Automated Spinal MRI Labelling from Reports Using a Large Language Model

Robin Y. Park, Rhydian Windsor, Amir Jamaludin, and Andrew Zisserman

Visual Geometry Group, Department of Engineering Science,
University of Oxford, Oxford, UK
robinpark@robots.ox.ac.uk

Abstract. We propose a general pipeline to automate the extraction of labels from radiology reports using large language models, which we validate on spinal MRI reports. The efficacy of our method is measured on two distinct conditions: spinal cancer and stenosis. Using open-source models, our method surpasses GPT-4 on a held-out set of reports. Furthermore, we show that the extracted labels can be used to train an imaging model to classify the identified conditions in the accompanying MR scans. Both the cancer and stenosis classifiers trained using automated labels achieve comparable performance to models trained using scans manually annotated by clinicians[1].

Keywords: Radiological reports · Cancer · Metastasis · Stenosis.

1 Introduction

Labelling medical image datasets can be time-consuming and requires expert annotators, whose time is limited and expensive. This is compounded by a large number of medical conditions that can occur in any given image and often large inter-reader variability leading to noisy labels. This means researchers applying deep learning to medical imaging problems usually settle for small-scale datasets compared to other areas in machine learning, spend large amounts of funding on data collection, or restrict themselves to a few publicly-available datasets covering only a few conditions and modalities. A possible solution to this problem is to extract labels directly from radiological reports – free-text descriptions written by radiologists describing the findings of an imaging investigation. These reports can be downloaded in bulk from a hospital's electronic database; if the extraction can be automated, it would significantly reduce the annotation bottleneck and unlock much larger training datasets for solving medical imaging problems.

However, automated extraction of structured information from clinical reports is not a new idea and has proved to be challenging [18], due to large variability in reporting styles, heavy use of domain-specific vocabulary and frequently assumed knowledge (e.g. a report describing an investigation into the presence of metastasis implies the existence of a primary cancer). We propose

[1] Code can be found at https://github.com/robinyjpark/AutoLabelClassifier.

a general method for extracting structured labels for vision models from clinical reports. This is achieved by adapting general-purpose large language models (LLMs) by asking the model to summarise the report with a target condition in mind and produce a binary label based on the summary. Crucially, to obtain labels for a new medical condition, all that is required is the class name and a definition but no further fine-tuning.

To test the efficacy of this method, we apply it to spinal magnetic resonance imaging (MRI) radiological reports, aiming to label spinal cancer and stenosis. Adapting Zephyr (7B) and Llama3 Instruct (8B), two open-source question-answering LLMs, our method achieves balanced accuracy and F1 scores equal to or exceeding that of GPT-4 for both conditions. We then use the extracted labels to train a vision model to detect these conditions in spinal MRI. The classification networks trained using the automated labels are able to match the performance of existing classifiers trained using large volumes of expert-annotated images.

1.1 Related Work

Rule-based report parsers have shown strong performance in many settings (e.g. CheXpert [7], DeepSpine [11]). However, developing these parsers requires domain knowledge and cannot be easily adapted to new conditions, which conflicts with our aim of reducing reliance on expert input and manual annotations.

LLMs have become increasingly accessible to researchers with the emergence of open-source models such as Llama [19], Alpaca [17] and Mistral [8]. Methods like instruction fine-tuning [15] can improve models' alignment with question-answering tasks, and low rank adaptation (LoRA) [6] allows training large models with limited compute. Many existing models have been further pre-trained on biomedical data to increase the base model's familiarity with biomedical vocabulary and syntax [5, 10]. However, these are often trained using clinical abstracts and papers, which have a different form to radiological reports [9, 10, 12] or use reports from a single modality, e.g. Chest X-rays [1, 2]. RadBERT [22] is an example of a model pre-trained using a large corpus of diverse radiology reports; however, this is a 110M parameter BERT-based model and is thus unlikely to adapt to new tasks as well as much larger publicly-available general models.

Accordingly, as more powerful LLMs continue to emerge, there has been a growing focus on the development of generalisable methods that do not require specialised training on the task at hand. Steering GPT-4 by asking the model to come up with its own chain-of-thought has achieved state-of-the-art results, demonstrating significant gains in accuracy on medical question-answering over specialist fine-tuned models like Med-PaLM 2 [14]. While GPT-4 achieves excellent performance across many specialist domains, costs can be high when processing large datasets. Furthermore, closed-source LLMs accessible through APIs like GPT-4 require data to be uploaded to remote servers for processing, which poses a privacy risk, especially for medical reports containing potentially sensitive information. Finally, since weights are not publicly available, the model cannot be customised or incorporated into vision-language models downstream.

2 Extracting Structured Labels from Radiological Reports Using a Large Language Model

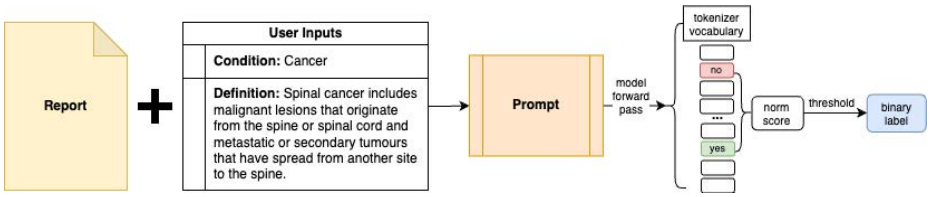

Fig. 1. Radiological report labelling pipeline: The prompt step formats the user inputs as shown in Figure 2 to summarise the report based on the target condition. Based on this summary, we extract the binary label using the normalised scores from a chosen set of two unique tokens ("yes" and "no") in the vocabulary.

Figure 1 shows an overview of our method to extract labels from clinical reports. The user specifies the condition for classification along with a definition to reduce ambiguity. These are input into a general-purpose prompting template (see Figure 2). Our method has two steps: (1) asking the model to generate a summary of the report based on the target condition, and (2) using the summary to assign a binary label. We also perform self-supervised fine-tuning to familiarise the model with the summary generation task, which is described in Section 2.2.

2.1 Model Prompting

We tested our prompting method using Zephyr-7B and Llama-8B Instruct, two instruction-fine-tuned language models that can support a wide range of use cases. To prompt the models, we provide a definition of the condition of interest. We tested two methods of prompting: (1) asking the model directly whether the patient has the condition based on the report, and (2) requesting that the model generates a summary of the report based on the condition of interest and decide whether the patient has the condition based on that summary. We formulated the prompt to extract binary labels on a given condition as seen in Figure 2. To generate these labels, we used the softmaxed logits of two unique tokens in the tokenizer's vocabulary: "yes" and "no"; see Figure 1. Using the token scores ensures that the model will produce a binary answer to every question.

2.2 Domain Adaptation By Summary Fine-Tuning

Radiology reports often include summary sections, which give overviews of the findings. These summaries typically include the most pertinent information, including potential diagnoses and descriptions of disease progression [7, 16]. Thus, the summary is especially relevant to consider when extracting clinical labels.

Direct Query Prompt	Summary + Query Prompt
<|begin_of_text|><|start_header_id|>system<|end_header_id|> You are a radiologist. your job is to diagnose **{{condition}}** using a medical report. Tell the truth and answer as precisely as possible.<|eot_id|> Report: **{{report}}** <|start_header_id|>user<|end_header_id|> **{{definition}}**. Based on the report, does the patient have **{{condition}}**? Answer 'yes' for yes, 'no' for no. Only output one token after 'ANSWER: '<|eot_id|> <|start_header_id|>assistant<|end_header_id|> ANSWER:	<|begin_of_text|><|start_header_id|>system<|end_header_id|> You are a radiologist. your job is to diagnose **{{condition}}** using a medical report. Tell the truth and answer as precisely as possible.<|eot_id|> Report: **{{report}}** <|start_header_id|>user<|end_header_id|> Can you write a summary for the report with the goal of diagnosing **{{condition}}**?<|eot_id|> <|assistant|> *[model-generated summary]*<|*eot_id*|> <|start_header_id|>user<|end_header_id|> **{{definition}}**. Based on the summary, does the patient have **{{condition}}**? Answer 'yes' for yes, 'no' for no. Only output one token after 'ANSWER: '<|eot_id|> <|start_header_id|>assistant<|end_header_id|> ANSWER:

Fig. 2. Model prompting strategies: The direct query method (left) asks the model to extract the label based on the report. The summary and query method (right) asks the model to generate a summary focused on the condition, which it uses as additional input to annotate the report. Words in bold indicate user inputs to be modified.

To ensure that the model would generate clinically relevant summaries, we fine-tuned the linear layers of Zephyr using LoRA [6] to perform next token prediction on spinal reports from a local hospital system. To do this, we identified the summary sections of 56,924 reports using regular expression matches for the following words (case-insensitive): conclusion, impression, findings, and summary (of the 124,771 reports available in the dataset, only 56,924 reports had matches for one of these words). While we fed the whole report to the model as context, the next-token-loss was computed only using the summary section of the report.

3 Supervised Learning Using Automated Labels

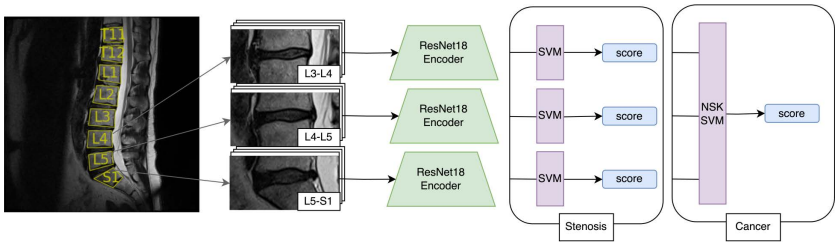

Fig. 3. MRI classification network: SpineNetV2 is used to detect IVDs. Each IVD is encoded using ResNet18. For stenosis, we use a SVM to get a score per IVD. For cancer, we aggregate IVD encodings and use NSK-SVM to get a score per scan.

To demonstrate that the labels automatically generated by our pipeline can be used to train a vision model, we used the generated labels to train a classifier to detect cancer and stenosis in the MR scans paired to the reports. SpineNetV2 [21] was used to automatically detect the vertebrae and extract the intervertebral discs (IVDs) in the T2 sagittal scans. Each IVD is of dimension

slice x height x width ($9 \times 112 \times 224$). We reshaped each slice to 224×224 to input them into a modified ResNet18 (pre-trained on ImageNet) without the fully connected layer to get slice-level embeddings. These were averaged across slices to get the volume-level representations. The embeddings were used as features to train support vector machines (SVM) with a linear kernel to perform binary classification for our conditions of interest. For stenosis, we could generate level-specific labels, so we detected stenosis at each IVD level. For cancer, the labels are provided at the spine-level but image samples are individual IVDs, so we used multiple-instance learning (MIL) using the Normalised Set Kernel method (NSK-SVM) to get the average representation across IVDs in a spinal scan [4]. Section 4 provides more information on the granularities of the report-generated labels. Figure 3 provides an illustration of the full pipeline.

4 Datasets

We use three datasets: **CancerData**, **StenosisData**, and **GeneralReports**. The first two contain report-image pairs while the last one has reports only.

Label Extraction: A subset of **CancerData** and **StenosisData** was manually labelled for prompt engineering experiments and testing our labelling pipeline. **CancerData** were provided by National Consortium of Intelligent Medical Imaging (NCIMI) and include reports and the associated MRIs of confirmed or potential cancer cases from six different NHS Trusts across the UK. **StenosisData** were collected as part of the Oxford Secondary Care Lumbar MRI Cohorts (OSCLMRIC) study and includes clinical MRI studies and reports of patients with lower back pain, sourced from a local hospital system. Each manually annotated dataset were a randomly chosen subset of the full source data, which we split into a calibration set (to develop our prompting strategy and determine a model calibration threshold) and a test set, using stratified sampling. Manual labels for **CancerData** were provided by a radiology registrar. Manual labels for **StenosisData** were provided by an orthopaedic surgeon. **GeneralReports** were also provided by OSCLMRIC and contain 56,924 unpaired spinal MRI reports with diverse conditions and control examples; it was used to fine-tune the model on the causal language modeling task to generate the summary. See Table 2 in Supplementary Materials for data splits for each task.

The reports were preprocessed based on use case. Spinal reports are often split into sections (see Supplementary Materials for an example). Since spinal cancers are usually metastases from a primary site elsewhere, clinical histories in **CancerData** frequently mention the presence of cancer in a non-imaged site and includes query words (i.e. "?cancer" to check for spinal cancer in the image). To correctly identify spinal cancer (i.e. the condition that would be visible in the paired image), we excluded clinical history when inputting reports from **CancerData** into the pipeline. As stenosis is specific to the spine, we did not perform any additional preprocessing on **StenosisData**. For the summary fine-tuning task, we identified the summary sections of the reports such that the rest of the report could be masked for loss computation (see Section 2.2).

Data	Split	Patients	Studies	IVDs		
				Total	Positives	Negatives
CancerData	Train (Scans)	1,223	1,256	14,167	9,012	5,155
	Validation (Scans)	324	327	3,612	2,149	1,463
	Test (Scans)	450	451	4,918	3,120	1,798
StenosisData	Train (Scans)	1,375	1,946	5,827	2,977	2,850
	Validation (Scans)	153	217	649	337	312
	Test (Scans)	117	123	368	139	229

Table 1. Summary of splits for MRI classification. Positive/negative labels on the training and validation sets were assigned using our labelling method described in section 5.1 whereas test set labels were manually annotated as described in Section 4. Test (Scans) in each dataset consists of studies from both Calibration and Test sets in Table 2 in Supplementary Materials where the IVDs were successfully extracted.

MRI Classification: For the MRI classification task, we used T2 sagittal images in a valid report-image pair in **CancerData** and **StenosisData**. **StenosisData** had only one sequence per study. **CancerData** could have multiple; in these cases, we chose the latest whole spine sequence in the study.

For stenosis, we generated labels for each IVD using our pipeline. This was feasible for stenosis as (1) reports almost always contained information about the level at which stenosis is present, and (2) report-level expert annotations were level-specific (for lumbar IVDs T12-S1). We only used the last three lumbar spine IVDs (L3-L4, L4-L5 and L5-S1) as stenosis examples above L3 are rare. It was less common for the cancer reports to list specific levels at which cancer is present, especially if present at multiple levels, which is common for both metastases and myeloma. As a result, we follow the approach in [20] and employ multiple instance learning, treating the entire spine as a bag which is labelled as positive if any of the vertebrae show cancer, or negative otherwise. We included IVDs from cervical to lumbar (18 IVDs C7-S1). Table 1 summarises the splits.

5 Results

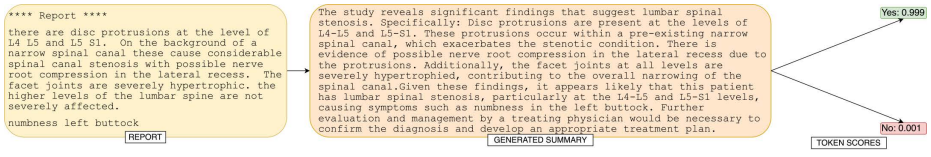

Fig. 4. Real example report from **StenosisData** with the summary and scores generated using our pipeline. Any dates, names or location information were removed from the report. Further examples can be found in the extended version of the paper.

5.1 Assessment of Labelling Accuracy

We evaluated the model's performance on label extraction by measuring area under the receiver-operator curve (AUROC) and equal error rate (EER) using the normalised scores from the model. To compute balanced accuracy and F1 score, we used the EER threshold from the calibration set.

Condition	Prompt Method	Summary Model	Query Model	Bal. Acc.	EER	AU ROC	F1
Cancer	Direct Query	-	GPT-4	**1.000**	-	-	**1.000**
	Direct Query	-	Zephyr	0.993	**0.000**	**1.000**	0.992
	Direct Query	-	Llama3	0.978	0.014	0.999	0.977
	Summary-Query	Zephyr	Zephyr	0.985	**0.000**	**1.000**	0.985
	Summary-Query	Z-SFT	Zephyr	0.994	**0.000**	**1.000**	0.993
	Summary-Query	Llama3	Llama3	**1.000**	**0.000**	**1.000**	**1.000**
Stenosis	Direct Query	-	GPT-4	**0.951**	-	-	**0.949**
	Direct Query	-	Zephyr	0.945	**0.037**	0.981	0.950
	Direct Query	-	Llama3	**0.963**	**0.000**	**0.987**	**0.962**
	Summary-Query	Zephyr	Zephyr	0.945	0.111	0.985	0.950
	Summary-Query	Z-SFT	Zephyr	0.933	**0.037**	0.943	0.937
	Summary-Query	Llama3	Llama3	0.945	**0.037**	**0.995**	0.950

Table 2. Scan-level report labelling performance on **CancerData** test set (n=145) and **StenosisData** test set (n=68). Normalised scores are used to compute EER and AUROC. Calibrated binary labels are used to compute balanced accuracy and F1. *Zephyr* refers to the base model without fine-tuning, whereas *Z-SFT* indicates the summary fine-tuned model (see §2.2). The top two best values per metric are bolded.

Table 2 shows the results of labelling each condition on the respective test sets. All methods exceed 0.9 balanced accuracy and F1. For cancer, the summary-query method using Llama3 achieves perfect balanced accuracy, AUROC and F1. For stenosis, the summary-query method using Llama3 achieves the highest AUROC; the direct query method using Llama3 achieves higher balanced accuracy and F1. In both tasks, our best methods outperform a direct query using GPT-4, prompted using the same format and input as the direct query strategy in Figure 2 but with the tags (e.g. user, system, assistant) formatted for GPT-4.

IVD Level	Bal. Acc.	EER	AUROC	F1
L3-L4	0.896	0.073	0.968	0.815
L4-L5	0.908	0.108	0.978	0.900
L5-S1	0.855	0.150	0.945	0.830

Table 3. IVD-level report labelling performance on expert-annotated **StenosisData** (n=68 patients, 204 IVDs). Normalised scores are used to compute EER and AUROC. Calibrated binary labels are used to calculate balanced accuracy and F1.

Since it achieved the highest AUROC across both tasks, the summary-query prompting method with Llama3 for summary generation and Q-A was used as the general method to extract binary labels for both MRI classification tasks. Figure 4 shows an example of the summary and token scores generated from our pipeline. For inference across the full datasets, we extracted scan-level labels for cancer and IVD-level labels for stenosis, as described in section 4. Table 3 shows the performance of our labelling pipeline for stenosis at each IVD.

5.2 Assessment of MRI Classification Accuracy

	Data	Model	Bal. Acc.	EER	AUROC	F1
(1)	**CancerData**	ResNet18+NSK-SVM	0.763	0.215	0.851	0.773
(2)	**StenosisData**	SpineNetV2 [21]	0.679	-	-	0.574
(3)	**StenosisData**	SpineNetV2 [21]+SVM	0.780	0.218	0.857	0.727
(4)	**StenosisData**	ResNet18+SVM	0.775	0.227	0.836	0.722

Table 4. MRI Classifier Performance on **CancerData** and **StenosisData** test sets. Calibrated binary labels are used to calculate balanced accuracy and F1. Only the first and fourth rows report results of models fully trained using our labels. (2) performs inference using SpineNetV2, which is trained using human annotations, and (3) uses SpineNetV2 to extract encodings and train an SVM using our report-generated labels. We treat (2) and (3) as baselines for our stenosis classification method (4).

We evaluated the model's performance on image classification by measuring area under the receiver-operator curve (AUROC) and equal error rate (EER) using the normalised scores from the model. We used the EER threshold from the validation set to derive binary labels from the scores on the test set, which were used to compute balanced accuracy and F1 score.

Stenosis: SpineNetV2 [21] has an existing grading model trained to classify three kinds of stenosis (central canal, foraminal right and left). We compare our method (ResNet18+SVM) to (1) an aggregated stenosis label from SpineNetV2 and (2) SVM using SpineNetV2 encodings (SpineNetV2+SVM). Our stenosis classifier outperforms aggregated SpineNetV2 stenosis scores and achieves a similar AUROC, balanced accuracy and F1 score after calibration for either choice of encoding with the SVM trained on automated labels (see Table 4).

Cancer: Considering IVD-level embeddings in aggregate, our classifier achieves a balanced accuracy of 0.763 in classifying cancer at the scan-level. SpineNetV2 does not classify cancer, so we are unable to provide a comparable baseline on the same dataset. A deep learning model to detect cancer in CT images achieved an F1 score of 0.72 [13], which we surpass (0.773).

5.3 Discussion and Limitations

There are several training and prompting methods we tested but did not include in our final methodology. While adding few-shot examples of desired Q-A pairs

when prompting LLMs is a common method to improve performance [3, 14], we found results to be highly sensitive based on the reports selected as examples. Futhermore, adding longer example medical reports often exceeds the models' context window (512 tokens for Zephyr, 8,000 tokens for Llama3). Interestingly, we found that fine-tuning Zephyr on report summmaries was only beneficial for the cancer labelling task (Table 2). We currently force the pipeline to output a positive or negative label without uncertainty. In future work, we plan to work directly with probabilities from our locally-run LLM to predict uncertainty.

6 Conclusion

We propose a general method that can be adapted to extract labels from radiological reports without additional model training. We show that this surpasses a strong GPT-4 baseline when applied to spinal MR reports, with the additional advantages that: (1) we use a locally-run open-source model that is privacy-preserving and cheap, and (2) we have direct access to the raw token-level scores, which can be used for model calibration to produce more accurate binary labels. We also demonstrate that the extracted labels can be used to train a classifiers with similar performance to models trained with expert-annotated scans. In the extended version of the paper, we include labelling results for two additional conditions: cauda equina and herniation.

Acknowledgments. We thank our clinical collaborators for providing annotations and advice: Prof. Jeremy Fairbank, Dr. Sarim Ather, Prof. Iain McCall, Dr. Mark Kong (in no particular order). We are also grateful to our funders: EPSRC CDT in Health Data Science (EP/S02428X/1), Cancer Research UK via the EPSRC CDT in Autonomous Intelligent Machines and Systems (EP/S024050/1), EPSRC programme grant Visual AI (EP/T025872/1), and the Oxford Big Data Institute. OSCLMRIC was collected as part of a Health Research Authority study (IRAS Project ID 207858), while **CancerData** was collected by NCIMI (Project ID 010A).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Bibliography

[1] Bannur, S., Hyland, S., Liu, Q., Pérez-García, F., Ilse, M., Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., Schwaighofer, A., Wetscherek, M., Lungren, M.P., Nori, A., Alvarez-Valle, J., Oktay, O.: Learning to exploit temporal structure for biomedical vision–language processing. In: CVPR (2023)

[2] Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., Poon, H., Oktay, O.: Making the most of text semantics to improve biomedical vision–language processing. In: ECCV (2022)

[3] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: NeurIPS (2020)

[4] Gärtner, T., Flach, P.A., Kowalczyk, A., Smola, A.J.: Multi-instance kernels. In: ICML (2002)

[5] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) **3**(1), 1–23 (2021)

[6] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022)

[7] 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)

[8] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7B. arXiv preprint arXiv:2310.06825 (2023)

[9] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)

[10] Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day (2023)

[11] Lu, J.T., Pedemonte, S., Bizzo, B., Doyle, S., Andriole, K.P., Michalski, M.H., Gonzalez, R.G., Pomerantz, S.R.: Deep spine: Automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning. In: Machine Learning for Healthcare Conference. pp. 403–419. PMLR (2018)

[12] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., Liu, T.Y.: BioGPT: generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics **23**(6) (2022)
[13] Motohashi, M., Funauchi, Y., Adachi, T., Fujioka, T., Otaka, N., Kamiko, Y., Okada, T., Tateishi, U., Okawa, A., Yoshii, T., et al.: A new deep learning algorithm for detecting spinal metastases on computed tomography images. Spine **49**(6), 390–397 (2024)
[14] Nori, H., Lee, Y.T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., King, N., Larson, J., Li, Y., Liu, W., et al.: Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. Medicine **84**(88.3), 77–3 (2023)
[15] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. In: NeurIPS (2022)
[16] Peng, Y., Wang, X., Lu, L., Bagheri, M., Summers, R., Lu, Z.: NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. AMIA Summits on Translational Science Proceedings **2018**, 188 (2018)
[17] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford Alpaca: An instruction-following LLaMA model. Tech. rep. (2023)
[18] Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W.: Large language models in medicine. Nature medicine **29**(8), 1930–1940 (2023)
[19] 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. Tech. rep. (2023)
[20] Windsor, R., Jamaludin, A., Kadir, T., Zisserman, A.: Context-aware transformers for spinal cancer detection and radiological grading. In: International Conference on Medical Image Computing and Computer Assisted Intervention (2022)
[21] Windsor, R., Jamaludin, A., Kadir, T., Zisserman, A.: SpineNetV2: automated detection, labelling and radiological grading of clinical MR scans. Tech. rep. (2022)
[22] Yan, A., McAuley, J., Lu, X., Du, J., Chang, E.Y., Gentili, A., Hsu, C.N.: RadBERT: Adapting transformer-based language models to radiology. Radiology: Artificial Intelligence **4**(4), e210258 (2022)