Exploring the Missing Medical Context in Generated Radiology Reports

Karan Bania[1,†], Harshvardhan Mestha[2,†] and Tanmay Tulsidas Verlekar[1,*,†]

[1] *Department of CSIS, BITS Pilani, K. K. Birla, Goa Campus, India*
[2] *Department of EE, BITS Pilani,K. K. Birla, Goa Campus, India*

Abstract

Recent advancements in multimodal LLMs have allowed its use in radiology, where, given an X-ray image, the report can be generated automatically. The state-of-the-art explores LLMs for this task through prompting or fine-tuning. The emphasis in such cases is to improve the syntax and context of the report with respect to natural language. This paper proposes that LLMS are unable to understand the medical context in images and in reports. Thus, the accuracy of the generated reports' medical context requires further evaluation. This paper uses a pre-trained GPT-4o, Qwen2-VL-7B, and a fine-tuned LLaMA-2 model to demonstrate that these LLMs can identify the input images as chest X-rays but are poor at identifying pathologies, even when fine-tuned. It then attempts to address this problem by proposing a pipeline that allows the LLMs access to a discriminative model that can classify pathologies present in a chest X-ray. The additional context of pathology labels allows the LLMs to generate more accurate reports. A second contribution of the paper involves the assessment of the generated reports. It illustrates that traditional metrics, such as the BERTScore, are ineffective in assessing generated medical reports. It then presents an LLM-as-a-judge that successfully compares generated and ground truth reports.

Keywords

Radiology Report, LLM-as-a-judge, Report Generation, LLM for health

1. Introduction

The field of medicine has traditionally employed discriminative models such as convolutional neural networks (CNN) for tasks such as pathology classification using X-ray images [1]. With the recent advancements in generative AI, Large Language Models (LLM) can now process text and images, understand complex instructions and generate detailed responses in natural language [2, 3, 4]. Thus, there is a growing interest in the research community in performing tasks such as medical report generation, medical visual question answering, and disease identification using LLMs [5, 6]. The field of medicine has been challenging for LLMs as they need to distinguish subtle differences in images or even parts of images to provide accurate text descriptions. A second issue with the LLMs is that the natural language on which it is trained can have a broad range of statements to communicate the same meaning. Medical language, on the other hand, is highly specific and comprehensive. Thus, the learned mappings between natural images and texts are ineffective in the medical domain, where, for instance, medical reports are generated using chest X-rays.

Recently, LLMs, such as Flamingo [7], have been explored to assess their ability to learn from a few examples in real-time for the task of medical visual question answering. The results indicated that while it returned a good BERT score, the exact match with the ground truth was extremely poor. The ability of a multimodal LLM, such as GPT-4o, to answer image-rich diagnostic radiology exam questions through prompting is assessed in [8]. The results suggest that there is a large variability in answers obtained from GPT-4o, which highlights its poor medical image interpretation abilities. It also demonstrated that

SLM4Health: Improving healthcare with Small Language Models, 26th June 2025
[*] Corresponding author.
[†] These authors contributed equally.
tanmayv@goa.bits-pilani.ac.in (T. T. Verlekar)
https://www.bits-pilani.ac.in/goa/tanmay-tulsidas-verlekar/ (T. T. Verlekar)
0009-0007-2496-4399 (K. Bania); 0009-0002-2373-7876 (H. Mestha); 0000-0001-5534-4351 (T. T. Verlekar)

GPT-4o is far more effective in answering text-only questions.

The XrayGPT [9] explored the route of fine-tuning an LLM on medical data and aligning it with a vision encoder. It allowed the GPT model to answer open-ended questions about chest X-ray images. However, the results indicated that it lacks the ability to identify specific pathologies. The work presented [6] also fine-tuned the LLaMA-2 language model on heterogeneous radiological images, encompassing X-rays, CT scans, and MRIs for tasks such as disease identification, medical visual question answering, and the generation of medical reports. It presented a grounding technique to allow the integration of spatial locations into the text fed into the language model. While the results appear promising, they were evaluated using the BERT score, which is ineffective in comparing the generated reports with the ground truth.

The current state-of-the-art models struggle to understand subtle differences in medical images, which appear homogeneous when compared with natural images. Thus, to generate reports that are accurate in terms of reporting pathologies that are present in chest X-rays, the paper proposes a two-step pipeline. To capture these subtle differences in the chest X-ray images associated with pathologies, the proposal uses a discriminative model called DenseNet 121, which performs the task of multi-label pathology classification. The classified pathology labels can then be communicated to the LLM through prompting, along with the chest X-ray. The ability of LLMs to generate natural language reports, coupled with the discriminative model's ability to classify pathologies accurately, leads to more accurate reports. It also demonstrates that the reports generated using state-of-the-art models, while being correct syntactically, are poor at communicating medical context. It discusses the drawbacks of BERT scores in capturing medical context. To address this issue, the paper presents the use of LLM as a judge, which evaluates the generated report in a structured manner while comparing it with the ground truth report.

2. Proposed pipeline

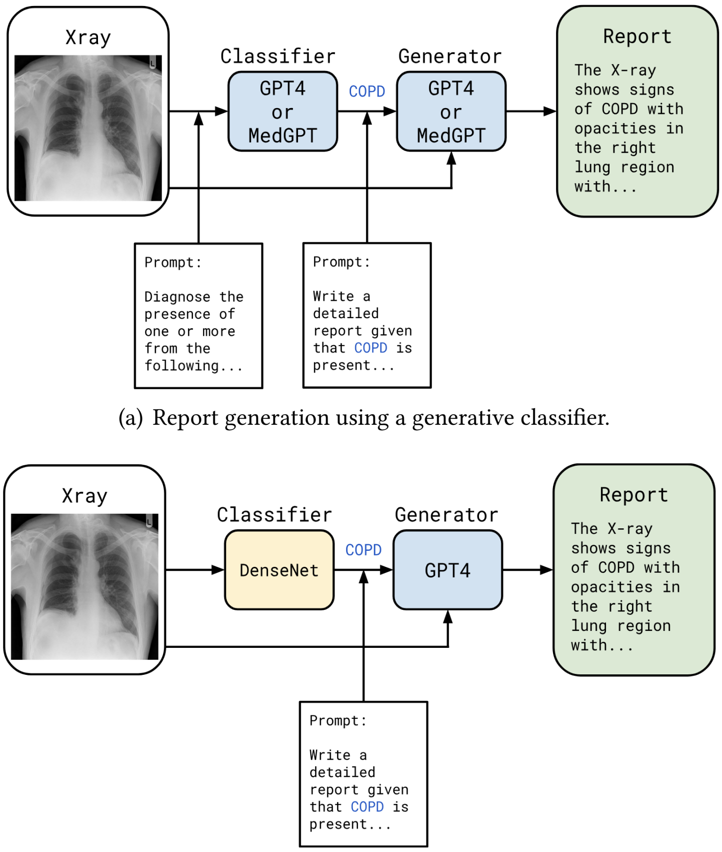
The default pipeline, where an LLM is directly prompted to generate a report for a given chest X-ray image. For the default pipeline, while the generated report appears natural, it misses a lot of medical context. It is observed that the missing context is usually about the pathologies that affect a small part of the whole image -see section 3. Since the LLMs cannot implicitly capture the context of the pathologies, the paper proposes a pipeline where the possible pathologies are explicitly communicated to the LLM as additional context, see Figure 1. It is done by allowing the LLM access to the output of a classifier that can successfully classify chest X-ray images according to the pathologies present in them. It leads to the generation of reports that can accurately communicate the correct medical diagnosis. To evaluate the pipeline, the paper considers three LLMs: the GPT-4o [10], Qwen2-VL [11] and the MiniGPT-Med [6] and a discriminative model: DenseNet 121 [12].

2.1. GPT-4o

The Generative Pre-Trained Transformer (GPT)-4o is a decoder-only transformer model trained to predict the next token in a document [10]. It is also multi-modal and can process text, images, and audio. It achieves this through a post-training alignment process. It can follow a variety of instructions, making it useful for classification, generation and logical analysis. To perform classification and report generation for the proposed pipeline, the default values are set for all parameters except `max_tokens`, which is set to 500. For the evaluation, to use GPT-4o as the LLM-as-a-judge, the `seed` value is set to 42 and the `temperature` is set to 0.

2.2. MiniGPT-Med

The MiniGPT-Med is composed of a visual backbone, a linear projection layer, and an LLM. The visual backbone is a specialized (Contrastive Language-Image Pretraining) CLIP model called EVA-CLIP [13] that is kept frozen during the training of MiniGPT-Med, while the LLM is finetuned.


Xray
Classifier
GPT4 or MedGPT
COPD
Generator
GPT4 or MedGPT
Report
The X-ray shows signs of COPD with opacities in the right lung region with...
Prompt: Diagnose the presence of one or more from the following...
Prompt: Write a detailed report given that COPD is present...
Xray
Classifier
DenseNet
COPD
Generator
GPT4
Report
The X-ray shows signs of COPD with opacities in the right lung region with...
Prompt: Write a detailed report given that COPD is present...

(a) Report generation using a generative classifier.

(b) Report generation using a discriminative classifier.

Figure 1: Architecture of the proposed pipeline.

It results in an association between visual concepts and text. The MiniGPT-Med uses Llama2-chat(7B) [14] for classifying pathologies and generating reports. It is trained on extensive medical knowledge and is fine-tuned for dialogue use cases. To perform classification, the mode is set to vqa with `temperature` set to 0.0 and a `top_p` set to 1.0. For report generation, the mode is set to `caption` with a `temperature` of 0.9 and a `top_p` of 0.9, as instructed in [13].

2.3. Qwen2-VL

The Qwen-VL-7B-Instruct [11] is an open-source LLM capable of image-text understanding and reasoning. The language component of Qwen-VL-7B consists of the Qwen-7B base model. The vision encoder of the model is a vision transformer. It transforms images into a variable number of visual tokens that the LLM can process. This allows Qwen to process images of arbitrary resolutions. While this paper uses the model's image processing capabilities, it can also handle videos through its multimodal rotary position embedding. To evaluate the ability of the Qwen-VL-7B-Instruct in classification pathologies and report generation, its `temperature` is set to 0.7, `top_p` to 0.8 and the `max_tokens` to 512.

DenseNet 121: DenseNet 121 is a CNN presented in [12] that performs classification of images across 1000 different categories. It consists of 121 layers organised into four dense blocks and separated by transition layers. Within a dense block, feature maps of a layer are concatenated with the feature maps of the preceding layer to maximise information flow. This paper considers a DenseNet 121 model that is pre-trained on the CheXpert [15] dataset. CheXpert is a large public dataset consisting of 224,316 chest X-ray images of 65,240 patients labelled across 13 different pathologies. Since the pathologies are different from the Radiopedia dataset [16] considered for evaluation in section **??**. The pre-trained DenseNet 121 model is then fine-tuned on the Radiopedia dataset to perform multilabel classification across 10 pathologies present in the Radiopedia dataset.

3. Evaluation

To evaluate the proposed pipeline and compare it with the state-of-the-art, the paper samples 200 X-ray images along with the corresponding reports from the Radiopedia dataset [16]. Each report in the dataset communicates the presence of one or more of the following pathologies: Atelectasis, Cardiomegaly, Calcifications, Pleural Effusion, COPD, Lung Nodules, Mesothelioma, Tuberculosis, Pneumothorax, Pneumonia, making this a multi-label classification problem. The subset is selected such that there is a minimal class imbalance, as illustrated in Figure 2.

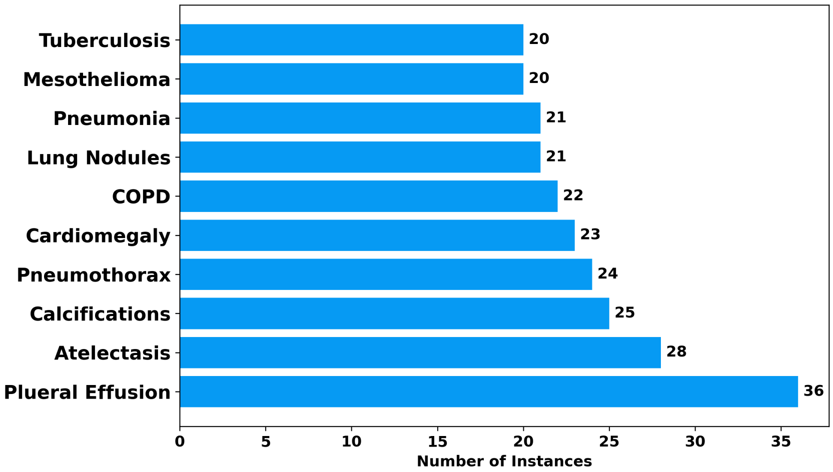


Figure 2: Distribution of classes in the curated dataset.

Algorithm 1 Pathology Classification and Report Generation

```
Input: $I$: Chest X-Ray image ,
$Pathologies$: List of possible pathologies ,
Classifier: Model for pathology classification,
Generator: Model for report generation.
Output: $Report$: The generated report
Let $List\_of\_classified\_pathologies := \phi$
for $Pathology$ in $Pathologies$ do
    if Classifier$(I, Pathology) = True$ then
        $List\_of\_classified\_pathologies := List\_of\_classified\_pathologies \cup Pathology$
    end if
end for
Report := Generator$(I, List\_of\_classified\_pathologies)$
return Report
```

For fine-tuning the DenseNet 121 model, a five-fold cross-validation is adopted over the 200 chest X-ray images. The model is fine-tuned using a learning rate of 0.0005 for 75 epochs, with a batch size of 8 using Adam optimizer and binary cross-entropy loss. Thus, each image is used for validation across the five folds. The experiments are performed on a single NVIDIA A100 GPU with 40GB of main memory.

3.1. Multi-label classification results

Both generative and discriminative models are evaluated for multi-label classification. When using the discriminative model, the Classifier outputs a set of labels representing pathologies. The generative

```
Prompt for classification:

Pretend you are a radiologist
Given is the hypothetical X-ray image.
Your task is to output an analysis of
the given X-ray with respect {cls}.
You shall analyse this image closely
for its attributes.
You must provide an output strictly in
the following format :

"Present" - Yes/No
```

```
Prompt for generation:

Pretend you are a radiologist
Given is the hypothetical X-ray image.
This is the initial diagnosis of an
X-Ray by a radiologist: {pred}.
Your output should be in the following
format :

"Explanation" - A short passage
describing the contents of the image
with respect to the ailment(s) above.
```

Figure 3: Prompts for the classification of pathologies and the generation of reports.

models are prompted for the presence of each of the ten diseases -see Algorithm 1. The models respond with a 'yes' or a 'no' response for each pathology. The list of pathologies identified as 'yes' makes up the output of the `Classifier`.

The performance of the considered models is reported in Table 1. To obtain the results with MiniGPT-Med, Qwen2-VL-7B and GPT-4o, they are prompted to classify the pathologies using the prompt in Figure 3. The accuracies, F1 score and the exact match ratio reported in Table 1 suggest that LLMs are not effective in capturing the context of pathologies in the Chest X-ray images.

The MiniGPT-Med and GPT-4o LLMs indicate the presence of most pathologies in every X-ray image. Meanwhile, the Qwen2-VL-7B reports the absence of pathologies in most X-ray images, resulting in an accuracy of 81.87% but an F1 score of 0.10. The DenseNet-121 has the best F1 score of 0.2466. To further highlight the advantage of using the DenseNet-121 in the proposed pipeline over the LLMs for the task of multi-label classification, Table 1 also reports the exact match ratio for each model. The exact match ratio accepts only those samples that have all their labels correctly classified. DenseNet 121's exact match ratio is 4.5 times higher than the considered LLMs.

It should be noted that the comparisons in Table 1 do not suggest that the DenseNet 121 is objectively a better classifier than the LLMs considered in this paper. But, the results do support the use of DenseNet-121 for the multi-label classification task in the proposed pipeline, as it provides the most accurate medical context for report generation.

Table 1
Metrics are weighted across ailments.

Model	Accuracy (%) ↑	F1 ↑	Exact match ratio (%) ↑
GPT-4o	60.52	0.1842	0.50
MiniGPT-Med	20.32	0.2184	0.00
Qwen2-VL-7B	**81.87**	0.1063	1.00
DenseNet	76.53	**0.2466**	**4.57**

3.2. Report generation results

The second step of the proposed pipeline uses the pathology labels obtained from the `Classifier` as additional context to prompt the `Generator` for report generation -see Algorithm 1. The `Generator` is an LLM that is instructed to produce a report using the prompt defined in Figure 3.

3.2.1. Evaluation using BERT similarity score

The generated report is evaluated by matching it with the ground truth using the BERT embeddings [17]. The embeddings are obtained by passing the report to a Bidirectional Encoder Representations from

Transformers (BERT) encoder-only model, which is capable of capturing deep bidirectional context between words. Thus, for evaluation, the generated report and the ground truth reports are converted into embeddings, and a similarity score is measured between them using cosine similarity as follows:

$$similarity = \frac{[\texttt{CLS}]_A \cdot [\texttt{CLS}]_B^T}{norm([\texttt{CLS}]_A) \odot norm([\texttt{CLS}]_B)}$$

where $A \, and \, B$ represents the generated and ground truth reports.

The BERT similarity score reported in Table 2 indicates that all the models have similar performance, with a similarity score of approximately 0.7 with the ground truth. However, it can be empirically assessed that while the BERT similarity score captures the semantics and grammatical consistencies, it fails to capture the logic and the medical context. As illustrated in Figure 4, a generated report predicting the presence of pathology when compared with the ground truth reporting its absence gets a high BERT embedding similarity score of 0.8850.

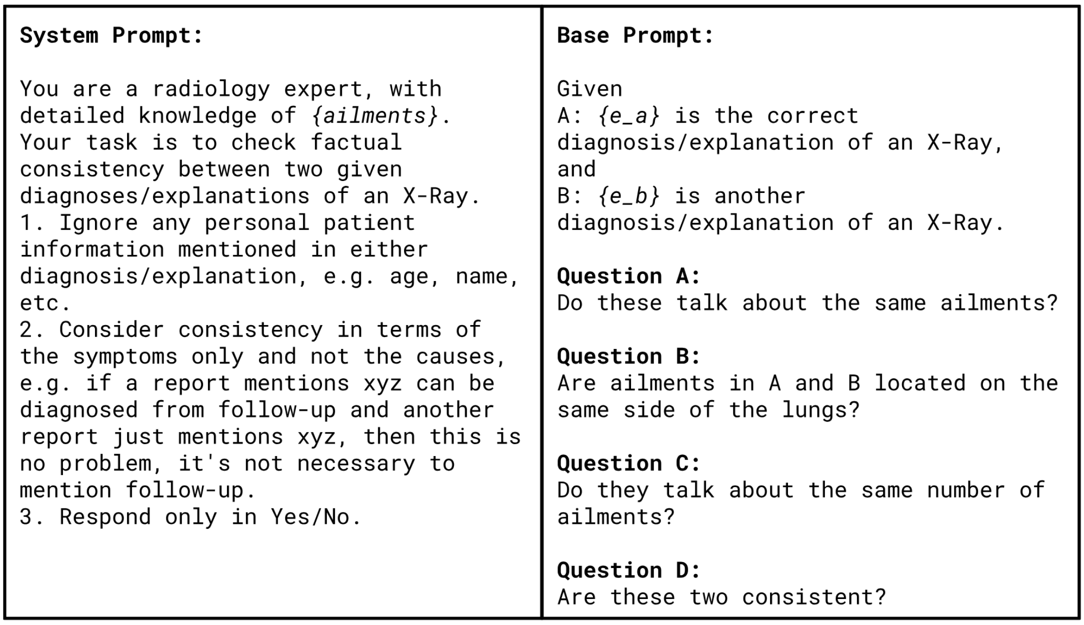
Ground Truth Report:	**Generated Report:**
The diagnosis is pleural calcification. The patient has a chronic cough and left-sided chest pain. The medical imaging shows dense irregular linear calcifications in the left lower hemithorax, deep to the left lateral rib cage, along with tenting of the left diaphragm. The lungs are clear.	The given X-ray image shows a well-defined radiographic view of the chest, including the spine, ribcage, clavicles, and a portion of the shoulder joints. The lung fields are visible and appear clear with no noticeable abnormalities. The heart and diaphragm can be seen and appear within normal size and positioning. **There is no presence of calcifications detected in the image.**
BERT Similarity **0.8850**	

Figure 4: BERT embedding similarity score for generated and ground truth reports for a scenario where the generated diagnosis differs from ground truth diagnosis. The maximum score is **1.00**.

3.2.2. Evaluation using GPT-4o as a Judge

To address the limitations of the BERT similarity score, the paper proposes to use GPT-4o [10] as a judge to evaluate the generated reports. Recent work in [18] reports that GPT-4o has a strong logical reasoning ability and a comprehensive natural language understanding. It also performs better than most other LLMs in logical reasoning tasks using datasets that are out-of-distribution for the GPT-4o. This motivates the use of GPT-4o to compare the generated reports with the ground truth. To evaluate the similarity between the generated report and ground truth, the GPT-4o is prompted with four questions, to which the GPT-4o responds with either a 'yes' or a 'no' response. The ground truth was compared with the generated reports manually, and it was observed that the ground truth significantly deviated from the generated reports when reporting: number of pathologies, consistency with the pathology description, the location and number of ailments. These observations motivated the questions for the LLM. The prompt for the LLM is presented in Figure 5. It follows the Tree-of-Thought [19] prompting strategy, where the depth of the tree is one. A normalised score is then calculated for the responses, reported in Table 2.

The MiniGPT-Med, a finetuned LLM, performs poorly in comparison to GPT-4o. The poor performance of the MiniGPT-Med can be attributed to the CLIP model, which is frozen during the fine-tuning process. When prompted with its classification labels, the results are poorer. The poor classification labels of the MiniGPT-Med mislead the model into generating bad reports. When prompted with the output from the DenseNet121, the results improve, suggesting that the discriminator model can



System Prompt:

```
You are a radiology expert, with
detailed knowledge of {ailments}.
Your task is to check factual
consistency between two given
diagnoses/explanations of an X-Ray.
1. Ignore any personal patient
information mentioned in either
diagnosis/explanation, e.g. age, name,
etc.
2. Consider consistency in terms of
the symptoms only and not the causes,
e.g. if a report mentions xyz can be
diagnosed from follow-up and another
report just mentions xyz, then this is
no problem, it's not necessary to
mention follow-up.
3. Respond only in Yes/No.
```

Base Prompt:

```
Given
A: {e_a} is the correct
diagnosis/explanation of an X-Ray,
and
B: {e_b} is another
diagnosis/explanation of an X-Ray.
```

Question A:
Do these talk about the same ailments?

Question B:
Are ailments in A and B located on the same side of the lungs?

Question C:
Do they talk about the same number of ailments?

Question D:
Are these two consistent?

Figure 5: Prompt used for evaluation of the generated reports.

correct the LLM following the proposed pipeline. The proprietary GPT-4o model performs much better, especially when operating with the GPT-4o as the classifier in the proposed pipeline. It allows the LLM to break down the problem into first classifying the pathology and then using it as context to generate the reports. It improves the results over direct use of GPT-4o by 0.018. The use of DenseNet 121 as a classifier further improves the results by 0.06. The Qwen-VL-7B model also sees a similar improvement in performance when using the DenseNet as the classifier in the proposed pipeline. When the Qwen-VL-7B model uses another Qwen-VL-7B model as a classifier in the proposed pipeline, the generated reports appear to be worse than when generated without the context of the pathologies. This further supports the fact that Qwen-VL-7B model is a poor classifier despite good reported accuracy in Table 1. Table 2 reports a question-wise evaluation of the two pipelines. To assess the capability of the GPT-4o as a judge, the ground truth is matched with itself using the GPT-4o. For questions A, B and D, the GPT-4o returns a score of 1.0. It suggests that GPT-4o understands the context of the question and scores it correctly. Question C returns a score of ≈ 0.7, which suggests that GPT-4o is unable to understand the context and may require further refining. For a preliminary assessment, however, this is considered adequate. The question-wise score for the proposed pipeline suggests that the context obtained from the discriminative models is meaningful in improving the quality of the generated reports. However, the low absolute scores for each question suggest that the LLMs require significant improvement before the reports can be trusted for their medical credibility.

The work presented in this paper is available at https://github.com/Harshvardhan-Mestha/DETerGENt/.

4. Conclusion

The paper assesses a pre-trained GPT-4o, a fine-tuned MiniGPT-Med model and an Qwen-VL-7B open source LLM for their ability to generate reports for chest X-ray images. It concludes that the poor report generation can be attributed to the inability of the models to capture the subtle context of pathologies in the X-ray images. To address this issue, the paper proposes a pipeline that combines an LLM and discrimination models. The discriminative model identifies the pathologies with significantly better accuracy than the LLMs. The classification results are used by the LLMs as added context to generate the report. The generated reports are compared with the ground truth using the BERT embedding

Table 2
Comparing the performance of LLMs in generating reports following the default and the proposed pipeline. *Evaluation results when the ground truth is matched with itself.

Model		**BERT Sim. ↑**	**LLM-as-a-judge**	**Question-wise scores**			
Classifier	*Generator*		**Score ↑**	*A*	*B*	*C*	*D*
None	MiniGPT-Med	0.739	0.0462	0.0200	0.1400	0.0000	0.0200
MiniGPT-Med	MiniGPT-Med	0.708	0.0112 **(-0.035)**	0.0050	0.0350	0.0000	0.0050
DenseNet	MiniGPT-Med	0.722	0.0685 **(+0.022)**	0.0863	0.1269	0.0558	0.0102
None	GPT-4o	0.840	0.0437	0.0600	0.0900	0.0000	0.0300
GPT-4o	GPT-4o	0.779	0.0625 **(+0.0188)**	0.0900	0.0800	0.0500	0.0250
DenseNet	GPT-4o	0.783	0.1053 **(+0.0616)**	0.1320	0.1269	0.0863	0.0609
None	Qwen2-VL-7B	0.782	0.0638	0.0650	0.0800	0.0050	0.0250
Qwen2-VL-7B	Qwen2-VL-7B	0.840	0.0475 **(-0.0163)**	0.0900	0.0900	0.0500	0.0250
DenseNet	Qwen2-VL-7B	0.783	0.1079 **(+0.0441)**	0.1472	0.1472	0.0660	0.0711
Ground Truth reference*		1.0000	0.925	1.0000	0.9700	0.7050	1.0000

score. While it checks for semantic similarity, it fails to assess the logic and medical context in the reports. Thus, the paper proposes to use LLM as a judge. It uses GPT-4o to evaluate the generated reports across four parameters to generate a normalised score. It checks if the pathologies match the ground truth, if the pathologies are located in the same place, if the number of pathologies is consistent and if they match their descriptions. The results suggest that the proposed two-step pipeline generates better reports than the default pipeline.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] J. P. Cohen, J. D. Viviano, P. Bertin, P. Morrison, P. Torabian, M. Guarrera, M. P. Lungren, A. Chaudhari, R. Brooks, M. Hashir, H. Bertrand, TorchXRayVision: A library of chest X-ray datasets and models, in: Medical Imaging with Deep Learning, 2022. URL: https://github.com/mlmed/torchxrayvision.

[2] OpenAI, Gpt-4 technical report, 2023. arXiv:2303.08774.

[3] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, 2023. URL: https://arxiv.org/abs/2304.08485. arXiv:2304.08485.

[4] Anthropic, Introducing the next generation of claude: Claude 3 model family, 2024. URL: https://www.anthropic.com/news/claude-3-family, accessed: 2024-10-21.

[5] S. Lee, W. J. Kim, J. Chang, J. C. Ye, Llm-cxr: Instruction-finetuned llm for cxr image understanding and generation, 2024. URL: https://arxiv.org/abs/2305.11490. arXiv:2305.11490.

[6] A. Alkhaldi, R. Alnajim, L. Alabdullatef, R. Alyahya, J. Chen, D. Zhu, A. Alsinan, M. Elhoseiny, Minigpt-med: Large language model as a general interface for radiology diagnosis, 2024. URL: https://arxiv.org/abs/2407.04106. arXiv:2407.04106.

[7] J.-B. Alayrac, J. Donahue, P. L. et al., Flamingo: a visual language model for few-shot learning, 2022. URL: https://arxiv.org/abs/2204.14198. arXiv:2204.14198.

[8] D. L. Payne, K. Purohit, W. M. Borrero, K. Chung, M. Hao, M. Mpoy, M. Jin, P. Prasanna, V. Hill,

Performance of GPT-4 on the american college of radiology in-training examination: Evaluating accuracy, model drift, and fine-tuning, Acad. Radiol. 31 (2024) 3046–3054.

[9] O. Thawkar, A. Shaker, S. S. Mullappilly, H. Cholakkal, R. M. Anwer, S. Khan, J. Laaksonen, F. S. Khan, Xraygpt: Chest radiographs summarization using medical vision-language models, 2023. URL: https://arxiv.org/abs/2306.07971. arXiv:2306.07971.

[10] OpenAI, Gpt-4o system card, 2023. arXiv:2303.08774.

[11] P. Wang, S. Bai, e. a. Tan, Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, arXiv preprint arXiv:2409.12191 (2024).

[12] G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger, Densely connected convolutional networks, 2018. URL: https://arxiv.org/abs/1608.06993. arXiv:1608.06993.

[13] Q. Sun, Y. Fang, L. Wu, X. Wang, Y. Cao, Eva-clip: Improved training techniques for clip at scale, 2023. URL: https://arxiv.org/abs/2303.15389. arXiv:2303.15389.

[14] L. M. e. a. Hugo Touvron, Llama 2: Open foundation and fine-tuned chat models, 2023. URL: https://arxiv.org/abs/2307.09288. arXiv:2307.09288.

[15] J. Irvin, P. Rajpurkar, M. K. et al., Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019. URL: https://arxiv.org/abs/1901.07031. arXiv:1901.07031.

[16] Radiopaedia, A collaborative educational web resource for radiology, 2007. Retrieved October 23, 2024, from https://radiopaedia.org.

[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL: https://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[18] H. Liu, R. Ning, Z. Teng, J. Liu, Q. Zhou, Y. Zhang, Evaluating the logical reasoning ability of chatgpt and gpt-4, 2023. URL: https://arxiv.org/abs/2304.03439. arXiv:2304.03439.

[19] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, K. R. Narasimhan, Tree of thoughts: Deliberate problem solving with large language models, in: Thirty-seventh Conference on Neural Information Processing Systems, 2023. URL: https://openreview.net/forum?id=5Xc1ecxO1h.