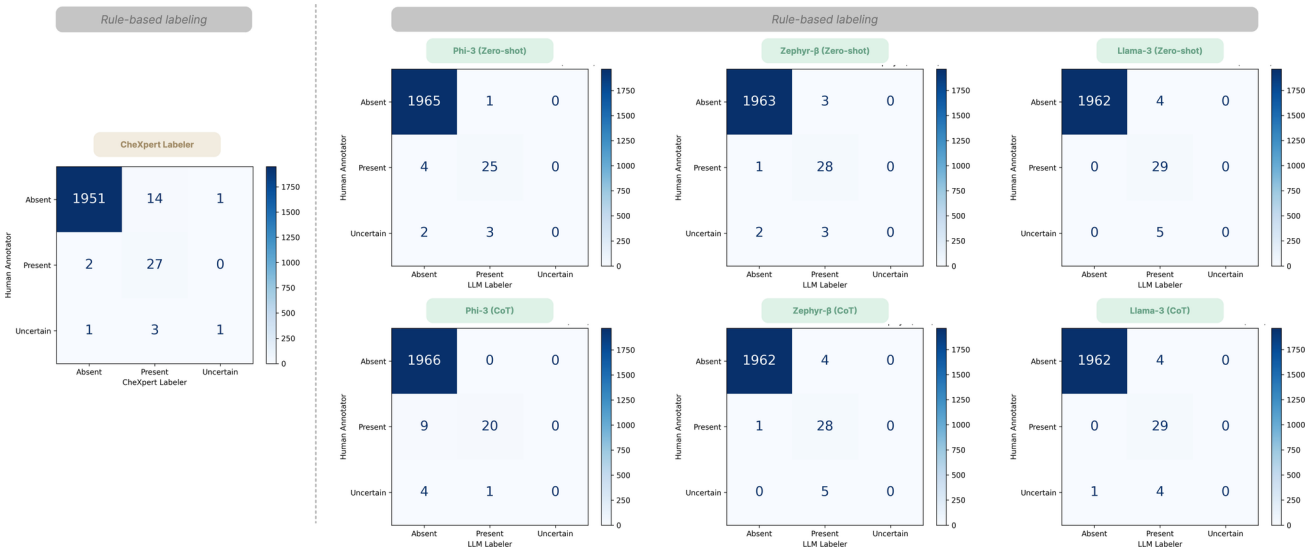

Fig. 1 Confusion matrices of different labeling techniques for extracting the pneumothorax label from a subset of 2000 reports from the MIMIC-CXR dataset

evaluated on the CANDID-PTX dataset. The area under the receiver operating characteristic curve (AUROC) values for pneumothorax and rib fracture detection are summarized in Figs. 3 and 4. Area under the precision-recall curve (AUPRC) values for these two pathologies are presented in Figures E1 and E2.

All models performed well on the downstream classification task, with LLM-based models consistently outperforming the CheXpert model when compared to human annotations on the CANDID-PTX dataset (human labeling > LLM-labeling > CheXpert-labeling; p-value < 0.01 in all instances). However, differences in model size and use of CoT prompting were less pronounced in this setting compared to the impact of the labeling schema. The patterns were reversed in some cases, with the smaller Phi- 3 model performing better than the larger models.

A key finding from this analysis is that the choice of labeling schema for the test set can significantly impact the

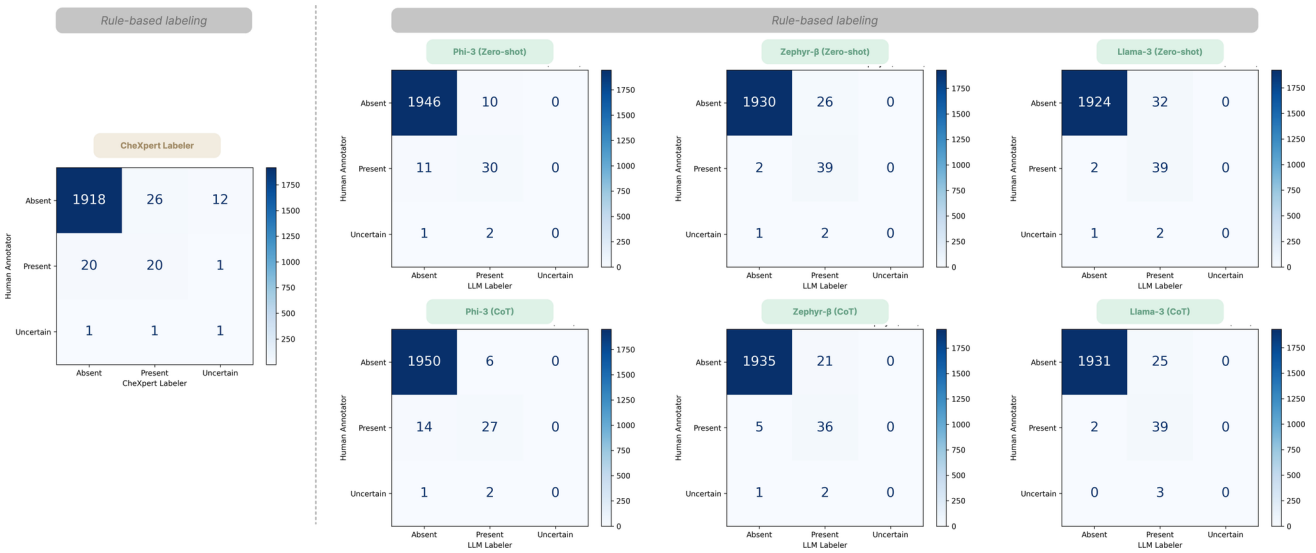

Fig. 2 Confusion matrices of different labeling techniques for extracting the pneumothorax label from a subset of 2000 reports from the MIMIC-CXR dataset

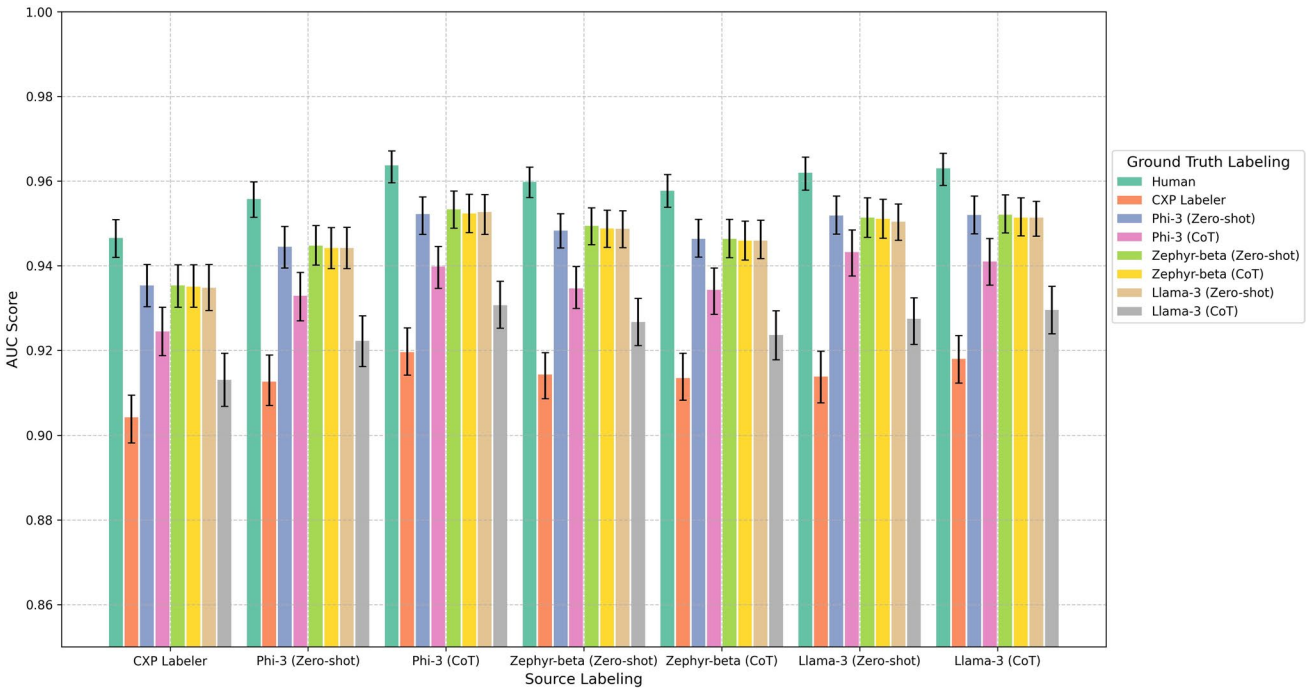

Fig. 3 Area under receiver operating curve (AUC) scores of image-based pneumothorax detection models, evaluated against different ground truth labeling schemas. Abbreviations: CXP Labeler, CheXpert Labeler; CoT, Chain-of-Thought prompting

reported performance of the classification models. When using human annotations as the ground truth, the LLM-based labeling schemas yielded the closest performance to the human evaluations. Conversely, employing weak labeling techniques like CheXpert to annotate the test set underestimated of the classification models'performance. For example, when a classifier was trained on Llama- 3 with CoT labels and evaluated on human annotations, it yielded an AUC of 0.96 and 0.84 for pneumothorax and fracture detection respectively. However, when the same model was evaluated on a test set annotated with the CheXpert labeler, we observed AUCs of 0.91 and 0.73, respectively, which were significantly lower (p-value < 0.01) than their actual real-world performance (derived from the human annotations).

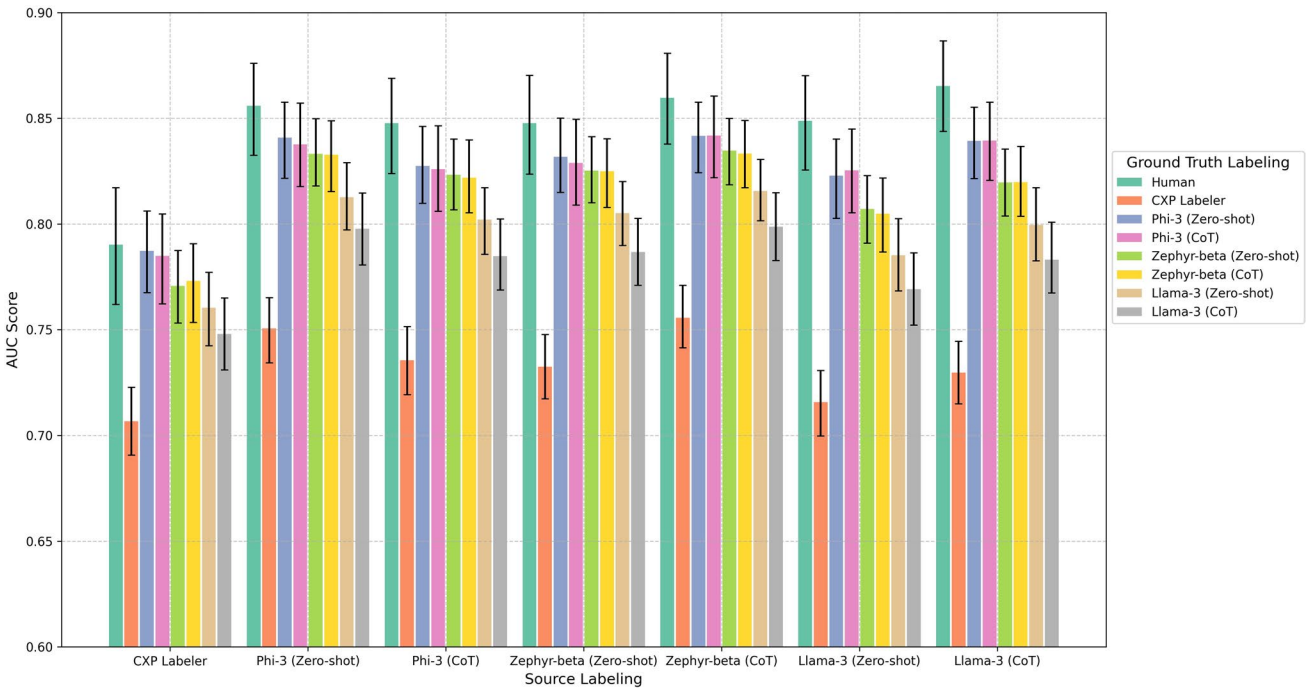

Fig. 4 Area under receiver operating curve (AUC) scores of image-based rib fracture detection models, evaluated against different ground truth labeling schemas. Abbreviations: CXP Labeler, CheXpert Labeler; CoT, Chain-of-Thought prompting

Inference Speed

On the MIMIC-CXR dataset, which contains generally longer reports, the Phi- 3 model with zero-shot prompting was the fastest, processing 1000 reports in 1.1 min. Zephyr-beta and LLama- 3 with zero-shot prompting took 1.2 min each. Applying chain-of-thought (CoT) prompting significantly increased the processing time, with Phi- 3 taking 4.7 min, Zephyr-beta taking 5.9 min, and LLama- 3 taking 6.5 min per 1000 reports. The inference times per 1000 reports for the different models and prompting strategies on the MIMIC-CXR and CANDID-PTX datasets are summarized in Table 3. Parallelizing the requests led to a reasonable inference time for a large dataset size.

Discussion

This study demonstrates the value of open-source large language models (LLMs) for extracting labels from a large-scale radiology dataset, providing valuable insights into factors influencing the quality of generated labels and their usefulness for training deep learning models in medical image analysis. Our findings highlight the importance

Table 3 Inference time of models with and without Chain of Thought (CoT) prompting for both pneumothorax and fracture labels combined (all numbers are minutes)

Model	Total Inference time on MIMIC-CXR ($n = 227{,}827$)	Total Inference time on CANDID-PTX ($n = 19{,}237$)	Inference time for 1000 reports[a]
Zero-shot prompting			
Phi- 3	265	20	1.1
Zephyr-beta	274	19	1.2
LLama- 3	281	21	1.2
Chain of thought (CoT) prompting			
Phi- 3	1071	86	4.7
Zephyr-beta	1358	112	5.9
LLama- 3	1488	122	6.5

[a]Please note that the reference calculation is the MIMIC-CXR reports as those reports are generally longer than CANDID-PTX reports, which are extended impressions

of model size and prompting strategies in label extraction accuracy and downstream task performance.

Quality labels are instrumental for training and evaluating deep learning models [38]. Although there is increased interest in unsupervised and self-supervised learning which usually do not rely on extracted labels, supervised learning remains the most data-efficient approach and is often necessary for fine-tuning pre-trained models. Radiology presents a unique opportunity for label extraction, as expert annotations are readily available in the form of radiology reports, providing a rich source of information.

Previous works have explored various approaches to extracting labels from radiology reports. Cid et al. developed and validated a BERT-based NLP model trained from scratch using 23,000 manually annotated reports, achieving a sensitivity of 94% for detecting pneumothorax and 98% for detecting fractures [8]. While such non-LLM-based models can achieve high accuracy, they require retraining for each new pathology. In contrast, LLMs can be used out-of-the-box and have a larger context length, enabling in-context learning with few-shot prompting [18].

Recent studies have investigated the potential of proprietary LLMs for extracting information from radiology reports. Fink et al. used GPT- 4 and ChatGPT to extract data from 424 CT reports on lung cancer, achieving an accuracy of 96% and 67%, respectively [17]. Similarly, Lehnen et al. employed GPT- 4 to extract 2,800 data points from reports on mechanical thrombectomy in acute ischemic stroke, achieving 94% accuracy and outperforming GPT- 3.5 [15]. These studies demonstrate the effectiveness of proprietary LLMs in extracting information from radiology reports.

While proprietary LLMs have demonstrated impressive performance in extracting information from radiology reports, their use raises concerns about data privacy and security, as sensitive patient information may need to be shared with third-party providers. In contrast, open-source LLMs can be deployed locally, ensuring that sensitive data remains within the institution's secure network and complies with privacy regulations such as HIPAA.

The use of open-source LLMs has also shown promise. Le Guellec et al. used an open-source LLM to extract information from 2,400 MRI reports of headache patients, achieving a sensitivity of 98% [18]. Additionally, Mukherjee et al. explored the feasibility of using the privacy-preserving open-source LLM Vicuna- 13B for labeling radiology reports, demonstrating high concordance between LLM-extracted and human-extracted labels [19]. They also investigated the impact of the temperature parameter in producing structured outputs, a technique we adopted in our study. Based on their findings, we used a temperature of zero to achieve the best possible results and ensure output determinism.

Our study adds to the existing literature through several key findings. First, we observed that increasing model size generally leads to better performance in label extraction tasks. Second, the impact of CoT prompting on model performance was equivocal, with some improvements noted in certain scenarios. One interesting observation was that especially in the case of smaller models, CoT, decreased the overall performance. This observation aligns with recent studies indicating that that "overthinking" negatively impacts model performance on simple tasks [39, 40]. This suggests there is no one-size-fits all prompting strategy, and plain zero-shot prompting can be more beneficial in easier tasks. When using these labels to train an image classifier and test it on an external dataset with image-level radiologist annotations, we found that deep learning image classification models were remarkably tolerant to label noise, with no significant performance difference between different labeling schemas [5, 41]. However, LLM-based models still outperformed rule-based models, highlighting the importance of the dataset labeling method [8]. Our results emphasize the significance of evaluation methods, with the best approach being a comparison against human annotations. When this is not feasible due to large dataset sizes, employing more reliable tools such as LLMs for labeling can provide a better estimate of the final model's performance.

One finding of our study is the lack of significant improvement when using CoT prompting, both in smaller and larger models, as opposed to previous studies [42, 43]. We hypothesize that this finding might be because CoT potentially causes the models to "overthink" their responses and change them to incorrect answers. We think that zero-shot prompting might be sufficient for simple reasoning tasks such as determining the presence or absence of a finding in a radiology report, as LLMs have been trained on vast amounts of data and can comprehend these language nuances. However, for more reasoning-intensive tasks, such as deducing a pathology based on a constellation of symptoms, CoT prompting might be beneficial. This finding is in line with recent work on general-purpose tasks but remains to be further explored in the medical domain [44].

Our results should be interpreted with some limitations considered. First, the largest model we evaluated had 8 billion parameters. Exploring even larger models might further improve performance, but our goal was to demonstrate a cost-efficient approach, and we showed that even the 3-billion-parameter Phi- 3 model performed better than traditional methods for labeling data. Second, our comparison across different model families (Phi- 3, Zephyr-beta, and Llama- 3) captures nuances beyond mere parameter count, as these models were developed by different organizations using varying data sources. A more controlled assessment would involve comparing models within the same family (e.g., different sizes of Phi- 3 or

Llama- 3 models). However, at the time of our study, such comprehensive size ranges within the same open-source model family were not readily available based on our computational resources. Third, when comparing our results to the CANDID-PTX dataset, we relied on radiologist labels, which are at the image level. This poses a limitation as occasionally the report does not mention fracture or a small pneumothorax, but it was annotated by the radiologist on the image, making the report an imperfect surrogate for image-level labels. Finally, we did not finetune our LLMs which has been shown to enhance the model's task-specific performance [45, 46]. Future studies should investigate the impact of fine tuning on label extraction and downstream tasks.

In conclusion, our study demonstrates the effectiveness of open-source LLMs for large-scale label extraction in radiology datasets, showing that heavily pre-trained LLMs, even with a smaller number of parameters, can be highly useful for training or fine-tuning downstream models. However, the choice of model size and prompting strategy should be guided by the specific task at hand. Our findings suggest that larger models may be necessary when directly using extracted labels for patient selection or inclusion/exclusion from a dataset, and appropriate metrics for evaluation should be chosen based on the desired outcome [47]. In situations where labels will be used without further human review, employing state-of-the-art models and improved prompting strategies may be warranted.

Our findings provide valuable insights for researchers and practitioners in the field of medical image analysis, emphasizing the importance of model size, prompting strategies, and evaluation methods in the context of label extraction and downstream model performance. Future work should explore the application of even larger models, fine-tuning, and the extension of these strategies to more complex reasoning tasks, as well as investigating the impact of these parameters on specific clinical use cases, such as patient selection for clinical trials or cohort studies.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10278-025-01505-7.

Funding This work was supported by the National Institute on Minority Health and Health Disparities [1R21MD019360 to B.K., B.J.E.]; the Robert Wood Johnson Foundation Harold Amos Medical Faculty Development Program [to J.W.G., 2022]; the Radiological Society of North America (RSNA) Health Disparities grant [EIHD2204 to J.W.G.]; Lacuna Fund [67 to J.W.G.]; Gordon and Betty Moore Foundation [to J.W.G.]; National Institute of Biomedical Imaging and Bioengineering (NIBIB) MIDRC grants [75 N92020 C00008, 75 N92020 C00021 to J.W.G.]; and the National Heart, Lung, and Blood Institute [R01HL167811 to J.W.G.].

Declarations

Competing Interests The authors declare no competing interests.

References

1. Khosravi B, Rouzrokh P, Maradit Kremers H, Larson DR, Johnson QJ, Faghani S, et al. Patient-specific hip arthroplasty dislocation risk calculator: An explainable multimodal machine learning-based approach. Radiol Artif Intell. 2022 Nov;4(6):e220067.
2. Faghani S, Khosravi B, Moassefi M, Conte GM, Erickson BJ. A comparison of three different deep learning-based models to predict the MGMT promoter methylation status in glioblastoma using brain MRI. J Digit Imaging. 2023 Jun 5;36(3):837–46.
3. Condon JJJ, Trinh V, Hall KA, Reintals M, Holmes AS, Oakden-Rayner L, et al. Impact of Transfer Learning Using Local Data on Performance of a Deep Learning Model for Screening Mammography. Radiol Artif Intell. 2024 May 8;e230383.
4. Dai C, Xiong Y, Zhu P, Yao L, Lin J, Yao J, et al. Deep Learning Assessment of Small Renal Masses at Contrast-enhanced Multiphase CT. Radiology. 2024 May;311(2):e232178.
5. Karimi D, Dou H, Warfield SK, Gholipour A. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Med Image Anal. 2020 Oct;65:101759.
6. 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 [Internet]. arXiv [cs.CV]. 2019. Available from: http://arxiv.org/abs/1901.07031
7. McDermott MBA, Hsu TMH, Weng WH, Ghassemi M, Szolovits P. CheXpert++: Approximating the CheXpert labeler for Speed,Differentiability, and Probabilistic Output [Internet]. arXiv [cs.LG]. 2020. Available from: http://arxiv.org/abs/2006.15229
8. Cid YD, Macpherson M, Gervais-Andre L, Zhu Y, Franco G, Santeramo R, et al. Development and validation of open-source deep neural networks for comprehensive chest x-ray reading: a retrospective, multicentre study. Lancet Digit Health. 2024 Jan;6(1):e44–57.
9. Kim W. Seeing the Unseen: Advancing Generative AI Research in Radiology. Radiology. 2024 May;311(2):e240935.
10. Doshi R, Amin KS, Khosla P, Bajaj SS, Chheang S, Forman HP. Quantitative Evaluation of Large Language Models to Streamline Radiology Report Impressions: A Multimodal Retrospective Analysis. Radiology. 2024 Mar;310(3):e231593.
11. Amin KS, Davis MA, Doshi R, Haims AH, Khosla P, Forman HP. Accuracy of ChatGPT, Google Bard, and Microsoft Bing for Simplifying Radiology Reports. Radiology. 2023 Nov;309(2):e232561.
12. Li D, Gupta K, Chong J. Evaluating Diagnostic Performance of ChatGPT in Radiology: Delving into Methods. Radiology. 2023 Sep;308(3):e232082.
13. Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of Breast Cancer Prevention and Screening Recommendations Provided by ChatGPT. Radiology. 2023 May;307(4):e230424.
14. Goel A, Gueta A, Gilon O, Liu C, Erell S, Nguyen LH, et al. LLMs Accelerate Annotation for Medical Information Extraction [Internet]. Hegselmann S, Parziale A, Shanmugam D, Tang S, Asiedu MN, Chang S, et al., editors. arXiv [cs.CL]. 2023. p. 82–100. (Proceedings of Machine Learning Research). Available from: https://proceedings.mlr.press/v225/goel23a/goel23a.pdf
15. Lehnen NC, Dorn F, Wiest IC, Zimmermann H, Radbruch A, Kather JN, et al. Data Extraction from Free-Text Reports on Mechanical Thrombectomy in Acute Ischemic Stroke Using ChatGPT: A Retrospective Analysis. Radiology. 2024 Apr;311(1):e232741.
16 Cozzi A, Pinker K, Hidber A, Zhang T, Bonomo L, Lo Gullo R, et al. BI-RADS Category Assignments by GPT-3.5, GPT-4, and Google Bard: A Multilanguage Study. Radiology. 2024 Apr;311(1):e232133.

17. 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. Radiology. 2023 Sep;308(3):e231362.
18. Le Guellec B, Lefèvre A, Geay C, Shorten L, Bruge C, Hacein-Bey L, et al. Performance of an Open-Source Large Language Model in Extracting Information from Free-Text Radiology Reports. Radiol Artif Intell. 2024 May 8;e230364.
19. Mukherjee P, Hou B, Lanfredi RB, Summers RM. Feasibility of Using the Privacy-preserving Large Language Model Vicuna for Labeling Radiology Reports. Radiology. 2023 Oct;309(1):e231147.
20. Johnson A, Lungren M, Peng Y, Lu Z, Mark R, Berkowitz S, et al. MIMIC-CXR-JPG - chest radiographs with structured labels [Internet]. PhysioNet; 2024. Available from: https://physionet.org/content/mimic-cxr-jpg/2.1.0/
21. Johnson AEW, Pollard TJ, Greenbaum NR, Lungren MP, Deng CY, Peng Y, et al. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs [Internet]. arXiv [cs.CV]. 2019. Available from: http://arxiv.org/abs/1901.07042
22. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals: Components of a new Research Resource for Complex Physiologic Signals. Circulation. 2000 Jun 13;101(23):E215–20.
23. Feng S, Azzollini D, Kim JS, Jin CK, Gordon SP, Yeoh J, et al. Curation of the CANDID-PTX Dataset with Free-Text Reports. Radiol Artif Intell. 2021 Nov;3(6):e210136.
24. AI@Meta. Llama 3 Model Card [Internet]. 2024. Available from: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
25. Tunstall L, Beeching E, Lambert N, Rajani N, Rasul K, Belkada Y, et al. Zephyr: Direct Distillation of LM Alignment [Internet]. arXiv [cs.LG]. 2023. Available from: http://arxiv.org/abs/2310.16944
26. Abdin M, Jacobs SA, Awan AA, Aneja J, Awadallah A, Awadalla H, et al. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone [Internet]. arXiv [cs.CL]. 2024. Available from: http://arxiv.org/abs/2404.14219
27. Kwon W, Li Z, Zhuang S, Sheng Y, Zheng L, Yu CH, et al. Efficient Memory Management for Large Language Model Serving with PagedAttention [Internet]. arXiv [cs.LG]. 2023. Available from: http://arxiv.org/abs/2309.06180
28. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models [Internet]. arXiv [cs.CL]. 2022. Available from: http://arxiv.org/abs/2201.11903
29. Khosravi B, Li F, Dapamede T, Rouzrokh P, Gamble CU, Trivedi HM, et al. Synthetically enhanced: unveiling synthetic data's potential in medical imaging research. EBioMedicine. 2024 May 30;104:105174.
30. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med. 2021 Dec;27(12):2176–82.
31. Yun S, Han D, Oh SJ, Chun S, Choe J, Yoo Y. CutMix: Regularization strategy to train strong classifiers with localizable features. ICCV. 2019 May 13;6022–31.
32. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. mixup: Beyond Empirical Risk Minimization [Internet]. arXiv [cs.LG]. 2017. Available from: http://arxiv.org/abs/1710.09412
33. The MONAI Consortium. Project MONAI [Internet]. Zenodo; 2020. Available from: https://zenodo.org/record/4323059
34. Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s [Internet]. arXiv [cs.CV]. 2022 [cited 2023 May 20]. Available from: http://arxiv.org/abs/2201.03545
35. Wightman R, Raw N, Soare A, Arora A, Ha C, Reich C, et al. rwightman/pytorch-image-models: v0.8.10dev0 Release [Internet]. Zenodo; 2023. Available from: https://zenodo.org/record/7618837
36. Chen X, Liang C, Huang D, Real E, Wang K, Liu Y, et al. Symbolic discovery of optimization algorithms [Internet]. arXiv [cs.LG]. 2023. Available from: http://arxiv.org/abs/2302.06675
37. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Med Res Methodol. 2013 Apr 29;13(1):61.
38. 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 Sep;4(5):e210290.
39. Chen X, Xu J, Liang T, He Z, Pang J, Yu D, et al. Do NOT think that much for 2+3=? On the overthinking of o1-like LLMs [Internet]. arXiv [cs.CL]. 2024. Available from: http://arxiv.org/abs/2412.21187
40. Cuadron A, Li D, Ma W, Wang X, Wang Y, Zhuang S, et al. The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks [Internet]. arXiv [cs.AI]. 2025. Available from: http://arxiv.org/abs/2502.08235
41. Rueckel J, Huemmer C, Fieselmann A, Ghesu FC, Mansoor A, Schachtner B, et al. Pneumothorax detection in chest radiographs: optimizing artificial intelligence system for accuracy and confounding bias reduction using in-image annotations in algorithm training. Eur Radiol. 2021 Oct;31(10):7888–900.
42. Gramopadhye O, Nachane SS, Chanda P, Ramakrishnan G, Jadhav KS, Nandwani Y, et al. Few shot chain-of-thought driven reasoning to prompt LLMs for open ended medical question answering [Internet]. arXiv [cs.CL]. 2024. Available from: http://arxiv.org/abs/2403.04890
43. Patel D, Raut G, Zimlichman E, Cheetirala SN, Nadkarni GN, Glicksberg BS, et al. Evaluating prompt engineering on GPT-3.5's performance in USMLE-style medical calculations and clinical scenarios generated by GPT-4. Sci Rep. 2024 Jul 28;14(1):17341.
44. Liu R, Geng J, Wu AJ, Sucholutsky I, Lombrozo T, Griffiths TL. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse [Internet]. arXiv [cs.LG]. 2024. Available from: http://arxiv.org/abs/2410.21333
45. Yang Q, Wang R, Chen J, Su R, Tan T. Fine-Tuning Medical Language Models for Enhanced Long-Contextual Understanding and Domain Expertise [Internet]. arXiv [cs.CL]. 2024. Available from: http://arxiv.org/abs/2407.11536
46. Kanemaru N, Yasaka K, Fujita N, Kanzawa J, Abe O. The Fine-Tuned Large Language Model for Extracting the Progressive Bone Metastasis from Unstructured Radiology Reports. J Imaging Inform Med [Internet]. 2024 Aug 26; Available from: https://doi.org/10.1007/s10278-024-01242-3
47. Faghani S, Khosravi B, Zhang K, Moassefi M, Jagtap JM, Nugen F, et al. Mitigating bias in radiology machine learning: 3. Performance metrics. Radiol Artif Intell. 2022 Sep 24;4(5):e220061.