Structured Radiology Intelligence: Extracting Structured Data from MRI Reports Using LLMs

Sushvin Marimuthu, Parameswari Krishnamurthy, Dipti Misra Sharma
Goldwin H, Anu Eapen, Betty Simon, Anuradha Chandramohan
IIIT Hyderabad, IIIT Hyderabad, IIIT Hyderabad,
CMC Vellore, CMC Vellore, CMC Vellore, CMC Vellore
sushvin.m@research.iiit.ac.in, param.krishna@iiit.ac.in, dipti@iiit.ac.in,
change2.dm@gmail.com, anuepn@yahoo.com, drbettysimon@gmail.com,
Anuradha.Chandramohan@cmcvellore.ac.in

Abstract

This study presents efforts focused on extracting and structuring doctor notes, specifically Magnetic Resonance Imaging (MRI) reports, into a standardized format using large language models (LLMs). We introduce a novel benchmark dataset comprising of 55 clinically relevant variables given by doctors, making it the first of its kind in the automated processing of unstructured medical texts. The annotations to the dataset were generated using a systematic prompt-tuning approach that was manually validated. It was then evaluated across three experimental stages: baseline, intermediate, and fine-tuned. Each stage assessed the impact of different prompt strategies on the performance of various LLMs (LLaMA, Qwen, and DeepSeek). Among the models tested, LLaMA 3.1 8B Instruct consistently achieved the highest composite score in both the intermediate and final phases, resulting in an 18.42% improvement in performance.

Keywords: Unstructured to Structured, MRI Report, LLM, schema-guided extraction

1. Introduction

In healthcare, a Magnetic Resonance Imaging (MRI) radiology report is a detailed medical document written by a radiologist after analyzing MRI scan images (Hashemi et al., 2012). It describes the internal structures of the body, highlights any abnormalities, and provides a professional interpretation to guide further medical care. These reports are essential because they convert complex MRI images into meaningful insights that help the referring doctors to make accurate diagnoses, plan treatments, or monitor recovery. The report typically includes sections such as clinical history, findings, and impression, and it plays a key role in the patient's overall medical management (John and Mittal, 2025). While most MRI reports follow a general format, the doctor notes based on these reports are usually written in a semi-structured or unstructured manner, meaning the content is expressed in free-form language rather than fixed templates.

The objective of this study is to develop a data extraction system[1] that converts such semi-structured or unstructured doctor notes on MRI reports into a structured and standardized format, such as JSON. The system will identify and extract key clinical elements from the notes and organize them in a consistent structure. This structured data can then be used for various downstream applications, including integration with hospital information systems, support for clinical decision-making, large-scale data analysis, and the generation of tabular or other standardized report formats.

To develop this system, we began by constructing a novel benchmark dataset consisting of 55 clinically relevant variables defined by medical professionals. The details of the base dataset, along with its annotation process and validation using multiple LLMs under various settings, is described in detail in Section 2.

1.1. Related Works

Multiple efforts in the recent have focused on converting semi-structured and unstructured medical texts into structured data formats. For instance, (Ntinopoulos et al., 2025) investigates the use of large language models (LLMs) to automate entity extraction and classification from health records, enabling faster generation of structured data from clinical narratives. Similarly, (Ying et al., 2025) introduced GENIE, a Generative Note Information Extraction System that leverages LLMs to transform health records into standardized data formats. These two studies differ notably in the number of variables extracted and the LLMs employed. Additionally, (Skyles et al., 2025) explored structured data extraction from oncological notes, targeting six primary clinical questions using various prompt strategies.

In addition, a couple of efforts have been made in the domain of radiology report extraction and structuring. Among them, (Steinkamp et al., 2019) proposed a prototype system using traditional ma-

[1]The model is made available through this link: SRI

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 268–280
12 May 2026.

chine learning techniques to extract useful information from abdominopelvic radiology reports. More recently, (Di Palma et al., 2025) applied LLMs to extract key radiological features specifically from prostate MRI reports.

Our work however, focuses on extracting and structuring doctor notes based on the MRI reports. We extract all clinically relevant information, including details that may not be explicitly stated, resulting in a comprehensive, structured dataset that goes beyond the scope of previous works in both scale and depth. We utilized LLMs in our work due to their high accuracy compared to traditional rule-based or machine learning models. Additionally, they significantly reduce the time and effort required for manual annotation.

By designing diverse prompt strategies, namely, one-question, multi-question, and all-question prompting, combined with various prompting settings such as zero-shot, one-shot, and few-shot, we guided the LLMs to simulate structured data from unstructured notes on MRI reports provided by doctors. This approach enabled the creation of a high-quality synthetic dataset that closely mirrors real clinical annotations.

We began our experiments using a basic instruction prompt with the Llama 3.1 8B Instruct model (Grattafiori et al., 2024), which achieved a composite score of **61.16%**. Building upon this baseline, we enhanced our prompt design and expanded the evaluation to include 6 other LLMs. Among them, Qwen3 4B Instruct-2507 (Team, 2025) achieved the highest composite score at **64.49%**, followed closely by the Qwen2.5 7B Instruct at **63.84%**. While DeepSeek R1 Distill Llama 8B (DeepSeek-AI, 2025) and DeepSeek R1 Distill Qwen 7B (DeepSeek-AI, 2025) attained accuracies of **63.01%** and **53.94%**, respectively, Llama 3.2 3B Instruct (Grattafiori et al., 2024) and MMed Llama 3 8B (Qiu et al., 2024) recorded the lowest performance, with accuracies of **35.56%** and **34.86%**.

In our intermediate set of experiments, we enhanced the prompt and applied different prompt strategies as well as the settings on all models. Subsequently, in the final phase, we fine-tuned all models, with Llama 3.1 8B Instruct, achieving the highest composite score of **87.36%**

Thus, through this multi-level experimental setup, we created a structured dataset that can be effectively utilized to train and evaluate information extraction models, particularly for converting unstructured medical text into structured formats.

2. Dataset Creation

A major challenge in building an unstructured-to-structured data extraction system is the lack of publicly available labeled datasets containing doctor notes on MRI reports and their corresponding structured outputs. We used a total of 630 unstructured doctor notes on MRI reports collected from a large academic health center. The Institutional Review Board (IRB) of the institution granted a waiver of informed consent for the use of de-identified radiology reports.[2]

To convert these notes into a structured format, we extracted 55 variables from each report using the Llama 3.1 8B Instruct model, guided by a basic instruction prompt developed by the specific requirements provided by the medical team. The resulting structured data was labeled as V1 (Version 1).

To validate the composite score and clinical relevance of the extracted attributes, doctors manually reviewed and corrected the outputs for the first 60 reports. These validated annotations were recorded as V2 (Version 2). When there was a disagreement between the extracted data and the clinicians, we refined the prompts in consultation with them to align with clinical requirements.

Using the V2 data as a reference, we further experimented with various prompting strategies across multiple models to generate structured outputs for an additional 65 reports. This output was labeled as V3 (Version 3).

Subsequently, doctors reviewed and validated the V3 outputs, resulting in the final refined dataset, referred to as Version 4 (V4). In this dataset, the first 60 reports served as the gold standard test set for evaluating the performance of all prompting strategies and models. The remaining 65 reports were used as example data for constructing the various prompt strategies used in the study.

Additionally, we took another 300 samples (from 630 notes) for training the models. This was referred to as Version 5 (V5) and was also validated by the doctors. The initial round of all validations was performed by us, followed by a second round conducted by the doctors. To facilitate easy validation, we developed a web application that was used by both the clinicians and us. Below, we describe the various prompt-tuning strategies we employed to optimally extract all variables for the construction of the dataset.

2.1. Prompt-Strategies

We tested three main types of prompt settings: zero-shot, one-shot, and few-shot. Within each of these categories, we further experimented with three prompt strategies: one-question prompt, multi-question prompt, and all-question prompt, to extract the required variables from the reports.

[2]However, due to privacy concerns, we are not releasing the dataset as a part of this paper.

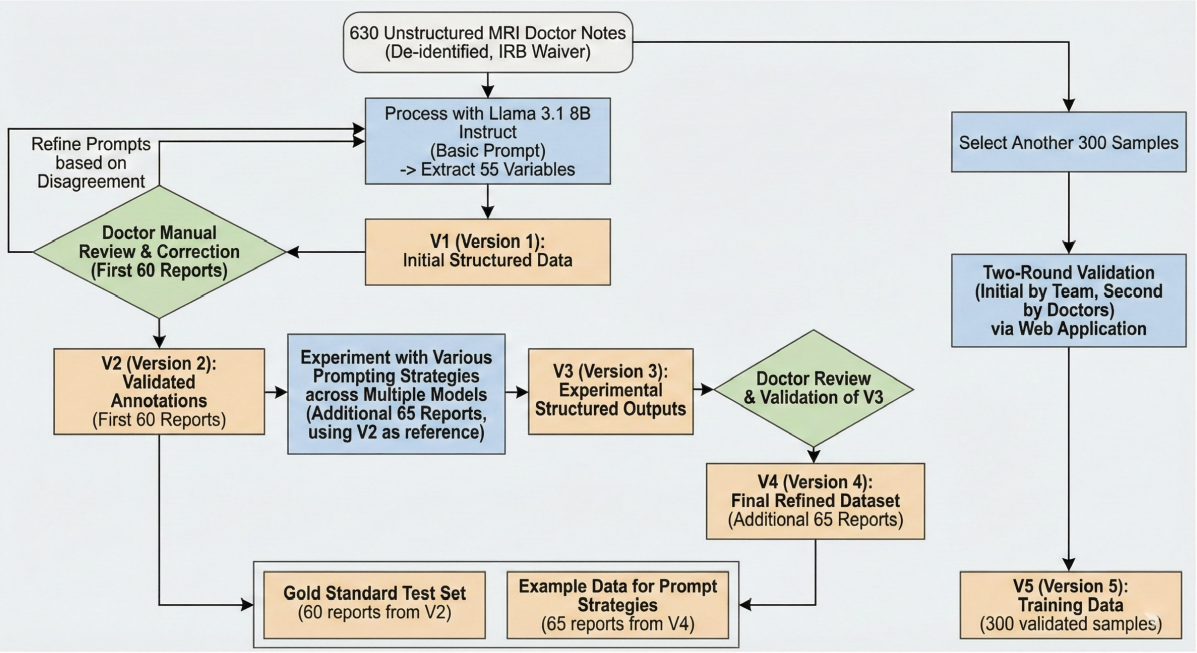

Figure 1: Dataset creation Flow-Chart

Dataset version	Size
Version 1 (V1)	630
Version 2 (V2)	60
Version 3 (V3)	65
Version 4 (V4)	125
Version 5 (V5)	300

Table 1: Summary of dataset versions and corresponding sample sizes

Initially, we planned to use the LLM to generate responses for all 55 variables by asking one question per prompt (shown in Baseline Prompt (Appendix A)). The idea was that generating the answers first using the LLM would make the annotation process easier and more efficient. For this, we used the Llama 3.1 8B Instruct model to generate the initial version of the data, which we referred to as V1. Then, doctors reviewed and refined only 60 reports in the initial data, resulting in a validated version referred to as V2. When we compared V1 with V2, the overall match composite score was **61.16%**. Then we experimented with multiple models using the same baseline prompt.

In this comparison, we observed several issues in the model-generated responses. The model often repeated numeric values unnecessarily, marked certain fields as 'not mentioned' even though they were clearly present in the report, and provided inconsistent answers for binary questions such as 'yes' or 'no'. We further extended our experiments to 6 other models (results summarized in Table 2). Across these models, we continued to observe similar issues such as repeated numeric values, incorrect 'not mentioned' labels, inconsistent answers to binary questions, and hallucinations.

To address these issues and improve the quality of the generated responses, we refined the prompting approach and adopted three different prompt strategies. They are discussed below:

1. **One-question prompt:** refers to a strategy where each of the 55 variables is derived from a separate, individual prompt. In this approach, we created a distinct question for each variable, resulting in one variable per prompt. Each prompt was designed to include all relevant information required to generate an accurate response, including contextual dependencies and supporting details related to that specific variable. This ensured that the model had sufficient context to produce a meaningful and precise answer for each prompt individually. The key distinction between the baseline prompt and the enhanced one-question prompt lies in the level of specificity provided for variable extraction. In the enhanced version, we explicitly defined how each variable should be identified and extracted, including cases where the information was implicitly stated rather than directly mentioned. Example: Enhanced One Question Prompt (Appendix A)

2. **Multi-question prompt:** refers to a strategy where related variables were grouped based on patterns and dependencies observed in the report responses. Instead of querying each variable individually, we combined several logi-

cally connected questions into a single prompt. These groupings were based on the nature of the answers, such as numeric values or binary (yes/no) responses, and their contextual interdependence. For instance, variables that are often mentioned together in the reports or rely on shared information were grouped to ensure the model had a broader context while still keeping the prompt concise. Example: Multi-Question Prompt (Appendix A)

3. **All-question prompt:** involves presenting all 55 questions within a single, comprehensive prompt. Each question in the prompt is designed to extract one specific variable, covering the entire set of 55 variables in one go. In this approach, we provided all the relevant information and instructions in a single prompt to ensure the model had complete context and access to all interdependent data points.

So far, all the prompt strategies were only applied in the zero-shot prompt setting, where we provided only variable extraction-specific instructions without including any examples for inputs or outputs. Therefore, we wanted to experiment with one-shot and few-shot prompt settings to assess the models' ability to generate accurate and relevant responses under varying levels of context and examples.

1. **One-shot setting:** For one-shot setting,
   - in the **one-question prompt** strategy, we provided one example for each prompt where the target variable was explicitly mentioned in the report. This helped the model understand how to extract that specific variable based on a concrete example.
   - in the **multi-question prompt** strategy, each prompt was accompanied by one example in which most of the grouped variables were clearly present in the report. This allowed the model to understand how to handle interdependent variables within a shared context.
   - in the **all-question prompt** strategy, we included a single comprehensive example containing a report where the majority of the 55 variables were explicitly mentioned. This example served as a reference for the model to understand how to extract all variables from a complete and information-rich report.
2. **Few-shot setting**: For the few-shot setting, we provided three examples for each prompt to help the model learn from a variety of scenarios and improve its generalization capability across different levels of report completeness
   - In the **one-question prompt** strategy, each prompt included:
     (a) An example where the target variable was explicitly mentioned in the report.
     (b) An example where the target variable was not mentioned in the report.
     (c) An example of the complete radiology report.
   - In the **multi-question prompt** strategy, each prompt included:
     (a) An example in which most of the grouped variables were explicitly mentioned in the report.
     (b) An example where some of the grouped variables were either not mentioned or marked as not applicable.
     (c) An example of the complete radiology report.
   - In the **all-question prompt** strategy, each prompt included:
     (a) An example radiology report where most of the 55 variables were mentioned.
     (b) An example where a subset of variables was either missing or not applicable.
     (c) An example of the complete radiology report.

The intermediate results are summarized in Table 3.

3. Finetuning

Despite experimenting with different prompt strategies across various prompt settings, the best-performing model, Llama 3.1 8B Instruct, achieved a maximum composite score of only **77.69%**. Therefore, to further improve performance, we proceeded with fine-tuning the models to determine whether additional gains in composite score could be achieved.

The V4 and V5 versions of the dataset, consisting of 125 + 300 notes, were utilized in the fine-tuning of the models. From these combined 425 notes, 345 were allocated for training, 20 for validation, and 60 for testing.

Fine-tuning was performed using each model's LoRA configuration and employed the one-question prompt strategy.

Overall, fine-tuning led to a clear improvement in performance, demonstrating measurable gains in composite score compared to the intermediate experiment. The detailed results are presented in Table 4.

4. Evaluation

To evaluate model performance, we applied 2 different metrics based on the type of the variable. For variables with numerical outputs, we computed an exact match score. The model's prediction for each variable was compared against the corresponding value in the gold-standard data, assigning a score of 1 for an exact match and 0 otherwise. The report-level score was then obtained by summing these scores across all numerical variables and dividing them by the total number of variables. For variables involving multi-label classifications, we calculated the F1 score using the same ground-truth annotations as the reference. We then calculate the average of the exact match and the F1 scores and report it as the **composite score**. Thus, this composite score forms the final result on the basis of which models are compared, and the best-performing model is noted.

$$\text{CS} = \frac{1}{2}\left(\frac{1}{N_{\text{num}}}\sum_{i=1}^{N_{\text{num}}}\mathbf{1}\left(\hat{y}_i = y_i\right) + \frac{2PR}{P+R}\right)$$

5. Results

We assessed the performance of the LLMs based on the type of prompt provided at each stage. For the baseline experiment, we evaluated the models using a simple instruction prompt (one-question prompt). The results of the baseline evaluation are summarized in Table 2.

In our intermediate experiment, we systematically evaluated the performance of the models across all three prompt strategies: one-question, multi-question, and all-question under zero-shot, one-shot, and few-shot settings. The detailed results of this evaluation are presented in Table 3.

Our intermediate experiment results show that:

- Under zero-shot setting, across all three prompting strategies, the Llama 3.1 8B Instruct model performed better than the other models. Using the one-question, multi-question, and all-question strategies, it achieved 73.19%, 72.96%, and 74.68%, respectively.
- Under one-shot setting, across all three prompting strategies, the Llama 3.1 8B Instruct model performed better than the other models. Using the one-question, multi-question, and all-question strategies, it achieved 73.98%, 71.94%, and 75.56%, respectively. However, Qwen2.5 7B Instruct performed comparably to Llama 3.1 8B Instruct on the one-question prompt under the one-shot setting.
- Under few-shot setting, the Llama 3.1 8B Instruct model outperformed the other models on the one-question and all-question strategies, achieving 74.91% and 77.69%, respectively. However, the MMed-Llama 3 8B model achieved 72.22%, and Llama 3.1 8B Instruct performed nearly the same as MMed-Llama 3 8B on the multi-question prompt under the few-shot setting.

LlaMA 3.1 8B Instruct achieved the highest composite score in both the baseline and intermediate experiments.

After fine-tuning, Llama 3.1 8B Instruct again outperformed all other models. Using the one-question, multi-question, and all-question strategies, it achieved 85.09%, 86.71%, and 87.36%, respectively. The overall report-level composite score reached **87.36%**, representing an improvement of nearly 18.42% over the baseline results.

Further, we also observed that the hallucination rate after finetuning was 0.

6. Discussion

We examine several errors observed in the models' performance across three different stages: baseline, intermediate, and fine-tuned. In the baseline setting, the models are provided with a simple instruction prompt and asked to extract both explicit and implicit variables. In the intermediate stage, the prompt is enhanced with more detailed guidelines along with examples to support variable extraction. And in the fine-tuned stage, the enhanced prompt is retained but the examples are removed, and the models are asked to extract the variables using only the refined instructions.

6.1. Baseline Stage

1. We observed that in the single-question prompt, the Llama models frequently returned "not mentioned" for certain variables even when answers were explicitly stated. For example, although the report explicitly mentioned DWI (Diffusion-weighted imaging) as 'dwi – restricted', the model still responded with 'not mentioned'.
2. The models also incorrectly responded "yes" for variables that were not affected by the tumour. For example, the report states: 'PROSTATE AND SEMINAL VESICLES – Normal,' indicating that the prostate and seminal vesicles are not affected by the tumour; however, the model incorrectly returned the response 'yes', implying involvement.
3. For numeric variables such as distance from the anal verge and distance from the ano-rectal junction, the models often repeated the 'tumour length' value in unrelated fields. For

Model	Single	Intermediate	All
Llama 3.2 3B Instruct	35.56%	47.78%	45.23%
Llama 3.1 8B Instruct	61.16%	64.81%	**68.94%**
MMed Llama 3 8B	34.86%	56.71%	51.25%
Qwen3 4B Instruct-2507	**64.49%**	65.79%	62.50%
Qwen2.5 7B Instruct	63.84%	**67.96%**	63.01%
DeepSeek R1 Distill Llama 8B	63.01%	61.02%	67.04%
DeepSeek R1 Distill Qwen 7B	53.94%	55.14%	43.47%

Table 2: Baseline Result: Comparison of Model composite score on baseline prompts

Prompt strategy	Composite Score		
	zero-shot	one-shot	few-shot
Llama 3.2 3B Instruct			
One-question	45.97%	44.31%	42.18%
multi-questions	34.68%	31.94%	35.00%
All-Questions	39.12%	33.38%	31.62%
Llama 3.1 8B Instruct			
One-question	**73.19%**	**73.98%**	**74.91%**
multi-questions	**72.96%**	**71.94%**	72.18%
All-Questions	**74.68%**	**75.56%**	**77.69%**
MMed-Llama 3 8B			
One-question	54.81%	51.94%	49.31%
multi-questions	71.02%	70.09%	**72.22%**
All-Questions	70.00%	71.44%	72.36%
Qwen 4B			
One-question	59.44%	58.94%	54.77%
multi-questions	68.15%	69.44%	71.16%
All-Questions	65.69%	63.10%	65.69%
Qwen 7B			
One-question	56.39%	58.89%	54.17%
multi-questions	68.80%	71.02%	71.81%
All-Questions	64.81%	64.63%	62.08%
DeepSeek Llama 8B			
One-question	69.54%	67.87%	70.46%
multi-questions	71.57%	67.73%	70.19%
All-Questions	70.23%	71.25%	69.17%
DeepSeek Qwen 7B			
One-question	53.94%	54.86%	53.61%
multi-questions	61.02%	63.01%	62.78%
All-Questions	41.94%	39.86%	44.49%

Table 3: Intermediate Result: Comparison of Model composite score on one-question, multi-question, and all-question under zero-shot, one-shot, and few-shot settings.

example, although the report explicitly stated ‘tumour length: 12 mm’, the model incorrectly returned ‘distance from the anal verge: 12 mm’ and ‘distance from the ano-rectal junction: 12 mm’.

4. In the multi-question prompt, the models returned "not mentioned" for some numeric variables while the correct answer should have been "nil significant."

5. Similar to the single-question prompting, the models also frequently returned “not mentioned” for most variables under the all-question prompting approach.

6. In both the single-question and multi-question prompts, the Qwen models frequently returned "not mentioned" for gender-related variables and stage-related variables. However, they produced appropriate outputs for numeric variables, correctly reporting the values along with

Model	Single	Multi	All
Llama 3.2 3B Instruct	45.42%	40.37%	47.59%
Llama 3.1 8B Instruct	**85.09%**	**86.71%**	**87.36%**
MMed-Llama 3 8B	79.26%	81.30%	80.79%
Qwen 3 4B Instruct	56.71%	65.79%	64.17%
Qwen 2.5 7B Instruct	62.31%	68.84%	70.32%
DeepSeek R1 Distill Llama 8B	73.89%	76.20%	75.93%
DeepSeek R1 Distill Qwen 7B	71.62%	74.58%	73.38%

Table 4: Finetuned Result: Comparison of Model composite score after fine-tuning.

their respective metrics.

7. On the other hand, DeepSeek models, particularly the DeepSeek version of the Llama model performed nearly as well as the Llama 3.1 8B Instruct model. However, both models occasionally produced Chinese characters in the output. Additionally, they often converted numbers and their corresponding units, especially for tumour length; for example, converting 1.2 cm to 11 mm or 125 mm to 12.5 cm. It was also observed that the DeepSeek models exhibited ambiguity when distinguishing among the following variables: tumour length, distance from the anal verge, and distance from the ano-rectal junction. They often returned the tumour length when asked for the distance from the anal verge or the distance from the ano-rectal junction, even when these values were explicitly stated. The models also consistently confused report types, including the first MRI report, restaging MRI, and others. For example, when the report was a first MRI, the other report types should have been marked as "no," but the models returned "yes" or "not mentioned." Similarly, when the report was a post-treatment MRI, they sometimes incorrectly marked it as a "first MRI report".

   These were some major reasons for a lower score in the baseline stage.

6.2. Intermediate

1. Although the prompts were enhanced, the Llama 3.1 3B Instruct model did not demonstrate any significant improvement across the three prompting settings (zero-shot, one-shot, and few-shot). In contrast, the MMed-LLaMA 8B model performed comparably to the Llama 3.1 8B Instruct model (which performed the best). The primary issue observed was that the models consistently returned “not mentioned” for MRI report stages. A similar pattern was identified for Tumour Regression Grade and treatment response, where the models frequently defaulted to “not mentioned.”

2. Qwen performed better than the other models including Llama 3.1 8B Instruct model when using the multi-question prompt. It produced more accurate outputs, particularly for tumour staging variables such as T-stage, N-stage, and M-stage. Additionally, it correctly identified CRM involvement, EMVI, tumour deposit (TD), and T4a status.

3. DeepSeek Llama performed well overall. However, it frequently generated unnecessary additional text, including explanations and repetitive content. Further, DeepSeek Qwen often reproduced the example outputs provided in the one-shot and few-shot prompts, rather than generating responses based on the actual input. Despite major improvements, the models still demonstrated confusion among the following variables: tumour length, distance from the anal verge, and distance from the ano-rectal junction. These values were interchanged, although they were explicitly stated in the report. Furthermore, when lymphadenopathy was clearly reported as “nil significant,” the models should have returned “nil significant” for internal iliac, obturator, inguinal, external iliac, common iliac, and para-aortic nodes. Instead, they often responded with “not mentioned” or, at times, generated incorrect random values from the examples.

6.3. Fine-tuned

1. After fine-tuning, not all the models showed substantial improvement. Although MMed-LLaMA 8B improved to some extent, its performance was still inferior to Llama 3.1 8B Instruct. Infact, Qwen showed a decline in output quality after fine-tuning, frequently returning “not mentioned” for most variables. Although the multi-question prompt continued to produce relatively good outputs, it did not demonstrate

significant improvement compared to the intermediate stage. In contrast, the all-question prompt showed slight improvement after fine-tuning, particularly in generating accurate numeric values.

2. Llama 3.1 8B Instruct produced the most accurate outputs across single-question, multi-question, and all-question prompting strategies. However, all the versions of Llama models still returned "not mentioned" for certain variables that were explicitly stated in the text. We infer that this issue may be related to training data imbalance. In our dataset, some variables such as levator ani, piriformis, and obturator (internus and externus) were predominantly labeled as "not mentioned" and occurred only rarely. Nevertheless, the models appeared to default to "not mentioned" even when these anatomical structures were clearly described in the report.

3. It was also observed that the DeepSeek models produced more accurate outputs for anal sphincter complex variables when compared to the intermediate stage. Moreover, the generation of unnecessary extra content was reduced drastically, and the responses became more concise, containing only minimal additional text. Metric and unit conversion errors were corrected, and hallucinations were largely eliminated. They also performed poorly in identifying post-treatment changes, such as hypointense changes, residual tumour with the same signal as the first MRI, and mucin reaction with a T2 hyperintense focus within a hypointense post-treatment change.

7. Conclusion

This study demonstrates the effectiveness of LLMs, particularly LLaMA 3.1 8B Instruct, in extracting structured information from unstructured MRI reports. A key contribution of this work is the creation of a novel dataset comprising of 55 clinically relevant variables, specifically tailored for doctor notes of MRI reports focused on tumors. Through the evaluation of prompt-tuning strategies across three experimental stages, we showcase the potential of LLMs to automate clinical data extraction with a high composite score. This dataset, along with our experimental findings, provides a valuable resource and foundation for future research in medical information extraction and supports the development of more reliable, scalable, and efficient NLP systems in healthcare.

8. Future Work

As a part of future work, we plan to extend the analysis to additional report types, including CT, biopsy, CEA, and surgical histopathology reports. Furthermore, we intend to incorporate a larger number of reports and curate additional cases in which most of the relevant variables are explicitly documented, to improve data balance and overall model performance.

9. Limitations

A significant proportion of the dataset contained variables labeled as "not mentioned," which introduced class imbalance across the variables in the dataset. Additionally, certain variable outputs were unevenly distributed. For example, within the T-Stage variable, the categories T3B and T4B appeared far more frequently than T0, T1, and T2, further contributing to class imbalance within the dataset. These were the only notable limitations identified.

10. Acknowledgements

We would like to express our sincere gratitude to Aisha Lakhani, Anju John, Joann Martha Mathew, Krishna Bharati, Lakshmi Meenakshi, Mac Win S, Mohanapriya A, Niveda Bharrath, Saloni Yadav, Shobiga Natarajan, and Sneha H, from CMC Vellore, who contributed their time and expertise to the validate the annotated data.

11. Bibliographical References

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Luca Di Palma, Fatemeh Darvizeh, Marco Alì, and Deborah Fazzini. 2025. Structured transformation of unstructured prostate mri reports using large language models. *Tomography*, 11(6):69.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret,

Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh

Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Ray Hashman Hashemi, William G Bradley, and Christopher J Lisanti. 2012. *MRI: the basics: The Basics*. Lippincott Williams & Wilkins.

Ajith John and Rohin Mittal. 2025. Role of the multidisciplinary team (mdt) meetings in rectal cancer management. *Journal of Gastrointestinal and Abdominal Radiology*, 08.

Vasileios Ntinopoulos, Hector Rodriguez Cetina Biefer, Igor Tudorache, Nestoras Papadopoulos, Dragan Odavic, Petar Risteski, Achim Haeussler, and Omer Dzemali. 2025. Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation. *BMJ Health Care Inform*, 32(1).

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine.

Ty J Skyles, Isaac J Freeman, Georgewilliam Kalibbala, David Davila-Garcia, Kendall Kiser, Silpa Raju, and Adam Wilcox. 2025. Exploring chatgpt 3.5 for structured data extraction from oncological notes. *AMIA Summits on Translational Science Proceedings*, 2025:518.

Jackson M Steinkamp, Charles Chambers, Darco Lalevic, Hanna M Zafar, and Tessa S Cook. 2019. Toward complete structured information extraction from radiology reports using machine learning. *Journal of digital imaging*, 32(4):554–564.

Qwen Team. 2025. Qwen3 technical report.

Huaiyuan Ying, Hongyi Yuan, Jinsen Lu, Zitian Qu, Yang Zhao, Zhengyun Zhao, Isaac Kohane, Tianxi Cai, and Sheng Yu. 2025. Genie: Generative note information extraction model for structuring ehr data.

A. Prompt Appendix

Baseline Prompt

You are an expert radiology information extraction model. Respond only in JSON.
Extract the n-stage from the radiology report.

OUTPUT FORMAT:
{'value': 'N0, N1, N1c, N2, or not mentioned'}

REPORT: {REPORT}

Enhanced One Question Prompt

You are an expert radiology information extraction model. Your task is to extract the current length of the tumour from the given radiology report text based on the detailed instructions below and only respond in JSON format.

Instructions:
You must follow these exact guidelines to generate answer:
1. If the radiology report clearly states phrases such as:
   - "No foci of T2 intermediate signal intensity or restricted diffusion in this segment"
   - "Likely post-RT changes"
   - "Post-treatment changes in the low rectum with no evidence of residual/recurrent tumor"
   → This indicates a complete radiologic response. In this case, return: "Not Relevant"
2. If tumour length is explicitly mentioned using terms such as:
   - "length"
   - "tumour size"
   - "tumour length"
   → Extract the numeric value of length of the tumour and its associated unit, ensuring there is a space between the number and the unit (e.g., "46 mm", "3.2 cm").
   → Always use numerals instead of words (e.g., "4 mm", not "four mm").
3. If the report is a post-treatment complete radiology report, and it only mentions scar size but not tumour size, extract the scar size as the current tumour length.
4. If neither tumour size nor scar size is mentioned, but the report qualifies as complete (based on the text), return: "Not Relevant"
5. If the tumour length is not mentioned, and the report is not complete, return: "Not Mentioned"

Multi Question Prompt

You are an information extraction system that can extract information from a given text and only respond in JSON format.
for 'crm_or_mrf_involvement':
1. If crm or mrf involvement is explicitly mentioned using terms such as:
   - "CRM - involved"
   - "CRM - 0 mm"
   - "mesorectal fascia - 9 mm"
2. return "Yes" if the distance to the mesorectal fascia is explicitly reported as less than 1 mm.
3. return "No" if the distance is explicitly reported as greater than or equal to 1 mm.
4. Return "Not Mentioned" if the distance to the mesorectal fascia is not mentioned in the text.
for 'emvi':
1. If EMVI is explicitly mentioned using terms such as:
   - "EMVI - present"
   - "EMVI - nil"
   - "EMVI - absent"
2. return "Yes" if EMVI is present and not mentioned as fibrosed, healed.

3. return "No" if EMVI is mentioned as fibrosed, healed, resolved, absent, or nil.
4. return "Not Mentioned" if EMVI is not explicitly mentioned.
for 'tumour_deposit':
1. return "Yes", if the report explicitly mentions the presence of tumor deposits, or uses terms like:
    - "tumor deposit (TD) - present"
    - "ENTD (extranodal tumor deposit) - present"
2. return "No" if the report explicitly mentions the absence of tumor deposits, or uses terms like:
    - "tumor deposit (TD) - absent"
    - "ENTD (extranodal tumor deposit) - absent"
3. return 'Not Relevant' if the input text is a complete radiology report.
4. return "Not Mentioned" if tumor deposit (TD) or ENTD (extranodal tumor deposit) are not explicitly mentioned.
for 'is_t4a':
1. return "Yes", if tumor invasion of the pelvic peritoneal reflection is explicitly mentioned.
2. return "No", if the pelvic peritoneal reflection is not involved, or confirms that the tumor is below the peritoneal reflection.
3. return "Not Mentioned", if the pelvic peritoneal reflection is not mentioned.

B. Example Appendix

Semi-Structured or Unstructured Report Example

MRI ABDOMEN AND PELVIS
FINDINGS:
RECTUM: Residual circumferential thickening of the low rectum with peri-rectal fat stranding , inferiorly extending to anorectal junction. Shows focus of T2 intermediate signal intensity with small focus of restricted diffusion in left lateral wall
- length - 15mm vs 20 mm previously
- extramural spread - 2.4 mm
- CRM - involved at 12 0clock position by tumor
- CRM - 0 mm - EMVI - nil
- highest margin of the tumor is below the peritoneal reflection
- no infiltration of puborectalis or anal sphincter complex
- no infiltration of vagina, uterus, cervix or bladder
- mesorectal nodes: nil significant
LYMPHADENOPATHY: para-aortic - Nil significant internal iliac - Nil significant external iliac - Nil significant inguinal - Nil significant
LIVER - No focal lesions in the liver, no IHBRD. SPLEEN - Normal GB - Distended, no wall thickening or calculi. CBD is of normal calibre. PANCREAS - Normal ADRENALS - Normal KIDNEYS - Normal BOWEL, MESENTERY, OMENTUM - Rectum as described.
Rest nil significant FLUID - nil
BLADDER - Minimally distended, normal
UTERUS - multiple fibroids, largest 15x13mm in the anterior wall
OVARIES - Nil significant
INGUINAL ORIFICES - Nil significant
ABDOMINAL WALL - Divarication of rectus noted.
BLOOD VESSELS -Nil significant
VISUALISED LUNG BASES -Nil significant
VISUALISED BONES -Nil significant
IMPRESSION: 70 year old lady with carcinoma rectum, post treatment on follow up, MRI shows
1. Residual circumferential thickening of the low rectum, extending to anorectal junction with T2 intermediate signal intensity and small focus of restricted diffusion CRM =0 EMVI - nil
2.No significant lymphadenopathy.
No significant change from previous MRI dated Feb 2023.
Suggest clinical evaluation and close follow up.

Structured Report Example


```
{
    "report_type": "MRI",
    "bio": {
        "age": "70",
        "gender": "Female"
    },
    "mri_numeric": {
        "current_length_of_tumour": "15 mm",
        "distance_from_anal_verge": "Not Mentioned",
        "distance_from_anorectal_junction": "0 mm",
        "extramural_spread_size": "2.4 mm",
        "mr_crm_distance": "0 mm",
        "number_of_mesorectal_nodes": "Nil Significant",
        "internal_iliac_nodes": "Nil Significant",
        "size_of_obturator_nodes": "Not Mentioned",
        "size_of_inguinal": "Nil Significant",
        "size_of_external_iliac": "Nil Significant",
        "size_of_common_iliac": "Not Mentioned",
        "size_of_para_aortic": "Nil Significant"
    },
    "mri_stage": {
        "is_first_mri_report": "No",
        "is_restaging_mri_report": "Yes",
        "is_mri_report_after_neoadjuvant": "Not Mentioned",
        "is_post_treatment_mri_report": "Yes",
        "is_post_operative_mri": "Not Mentioned"
    },
    "location_of_the_tumour: low"
    "t2_signal_intensity": "Intermediate",
    "dwi": "Restricted diffusion",
    "morphology": "Not Mentioned",
    "post_treatment_change": {
        "is_thick_t2_hypointense_band": "Not Mentioned",
        "is_thin_t2_hypointense_band": "Not Mentioned",
        "is_residual_tumor_as_first_mri": "Not Mentioned",
        "is_mucin_reaction_t2_hyper_hypo_post_treatment": "Not Mentioned"
    },
    "rectal_perforation": {
        "obstruction": "Not Mentioned",
        "perforation": "Not Mentioned"
    },
    "radial_extent": "Annular or Circumferential",
    "crm_or_mrf_involvement": "Yes",
    "emvi": "No",
    "tumour_deposit": "Not Mentioned",
    "is_t4a": "No",
    "adjacent_structures_t4b": {
        "female": {
            "uterus": "No",
            "vagina": "No",
            "ovaries": "No"
        },
        "male": {
            "prostate": "Not Applicable",
            "seminal_vesicles": "Not Applicable"
        },
        "male_and_female": {
            "levator_ani": "Not Mentioned",
            "puborectalis": "No",
            "piriformis": "Not Mentioned",
            "obturator_internus": "Not Mentioned",
            "obturator_externus": "Not Mentioned"
        }
    },
    "anal_sphincter_complex": {
        "internal_sphincter": "No",
        "t4b": {
            "external_sphincter": "No",
            "inter_sphincteric_plane": "No",
            "ischiorectal_foss": "No",
            "fistula_in_ano": "No"
        }
    },
    "t_stage": "T3b",
    "n_stage": "N0",
    "m_stage": "M0",
    "mr_trg": "MR TRG 2",
    "response": "Near Complete"
}
```