MAGDA: Multi-agent guideline-driven diagnostic assistance

David Bani-Harouni[1,2], Nassir Navab[1,2], and Matthias Keicher[1,2]

[1] Computer Aided Medical Procedures, School of Computation, Information and Technology, Technical University of Munich, Garching, Germany
[2] Munich Center for Machine Learning (MCML), Munich, Germany
david.bani-harouni@tum.de

Abstract. In emergency departments, rural hospitals, or clinics in less developed regions, clinicians often lack fast image analysis by trained radiologists, which can have a detrimental effect on patients' healthcare. Large Language Models (LLMs) have the potential to alleviate some pressure from these clinicians by providing insights that can help them in their decision-making. While these LLMs achieve high test results on medical exams showcasing their great theoretical medical knowledge, they tend not to follow medical guidelines. In this work, we introduce a new approach for zero-shot guideline-driven decision support. We model a system of multiple LLM agents augmented with a contrastive vision-language model that collaborate to reach a patient diagnosis. After providing the agents with simple diagnostic guidelines, they will synthesize prompts and screen the image for findings following these guidelines. Finally, they provide understandable chain-of-thought reasoning for their diagnosis, which is then self-refined to consider inter-dependencies between diseases. As our method is zero-shot, it is adaptable to settings with rare diseases, where training data is limited, but expert-crafted disease descriptions are available. We evaluate our method on two chest X-ray datasets, CheXpert and ChestX-ray 14 Longtail, showcasing performance improvement over existing zero-shot methods and generalizability to rare diseases.

Keywords: Clinical guidelines · Large Language Models · Zero-shot classification.

1 Introduction

Radiology holds a critical position in contemporary healthcare, being integral to the treatment and management of most patients. However, the healthcare sector is currently grappling with what has been termed the "radiologist shortage" [12]. In the UK, this shortage stands at 29% and is predicted to worsen, reaching 40% within the next four years [15]. This effect is exacerbated in rural hospitals or clinics in less developed regions of the world, where the population per radiologist is much greater [11, 18]. When there is a lack of radiologists, the clinicians

with patient contact have to either miss out on valuable radiological information or evaluate that information themselves without proper training. Large Language Models (LLMs) have recently demonstrated remarkable potential for reasoning and solving complex problems, presenting an opportunity to address this challenge [14]. However, in a clinical context, deterministic models strictly adhering to evidence-based medical guidelines are preferred over creative but unpredictable LLM outputs. Moreover, generalist LLMs like GPT-4 [1] may lack domain-specific knowledge required for accurate diagnosis or have outdated medical insights. Consequently, providing LLMs access to relevant clinical knowledge sources, such as guidelines encapsulating the medical community's consensus, is critical for effective diagnostic processes, particularly for rare diseases with limited data. In contrast to visual instruct tuning [7], prompting Vision-Language Models (VLMs) is an intriguing approach to enabling LLMs to understand the content of images without the need to retrain. Recent explorations have successfully employed contrastive language-image pertaining (CLIP) [10] for few- and zero-shot classification of common diseases in chest X-rays [2, 13, 16, 20, 22, 23]. Building on this, Xplainer [9] introduced a classification-by-description approach, querying a vision-language model for image observations indicative of a disease, providing inherent explainability. However, it naively averages concept probabilities, failing to account for dependencies between these concepts. To address those limitations, we propose MAGDA (Multi-Agent Guideline-driven Diagnostic Assistance), a multi-agent framework that unifies the incorporation of clinical guidelines as knowledge sources, dynamic prompting of a vision-language model for LLM understanding of radiology images, and a transparent diagnosis reasoning following the domain-specific knowledge provided by clinical guidelines. We show that this approach achieves state-of-the-art performance on zero-shot classification tasks of pathologies in the CheXpert dataset [5] and rare diseases in the ChestXRay 14 Longtail dataset [4]. Our key contributions are:

- An end-to-end guideline-driven approach that requires only a clinical guideline and a medical image as input to perform zero-shot diagnosis
- Novel dynamic prompting of vision-language models to enable LLMs to screen medical images for unseen diseases without the need for fine-tuning
- A transparent reasoning process through chain-of-thought reasoning, providing insights into the diagnostic decision-making

2 Methodology

2.1 Model Overview

We propose MAGDA, a multi-agent zero-shot method that can work with expert-crafted disease descriptions to provide transparent decision support. A general overview of the method is shown in Fig. 1. The LLMs used are not fine-tuned and all adaptions to the tasks are performed in-context. The multi-agent system consists of three agents that take over different tasks in the diagnosis procedure:

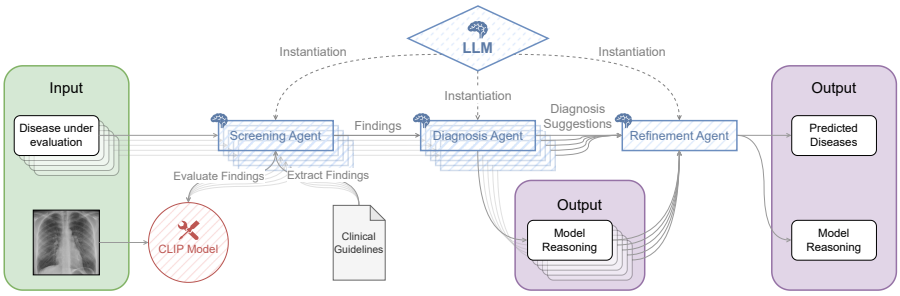

Fig. 1. Schematic overview of the proposed method MAGDA.

1. **Screening agent $\mathcal{S}$:** This agent handles the image analysis. As an Augmented Language Model [8] with tool-using capabilities, it can prompt a CLIP model [10] to evaluate fine-grained image findings according to given diagnosis guidelines.
2. **Diagnosis agent $\mathcal{D}$:** This agent is given the image findings from the screening agent and is tasked to reason about these findings to reach a diagnosis for the patient
3. **Refinement agent $\mathcal{R}$:** This agent is responsible for refining the predictions of the diagnosis agent by considering inter-dependencies of diseases and evaluating the quality of the reasoning. It then gives the final diagnosis prediction for the patient at hand.

2.2 Screening Agent

When diagnosing a specific patient $p \in P$, this agent is run once for every disease $d \in D$ that we want to evaluate. It is presented with expert-crafted fine-grained image findings and returns the positive or negative findings present in the image following the given diagnosis guidelines.

$$\mathcal{S}_d^p(G_d, d) \rightarrow F_d^p,$$

where G_d are the disease guidelines, d is the condition under evaluation, and F_d^p are the patient and disease-specific positive and negative image findings. The guidelines G_d can be provided in either an already expert-crafted fine-grained list of disease-specific image findings or an unstructured disease description from which the model can extract these fine-grained image findings. For example, in the case of an enlarged cardiomediastinum, one image finding may be described as "Abnormal contour of the heart border". In order to screen the image for the presence or absence of these image findings, we augmented the screening agent with the ability to prompt a CLIP model [10]. Following established works of classification-by-description [9, 13, 16], we task the agent to use contrastive prompting, i.e., prompting the model with both a positive and a negative description. This has shown to be superior to just evaluating the similarity between

the image embedding and the text embedding of the positive description. Since the findings are provided only in positive form, the model is tasked with creating negations following grammar rules and ensuring sensibility beyond simply appending the word "no" in front of the description. As we are not fine-tuning the LLM to be able to use the CLIP-tool, we provide an in-context description of the tool together with instructions on how to use it. Through one example, it is further tasked with following the report style template from Xplainer [9]. During inference, the tool can be called using the call:

CLIP: [positive description] / [negative description] ->

Once the descriptions have been extracted from the model output, we provide them to the CLIP model. Specifically, we employ the image and text encoder from BioVil-T [2]. After computing the cosine similarity of the image embedding with the positive and negative text embeddings, we calculate the softmax over these two similarities to get the final probabilities of each finding. Initial results on the validation set showed that the BioVil-T model tends to over-predict positive findings, so we only count a positive finding if its probability exceeds a threshold ψ. Finally, depending on whether the tool returns a positive or negative result for the given description, we append "Positive" or "Negative" to the inference text and continue the LLM inference from there. After all given descriptions have been evaluated, the collected positive and negative disease descriptions are passed to the diagnosis agent.

2.3 Diagnosis Agent

The diagnosis agent is again run once per patient and condition under evaluation. It is given the list of findings extracted from the image by the screening agent and returns a positive or negative prediction, including the reasoning for that decision.

$$\mathcal{D}_d^p(F_d^p, d) \rightarrow p_d^p, r_d^p,$$

where p_d^p is the binary disease prediction for patient p and disease d, and r_d^p is the reasoning for that prediction. It has been shown that LLM reasoning can be significantly improved by chain-of-thought prompting [21], a prompt engineering technique where the model is asked to provide step-by-step reasoning before answering a question. Additionally, to an increase in reasoning capabilities, this reasoning makes the method inherently explainable. As the model provides explanations for its predictions, clinicians can use these explanations to evaluate the decision process and increase trust in the model output. Specifically, we ask the model to provide reasoning before answering the question "Does the patient have [d]?". We prompt the model to use a specified format to make parsing the model output possible, ending the reasoning process with the sentence: "Therefore, my answer is: [yes/no]." Once the various predictions and reasonings have been collected, they are passed to the refinement agent for the final patient diagnosis.

2.4 Refinement Agent

The refinement agent is run once per patient. It is presented with all positive disease predictions and the diagnosis agent's reasonings for these predictions. It returns the final patient prediction.

$$\mathcal{R}^p(\{(p_d^p, r_d^p) | d \in D, p_d^p \text{ is positive}\}) \rightarrow \{\hat{p}_d^p | d \in D\}$$

The refinement agent is tasked with evaluating the provided reasoning. So far, every disease has been evaluated on its own in order to not overload the agents. At this step in the diagnosis process, inter-dependencies between diseases can be considered. The refinement agent is queried for every disease under evaluation if that disease is present or not and again asked to provide chain-of-thought reasoning for that decision. From the model replies we parse the final patient predictions $\hat{p}^p$. "No Finding" is predicted if all other disease predictions are negative.

3 Experimental setup

Datasets and evaluation metrics We evaluate our method on two chest X-ray datasets, CheXpert [5] and ChestXRay 14 Longtail [4, 19]. The CheXpert dataset includes manually annotated validation and test sets comprising 200 and 500 patients, respectively. It encompasses 14 different categories, featuring "No Finding", 12 pathology labels, e.g., "Pneumonia", and a class "Support Devices". On CheXpert, we perform multi-label classification. Most comparable methods evaluate using the Area Under the ROC-curve (AUC) metric. As our method generates discrete predictions, threshold-independent metrics, like AUC, cannot sensibly be evaluated. Instead, we report micro and macro F1-score, precision, and recall. The CLIP finding probability threshold ψ, which is used to combat the over-prediction of the CLIP model, is set to 0.55 based on experiments on the validation set.

The ChestXRay14 Longtail dataset is an extension of the common ChestXRay 14 dataset by adding five additional disease findings, expanding the classification to 20 categories. These are divided into 7 head classes (most common), 10 medium classes (moderately common), and 3 tail classes (least common). The dataset includes a balanced validation and test set, each offering 15 or 30 images per class, respectively, to ensure comprehensive coverage and evaluation capabilities across the spectrum of conditions. We evaluate on this balanced test set with equal number of cases per class. Here, we perform single-label classification and report the accuracy on the three tail classes. In this setting, we prompt the CLIP model without description negation. As the screening and diagnosis agents always perform multi-label classification, we further adapt our refinement agent to decide on exactly one positive prediction.

Implementation details The backbone of our method lies in a powerful LLM instantiated in different ways as the various agents. Unless stated otherwise, we

Table 1. Test set results for zero-shot classification on the CheXpert dataset. MAGDA (nG) is our proposed method without the use of guidelines, relying on LLM knowledge.

	F1-score		Precision		Recall	
Method	micro	macro	micro	macro	micro	macro
CheXzero	35.69	33.50	25.58	**37.72**	58.98	64.88
Xplainer	45.33	39.27	31.36	33.74	81.74	83.27
MAGDA (nG)	42.94	36.50	29.39	30.82	79.62	82.07
MAGDA	**46.18**	**39.58**	**31.93**	33.43	**83.43**	**83.47**

Table 2. Test set results on the ChestXray 14 Longtail dataset. Accuracy is the classification accuracy on the rare tail classes. Methods above the line are fully supervised.

Method	Zero-shot	Accuracy
ResNet-50 [4]	×	1.7
Decoupling-cRT [4]	×	30.0
CheXzero	✓	12.3
Xplainer	✓	8.2
MAGDA	✓	**18.5**

employ the Mixtral 8x7B instruct model from Mistral AI [6]. This model provides a good trade-off between memory efficiency, inference speed, and model capabilities. We use a 4-bit GPTQ quantized version of the model [3]. This reduces the memory requirements while keeping the loss in model accuracy minimal. Thus processed we are able to run text generations on a single Nvidia A40 GPU using a temperature of 0.8. Where available, we used the image findings from the public Xplainer repository [9] as guidelines to ensure fair comparison. Where not available, we used a similar approach to Xplainer of prompting GPT-4 to generate candidate findings, and correcting them based on text book knowledge.

4 Results and discussion

In Table 1, we compare with state-of-the-art zero-shot classification methods CheXzero [16] and Xplainer [9] on the CheXpert test dataset. Because CheXzero is only evaluated on the six competition pathologies in the original paper, we use their public code and model to evaluate on all CheXpert classes. Most state-of-the-art methods only report the AUC, we can therefore only compare with methods with published code and calculate their respective F1-score, precision, and recall. For example, a comparison with Seibold et al. [13] or ELIXR [24] was not possible for that reason. We outperform all comparable methods on zero-shot classification on all metrics except macro precision, where CheXzero has a better score at the expense of a much lower recall. Here, we also compare the guideline-driven approach with the generation of findings by the model itself, showing that the provision of guidelines to our method increases performance. Table 2 shows

Table 3. Comparison of zero-shot classification of the diagnosis agent, i.e., before refinement with the refinement agent, using different LLMs as a reasoning backbone on the test set of the CheXpert dataset.

	F1-score		Precision		Recall	
Reasoning Model	micro	macro	micro	macro	micro	macro
GPT-4	**46.87**	**41.10**	**32.17**	**34.17**	86.31	85.13
Llama2 70B chat	46.36	40.72	31.62	33.48	**86.87**	**88.01**
Mixtral 8x7B instruct	45.31	39.70	30.85	33.41	85.25	86.62

a comparison on the ChestXray14 Longtail dataset with the same zero-shot methods as before and additionally with two fully supervised methods trained on that dataset [4]. Here, we again outperform both zero-shot methods on the tail classes. Notably, we even reach a higher accuracy than a simple supervised method trained on the 68,058 training samples. Only a highly tuned method employing decoupled training [4] achieves a higher accuracy. These results show that the provision of detailed guidelines describing the diagnosis of rare and lesser-known diseases can help with diagnostic accuracy. In Table 3, we compare exchanging our Mixtral 8x7B instruct model for other well-known LLMs, namely Llama 2 70B chat [17] and GPT-4 [1]. While both alternative models reach higher performance, this comes at the cost of higher computational needs in the case of Llama2 70B chat, or dependence on a proprietary API.

Ablation studies We now want to look at the benefits of different aspects of our method. First, in Table 4, we compare the rule-based negation by simply appending a "no" before the finding description with the LLM-created finding negation done by the screening agent. We see that the latter results in better performance, highlighting the benefit of descriptions that are more aligned with natural language and thus the style of radiology reports. In Table 5, different approaches for the refinement agent are compared. We compare combinations of chain-of-thought reasoning and including the CheXpert disease graph in textual form [5]. This disease graph models the dependencies between classes, e.g. "Enlarged Cardiomediastinum" being a sign of "Cardiomegaly". However, both in the case of using chain-of-thought reasoning and not using it, the inclusion of the disease graph decreases the performance.

Qualitative results In Fig. 2, we show a qualitative example of the reasoning provided by the diagnosis agent. The agent is presented with conflicting findings regarding the diagnosis of an enlarged cardiomediastinum. Instead of naively aggregating them, giving each finding the same importance, the diagnosis agent can differentiate between more and less relevant findings and can employ reason to reach a diagnosis. This example also shows the high dependence on the correctness of the image findings identified by the CLIP model. Even perfect reasoning cannot reach the correct conclusion if it works with wrong information.

Table 4. Comparison of naive finding negation with the improved negation by the Screening agent. Results are reported before the refinement by the refinement agent on the CheXpert validation set.

	F1-score	
CLIP prompting	micro	macro
Naive negation	46.79	41.75
LLM negation	**48.00**	**42.10**

Table 5. Comparison of different refinement approaches on the CheXpert validation set. CoT = chain-of-thought reasoning, DG = disease graph.

		F1-score	
CoT	DG	micro	macro
×	×	48.10	41.56
×	✓	47.79	40.39
✓	×	**49.17**	**42.05**
✓	✓	47.52	40.70

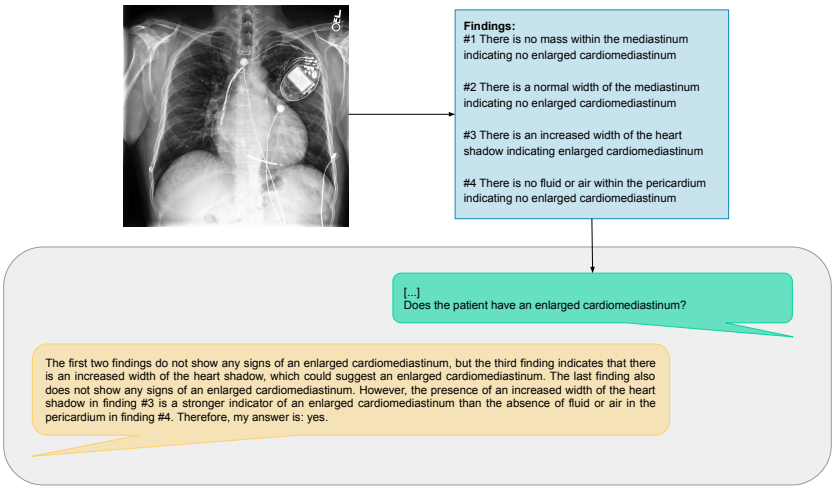

Fig. 2. A qualitative example of the model reasoning.

Therefore, improvements in the performance of VLMs can directly translate into further improvement of our method.

5 Conclusion

In this paper, we presented MAGDA, a novel multi-agent approach that integrates clinical guidelines, dynamic vision-language model prompting, and large language model reasoning to address the challenges of diagnostic assistance using LLMs. Our approach leverages the strengths of both LLMs and VLMs, enabling the zero-shot classification of diseases without the need for model retraining or fine-tuning. This guideline-driven methodology not only facilitates accurate diagnoses from medical images but also introduces a transparent reasoning process that enhances the explainability and trustworthiness of the diagnostic outcomes. Our evaluation on the CheXpert and ChestXray14 LT datasets demonstrates the

effectiveness of our approach, particularly in scenarios involving rare diseases where traditional diagnostic methods are often hindered by data scarcity. By incorporating domain-specific knowledge through clinical guidelines and employing dynamic prompting techniques, we improve diagnostic accuracy and model trustworthiness.

Acknowledgments. The authors gratefully acknowledge the financial support by the Bavarian Ministry of Economic Affairs, Regional Development and Energy (StMWi) under project ThoraXAI (DIK-2302-0002).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., et al.: Learning to exploit temporal structure for biomedical vision-language processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15016–15027 (2023)
3. Frantar, E., Ashkboos, S., Hoefler, T., Alistarh, D.: Gptq: Accurate post-training quantization for generative pre-trained transformers. arXiv preprint arXiv:2210.17323 (2022)
4. Holste, G., Wang, S., Jiang, Z., Shen, T.C., Shih, G., Summers, R.M., Peng, Y., Wang, Z.: Long-tailed classification of thorax diseases on chest x-ray: A new benchmark study. In: MICCAI Workshop on Data Augmentation, Labelling, and Imperfections. pp. 22–32. Springer (2022)
5. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)
6. Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., Casas, D.d.l., Hanna, E.B., Bressand, F., et al.: Mixtral of experts. arXiv preprint arXiv:2401.04088 (2024)
7. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 34892–34916. Curran Associates, Inc. (2023)
8. Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., et al.: Augmented language models: a survey. arXiv preprint arXiv:2302.07842 (2023)
9. Pellegrini, C., Keicher, M., Özsoy, E., Jiraskova, P., Braren, R., Navab, N.: Xplainer: From x-ray observations to explainable zero-shot diagnosis. arXiv preprint arXiv:2303.13391 (2023)
10. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
11. Ramli, N.M., Zain, N.R.M.: The growing problem of radiologist shortage: Malaysia's perspective. Korean Journal of Radiology **24**(10), 936 (2023)

12. Rimmer, A.: Radiologist shortage leaves patient care at risk, warns royal college. BMJ: British Medical Journal (Online) **359** (2017)
13. Seibold, C., Reiß, S., Sarfraz, M.S., Stiefelhagen, R., Kleesiek, J.: Breaking with fixed set pathology recognition through report-guided contrastive training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 690–700. Springer (2022)
14. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al.: Large language models encode clinical knowledge. Nature **620**(7972), 172–180 (2023)
15. The Royal College of Radiologists: Clinical radiology census report 2022. The Royal College of Radiologists (Online) (2022)
16. Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. Nature Biomedical Engineering **6**(12), 1399–1406 (2022)
17. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
18. Vu, L.D., Nguyen, H.T.T., Nguyen, T.N., Pham, T.M.: The growing problem of radiologist shortage: Vietnam's perspectives. Korean Journal of Radiology **24**(11), 1054 (2023)
19. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)
20. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022)
21. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems **35**, 24824–24837 (2022)
22. Windsor, R., Jamaludin, A., Kadir, T., Zisserman, A.: Vision-language modelling for radiological imaging and reports in the low data regime. In: Medical Imaging with Deep Learning (2023)
23. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21372–21383 (2023)
24. Xu, S., Yang, L., Kelly, C., Sieniek, M., Kohlberger, T., Ma, M., Weng, W.H., Kiraly, A., Kazemzadeh, S., Melamed, Z., et al.: Elixr: Towards a general purpose x-ray artificial intelligence system through alignment of large language models and radiology vision encoders. arXiv preprint arXiv:2308.01317 (2023)