LEAVS: An LLM-based Labeler for Abdominal CT Supervision

Ricardo Bigolin Lanfredi[1 [0000−0001−8740−5796]], Yan Zhuang[2,3 [0000−0003−1756−0277]], Mark Finkelstein[2], Praveen Thoppey Srinivasan Balamuralikrishna[1], Luke Krembs[4], Brandon Khoury[4 [0000−0002−9625−9353]], Arthi Reddy[2 [0009−0001−7645−5613]], Pritam Mukherjee[1 [0000−0002−9975−9994]], Neil M. Rofsky[2], and Ronald M. Summers[1 [0000−0001−8081−7376]]

[1] National Institutes of Health Clinical Center, Bethesda, MD 20892, USA
bigolinlanfrer2@cc.nih.gov rsummers@cc.nih.gov
[2] Department of Diagnostic, Molecular and Interventional Radiology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[3] Windreich Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[4] Walter Reed National Military Medical Center, Bethesda, 20892, MD, USA

Abstract. Extracting structured labels from radiology reports has been employed to create vision models that detect several types of abnormalities simultaneously. However, existing works focus mainly on the chest region. Few works have investigated abdominal radiology reports due to the more complex anatomy and a wider range of pathologies in the abdomen. We propose LEAVS (Large language model Extractor for Abdominal Vision Supervision). This labeler can annotate the certainty of presence and the urgency of seven types of abnormalities for nine abdominal organs on CT radiology reports. To ensure broad coverage, we chose abnormalities that encompass most of the finding types from CT reports. Our approach employs a specialized chain-of-thought prompting strategy for a locally run LLM using sentence extraction and multiple-choice questions in a tree-based decision system. We demonstrate that the LLM can extract several abnormality types across abdominal organs with an average F1 score of 0.89, significantly outperforming competing labelers and humans. Additionally, we show that the extraction of urgency labels achieves performance comparable to that of human annotations. Finally, we demonstrate that the abnormality labels contain valuable information for training a vision model that classifies several organs as normal or abnormal. We release our code and structured annotations for a publicly available dataset containing over 1,000 CT volumes.

Keywords: Large language models · Abdominal CT · Medical reports · Abnormality labels · Annotation · Classification

1 Introduction

Several works have extracted and shared structured labels from medical reports to develop generalist vision models in radiology, with examples for chest X-

rays [25] and chest CTs [5]. Labeling has traditionally been done by rule-based algorithms [25,10,13,5,22] and supervised deep learning algorithms [19]. However, recent works have employed large language models (LLMs) and shown their superiority [3]. Work with LLMs on extracting labels from the long and diverse CT reports has been limited to specific findings [6,23,24,27].

We propose a prompt system named LEAVS (**L**LM **E**xtractor for **A**bdominal **V**ision **S**upervision). It uses LLMs to extract comprehensive findings from abdominal CT reports for several organs and represent them as structured labels, as shown in Figure 1. The LEAVS prompt system is inspired by MAPLEZ (Medical report Annotations with Privacy-preserving LLM using Expeditious Zero shot answers) [3], but provides several innovations when adapting it to CT reports:

- sentence filtration, because we hypothesize that it will allow the LLM to focus only on the parts of the long CT report that matter for its task;
- multiple-choice questions, so the LLM picks one type for each report finding;
- finding type definitions for abdominal CT, chosen to cover almost all findings from reports while having enough types for separating distinct findings;
- urgency assessment, which might be important for filtering findings and as additional information for supervision in vision models.

We demonstrate better scores than the average human for abnormality labeling and achieve the same level of scores for urgency labeling. We also employ the structured labels to train a CT vision model and demonstrate preliminary classification results, a potential step towards developing universal abnormality detectors for abdominal CT scans. Our code and annotations for the AMOS-MM dataset [12] can be found at `https://github.com/rsummers11/LEAVS`.

1.1 Related works

There have been few works that extract several abnormality types from abdominal CT reports. Islam Tushar et al. [22] employed a rule-based algorithm to extract labels of 5 different findings per organ for three organs in the chest and abdomen and train a vision model with a private dataset. Lea Draelos et al. [5] employ a similar approach with the SARLE (Sentence Analysis for Radiology Label Extraction) labeler, which annotates 83 abnormalities across 52 body regions, mainly targeting chest CT scans. Both works employ rule-based algorithms, which are less flexible than LLMs, as they require a new set of expertly crafted rules for every new abnormality or keyword they can identify. As we show in our work, the SARLE labeler does not perform as well in the generic task of extracting "any abnormality" presence in organs.

2 Methods

The proposed zero-shot prompt system is presented in Figure 2 and consists of four stages: sentence filtration, finding type assessment, finding uncertainty assessment, and urgency assessment. The first two stages are executed separately

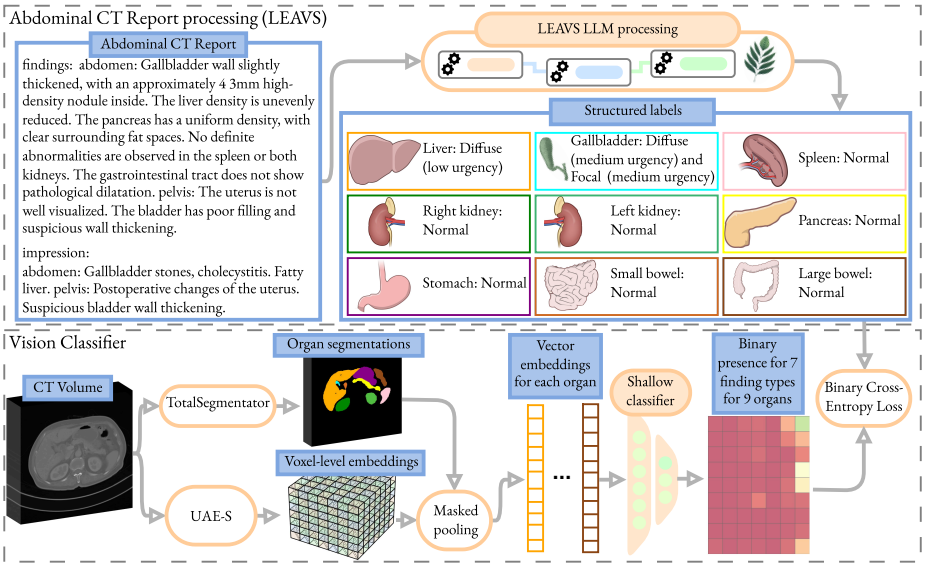

Fig. 1. We propose to employ LLMs to extract structured labels from CT reports to train vision models. We present an example of an input and output of our method. The report chosen for display is purposefully short. We include the pipeline we employed to classify the presence of several types of findings in several abdominal organs.

for each organ, and the remaining two stages are executed for each finding type in each organ. The nine organs we evaluate are presented in Figure 1. The prompt system, however, allows for the easy integration of any organ or body region.

Sentence filtration occurs in two steps: the LLM first lists informative sentences for each organ in a single answer and then reviews the remaining ones to identify any additional relevant content. We then join the informative sentences from both steps and provide only those in the subsequent prompt stages. For the finding type assessment, the LLM is asked, in a multiple-choice question, for the finding types mentioned in the report, as described in Table 1. A sentence or organ may include multiple distinct findings. The LLM categorizes the uncertainty of the identified findings among the choices shown in Figure 2. We classify the urgency of present or possible findings according to the definitions from Larson et al. [16], with an added level for non-actionable findings.

Sentences were added at the start of the prompt to ask the LLM to consider all information. Chain-of-thought prompting (CoT) [14] was used except for the first sentence filtration step. The LLM was instructed to interpret the medical meaning of each sentence before responding. For easier parsing, the model summarized its CoT answer. Full prompts can be found in our code.

We evaluated the extracted labels when supervising a vision model. As shown in Figure 1, we trained a model that utilizes a pre-trained CT embedder, UAE-S [2], for feature extraction, and the segmentation outputs from TotalSegmenta-

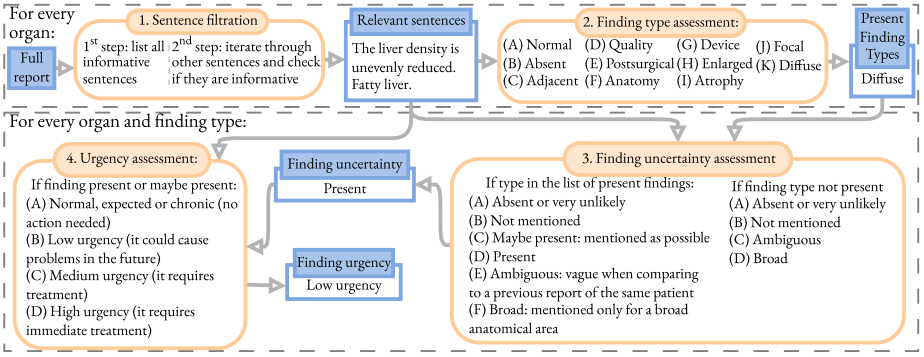

Fig. 2. Representation of all four steps of the LEAVS LLM processing. We show data outputs for the report shown in Figure 1 for the diffuse finding in the liver.

Table 1. Definition of employed finding types. The definitions were considered mutually exclusive, and the multiple-choice question helped the LLM to enforce that.

Finding Type	Description
Absent	{organ} is not present
Device	{organ} has support device
Postsurgical	{organ} has postsurgical changes
Enlarged	{organ} is enlarged
Atrophy	{organ} has atrophied
Anatomy	Uncommonly seen displacements, relative positionings, or shapes of the {organ}
Focal	{organ} has a finding that can be measured from its borders
Diffuse	{organ} has a finding without a well-defined border or shape for measurement or that affects large regions
Quality	Finding about the acquisition process for the {organ}
Adjacent	An adjacent, extrinsic finding for the {organ}
Normal	{organ} is normal

tor [11,26] to focus on the respective organ for each output. The only trainable part of the model was the shallow classifier. The classifier had seven outputs, one for each finding type among "Postsurgical"+"Absent", "Quality", "Anatomy", "Size": "Enlargement"+"Atrophy", "Device", "Diffuse", and "Focal".

3 Results

We validated LEAVS on an anonymized private dataset of 15 reports from the NIH Clinical Center, approved by the IRB (institutional review board). We evaluated many LLMs, including the Llama 3 family [7], Llama3-OpenBioLLM-70B [1], medllama3-v20 [18], QwQ-32B-Preview [21], Qwen2-72B-Instruct [28], and Qwen2.5-72B-Instruct [20]. Our experiments employed the best validation LLM in terms of F1-score for abnormality type labeling, Qwen2-72B-Instruct.

To test our labeler, we annotated 200 reports from the validation set of AMOS-MM [12], a publicly available dataset of abdominal CT volumes and reports with unspecified license. Humans annotated five finding types ("Quality", "Postsurgical"+"Absent", "Size", "Diffuse", and "Focal") for nine abdominal

Table 2. Scores of the proposed labeler (LEAVS), the MAPLEZ baseline [3], and humans. GBl: Gallbladder; RKid: Right Kidney; LKid: Left Kidney; LBow: Large Bowel; PS: postsurgical; H: Human; H Avg: average over the five humans; N: number of samples for score calculation (micro scores accumulate samples from non-micro rows, and humans have a reduced N as each human labeled 100-150 reports and cases with disagreement between other human labelers were excluded); N_+: number of positive samples; F1: F1 score; ^{ns}: $P \geq .05$; *: $P < .05$;**: $P < .01$; **: $P < .001$; LEAVS$_{sub}$F1: F1 score for LEAVS in the evaluation subset of the respective human (used for P values comparisons); Pre: Precision; Rec: Recall; Spe: Specificity; MCC: Matthews correlation coefficient. We display 95% confidence intervals for some metrics between parentheses.

Organ	Type	Labeler	N	N_+	F1	LEAVS$_{sub}$F1	Pre	Rec	Spe	MCC
Liver	Diffuse	LEAVS	200	31	.925(.844,.984)	-	.884	.970	.976	.912
Liver	Focal	LEAVS	200	91	.963(.931,.988)	-	.929	1.00	.936	.932
GBl	Diffuse	LEAVS	200	36	.817(.700,.907)	-	.903	.750	.982	.788
GBl	Focal	LEAVS	200	27	.732(.600,.835)	-	.588	.966	.896	.708
GBl	PS	LEAVS	200	11	.957(.824,1.00)	-	1.00	.917	1.00	.955
Spleen	Size	LEAVS	200	21	.978(.919,1.00)	-	.957	1.00	.994	.975
RKid	Focal	LEAVS	200	62	.912(.850,.957)	-	.935	.889	.972	.874
LKid	Focal	LEAVS	200	57	.901(.835,.952)	-	.862	.948	.939	.863
LBow	Focal	LEAVS	200	38	.759(.640,.851)	-	.735	.794	.933	.702
LBow	PS	LEAVS	200	20	.977(.909,1.00)	-	1.00	.955	1.00	.974
micro	micro	LEAVS	2000	394	.892(.869,.913)	-	.865	.921	.965	.865
micro	micro	MAPLEZ	2000	394	.827(.799,.851)**	.892(.869,.913)	.726	.960	.911	.789
micro	micro	H1	1124	236	.871(.837,.902)**	.923(.898,.946)	.950	.805	.989	.845
micro	micro	H2	843	193	.935(.906,.960)^{ns}	.927(.899,.952)	.973	.902	.992	.918
micro	micro	H3	476	125	.930(.893,.961)^{ns}	.940(.907,.968)	.967	.898	.989	.908
micro	micro	H4	543	117	.869(.819,.911)**	.935(.900,.964)	.885	.855	.970	.835
micro	micro	H5	451	107	.868(.811,.913)^{ns}	.898(.850,.936)	.918	.826	.977	.832
micro	micro	H Avg	-	-	.894(.876,.911)**	.924(.909,.938)	.938	.856	.983	.867

organs. Five annotators (two board-certified radiologists, two senior radiology residents, and one postdoctoral MD researcher) labeled 100 to 150 reports each, depending on annotator availability, on an internally developed interface that displayed one report at a time. In the interface, abnormality presence was selected through checkboxes for each organ/finding type, and urgency was selected from drop-down menus. Each report had three annotators. When evaluating LEAVS, the ground truth was the majority vote from humans for binary labels and the average of available human labels for urgency. We report results for organs/abnormalities with more than 10 positive cases in the test set, which removed some of the nine organs from our results. We calculated two-sided hypothesis tests of the difference in scores against the proposed labeler (LEAVS) and confidence intervals with 95% significance employing a paired bootstrap permutation test with 2,000 samples. When evaluating humans, we only considered cases when the two other humans who labeled that specific report agreed on the presence to avoid biasing the scores. Urgency labels considered only the urgency of the other two humans. When cases were filtered for human evaluation, we compared against the LEAVS scores in the same subset for fair comparison. We use micro-F1, except when aggregating over humans. For that case and other metrics, we perform macro aggregation.

Table 3. Evaluation of "any abnormality" ("Size"+"Focal"+"Diffuse"+"Postsurgical"+ "Absent") presence labeling in each organ for proposed LEAVS labeler, MAPLEZ [3] and SARLE [5] baseline, and the average of humans. Scores for the six organs with N_+>10 (liver, gallbladder, spleen, kidneys, pancreas, bowels) were aggregated using micro scores. Refer to Table 2 for table symbols.

Labeler	N	N_+	F1	LEAVS$_{sub}$F1	Pre	Rec	Spe	MCC
LEAVS	1200	381	.961(.946,.973)	-	.936	.987	.969	.942
MAPLEZ	1200	381	.938(.919,.954)**	.960(.946,.974)	.889	.992	.942	.909
SARLE	1200	381	.705(.673,.734)**	.960(.946,.974)	.570	.921	.678	.558
H Avg	-	-	.961(.952,.969)**	.976(.969,.983)	.975	.948	.984	.938

Table 4. Ablation study for LEAVS and comparison to the employment of other LLMs. "Finding type individually": employing individual finding type assessment questions for each finding type instead of multiple-choice questions; "Tree prompt (MAPLEZ)": employing the MAPLEZ prompt [3] for finding uncertainty assessment; "Fast sentence filtration": skipping the second step from sentence filtration; "RpH": reports processed per hour. All rows employed N=2000 and N_+= 394. Refer to Table 2 for table symbols.

Labeler	F1	RpH	Labeler	F1	RpH
LEAVS (Qwen 2 72B)	.892(.868,.913)	3.49	Llama 3.3 70B	.914(.893,.933)*	2.96
No CoT	.879(.853,.902)^{ns}	42.9	Qwen 2.5 72B	.877(.850,.899)^{ns}	2.80
Finding type individually	.874(.849,.898)**	3.57			
Tree prompt (MAPLEZ)	.879(.853,.901)^{ns}	3.73			
Fast sentence filtration	.893(.870,.915)^{ns}	10.3			
No sentence filtration	.866(.841,.891)*	4.62			

In Table 2, we evaluate the labeler. The MAPLEZ baseline employed the definitions of finding types from LEAVS. We provide Table 3 for evaluation against the rule-based labeler SARLE, which has a different label set. For SARLE, we considered only relevant finding types and included "Other" since it provided higher scores. An ablation study is presented in Table 4. It also shows the throughput of each method when using 2×A100 80 GB GPUs. We used the vLLM library [15] for inference, reducing the required time by 90%.

We employ the Kendall Tau-b correlation coefficient to evaluate urgency outputs in Table 5. Scores were calculated for the maximum urgency in each organ. Similarly to the AUC metric, differences in calibration do not impact these scores as long as the more urgent cases have a higher urgency than the less urgent cases. To evaluate differences in calibration, we present Table 6.

We trained the classifier by randomly splitting the AMOS-MM training set, which contained 1,287 reports and CT volumes, into training (80%) and validation (20%) sets, labeled by LEAVS. The final hyperparameters included a learning rate of 1e-3, a batch size of 512, the AdamW [17] optimizer, two sequential ResNet layers [8] with a dropout rate of 0.9 between them as the shallow classifier, binary cross-entropy loss, and the concatenation of maximum, minimum, and average pooling outputs. The model was validated every five epochs, and we employed the model with the best average validation AUC. The testing on the AMOS-MM validation set, labeled by the human labelers from reports, is presented in Table 7. In addition to data bootstrap, we included the variation of 5 random seeds in our statistics.

Table 5. Scores for urgency outputs from LEAVS and humans employing the Kendall Tau-b correlation coefficient (τ_b). For filtering results, instead of N_+, we use $N - N_{mode}$, where N_{mode} is the most common ground truth urgency. Refer to Table 2 for symbols.

Organ	Labeler	N	$N - N_{mode}$	τ_b	$LEAVS_{sub}\tau_b$
Liver	LEAVS	102	52	.612(.502,.705)	-
GBl	LEAVS	50	37	.635(.471,.759)	-
Kidneys	LEAVS	84	43	.632(.489,.734)	-
Bowels	LEAVS	45	35	.460(.214,.654)	-
Macro	LEAVS	-	-	.582(.503,.655)	-
Macro	H1	-	-	$.606(.500,.696)^{ns}$	.542(.409,.661)
Macro	H2	-	-	$.546(.446,.633)^{ns}$	.507(.393,.607)
Macro	H3	-	-	$.717(.621,.791)^{*}$	.556(.371,.682)
Macro	H4	-	-	$.581(.466,.674)^{ns}$	.492(.313,.645)
Macro	H5	-	-	$.336(.208,.444)^{**}$	.579(.444,.692)
-	H Avg	-	-	$.556(.505,.599)^{ns}$	.533(.453,.598)

Table 6. Prevalence of each urgency output. 0: normal/chronic/expected, 1: low urgency, 2: medium urgency, 3: high urgency. Refer to Table 2 for table symbols.

Labeler	$\%_0$	$\%_1$	$\%_2$	$\%_3$
LEAVS	4.6%	41.4%	38.4%	15.6%
H1	54.3%	20.8%	23.9%	1.0%
H2	65.1%	10.5%	24.1%	0.3%
H3	35.3%	36.3%	27.0%	1.4%
H4	36.2%	37.1%	22.6%	4.1%
H5	82.4%	8.2%	6.3%	3.1%

4 Discussion

LEAVS significantly surpasses the average human and beats two of the five human labelers in Table 2. It tends to show higher recall than humans, missing fewer mentions, but is less precise in applying medical definitions. The LLM achieves higher F1 scores when compared to humans (column $\text{LEAVS}_{sub}\text{F1}$) because the subsets on which humans are being evaluated are easier as they include only cases with agreement between the other two human labelers. LEAVS surpasses the LLM baseline, MAPLEZ, increasing the F1 score by 0.065 points.

The MAPLEZ evaluation used the type definitions we derived for LEAVS, which represent an additional contribution not reflected in the F1 score difference. Furthermore, LEAVS was validated in an anonymized private dataset, a domain different from the test AMOS-MM dataset, showing the potential of the prompt system to adapt to new domains. Matthews Correlation Coefficients (MCC) were included to evaluate whether F1 scores were inflated by class imbalance [4], but they did not reveal large gaps compared to the F1 scores.

Table 3 evaluates the labelers in an easier and less fine-grained task, as reflected in the higher scores. LEAVS significantly outperforms both baselines, SARLE and MAPLEZ. The SARLE performance is probably low because it does not account for all possible abnormalities with its rules and label set.

As shown in Table 4, sentence filtration and multiple-choice questions for finding-type assessment significantly improved results, whereas using CoT and multiple-choice questions for finding uncertainty assessment led to probable, but

Table 7. Scores of the vision classifier trained to predict several types of abnormalities for several abdominal organs. Refer to Table 2 for table symbols.

Organ	Type	N	N_+	AUC	Organ	Type	N	N_+	AUC
Liver	Diffuse	200	31	.755(.656,.840)	Spleen	Size	200	21	.927(.865,.968)
Liver	Focal	200	91	.678(.602,.744)	RKid	Focal	200	62	.602(.522,.684)
GBl	PS	200	11	.985(.953,1.00)	LKid	Focal	200	57	.508(.421,.591)
GBl	Diffuse	200	36	.758(.656,.850)	LBow	PS	200	20	.689(.583,.786)
GBl	Focal	200	27	.692(.576,.800)	LBow	Focal	200	38	.580(.473,.676)
					Macro	Macro	-	-	.716(.690,.743)

not statistically significant, improvements. Sentence filtration probably allows the model to focus on the important parts of the long reports: the average report in the AMOS-MM training set has 16 sentences and 1,400 characters. Although Llama 3.3 was not the best model in validation, it performed the best in the larger testing set and is a potential improvement to consider. This difference might be due to domain shifts between validation and test reports. This result also shows that the method is adaptable to at least one other LLM family.

The inference time is one limitation of LEAVS since 3.49 reports per hour can hinder use in large datasets or real-time applications. We prioritized labeling quality for this specific work. Table 4 shows one way to speed it up: fast sentence filtration. Speeding up inference with knowledge distillation [9] is a future effort.

The results from Table 5 show that the urgency labeling by the LEAVS method has approximately the same quality as the labeling from the average human labeler. However, human labelers vary greatly, with a Kendall Tau-b ranging from .336 to .717. Table 6 shows a considerable variation in calibration for humans, with the prevalence for the label "normal/chronic/expected" ranging from 35.3% to 82.4%. LEAVS deviated from human urgency calibration, labeling only 4.6% of cases as "normal" and assigning higher urgency levels more frequently. This bias suggests the model adopts a more cautious approach.

Table 7 shows that the vision model can learn to identify most evaluated finding types, with AUCs ranging from 0.508 to 0.985 and an average of 0.716. The AUCs have potential for future improvement, but we were able to show that the extracted information is learnable. Focal findings had the lowest scores, possibly due to coarse pooling across entire organs or the reduced resolution of UAE-S inputs ($2 \times 2 \times 2$ mm^3) and outputs ($4 \times 4 \times 4$ mm^3). The high performance for postsurgical findings in the gallbladder is likely due to the absence of organ in most of the postsurgical cases, resulting in embeddings full of zeros. Future work will investigate classification improvements from learnable embedding networks and pooling weights, as well as from training with new abdominal CT datasets that include associated reports. We also plan to explore visual attribution methods to check if models can weakly learn to localize focal findings.

5 Conclusion

The zero-shot use of LLMs with the LEAVS prompt system can successfully label abnormalities for several organs in abdominal CT reports, outperforming rule-

and LLM-based alternatives. A supervised vision model learned some information from the structured labels, showing potential for achieving general-purpose abnormality classification in abdominal CT.

Acknowledgments. This work was supported by the Intramural Research Programs of the NIH Clinical Center. Y.Z. is supported in part by the Eric and Wendy Schmidt AI in Human Health Fellowship Program at Icahn School of Medicine at Mount Sinai. This work utilized the computational resources of the NIH HPC Biowulf cluster. (`http://hpc.nih.gov`)

Disclosure of Interests. R.M.S.: Royalties from iCAD, ScanMed, Philips, Translation Holdings, PingAn, MGB; research support through a CRADA with PingAn. Other authors have no competing interests to declare relevant to this article's content.

Disclaimer. The views, information, or content, and conclusions presented do not necessarily represent the official position or policy of, nor should any official endorsement be inferred on the part of, the Clinical Center, the National Institutes of Health, or the Department of Health and Human Services.

References

1. Ankit Pal, M.S.: OpenBioLLMs: Advancing open-source large language models for healthcare and life sciences. `https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B` (2024)
2. Bai, X., Bai, F., Huo, X., et al.: UAE: Universal anatomical embedding on multi-modality medical images. Medical Image Analysis **103**, 103562 (2025). `https://doi.org/10.1016/j.media.2025.103562`
3. Bigolin Lanfredi, R., Mukherjee, P., Summers, R.M.: Enhancing chest X-ray datasets with privacy-preserving large language models and multi-type annotations: A data-driven approach for improved classification. Medical Image Analysis **99**, 103383 (2025). `https://doi.org/10.1016/j.media.2024.103383`
4. Chicco, D., Jurman, G.: The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics **21**(1), 6 (Jan 2020). `https://doi.org/10.1186/s12864-019-6413-7`
5. Draelos, R.L., Dov, D., Mazurowski, M.A., et al.: Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. Medical Image Analysis **67**, 101857 (2021). `https://doi.org/10.1016/j.media.2020.101857`
6. Fink, M.A., Bischoff, A., Fink, C.A., et al.: Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. Radiology **308**(3), e231362 (2023). `https://doi.org/10.1148/radiol.231362`, pMID: 37724963
7. Grattafiori, A., Dubey, A., Jauhri, A., other: The Llama 3 herd of models. CoRR **abs/2407.21783** (2024). `https://doi.org/10.48550/ARXIV.2407.21783`
8. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 770–778. IEEE Computer Society (2016). `https://doi.org/10.1109/CVPR.2016.90`
9. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. CoRR **abs/1503.02531** (2015). `https://doi.org/10.48550/arXiv.1503.02531`

10. Irvin, J., Rajpurkar, P., Ko, M., et al.: CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019. pp. 590–597. AAAI Press (2019). https://doi.org/10.1609/AAAI.V33I01.3301590
11. Isensee, F., Jaeger, P.F., Kohl, S.A.A., et al.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (Feb 2021). https://doi.org/10.1038/s41592-020-01008-z
12. Ji, Y., Bai, H., Yang, J., et al.: AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in neural information processing systems **35**, 36722–36732 (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/ee604e1bedbd069d9fc9328b7b9584be-Paper-Datasets_and_Benchmarks.pdf
13. Johnson, A.E.W., Pollard, T., Berkowitz, S.J., et al.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**(1), 317 (Dec 2019). https://doi.org/10.1038/s41597-019-0322-0
14. Kojima, T., Gu, S.S., Reid, M., et al.: Large language models are zero-shot reasoners. In: Koyejo, S., Mohamed, S., Agarwal, A., et al. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022), http://papers.nips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html
15. Kwon, W., Li, Z., Zhuang, S., et al.: Efficient memory management for large language model serving with pagedattention. In: Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles (2023). https://doi.org/10.1145/3600006.3613165
16. Larson, P.A., Berland, L.L., Griffith, B., et al.: Actionable findings and the role of IT support: Report of the ACR actionable reporting work group. Journal of the American College of Radiology **11**(6), 552–558 (2014). https://doi.org/10.1016/j.jacr.2013.12.016
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net (2019), https://openreview.net/forum?id=Bkg6RiCqY7
18. ProbeMedicalYonseiMAILab: medllama3-v20 (May 2024), https://huggingface.co/ProbeMedicalYonseiMAILab/medllama3-v20
19. Smit, A., Jain, S., Rajpurkar, P., et al.: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1500–1519. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.117
20. Team, Q.: Qwen2.5: A party of foundation models (September 2024), https://qwenlm.github.io/blog/qwen2.5/
21. Team, Q.: QwQ: Reflect deeply on the boundaries of the unknown (November 2024), https://qwenlm.github.io/blog/qwq-32b-preview/
22. Tushar, F.I., D'Anniballe, V.M., Hou, R., et al.: Classification of multiple diseases on body CT scans using weakly supervised deep learning. Radiology: Artificial Intelligence **4**(1), e210026 (2022). https://doi.org/10.1148/ryai.210026

23. Voltin, C., Dietlein, M., Kottlors, J., et al.: Evaluating GPT-4's performance in oncologic disease classification based on PET/CT reports of lymphoma patients: Are large language models the long-awaited 'magic bullet'? Nuklearmedizin **63**(02), P80 (Mar 2024), http://www.thieme-connect.com/products/ejournals/abstract/10.1055/s-0044-1782487
24. Vong, T., Rizer, N., Jain, V., et al.: Automated identification of incidental hepatic steatosis on emergency department imaging using large language models. Hepatology Communications **9**(3) (2025)
25. Wang, X., Peng, Y., Lu, L., et al.: ChestX-Ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 3462–3471. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.369
26. Wasserthal, J., Breit, H.C., Meyer, M.T., et al.: TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. Radiology: Artificial Intelligence **5**(5), e230024 (2023). https://doi.org/10.1148/ryai.230024
27. Wihl, J., Rosenkranz, E., Schramm, S., et al.: Data extraction from free-text stroke CT reports using GPT-4o and Llama-3.3-70B: the impact of annotation guidelines. European Radiology Experimental **9**(1), 61 (Jun 2025). https://doi.org/10.1186/s41747-025-00600-2
28. Yang, A., Yang, B., Hui, B., et al.: Qwen2 technical report (2024). https://doi.org/10.48550/arXiv.2407.10671