Integrating ChatGPT into Secure Hospital Networks: A Case Study on Improving Radiology Report Analysis

Kyungsu Kim* KSKIM@MGH.HARVARD.EDU
Department of Radiology, Massachusetts General Hospital and Harvard Medical School, USA
Junhyun Park* SEAN05071@DGIST.AC.KR
Department of Robotics and Mechatronics Engineering, DGIST, Republic of Korea
Saul Langarica SLANGARICA@MGH.HARVARD.EDU
Department of Radiology, Massachusetts General Hospital and Harvard Medical School, USA
Adham Mahmoud Alkhadrawi ADHAM.ALKHADRAWI@MGH.HARVARD.EDU
Department of Radiology, Massachusetts General Hospital and Harvard Medical School, USA
Synho Do† SDO@MGH.HARVARD.EDU
Department of Radiology, Massachusetts General Hospital and Harvard Medical School, USA
KU-KIST Graduate School of Converging Science and Technology, Korea University, Republic of Korea
Kempner Institute, Harvard University, USA

Abstract

This study demonstrates the first in-hospital adaptation of a cloud-based AI, similar to ChatGPT, into a secure model for analyzing radiology reports, prioritizing patient data privacy. By employing a unique sentence-level knowledge distillation method through contrastive learning, we achieve over 95% accuracy in detecting anomalies. The model also accurately flags uncertainties in its predictions, enhancing its reliability and interpretability for physicians with certainty indicators. Despite limitations in data privacy during the training phase, such as requiring de-identification or IRB permission, our study is significant in addressing this issue in the inference phase (once the local model is trained), without the need for human annotation throughout the entire process. These advancements represent a new direction for developing secure and efficient AI tools for healthcare with minimal supervision, paving the way for a promising future of in-hospital AI applications.

Data and Code Availability. Our study employs the MIMIC-CXR radiology report dataset (Johnson et al., 2019), accessible to the public. The replication code is available at GitHub and HuggingFace.

Institutional Review Board (IRB) Our research, utilizing only the publicly available MIMIC-CXR dataset, is exempt from Institutional Review Board (IRB) regulation. Access to this dataset has been approved by PhysioNet.

1. Introduction

The research explores the integration of artificial intelligence (AI), specifically large language models (LLMs) like ChatGPT into radiology within hospitals with an emphasis on maintaining security during implementation.

Despite the proven effectiveness of these AI tools in processing radiological reports (Wu et al., 2024; Mirza et al., 2024; Lee et al., 2023), their integration into hospital environments poses challenges due to the sensitive nature of patient data and the need for data confidentiality (Senbekov et al., 2020). The direct use of cloud-based LLMs like ChatGPT is limited by data security concerns, especially when considering healthcare regulations such as HIPAA (Gostin et al., 2009) and GDPR (Voigt and Von dem Bussche, 2017).

Our study addresses this by adapting these LLMs for secure, internal use within hospital radiology departments, transforming them into closed-network systems to comply with healthcare privacy standards. This approach aims to leverage the advanced capabilities of LLMs while safeguarding patient data privacy.

This paper delves into how radiology reports can be automatically classified as normal or abnormal using

* These authors contributed equally
† Correspondence to

© 2024 K. Kim, J. Park, S. Langarica, A.M. Alkhadrawi & S. Do.

Table 1: Comparison of our study with related ones to develop language models applied to electronic medical record (EMR) documents. Our is the first study aimed at reproducing the cloud model into a non-cloud/secure model (second column).

Study	Does it demonstrate the feasibility of knowledge distillation (KD) learning from a cloud-based model to an on-premises model?	Does it propose an advancement technique for KD and provide its reasons of improvement?	Model type (Cloud or On-premises type)?	Does it address model learning? / If yes, what kind of learning data the model use among public data (p), private data (i), cloud model's prediction result data (c)?	What is the showcase application of the model?	Is the public code available?
Li et al. (2023)	No	No	Cloud (without KD)	No	Summarization	No
Liu et al. (2023c)	No	No	Cloud or On-premises (without KD)	No	Generation	No
Ma et al. (2023)	No	No	Cloud (without KD)	No	Generation	Yes
Liu et al. (2023a)	No	No	On-premises (without KD)	Yes (p)	Generation	No
Liu et al. (2023b)	No	No	On-premises (without KD)	Yes (p)	Generation	(Param. only)
Zhong et al. (2023)	No	No	On-premises (without KD)	Yes (i)	Generation	No
Van Veen et al. (2023)	No	No	On-premises (without KD)	Yes (p)	Generation	Yes
Mukherjee et al. (2023)	No	No	On-premises (without KD)	Yes (p)	Classification	No
Bressem et al. (2020)	No	No	On-premises (without KD)	Yes (i)	Classification	Yes
Yan et al. (2022)	No	No	On-premises (without KD)	Yes (p)	Classification	(Param. only)
Our study	**Yes**	**Yes (i.e., Sentence-level KD)**	**On-premises (trained with KD from cloud model)**	**Yes (c)**	**Classification (Abnormal detection)**	**Yes**

cloud-based/high-performing LLMs like ChatGPT, with the goal of adapting these models for secure and internal use within hospital networks. This approach aims to enhance hospital workflows by streamlining the analysis of radiology findings, potentially leading to more efficient and accurate medical diagnostics and patient care management.

This investigation is important for enhancing the practical utility of AI in radiology, ensuring both technological advancement and adherence to the paramount principle of patient confidentiality. Our contribution is three-fold:

- Successfully adapted a cloud-based model like ChatGPT into an on-site version with over 95% accuracy for detecting anomalies in radiology reports, offering a secure method for local data processing (Table 2).
- Demonstrated that sentence-level knowledge distillation outperforms traditional document-level methods in improving model replication by better identifying rare abnormal findings, supported by analytical evidence (Figure 3).
- Improved model interpretability by adding an "uncertain" label to the usual "normal" and "abnormal" in sentence-level knowledge distillation. This allows the model to identify ambiguous cases in radiology reports, enhancing sentence-level accuracy and clarity (Figure 6). The provided code visualizes sentence-based predictions, helping physicians focus on critical findings during review by clearly marking uncertain sentences.

2. Related Works

LLMs, like ChatGPT, are used in radiology research to analyze reports and are classified as cloud-based or non-cloud-based. Rows 1-3 of Table 1 illustrate how cloud-based studies employ LLMs for report generation or summarization, utilizing prompt engineering without additional model training. However, this dependence on cloud storage raises concerns regarding data security. On the other hand, as shown in column 5 of Table 1, non-cloud (i.e., on-premises) approaches, which are described in rows 4–10 of the table, need human annotation for model training, which means a substantial amount of work for data preparation—more particularly, p- or i-type annotation.

In contrast to these previous studies that focus on either cloud type or non-cloud type, our study is the first to use both types, by incorporating knowledge distillation (KD) (Gou et al., 2021) into LLMs for radiology report analysis (see the second column in Table 1 and Fig. 1). Specifically, we utilize both types by replicating the cloud type as the non-cloud type through KD. This approach involves training a condensed non-cloud model (referred to as the student) to emulate the capabilities of a more extensive cloud model (known as the teacher), such as ChatGPT. Our approach stands out for utilizing automatically processed data from the cloud model to train

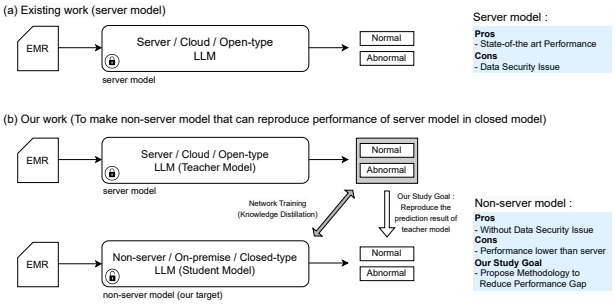


Figure 1: Our study is the first to test the feasibility of distilling knowledge from cloud model like ChatGPT into a non-cloud model for radiology report analysis

the non-cloud model (i.e., as c-type data), instead of relying on human-annotated data such as p- or i-type annotation data stated in rows 4-10 in column 5 of Table 1. Accordingly, our approach bypasses the requirement for data labeled by humans. Furthermore, our technique addresses the security concerns associated with evaluation data, by uploading only the limited, de-identified training data to the cloud model, thereby excluding the remaining/unlimited evaluation data that instead utilizes our trained non-cloud model without the security concerns.

In addition, we have developed an advanced technique for KD using a sentence-level approach. This method outperforms baseline document-level KD methods as shown in the third column of Table 1 and Fig. 2. This new approach greatly enhances the model's ability to identify anomalies in documents, especially in challenging scenarios when the document includes a lower number of abnormal sentences. Furthermore, the incorporation of contrastive learning loss into the KD process has enhanced the model's precision in recognizing the class with few training instances (i.e., normal class). Therefore, our incorporation of sentence-level KD and contrastive learning loss results in a notable improvement in the utilization of KD for analyzing radiological reports in language models.

3. Method

In this section, we introduce a KD approach for anomaly detection in radiology reports. This involves two primary methods: the baseline document-level KD (Sec. 3.1) and our proposed sentence-level KD (Sec. 3.2) approaches. We also introduce the KD

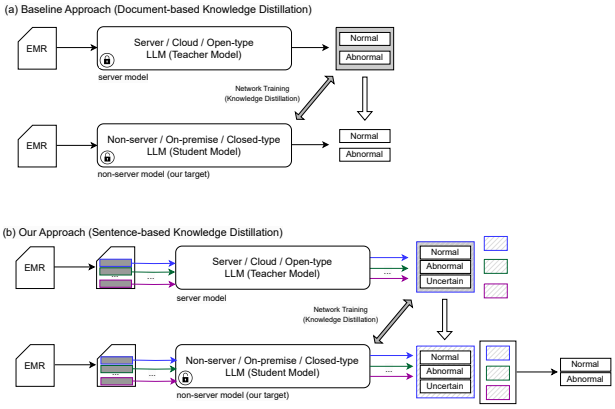


Figure 2: Improving knowledge distillation performance and interpretability, our approach incorporates sentence-level knowledge distillation and enhances reliability by introducing an additional label (uncertain) for the network explicitly to indicate uncertainty in prediction results

objective function employed for training both approaches (Sec. 3.3).

3.1. Baseline: Document-level Knowledge Distillation

In this section, we present a method for anomaly detection (AD) in radiology reports using document-level knowledge distillation (D-KD). Document-level input is standard approach in text document classification (Adhikari et al., 2020; Yao et al., 2019; Ranjan and Prasad, 2023). However, as there are no cases in which the KD technique has been applied to processing radiology reports, we proposed the baseline D-KD technique as a representative and straightforward method for this purpose. Sec. 3.1.1 details the process of extracting labels from the teacher model, ChatGPT, to create training data. Sec. 3.1.2 details the KD training of an on-premise LLM student model using data from Sec. 3.1.1, showing the subsequent testing of the model.

3.1.1. Label Extraction from Teacher Model

The radiology report was input into the teacher model (i.e., ChatGPT), which then determined if the report was normal or abnormal, using n for normal labels and a for abnormal. This process is defined by

the function f^d as follows:

$$f^d(x_i; c) =: y_i^d \in \{a, n\}. \quad (1)$$

Here, x_i represents the i-th radiology document, and c denotes the question prompt (see the details in Appendix) used in ChatGPT to generate predictions for anomaly detection by the binary label symbolized as y_i^d. Through this process, T number of KD-training data pairs for AD (i.e., $D = (x_i, y_i^d)_{i=1}^T$) were constructed.

3.1.2. Training and Inference for Student Model

- **Training Phase.** We updated the student model g_θ from the training data constructed in Sec. 3.1.1 with model parameters θ to minimize the objective below

$$\theta^* \leftarrow \min_\theta \mathbb{E}_{(i \in \{1:T\})} \left[\mathcal{L}_\theta \left(g_\theta(x_i), y_i^d \right) \right] \quad (2)$$

where θ^* is the trained parameter and $\mathcal{L}_\theta$ is our objective function for KD (see the details in Sec. 3.3).

- **Inference Phase.** Binary classification evaluation of radiology reports for AD was performed from the student model g_{θ^*} on which KD-learning was completed as follows:

$$p_a \leftarrow g_{\theta^*}(x^{te})_{\{a\}},$$
$$p_n \leftarrow 1 - p_a.$$

Here, x^{te} denotes the radiology report used for testing, and the student model $g_{\theta^*}(x^{te}) \mapsto (g_{\theta^*}(x^{te})_{\{a\}}, g_{\theta^*}(x^{te})_{\{n\}}) =: (p_a, p_n) \in [0,1]^2$ converts it into a binary probability vector (p_a, p_n) within $\mathbb{R}^2$. This vector's first element, p_a, reflects the model's estimated probability of the input document x^{te} being abnormal.

3.2. Proposed: Sentence-level Knowledge Distillation

In this section, we newly introduce S-KD, a sentence-level-based KD method, more advanced than D-KD in Sec. 3.1.

3.2.1. Label Extraction from Teacher Model

Unlike the baseline method, where the entire radiology report is input into the teacher model as shown in Eq. (1), our approach inputs individual sentences $s_{ij} \in \{s_{ij}\}_{j=1}^D$ from report x_i into ChatGPT. This yields ternary anomaly detection (AD) labels $\{a, n, u\}$, explicitly incorporating an "uncertain" label u alongside the existing binary labels $\{a, n\}$:

$$f^s(s_{ij}; c) =: y_{ij}^s \in \{a, n, u\}.$$

Here, x_i, s_{ij}, and y_{ij}^s denote the i-th radiology document, its j-th sentence, and its model prediction as ternary label, respectively. Accordingly, a total $\sum_{j=1}^T D_j$ of KD-training data pairs for AD were constructed, as $(s_{ij}, y_{ij}^s)_{(i,j)=(1,1)}^{(T,D_i)}$.

Table 2: Anomaly detection performance comparison between document-level and sentence-level KD approaches across various backbone student models

Model	Accuracy	Specificity	Sensitivity	AUC
RadBERT-Roberta -4m-document	85.52	0.858	0.84	0.901
RadBERT-Roberta -4m-sentence	**95.06** (+ 9.54)	**0.941** (+ 0.083)	**0.952** (+ 0.112)	**0.977** (+ 0.076)
BioMed-Roberta -document	86.12	0.82	0.869	0.877
BioMed-Roberta -sentence	**94.6** (+ 8.48)	**0.947** (+ 0.127)	**0.943** (+ 0.074)	**0.979** (+ 0.102)
BlueBERT -document	91.17	0.91	0.922	0.958
BlueBERT -sentence	**93.43** (+ 2.26)	**0.933** (+ 0.023)	**0.945** (+ 0.023)	**0.98** (+ 0.022)
Clinical BERT -document	90.15	0.888	0.961	0.968
Clinical BERT -sentence	**93.07** (+ 2.92)	**0.922** (+ 0.034)	**0.973** (+ 0.012)	**0.982** (+ 0.014)
BiomedBERT -document	92.76	**0.93**	0.916	0.926
BiomedBERT -sentence	**93.07** (+ 0.31)	0.926 (- 0.004)	**0.961** (+ 0.045)	**0.982** (+ 0.056)
BioBERT -document	90.5	0.905	0.906	0.959
BioBERT -sentence	**92.37** (+ 1.87)	**0.923** (+ 0.018)	**0.929** (+ 0.023)	**0.973** (+ 0.014)
p-value	0.002	0.04	0.002	0.002
Average ratio of sent./doc. performance	1.047	1.053	1.053	1.051

3.2.2. Training and Inference for Student Model

- **Training Phase.** Our KD training follows the same method as document-level KD training in Eq. (2), but differs only in the input/label data as sentence-level:

$$\theta^* \leftarrow \min_\theta \mathbb{E}_{(i \in T,\, j \in D_i)} \left[\mathcal{L}_\theta \left(g_\theta(s_{ij}), y_{ij}^s \right) \right]. \quad (3)$$

Table 3: Distribution comparison in medical reports: Analyzing abnormal, normal, and uncertain sentences between D-KD and S-KD on RadBERT-Roberta

Percentage(%)	GT	Abnormal	Normal	Uncertain	Doc. Count
Test Dataset	Abnor	**44.85 ± 22.1**	27.17 ± 20.1	27.97 ± 17.3	2394
	Normal	0.0 ± 0.0	63.61 ± 21.1	**36.39 ± 21.1**	438
D-KD Incorrect	Abnor	**29.14 ± 17.4**	42.17 ± 21.1	28.69 ± 18.0	340
	Normal	0.0 ± 0.0	61.86 ± 26.3	**38.14 ± 26.3**	70
S-KD Incorrect	Abnor	**21.28 ± 11.5**	51.10 ± 17.3	27.62 ± 18.3	114
	Normal	0.0 ± 0.0	58.79 ± 27.4	**41.21 ± 27.4**	26
D-KD Incorrect ∩ S-KD Correct	Abnor	31.46 ± 18.0	40.01 ± 20.9	28.53 ± 17.8	282
	Normal	0.0 ± 0.0	63.69 ± 24.5	36.31 ± 24.5	60

- **Inference Phase.** Therefore, the learned student model g_{θ^*} can provide ternary classification prediction results for individual sentences in the test report x^{te}, i.e., $g_{\theta^*}(s_j^{te}) \mapsto (p_a^j, p_n^j, p_u^j) \in [0,1]^3$ for $j \in \{1 : D_{x^{te}}\}$, where $(p_a^j, p_n^j, p_u^j) =: (g_{\theta^*}(s_j^{te})_{\{a\}}, g_{\theta^*}(s_j^{te})_{\{n\}}, g_{\theta^*}(s_j^{te})_{\{u\}})$ and $D_{x^{te}}$ is the total number of sentences in the report x^{te}. Here, p_a^j, p_n^j, and p_u^j represent the sentence-level (the j-th sentence's) probability for being abnormal, normal, and uncertain, respectively. Then, the final document-level abnormality probability p_a is driven as the highest sentence-level probability (togetherwith its inverse value as normal probability p_n) as follows:

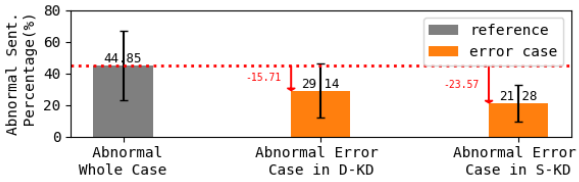
$$p_a \leftarrow \max_{(j \in \{1:D_{x^{te}}\})} \left[g_{\theta^*}(s_j^{te})_{\{a\}} \right], \quad (4)$$
$$p_n \leftarrow 1 - p_a.$$

This allows for an abnormal document classification if even one sentence is deemed abnormal.

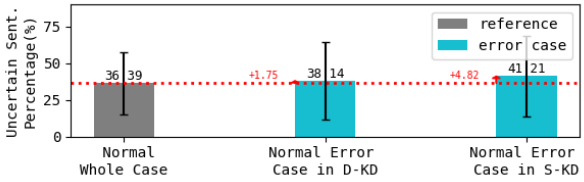
3.3. Objective Function for KD Training

Note the KD objective function $\mathcal{L}_\theta$ used in Eqs. (2) and (3), is defined as

$$\mathcal{L}_\theta(g_\theta(x), y) := \mathcal{L}_\theta^{cross}(g_\theta(x), y) + \lambda \cdot L_\theta^{cont}(g_\theta(x), y) \quad (5)$$



(a) Abnormal sentence distribution for abnormal documents



(b) Uncertain sentence distribution for normal documents

Figure 3: Comparison of AD performance between S-KD and D-KD: S-KD demonstrates superior detection in abnormal (or normal) documents with fewer abnormal (or uncertain) sentences, outperforming D-KD in identifying challenging AD cases

by adding the supervised contrastive loss L^{cont} (Khosla et al., 2020) to the cross-entropy loss $\mathcal{L}^{cross}$, where

$$\mathcal{L}_\theta^{cross}(g_\theta(x), y) := -\sum_k p_y[k] \cdot \log\Big(g_\theta(x)[k]\Big),$$

$$L_\theta^{cont}(g_\theta(x), y) := -\log \mathbb{E}_{(v \in B_y)} \left[\frac{e^{(\mathrm{sim}(z, z_v)/\tau)}}{\sum_{k \in B} e^{(\mathrm{sim}(z, z_k)/\tau)}} \right].$$

Here, p_y is the one-hot vector representation of y, $z := z_\theta(x)$ and $z_v := z_\theta(x_v)$ represent the latent feature vectors of g_θ for the target input x and another x_v (where y_v is its label), B is the batch set of training data (and $B_y := \{v \in B | y_v = y\}$ is its subset whose labels are our target label y), and $\mathrm{sim}(\cdot)$ is the similarity metric. The addition of the contrastive loss L^{cont} aims to minimize distances in the latent space $z_\theta(\cdot)$ within the same class and maximize those between different classes, thereby enhancing KD performance by strengthening the balance of each class (refer to Secs. 4.4 and 4.5).

4. Results

In this section, we present the experimental results of our study. Specifically, we explain the superiority of

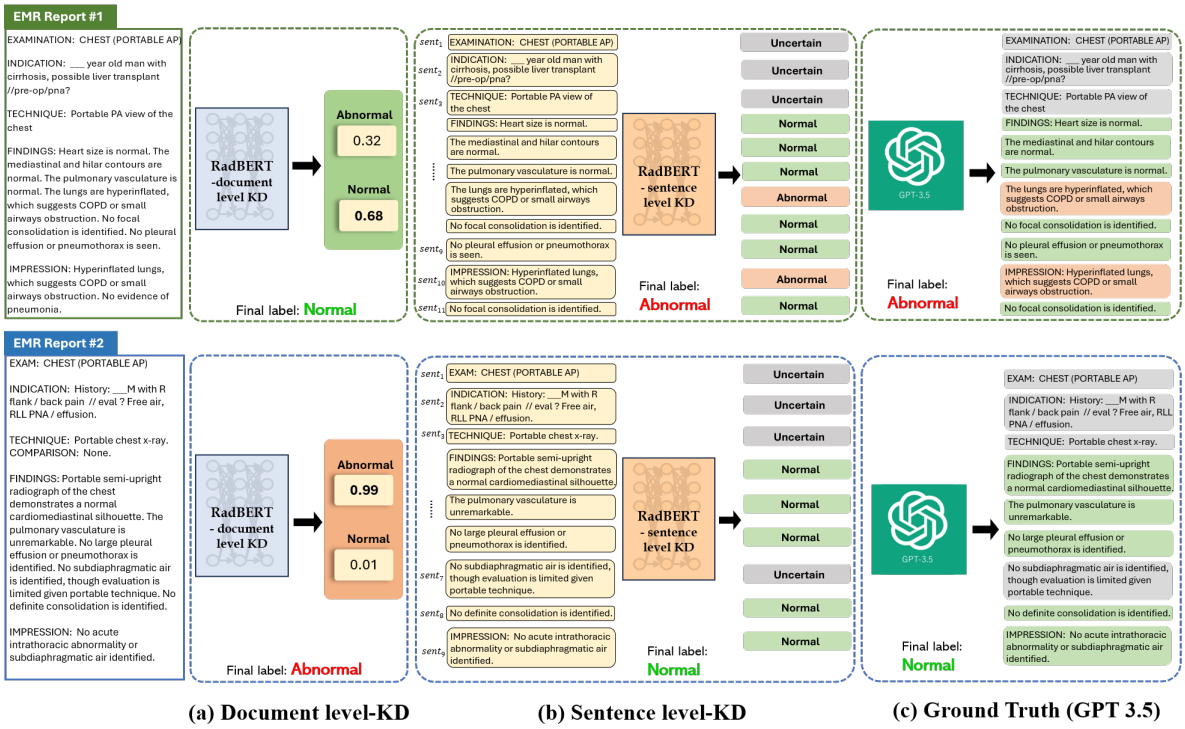


Figure 4: Sample cases: Document-level vs Sentence-level KD - Demonstrating instances where document-level KD fails and sentence-level KD succeeds in accurately predicting abnormal and normal medical reports

sentence-level knowledge distillation (S-KD) over the document-level KD (D-KD) (Sec. 4.2), along with the underlying reasons (Sec. 4.3). Additionally, we discuss the advantages of integrating a contrastive setup into KD training (Sec. 4.3) and the rationale behind this enhancement (Sec. 4.5). The setup details are given in Sec. 4.1.

4.1. Setup

We employed the MIMIC-CXR dataset (Johnson et al., 2019) for both our training and test datasets. For the training dataset, we used all of the p10 documents. For the test dataset, we only used the initial subset of the p11 documents. We employed GPT-3.5 to assign labels (normal n, abnormal a, and uncertain u) to each sentence in the datasets. Documents were labeled based on the presence of abnormal sentences; if any abnormal sentences were detected within a document, the document was classified as abnormal, otherwise, it was labeled as normal.

Using our high-confidence label selection method given in Appendix B, we extract documents and sentences with high confidence from the target dataset; 11,158 training documents (1,698 normal and 9,860 abnormal) and 2,832 testing documents (2,394 abnormal and 438 normal). These documents consist of 172,105 training sentences (51,568 normal, 64,715 abnormal, and 55,822 uncertain) and 40,779 testing sentences (12,105 normal, 15,655 abnormal, and 13,019 uncertain). Baseline D-KD training uses the document-level dataset, whereas our S-KD training employs the associated sentence-level dataset for a fair comparison. Other details are in Appendix A.

4.2. Performance Comparison between Proposed and Baseline KD Approach

In this section, we present a comparative analysis of the test performances of two knowledge distillation (KD) methods as outlined in Sec. 3: document-level KD (D-KD) and sentence-level KD (S-KD). These methodologies were applied to six pre-trained medical domain-specific BERT (Bidirectional Encoder Representations from Transformers) backbone models as for KD student models. The models assessed under KD training are RadBERT-Roberta (Yan et al., 2022), BioMed-Roberta (Gururangan et al., 2020),

Table 4: Anomaly detection performance comparison between with contrastive loss setup ($\lambda = 1$, CE + contrastive) and baseline loss setup ($\lambda = 0$, CE only)

Model	Accuracy	Specificity	Sensitivity	AUC
RadBERT-Roberta -4m-document **baseline loss**	85.17	0.832	**0.852**	0.846
RadBERT-Roberta -4m-document **contrastive loss**	**85.52** (+ 0.35)	**0.858** (+0.026)	0.840 (-0.012)	**0.901** (+0.055)
RadBERT-Roberta -4m-sentence **baseline loss**	91.53	0.910	0.936	0.962
RadBERT-Roberta -4m-sentence **contrastive loss**	**95.06** (+3.53)	**0.941** (+0.031)	**0.952** (+0.016)	**0.977** (+0.015)
BioMed-Roberta-document **baseline loss**	82.17	0.814	**0.861**	**0.889**
BioMed-Roberta-document **contrastive loss**	**86.12** (+3.95)	**0.869** (+0.055)	0.820 (-0.041)	0.877 (-0.012)
BioMed-Roberta-sentence **baseline loss**	94.42	0.910	**0.961**	0.976
BioMed-Roberta-sentence **contrastive loss**	**94.60** (+0.18)	**0.947** (+0.037)	0.943 (-0.018)	**0.979** (+ 0.003)
p-value	0.021	0.021	0.282	0.282

BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019), BiomedBERT (Gu et al., 2020), and BlueBERT (Peng et al., 2019).

During training phase, we utilized the contrastive loss setup with $\lambda = 1$ in Eq. (5) to specifically investigate the effect of S-KD compared to D-KD. Our results, depicted in Table 2, reveal the statistical significance of S-KD's performance superiority over the baseline D-KD across all evaluated models. This was corroborated by the Wilcoxon rank-sum test, indicating p-values less than 0.05 for various evaluation metrics. S-KD consistently showed improved performance in terms of accuracy, specificity, sensitivity, and AUC, with average enhancements of 1.047 times, 1.053 times, 1.053 times, and 1.051 times, respectively. These improvements suggest that S-KD is more effective in identifying anomalies in radiology reports, potentially leading to a lower rate of both false negatives (missed anomalies) and false positives (incorrectly identified anomalies) compared to D-KD. It is noteworthy that metrics other than AUC were measured at the optimal threshold on the AUC curve.

Particularly noteworthy is the performance of the recently introduced RadBERT model. In our evaluations, the anomaly detection accuracy of the S-KD method on RadBERT was recorded at 95.06%. This represents a substantial reduction in error rate, approximately threefold (i.e, S-KD accuracy: 95.06%, S-KD error rate: 4.94%, D-KD accuracy: 85.52%, D-KD error rate: 14.48%) compared to the D-KD method, which achieved an accuracy of 85.52%. These results underscore the superiority and efficacy of the S-KD technique in this context.

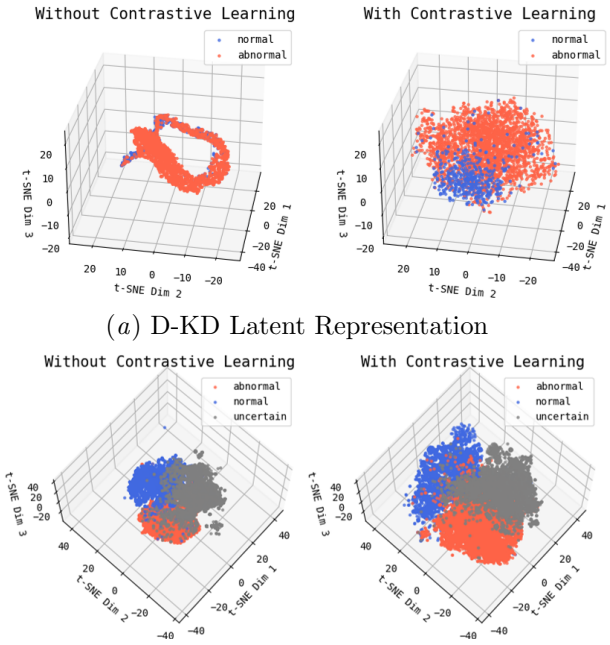


Figure 5: Comparison of latent vector distribution for each class depending on whether the contrastive setup is used ($\lambda = 1$) or not ($\lambda = 0$)

Table 5: Comparison of error distance depending on whether contrastive loss is used ($\lambda = 1$) or not ($\lambda = 0$)

KD Method With vs Without Cont.		**D-KD** Without Cont.	**D-KD** With Cont.	**S-KD** (ours) Without Cont.	**S-KD** (ours) With Cont.
Error Distance (Int./Ext. Dist.)	Normal	1.39	**0.55**	0.75	**0.67**
	Abnormal	1.30	**0.70**	0.66	**0.63**
	Uncertain	—	—	**0.68**	0.75

4.3. Analysis of Potential Cause for S-KD Advancement

Sec. 4.2 demonstrates that sentence-level knowledge distillation (S-KD) surpasses document-level knowl-

edge distillation (D-KD) in performance. This section delves into the underlying reasons for this improved performance and presents the following key findings: S-KD exhibits superior capabilities over D-KD in two critical aspects. Firstly, S-KD more accurately identifies documents as abnormal when they contain only a low presence of abnormal sentences, which is a more challenging scenario for anomaly detection. Secondly, S-KD effectively corrects misclassifications where documents are incorrectly identified as abnormal due to a high number of uncertain sentences, despite being normal as truth.

To arrive at these conclusions, we analyzed the distribution of normal, abnormal, and uncertain sentences in medical reports. This comprehensive analysis encompassed the document cases of the entire test dataset, incorrect classifications by D-KD, incorrect classifications by S-KD, and cases where D-KD failed but S-KD succeeded, as illustrated in Fig. 3.

In the document group where D-KD incorrectly classified cases from abnormal truth (Fig. 3(a)), the distribution of abnormal sentences was approximately 29.14%, compared to an average of 44.85% in all abnormal truth documents. This indicates D-KD's struggle with sparsely abnormal sentences. In contrast, S-KD's incorrect cases showed a 21.28% distribution of abnormal sentences, indicating higher accuracy with even fewer abnormal sentences. This suggests S-KD's enhanced capability in detecting sparser abnormal sentences by approximately 27%. Notably, in cases where D-KD was incorrect but S-KD was correct, the document count was 282, signifying that S-KD correctly addressed 82.94% of these D-KD incorrect cases (282 out of 340).

Regarding normal truth document cases (Fig. 3(b)), documents incorrectly classified by D-KD had a 38.14% distribution of uncertain sentences, higher than the test dataset's average of 36.39%. This suggests that D-KD tends to be incorrect for documents where the number of uncertain sentences is large. In contrast, documents incorrectly classified by S-KD showed a 41.21% distribution of uncertain sentences, effectively managing correct classifications with about 8.05% more uncertain sentences. This emphasizes the impact of our consideration in training for sentence-level uncertainty as an additional/explicit label in S-KD. In particular, in the intersection of D-KD's incorrect and S-KD's correct cases, the document count was 60, indicating that S-KD correctly resolved 85.71% of these D-KD incorrect document cases (60 out of 70).

Fig. 4 presents examples that illustrate the two primary reasons for the enhanced performance of S-KD. In the first document, a relatively small proportion, approximately 18.18% (i.e., 2 out of 11), consists of abnormal sentences. Under the D-KD approach, accurately detecting these sparse abnormal sentences poses a challenge, leading to incorrect classifications. In contrast, the S-KD approach effectively identifies these sparse abnormal sentences, resulting in correct classification. In the second report, uncertain sentences constitute 40% (i.e., 4 out of 10) of the document. Specifically, sentences 2 and 7 contribute to the ambiguity regarding abnormality within the D-KD framework, culminating in an erroneous classification. However, S-KD successfully navigates this ambiguity, accurately classifying the report even amidst a substantial presence of uncertain sentences.

4.4. Ablation Study for Contrastive Loss in KD Learning

In our study, we formulated a KD training objective that incorporates a contrastive learning objective, $\mathcal{L}^{cont}$, in addition to the conventional cross-entropy loss, $\mathcal{L}^{cross}$, as detailed in Eq. (5). Specifically, for our baseline model, which excludes contrastive learning, we set the parameter λ to 0. Conversely, in our enhanced KD training scheme that includes contrastive learning, we assigned a value of 1 to λ in Eq. (5).

Our ablation study, presented in Table 4, examines the effect of adding the contrastive learning objective. This study was performed using the two highest-performing backbone models for AD as identified in Table 2: RadBERT-Roberta and BioMED-Roberta. The results, as depicted in Table 4, demonstrate that incorporating the contrastive loss setup ($\lambda = 1$) significantly improves accuracy and specificity in AD testing for all student models, compared to the baseline loss setup ($\lambda = 0$) which does not include contrastive learning. However, the baseline loss setup exhibits a higher specificity, a discrepancy we attribute to the training data distribution. As discussed in Sec. 4.1, the training data contains a higher proportion of abnormal sentences and documents compared to normal instances, leading to a baseline bias towards abnormal labels. This results in elevated sensitivity but reduced specificity. The introduction of the contrastive loss setup helps to counteract this bias, enhancing specificity and thereby improving overall

accuracy relative to the baseline. Consequently, we applied contrastive learning in all experiments except this ablation study.

4.5. Analysis of Potential Cause for Contrastive Setting Advancement

In the previous section, our investigation centered on the enhanced capacity of our network to accurately identify the normal class, a minor category, through the application of contrastive learning. This approach resulted in a significant improvement in specificity, suggesting that contrastive learning contributes to more precise clustering of feature vectors within the same class. The current section aims to provide both qualitative and quantitative evidence to further substantiate the performance improvements attributable to the use of contrastive learning.

For visual validation, we extracted latent feature vectors from the point immediately preceding the linear layers in our trained network for each sample in the test dataset. These vectors were then visualized using t-SNE, as shown in Fig. 5. The resultant visualization indicates a more pronounced demarcation between class clusters when contrastive learning is employed. Specifically, in the D-KD scenario, we observed a distinct separation between the blue and red class clusters. In the S-KD context, there was a marked decrease in the instances of gray class samples overlapping with the red class region. These observations underscore the efficacy of contrastive learning in enhancing class discriminability in our network.

We defined c as the centroid for the samples of each class. The intra distance for a sample was calculated as the ℓ_2 distance from its class centroid c, while the extra distance was determined as the average distance from c to the centroids of other classes. We then computed an error distance for each sample, defined as the ratio of intra to extra distance, to gauge the proximity of a sample to its true class, with lower values indicating a closer alignment. The outcomes of these error distance calculations are in Table 5; the application of contrastive learning significantly reduces the mean error distance for test samples in both normal and abnormal classes in both D-KD and S-KD scenarios. This reduction in error distance markedly enhances the network's capability to detect even minor (i.e., normal) class, thereby improving the specificity and accuracy of the model's AD performance.

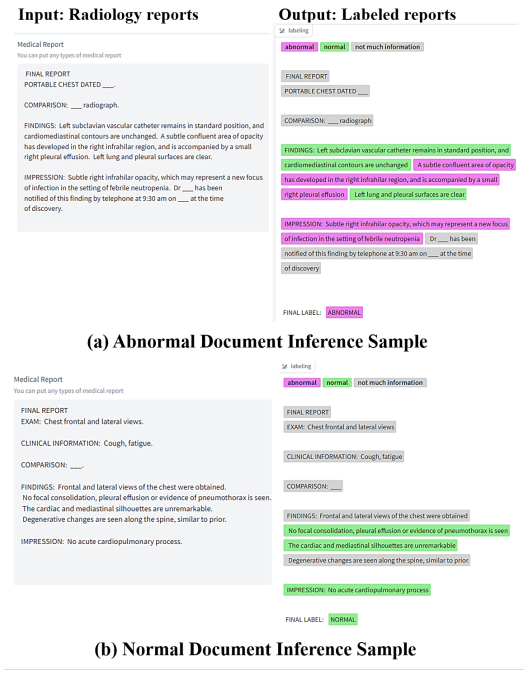


Figure 6: Sample results of our model deployed on HuggingFace - Demonstrating practical examples using color coding for enhanced interpretability

5. Discussion

In this section, we delineate the methodologies employed for extracting high-confidence labels from ChatGPT (Sec. 5.1), detail the process of model deployment along with its potential utility (Sec. 5.2), and discuss the limitations of our approach (Sec. 5.3).

5.1. Method to Extract High-Confidence Label from GPT 3.5

To ascertain high-confidence labels from GPT-3.5, our approach involves an ensemble methodology. We execute three independent label extractions (normal, abnormal, uncertain) for each sentence from GPT-3.5. If these labels exhibit consistency across extractions, we accept them; otherwise, we reject them. Comprehensive details are presented in Appendix B.

5.2. Model Deployment and Implication

Our RadBERT-Roberta model, which was trained using sentence-level knowledge distillation (S-KD) with

a contrastive setup, has been deployed and is accessible via HuggingFace.

Fig. 6 illustrates the model's functionality. When a clinician uploads a radiology report, the model highlights sentences indicative of normal and abnormal findings in green and purple, respectively, while sentences deemed uncertain are marked in gray. This feature enables radiologists or doctors to review reports more efficiently by focusing primarily on the text highlighted in green and purple, thereby potentially omitting the gray-marked uncertain content.

5.3. Limitation

First, our methodology was validated using the MIMIC-CXR dataset, a renowned public source, with distinct data separation for training. However, it lacks supplementary verification with varied public datasets. Future endeavors will expand this validation to encompass a broad spectrum of radiology reports.

Second, our study aims to address how to replicate the cloud-based model (i.e., ChatGPT-3.5) predictions in the non-cloud-based (i.e., secure) model. Therefore, GPT-3.5 was utilized for labeling radiologist reports without any human annotation load. However, our approach reveals a ground truth limitation: the absence of radiologist-confirmed ground truth in our process. Future work may focus on integrating radiologist-verified ground truth to enhance the accuracy of GPT-3.5's predictions. Nevertheless, our research demonstrates, for the first time, the potential of replicating ChatGPT within a secure model for radiologist report analysis (without any human manual annotation) and introduces advanced KD strategies tailored for this aim.

Lastly, our approach assumes the use of non-sensitive data (i.e., data without security issues) for training, as the training data still requires uploading to GPT-3.5. In practical applications, especially in hospital settings, this necessitates de-identification protocols to ensure data privacy. Despite this limitation, our method removes the need for human manual annotation (e.g., an indication of abnormal status per sentence), suggesting a promising direction for efficiently replicating cloud-based models like GPT-3.5 under in-hospital and secure environments. We also expect that the costs and efforts associated with de-identification are generally less than those required for manual annotation processes, thereby supporting the usefulness of our approach. Furthermore, our study is significant in that it utilizes only a limited portion of data as training material. This implies that all other unrestricted datasets can be used as evaluation data without security concerns, as test data do not need to be uploaded to GPT-3.5 but can be processed by our secure model. This capability presents a vital step forward in the practical application of AI in medical settings, offering a state-of-the-art (i.e., reproducing modern performance of cloud models like ChatGPT), efficient (i.e., without human annotation labor), and secure (i.e., implemented by non-cloud model) method for enhancing medical record processing.

6. Conclusion

This paper presents a novel approach to replicating cloud model like ChatGPT as non-cloud one for secure usage in radiology report processing at hospitals, eliminating the need for human annotation. Our method involves a unique knowledge distillation process from ChatGPT, ensuring data remains on-site while maintaining comparable performance. The effectiveness of this approach is demonstrated through anomaly detection in radiology reports, highlighting our model's ability in sentence-level knowledge distillation and explicit management of uncertainty (e.g., 95.06% accuracy achieved on the RadBERT using S-KD with contrastive setup). We also expect our approach's principle could extend to other report processing tasks such as question-answering tasks (e.g., for detection of individual disease), where our model's variant could adeptly identify and filter out low-confidence sentences in relation to the question. We expect that our research sets a precedent for developing secure and in-hospital LLM AI systems with minimal human supervision. By focusing on developing robust de-identification techniques utilized in the training procedure, we can further enhance the privacy aspects of our method. Ultimately, our study potentially heralds a new era in healthcare technology applications.

References

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, William L. Hamilton, and Jimmy Lin. Exploring the limits of simple learners in knowledge distillation for document classification with DocBERT. *Workshop on Representation Learning for NLP (RepL4NLP)*, pages 72–77, jul 2020.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, page 72–78, 2019.

Keno K Bressem, Lisa C Adams, Robert A Gaudin, Daniel Tröltzsch, Bernd Hamm, Marcus R Makowski, Chan-Yong Schüle, Janis L Vahldiek, and Stefan M Niehues. Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports. *Bioinformatics*, 36(21):5255–5261, 2020.

Lawrence O Gostin, Laura A Levit, Sharyl J Nass, et al. Beyond the HIPAA privacy rule: Enhancing privacy, improving health through research. 2009.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint:2007.15779*, 2020.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 8342–8360, 2020.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint:1503.02531*, 2015.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(160035), 2016. doi: https://doi.org/10.1038/sdata.2016.35.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(317), 2019.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *34th Conference on Neural Information Processing Systems (NeurIPS2020), Vancouver, Canada*, 2020.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234–1240, 2020.

Peter Lee, Sebastien Bubeck, and Joseph Petro. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New England Journal of Medicine*, 388(13): 1233–1239, 2023.

Hanzhou Li, John T Moon, Deepak Iyer, Patricia Balthazar, Elizabeth A Krupinski, Zachary L Bercu, Janice M Newsome, Imon Banerjee, Judy W Gichoya, and Hari M Trivedi. Decoding radiology reports: Potential application of OpenAI ChatGPT to enhance patient understanding of diagnostic reports. *Clinical Imaging*, 2023.

Zheng Liu, Yiwei Li, Peng Shu, Aoxiao Zhong, Longtao Yang, Chao Ju, Zihao Wu, Chong-Yi Ma, Jie Luo, Cheng Chen, Sekeun Kim, Jiang Hu, Haixing Dai, Lin Zhao, Dajiang Zhu, Jun Liu, W. Liu, Dinggang Shen, Tianming Liu, Quanzheng Li, and Xiang Li. Radiology-Llama2: Best-in-class large language model for radiology. *ArXiv*, abs/2309.06419, 2023a.

Zhengliang Liu, Aoxiao Zhong, Yiwei Li, Longtao Yang, Chao Ju, Zihao Wu, Chong Ma, Peng Shu, Cheng Chen, Sekeun Kim, et al. Radiology-GPT: A large language model for radiology. *arXiv preprint:2306.08666*, 2023b.

Zhengliang Liu, Tianyang Zhong, Yiwei Li, Yutong Zhang, Yi Pan, Zihao Zhao, Peixin Dong, Chao Cao, Yuxiao Liu, Peng Shu, et al. Evaluating large language models for radiology natural language processing. *arXiv preprint:2307.13693*, 2023c.

Chong Ma, Zihao Wu, Jiaqi Wang, Shaochen Xu, Yaonai Wei, Zhengliang Liu, Lei Guo, Xiaoyan Cai, Shu Zhang, Tuo Zhang, et al. ImpressionGPT: An iterative optimizing framework for radiology report summarization with chatGPT. *arXiv preprint:2304.08448*, 2023.

Fatima N Mirza, Oliver Y Tang, Ian D Connolly, Hael A Abdulrazeq, Rachel K Lim, G Dean Roye, Cedric Priebe, Cheryl Chandler, Tiffany J Libby, Michael W Groff, et al. Using ChatGPT to facilitate truly informed medical consent. *NEJM AI*, page AIcs2300145, 2024.

Pritam Mukherjee, Benjamin Hou, Ricardo B Lanfredi, and Ronald M Summers. Feasibility of using the

privacy-preserving large language model Vicuna for labeling radiology reports. *Radiology*, 309(1):e231147, 2023.

Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, pages 58–65, 2019.

Nihar M. Ranjan and Rajesh S. Prasad. A brief survey of text document classification algorithms and processes. *Journal of Data Mining and Management*, 8(1), 2023.

Maksut Senbekov, Timur Saliev, Zhanar Bukeyeva, Aigul Almabayeva, Marina Zhanaliyeva, Nazym Aitenova, Yerzhan Toishibekov, and Ildar Fakhradiyev. The recent progress and applications of digital technologies in healthcare: A review. *International journal of telemedicine and applications*, 2020, 2020.

Dave Van Veen, Cara Van Uden, Maayane Attias, Anuj Pareek, Christian Bluethgen, Malgorzata Polacin, Wah Chiu, Jean-Benoit Delbrouck, Juan Manuel Zambrano Chaves, Curtis P Langlotz, et al. RadAdapt: Radiology report summarization via lightweight domain adaptation of large language models. *arXiv preprint:2305.01146*, 2023.

Paul Voigt and Axel Von dem Bussche. The EU general data protection regulation (GDPR). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10 (3152676):10–5555, 2017.

Sean Wu, Michael Koo, Lesley Blum, Andy Black, Liyo Kao, Zhe Fei, Fabien Scalzo, and Ira Kurtz. Benchmarking open-source large language models, GPT-4 and Claude 2 on multiple-choice questions in nephrology. *NEJM AI*, page AIdbp2300092, 2024.

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. RadBERT: Adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4): e210258, 2022.

Liang Yao, Chengsheng Mao, and Yuan Luo. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Medical Informatics and Decision Making*, 79, apr 2019.

Tianyang Zhong, Wei Zhao, Yutong Zhang, Yi Pan, Peixin Dong, Zuowei Jiang, Xiaoyan Kui, Youlan Shang, Li Yang, Yaonai Wei, et al. Chatradio-valuer: A chat large language model for generalizable radiology report generation based on multi-institution and multi-system data. *arXiv preprint:2310.05242*, 2023.

Appendix A. Data Setup Details

In our study using the MIMIC-CXR dataset (Johnson et al., 2019), we initially utilized a training set of 22,195 p10 documents and a test set of 5,217 p11 documents. Post our high-confidence label filtering (Sec. B), the training set, originally with 188,827 sentences, was reduced to 172,105, and the test set from 44,516 to 40,779 sentences. Only documents with all sentences classified as high-confidence were retained, resulting in 11,158 training (from 22,195) and 2,832 test (from 5,217) documents as the final dataset in our study.

The training set comprised 1,698 normal and 9,460 abnormal documents (i.e., total 11,158 documents), with 51,568 normal, 64,715 abnormal, and 55,822 uncertain sentence (i.e., total 172,105 sentence). The test set included 2,394 abnormal and 438 normal documents (i.e., total 2,832 documents), with 12,105 normal, 15,655 abnormal, and 13,019 uncertain sentences (i.e., total 40,779 sentence). This dataset enabled a robust comparison of our S-KD approach against the baseline D-KD.

Appendix B. High-Confidence Label Extraction from GPT-3.5

Our research introduces a method for extracting high-confidence labels from GPT-3.5, as outlined in Fig. 7. We obtain independent sentence labels from GPT-3.5 three times to ensure label consistency. Labels are considered high-confidence when all three extractions match. Discrepancies lead to label dismissal unless a majority (two out of three) consistency is observed. In such cases, we calculate the cosine similarity between the input text and the GPT explanation for each label, followed by averaging the confidence scores of the consistent labels. A label is accepted if its average confidence score is higher than the score of the minority label. This method, crucial for extracting reliable labels from ChatGPT, serves as training data for knowledge distillation (KD), ensuring both label consistency and confidence based on GPT explanations.

Appendix C. Discussion

C.1. Validation of Consistency between GPT Results and Ground Truth Labels

The primary objective of this research is to investigate whether local models can successfully reproduce the results of the GPT model. Given the extensive validation of GPT models' high performance in numerous publications, experimental results demonstrating the validity of GPT model outputs as ground truth were not initially included in the main paper. However, to provide further justification for employing GPT for labeling purposes, an additional experiment was conducted to validate the consistency between the GPT ensemble labeling approach and ground truth labels.

Table 6: Accuracy comparison of labeling methods between GPT solo and GPT ensemble (our labeling method) using subset ground truth

Accuracy (%)	abnormal	normal	uncertain
GPT solo -trial 1	72.5% (226/312)	66.5% (214/322)	71.3% (261/366)
GPT solo -trial 2	72.5% (226/312)	67.3% (217/322)	71.9% (263/366)
GPT solo -trial 3	67.3% (210/312)	67.7% (218/322)	70.2% (257/366)
GPT ensemble (our method)	**100%** (204/204*)	**100%** (202/202*)	**99.2%** (253/255*)
number of sentences (Total 1000)	312	322	366

* The numbers 204, 202, and 253 refer to the subsets extracted by our GPT ensemble from each group of 312, 322, and 366 sentences per class, respectively. The extracted sentences are considered as truth, while the unextracted sentences are discarded.

Instead of relying solely on raw GPT model results, an ensemble technique was employed that considered GPT model outputs only when the results from three different GPT models were in agreement (refer to Appendix B). In this additional experiment, ground truth labels for the first 1000 sentences of the test data were collected through human annotation. Using these ground truth labels as a reference, the experiment aimed to verify whether the GPT model results obtained using the ensemble technique (Appendix B) exhibit higher consistency with the ground truth compared to the case where the ensemble technique is not employed. Furthermore, the absolute agreement between the ensemble-based GPT model results and the ground truth was determined, allowing for an assessment of the reliability of GPT model outputs as truth labels.

The results of this additional experiment are presented in Table 6. This result in this table validate the consistency between the GPT ensemble labeling approach and ground truth labels.

1) Dataset: We corrected truth labels for the first 1000 labeled data sentences from our test dataset, annotated by radiologists as abnormal (312), normal (322), and uncertain (366).

2) Methods: Using the truth labels as reference, we compare the accuracy of our GPT ensemble result (detailed in Appendix B) and GPT individual result. GPT solo presents the direct label predictions from GPT-3.5. GPT ensemble (our method) is leveraging GPT-3.5 to generate independent labels for each sentence three times. Labels are only accepted if: (1) All three predictions are identical. (2) In the absence of perfect agreement, a majority (2 out of 3) have higher cosine similarity score between the text embedding vectors of GPT explanation and the input text compared to minority (1 out of 3) similarity score.

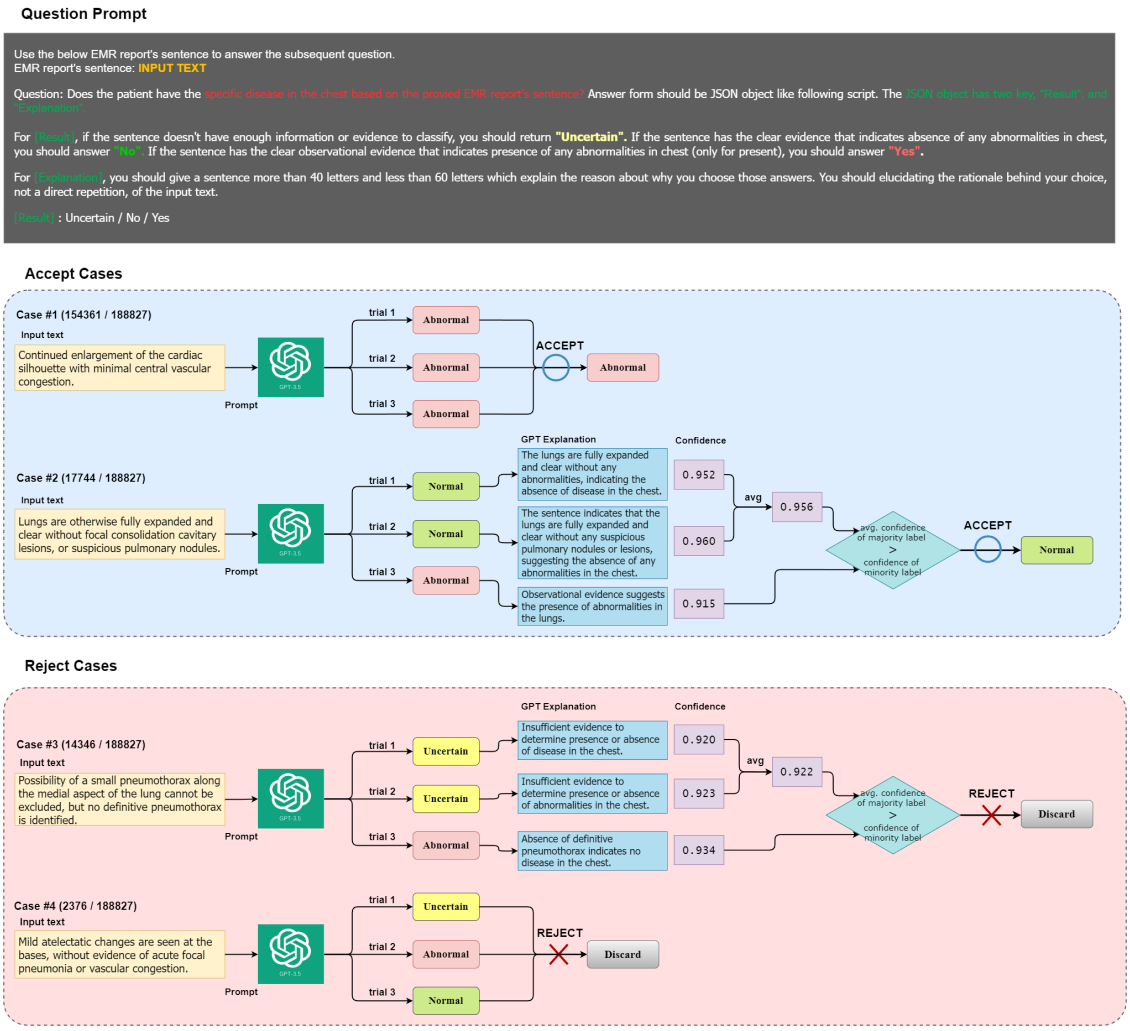


Figure 7: Our approach employs two key techniques to derive high-confidence sentence labels from GPT-3.5: (1) the prompt engineering that obliges the network to elucidate the rationale behind its outputs, and (2) the prediction ensemble methodology to extract the result ensuring all models concur in providing identical reasoning.

3) Findings: Our GPT ensemble labeling method achieved higher than 99% accuracy on the ground truth by eliminating sentences with low confidence scores (further details in the attached table). In contrast, GPT solo exhibited lower accuracy: approximately 70.8%, 67.2%, and 71.1% for abnormal, normal, and uncertain labels, respectively.

4) Implication: This experiment demonstrates that our GPT ensemble approach significantly improves accuracy compared to using a single GPT prediction, allowing us to obtain truth labels for each sentence with higher than 99% accuracy. Although the ground-truth subset may not completely generalize to the entire dataset due to the limited labeled sample size (e.g., 1000 sentences), this experiment validates the consistency between our ensemble-based GPT-3.5 labeling and ground truth labels.

C.2. Details for Data Privacy Limitations

To provide a detailed explanation of the data privacy aspects mentioned in the Limitations section, we have included Table 7 in this section, which outlines the advantages and limitations of our research from a data privacy perspective. The table is accompanied by a comprehensive description of these points.

1) Advantages on training phase

1. Eliminate the need for costly human (doctor) annotations: Our method eliminates the requirement for expensive doctor annotations.

Table 7: Advantages and limitations of our study for knowledge distillation of cloud AI model (e.g., GPT) into local AI model in hospital settings

	Training	Inference
Advantages	• No human annotation cost • Leverage knowledge from high-performance cloud models	• Preserve data privacy • Achieve comparable classification performance to the cloud model • No preprocessing or authorization required for inference • No pay-per-use charges for utilizing the cloud model
Limitations	• Potential data privacy concerns • Require additional data preprocessing required (e.g., de-identification) • Authorization needed for data upload	N/A

2. Leverage knowledge from high-performance cloud models: The training process can leverage any high-performance cloud model, such as OpenAI GPT-4 or Google DeepMind Gemini-1.5.

2) Limitations on training phase

1. Raise potential data privacy concerns: Uploading medical data to cloud models raises potential data privacy concerns.

2. Necessitate additional data preprocessing: Additional data preprocessing steps, such as removing direct identifiers (patient names, ID numbers), are required before uploading data due to health care regulations (e.g., HIPPA) and data privacy regulation (e.g., GDPR).

3. Require authorization for data upload: Authorization from a qualified expert is recommended to assess the risk of re-identification when uploading de-identified medical report data to cloud models.

3) Advantages on inference phase

1. Ensure data privacy: Inferences performed by the trained on-premise student model eliminate the need for further data upload to cloud models, thereby safeguarding patient data privacy.

2. Achieve performance comparable to the cloud model: The proposed sentence-level knowledge distillation method with cross-entropy and supervised contrastive loss achieves an accuracy of 95% using the radBERT-Roberta-4m student model, demonstrating performance comparable to the cloud model.

3. Require no preprocessing or authorization for model inference: Inference with the on-premise model avoids the need for additional data preprocessing steps like de-identification mandated by healthcare regulations and eliminates the requirement for expert authorization before upload to a cloud model.

Table 8: Pretraining Information for Student Models

Student Model	Initial Weight	Pre-trained Data
RadBERT -Roberta-4m	BioMed-RoBERTa	4.43M radiology reports
BioMED-Roberta	RoBERTa -base	2.68M Semantic Schocal scientific papers
BlueBERT	BERT-base	PubMed texts (4B words)
Clinical BERT	BioBERT	All notes from MIMIC-III
BiomedBERT	BERT-large	14M PubMed abstracts
BioBERT	BERT-base	4.5B PubMed abstracts 13.5B Words of PubMed Central full-text articles

• MIMIC-III is a wide health-related dataset containing information on over 40,000 patients admitted to critical care units at Beth Israel Deaconess Medical Center between 2001 and 2012 (Johnson et al., 2016).

• MIMIC-CXR is a radiology electronic health record dataset collected in the emergency department of BethIsrael Deaconess Medical Center between 2011 and 2016

4. Incur no pay-per-use charges for utilizing the cloud model: Unlike cloud models with pay-per-use charges based on token utilization (e.g., OpenAI GPT API), the on-premise model incurs no additional costs.

C.3. Pretraining information for Student Models

Table 8 explicitly presents the initial parameters and the pretraining data used for each student model in this study. All models do not utilize MIMIC-CXR for their initial parameters, we believe that the evaluation performance of our study for these models is free from related bias issues. RadBERT (i.e., RadBERT-Roberta-4m) was selected as the representative model for this study due to its recent introduction, the fact that it does not utilize MIMIC-CXR for obtaining its initial parameters, and its ability to leverage a wide range of internal radiology report data.

C.4. Performance Comparison Between Conventional Annealing-Based Soft KD Training and Our KD Training Method

This section compares our proposed KD training method with a conventional approach that utilizes annealing-based relaxation of cross-entropy for KD (Hinton et al., 2015). The loss term introduced in the conventional approach (Hinton et al., 2015) considers the uncertainty of the probability values of the teacher and student models during KD training using a temperature value based on an annealing technique. This loss term, defined as the Soft-KD loss (L^{soft}), is a soft relaxation version of the existing cross-entropy loss term for KD, taking into account the temperature value.

To investigate the effectiveness of incorporating L^{soft} into our proposed KD training loss, we define a new loss term as follows:

$$\mathcal{L}_\theta^{new}(g_\theta(x), y) := (1-\alpha)\cdot\mathcal{L}_\theta^{cross}(g_\theta(x), y) + \alpha\cdot L_\theta^{soft}(g_\theta(x), y) + \lambda\cdot L_\theta^{cont}(g_\theta(x), y) \quad (6)$$

Table 9: Anomaly detection performance comparison between different KD training methods (RadBERT-Roberta-4m model commonly used)

KD Training Method (Document level-KD)	Accuracy	Specificity	Sensitivity
CE only (i.e., $(\alpha, T, \lambda) = (1, 0, 0)$)	85.17	0.832	0.852
CE + Soft-KD (i.e., $(\alpha, T, \lambda) = (0.5, 1, 0)$)	82.13 (-3.04)	0.817	0.854
CE + Soft-KD (i.e., $(\alpha, T, \lambda) = (0.5, 10, 0)$)	80.33 (-4.84)	0.787	0.913
CE + Soft-KD (i.e., $(\alpha, T, \lambda) = (0.5, 100, 0)$)	79.94 (-5.23)	0.785	0.900
Ours: CE + Contrastive (i.e., $(\alpha, T, \lambda) = (1, 0, 1)$)	**85.52** (+0.35)	0.858	0.840

KD Training Method (Sentence level-KD)	Accuracy	Specificity	Sensitivity
CE only (i.e., $(\alpha, T, \lambda) = (1, 0, 0)$)	91.53	0.910	0.936
CE + Soft-KD (i.e., $(\alpha, T, \lambda) = (0.5, 1, 0)$)	93.75 (+2.22)	0.952	0.890
CE + Soft-KD (i.e., $(\alpha, T, \lambda) = (0.5, 10, 0)$)	91.70 (+0.17)	0.906	0.973
CE + Soft-KD (i.e., $(\alpha, T, \lambda) = (0.5, 100, 0)$)	92.09 (+0.56)	0.917	0.947
Ours: CE + Contrastive (i.e., $(\alpha, T, \lambda) = (1, 0, 1)$)	**95.06 (+3.53)**	0.941	0.952

where $\mathcal{L}_\theta^{soft}(g_\theta(x), y) := T^2 \cdot \mathrm{KL}\Big(sm_{log}\Big(\frac{g_\theta(x)}{T}\Big), p_y\Big)$, $sm_{log}(\cdot)$ is the log softmax function, and $\mathrm{KL}(\cdot)$ is the Kullback Leibler (KL)-divergence measurement. Compared to the loss function (Equation (5)) proposed in this study, a second term (L^{soft}) has been added.

To evaluate the performance impact of the second term, we conducted additional experiments using the RadBERT-Roberta-4m student model for both sentence-level and document-level KD tasks. The results are presented in Table 9, and the performance was evaluated using the same test dataset utilized in this study.

For the sentence-level KD task, the setup considering L^{soft} in addition to our originally proposed setup (cross-entropy and contrastive learning) exhibited lower anomaly detection accuracy on the evaluation data. Specifically, in document-level KD, exploiting the contrastive setting (ours) achieved 85.52% accuracy, while the best-performing accuracy using Soft-KD loss with different temperature values was 82.13%. Similarly, in sentence-level KD, using the contrastive setting achieved 95.06% accuracy compared to 93.75% for the best-performing accuracy using Soft-KD loss. Although using L^{soft} in addition to cross-entropy alone yielded better performance than using cross-entropy alone, its impact was less significant compared to the addition of contrastive learning.

These experimental results further demonstrate that the combination of the sentence-level KD task and contrastive learning proposed in this study achieves the highest performance improvement compared to related (soft) KD studies.

Table 10: Anomaly detection performance comparison results for training and test datasets

Model	Accuracy	Specificity	Sensitivity	AUC
Document-level KD: Baseline loss (CE only) **Train Dataset**	97.3	0.974	0.978	0.995
Document-level KD: Baseline loss (CE only) **Test Dataset**	85.17 (-12.13)	0.832	0.852	0.846
Document-level KD: Our loss (CE + Contrastive) **Train Dataset**	97.5	0.977	0.981	0.997
Document-level KD: Our loss (CE+Contrastive) **Test Dataset**	85.5 (-12.0)	0.858	0.840	0.901
Sentence-level KD (ours): Baseline loss (CE only) **Train Dataset**	98.0	0.979	0.989	0.998
Sentence-level KD (ours): Baseline loss (CE only) **Test Dataset**	91.53 (-6.47)	0.910	0.936	0.962
Sentence-level KD (ours): Our loss (CE+Contrastive) **Train Dataset**	**98.3**	0.991	0.984	0.997
Sentence-level KD (ours): Our loss (CE+Contrastive) **Test Dataset**	**95.1** (-3.2)	0.941	0.952	0.977

C.5. Performance Result on Training Dataset and Its Implication

In this section, we conducted additional measuring anomaly detection performance on our training dataset to analyze the training-test performance gap in our KD methods. The detailed results are presented in Table 10.

1) Experimental setting: We evaluated the training performance for each document-level and sentence-level KD training setup, using the same RadBERT-Roberta-4m student model.

2) Findings: Our analysis revealed a consistent pattern of a significant gap between training and test accuracy for document-level KD (e.g., without contrastive: 12.13% and with contrastive: 12.00%). Sentence-level KD exhibited a smaller gap (e.g., without contrastive: 6.47% and with contrastive: 3.2%).

These results suggest that document-level KD might struggle with generalization. Conversely, sentence-level KD demonstrates superior performance in generalizability (i.e., the minimum training-test performance gap of 3.2% is observed in the sentence-level KD setup with additional contrastive learning), potentially capturing more transferable knowledge from the teacher model. These findings further highlight the advantages of our research's key technical contribution: the combination of sentence-level KD and contrastive learning, which achieves the lowest training-test performance gap.