Artificial Intelligence-Powered Radiology

Tianshuai Gao
tg2935@columbia.edu
Columbia University, Department of Statistics
New York, NY, USA

ABSTRACT

In this work, we implemented an AI-powered radiology assistant designed to automate brain tumor diagnosis through MRI image classification, segmentation, and structured report generation. The system features a two-stage deep learning pipeline that combines TransUNet for tumor classification and 2D-UNet for precise tumor segmentation. To enhance clinical interpretability, structured information such as tumor size, location, and mass effect is extracted from the segmentation masks. A fine-tuned LLaMA language model, optimized with LoRA, is then used to generate coherent and context-aware diagnostic reports. Prompt engineering strategies—including few-shot, instruction-based, and role-based prompting—further improve the quality and structure of the generated content. Evaluation results demonstrate high classification accuracy (0.94), strong segmentation performance (Dice 0.85, IoU 0.76), and significant gains in BLEU, ROUGE, and BERTScore for report generation. Our system delivers a robust end-to-end automated pipeline while reducing computational cost and enhancing clinical utility. Future work will focus on integrating follow-up recommendation mechanisms, multi-modal learning, and an interactive user interface to facilitate real-world clinical deployment. My code can be found at https://github.com/tianshuai-gao/AI-Powered-Radiology-Assistant

ACM Reference Format:
Tianshuai Gao. 2025. Artificial Intelligence-Powered Radiology. In . ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Medical imaging plays a crucial role in diagnosis and treatment planning, especially when it comes to brain tumor MRI analysis. However, manual methods struggle with human limitations. Segmentation is time-consuming and labor-intensive, often leading to delays in critical decision-making. Radiologists rely on their experience to identify abnormalities, but in cases where they encounter something rare or that they've never seen before, there is a risk of missing important details. This is a potentially deadly consequence when it comes to brain tumors. There is also variability when it comes to the opinions of different radiologists about the same scan. AI-powered tools have the ability to address all of these limitations and enhance radiology-based medicine[1]. They have the unique power to leverage large datasets to recognize patterns beyond human capability and therefore ensure more accurate and consistent diagnosis. Faster AI-driven results mean patients can receive the treatment they need sooner, which will improve outcomes and reduce the burden on radiologists.

To address these challenges, in this project we have created an AI-powered radiology assistant designed to enhance radiologists' workflows by automating brain tumor classification and segmentation and using these results to generate detailed reports and treatment recommendations[2]. Our model consists of a deep-learning based image processing system that classifies and segments MRI scans, followed by a medical language model that interprets the findings and provides structured clinical insights[3]. Radiologists can use this to help them confirm their own interpretations or it can give them an indication to reconsider their possibilities if the model reaches different conclusions. By integrating AI into radiology analysis, we aim to increase efficiency, reduce diagnostic errors, and ultimately improve patient care.

2 METHODOLGY

To simulate and enhance the diagnostic capabilities of radiology experts, we designed and implemented an end-to-end AI-assisted diagnostic system that integrates medical image analysis with natural language generation. The system comprises two core modules: an image understanding module and a clinical report generation module.

In the image understanding stage, we constructed a two-stage deep learning pipeline. First, a TransUNet model is employed to classify brain MRI scans. This model combines convolutional neural networks (CNNs) for local feature extraction with a transformer architecture for global attention, enabling it to categorize the scans into four types: "No Tumor," "Glioma," "Meningioma," or "Pituitary Tumor." For scans identified as tumor-positive, a 2D UNet model is further applied to perform fine-grained tumor segmentation, generating high-resolution tumor masks that serve as structured visual evidence for subsequent analysis.

Once image analysis is complete, the structured classification and segmentation results are passed to a medical language model to generate more detailed and personalized diagnostic reports. We adopted the LLaMA architecture and fine-tuned it using LoRA (Low-Rank Adaptation), allowing the model to efficiently adapt to the medical language domain while maintaining strong language modeling performance with reduced computational cost.

To ensure the generated reports are professional and consistent, we incorporated a range of prompt engineering strategies, including few-shot examples, instruction-based prompts, and role-based prompting (e.g., "You are a radiologist"), guiding the model to produce reports with a clear structure, such as sections for Findings, Risk Assessment, and Treatment Recommendations. Through this

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-1-4503-XXXX-X/2025/06
https://doi.org/XXXXXXX.XXXXXXX

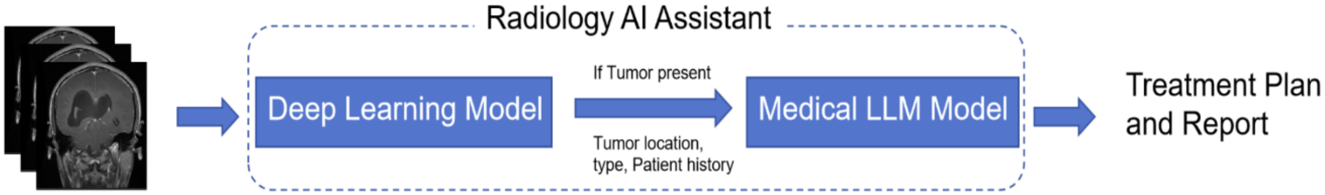


Figure 1: AI-Powered Radiology Assistant Workflow

multi-stage intelligent pipeline, we have built an AI system that emulates the diagnostic reasoning of radiology experts, delivering both accurate image analysis and semantically coherent, clinically valuable diagnostic reports.

3 DEEP LEARNING MODELS FOR CLASSIFICATION AND SEGMENTATION

The classification and segmentation of brain tumors in MRI images are crucial tasks in radiology. However, these manual methods are time-consuming and easily affected by human bias. Deep learning models, particularly those leveraging convolutional neural networks (CNNs) and transformer architectures, have shown promising results in automating and improving the accuracy of these tasks.

Our project employs a two-stage deep learning pipeline:

(1) **Tumor Classification with TransUNet** - Identifies whether a brain MRI scan contains a tumor and classifies it into one of four categories
(2) **Tumor Segmentation with 2D-UNet** - Generates precise tumor masks for cases where a tumor is detected

These models integrate CNN-based local feature extraction with transformer-based global attention mechanisms, giving robust performance on the medical imaging tasks.

3.1 Data Preparation

3.1.1 ***Dataset***.

We used the Kaggle Brain Tumor MRI Dataset, which contains 7,022 MRI scans categorized into four classes:

- No Tumor
- Pituitary Tumor
- Meningioma
- Glioma

These scans were collected from various medical sources, resulting in variations in image size, resolution, and noise levels.

3.1.2 ***Preprocessing***.

Because of the variations in the dataset, we created a preprocessing pipeline in order to ensure robust model performance. These steps included:

- Image Standardization
  - Resized all images to 256x256 pixels
  - Converted all images to grayscale (single channel -> 256x256x1)
- Normalization
  - Pixel values were normalized to [0,1] range
  - Histogram equalization was applied to enhance contrast
- Data Augmentation
  - Random flipping (horizontal and vertical) to increase variability
  - Gaussian noise injection
  - Morphological operations (erosion/dilation) to improve boundary detection

(should we add something here about training / testing split too?)

3.2 Model Architectures

TransUNet for Tumor Classification

For the classification task, we implemented TransUNet, a hybrid CNN-Transformer Model custom built in PyTorch. We chose to build this model because CNNs are really good at learning local features in images like edges, textures, and shapes, but they do not inherently understand relationships between distant regions in an image. This is a challenge in medical imaging because tumors often have complex, non-localized structures. Distinguishing between tumor and normal tissue requires understanding of the full context of the image. TransUNet addresses this limitation by combining the strengths of both CNNs and Transformers. [7]The CNN extracts fine-grained spatial features, while the Transformer captures global context and feature relationships across the entire image. Architecture:

- CNN Backbone: A ResNet backbone extracts local spatial features
- Transformer Encoder: The image is divided into small patches which are converted into embeddings. The self-attention mechanism then learns relationships between these patches, allowing the model to capture long-range dependencies and global context
- Feature Fusion: The CNN-extracted and Transformer-derived features are combined
- Fully Connected Classification Head: The model uses dense layers to classify the image as tumor type or non-tumor

2D-UNet for Tumor Segmentation

For the segmentation task, we implemented a customized U-Net model using TensorFlow/Keras. U-Net is one of the most widely used architectures for biomedical image segmentation due to its ability to capture both high-level context and fine-grained spatial details. The architecture is designed in a symmetric encoder-decoder structure, allowing it to learn hierarchical features while preserving spatial resolution through skip connections. We chose U-Net because brain tumor segmentation requires precise pixel-level delineation of lesions, which often have complex shapes and unclear boundaries. The encoder path extracts semantic features at multiple scales, while the decoder path reconstructs the original image resolution by combining low-level and high-level features. This design helps the model retain spatial information and achieve accurate segmentation even in the presence of small or irregular tumor regions. Architecture:

- Encoder: A series of convolutional layers with increasing depth and max pooling to capture hierarchical representations.
- Bottleneck: Deepest layer with high-level abstract features.
- Decoder: Transposed convolutional layers with skip connections to gradually recover spatial details.
- Output Layer**:** A 1×1 convolution followed by a sigmoid activation to produce binary masks for tumor regions.

The model was trained with a binary cross-entropy loss and evaluated using metrics such as Dice coefficient and Intersection over Union (IoU). To enhance robustness, dropout layers were included after each convolutional block. We also implemented model checkpointing based on validation Dice score to prevent overfitting.

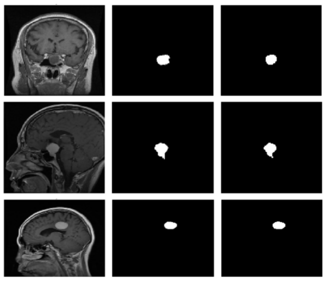


Figure 2: Examples of Brain MRI Slices masks and Corresponding Predicted Tumor Masks

3.3 Model Evaluation

TransUNet(for classification)
We achieved very good classification results from the TransUNet classification model. The following metrics were achieved:

- **Final Test Accuracy** : 94.2%
- **Precision** : 92.8%
- **Recall** : 94.0%
- **F1-Score** : 93.4%

This high test accuracy suggests that our model generalizes well to unseen data and makes reliable predictions. The precision and recall is also balanced. The high precision suggests that when the model predicts a tumor, it is correct most of the time. This means fewer false positives, which is important for avoiding unnecessary concern and further invasive tests. Recall shows the model's ability to correctly identify actual tumor cases, and its high value shows there are few false negatives. This means there are low chances of missing a malignant tumor, which is essential since missing a tumor could have deadly consequences. This is also important because the classification model must detect a tumor in order to trigger the segmentation model to run. The F1-score is also strong which shows that the model maintains a good balance and is not skewed towards either over or under-detecting tumors.

U-Net(for segmentation)
We achieved strong segmentation performance using the U-Net model. The following metrics were obtained on the validation set:

- **Validation Accuracy**: 99.0%
- **Dice Coefficient**: 0.85
- **IoU Score**: 0.76

These results show that the model effectively segments tumor regions with high spatial accuracy. The Dice coefficient indicates strong overlap between predicted and actual tumor areas, which is essential for capturing tumor shape and size. A high IoU score further supports the model's precision and consistency. While pixel-wise accuracy is high, Dice and IoU are more reliable for segmentation evaluation, especially in medical imaging. Overall, the U-Net model delivers accurate and consistent results, making it well-suited for clinical support tasks like tumor localization and volume estimation.

3.4 Extraction of Text Information from Predicted Segmentation Masks Results

In the segmentation module, the U-Net model performs pixel-level prediction on input brain MRI images to generate binary masks that delineate tumor regions. These masks clearly indicate all pixels that the model identifies as belonging to pathological areas. To further enhance the clinical interpretability of the segmentation output, we designed an information extraction pipeline that converts the mask into a structured textual report. The system analyzes the following tumor-related indicators:

- Tumor Size: By counting the number of foreground pixels in the predicted mask (i.e., those classified as tumor) and applying the known pixel spacing of the MRI (e.g., 0.2 mm/pixel), the system calculates the tumor area in square centimeters. Tumor size is a crucial quantitative metric for assessing the severity of the lesion and plays an important role in determining whether surgical intervention is needed or if continued observation is more appropriate.
- Tumor Location: The system computes the centroid of the predicted mask and compares its position relative to the center of the brain image to determine the anatomical quadrant in which the tumor resides. In the current design, the location is categorized into the Left/Right Parietal Region or Left/Right Temporal Region. This regional-level spatial information helps radiologists and neurosurgeons in planning the next steps for diagnosis or treatment.
- Mass Effect and Midline Shift: To assess mass effect, the system compares the tumor centroid's horizontal position with the brain midline. If the displacement exceeds a predefined threshold (e.g., 10 pixels), it is flagged as significant mass effect. This indicates that the tumor is exerting pressure on surrounding brain structures, which may result in elevated intracranial pressure or even brain herniation—conditions requiring urgent attention.
- Signal Characteristics & Growth Pattern: Based on the tumor type identified by the classification model (e.g., meningioma, pituitary tumor, glioma), the system appends predefined textual templates that describe the tumor's imaging features

in various sequences (e.g., isointense on T1, hyperintense on T2) and its anatomical growth pattern (e.g., extra-axial lesion with clear boundaries and focal expansion). These descriptions mimic the language used in radiological reports, enhancing the readability, completeness, and clinical utility of the generated output.

3.5 Deep Learning Part Program Process

We implemented a complete pipeline for automated brain tumor analysis from MRI scans, including tumor classification, region segmentation, and clinical report generation. The process is organized as follows:

(1) **Model Loading and Environment Setup**
The workflow begins by loading two pretrained models:
- A **TransUNet-based classification model** for tumor detection and type identification
- A **U-Net segmentation model** for delineating tumor regions
Required dependencies are imported, and hardware (CPU/GPU) configuration is set.

(2) **Image Upload and Preprocessing**
The user uploads a brain MRI image (in `.jpg` or `.png` format). The image is converted to grayscale and resized (typically to 256×256 pixels) to match the input requirements of the models.

(3) **Tumor Classification**
The classification model predicts whether the image contains a tumor and, if so, classifies it into one of four categories: **glioma**, **meningioma**, **pituitary**, or **notumor**.

(4) **Conditional Segmentation Execution**
If the tumor type is classified as "notumor", the pipeline stops and reports no abnormality. Otherwise, the U-Net segmentation model is triggered to produce a binary mask that highlights the tumor region at the pixel level.

(5) **Tumor Property Extraction**

(6) **Tumor Property Extraction**
Post-processing is applied to the binary mask to extract structured information, including tumor size, tumor location, mass effect, and signature characteristics and distribution pattern. This information is then combined with the tumor type predicted by the classification model to form a comprehensive textual report. The resulting text serves as an input for downstream large language models (LLMs), enabling the generation of more personalized, clinically relevant, and context-aware reports or training data for medical AI applications.

Post-processing is applied to the binary mask to extract structured information, including tumor size, tumor location, mass effect, and signature characteristics and distribution pattern. This information is then combined with the tumor type predicted by the classification model to form a comprehensive textual report. The resulting text serves as an input for downstream large language models (LLMs), enabling the generation of more personalized, clinically relevant, and context-aware reports or training data for medical AI applications.

4 MEDICAL LLM MODEL FOR STRUCTURED MRI REPORTS

The rapid advancement of Large Language Models (LLMs) has revolutionized various domains, including healthcare. In this project, we employ the **LLaMA (Large Language Model Meta AI)** architecture enhanced with **LoRA (Low-Rank Adaptation)** to generate structured MRI diagnostic reports.

Our approach addresses two major challenges in medical report generation:

- **Specialized Medical Language:** Medical reports often contain domain-specific terms, abbreviations, and diagnostic language patterns. Fine-tuning LLaMA with LoRA effectively adapts the model to the medical domain.
- **Coherent Report Structure:** Generating MRI reports requires structured content with clear sections (e.g., Findings, Risk Assessment, Treatment Plan). We address this through strategic **Prompt Engineering**.

The key innovations in our Medical LLM Model include:

- Integration of **Few-shot**, **Instruction-based**, and **Role-based** prompting strategies to improve content precision.
- Efficient fine-tuning using **LoRA**, reducing memory consumption while maintaining strong performance.

The following subsections describe the data preparation process, model design, and performance evaluation.

4.1 Data Preparation

The dataset used for our MRI report generation model was collected from the **Open-i API**, a platform developed by the *National Library of Medicine (NLM)*. This database was chosen for its extensive collection of MRI-related research publications, ensuring our model could effectively learn specialized medical language patterns.

To efficiently gather MRI-related data, we developed an automated data retrieval pipeline using Python’s `requests` library. The Open-i API imposes a limit of 10 records per request; thus, pagination logic was implemented to iteratively accumulate data until reaching the target of **100,000 records**.

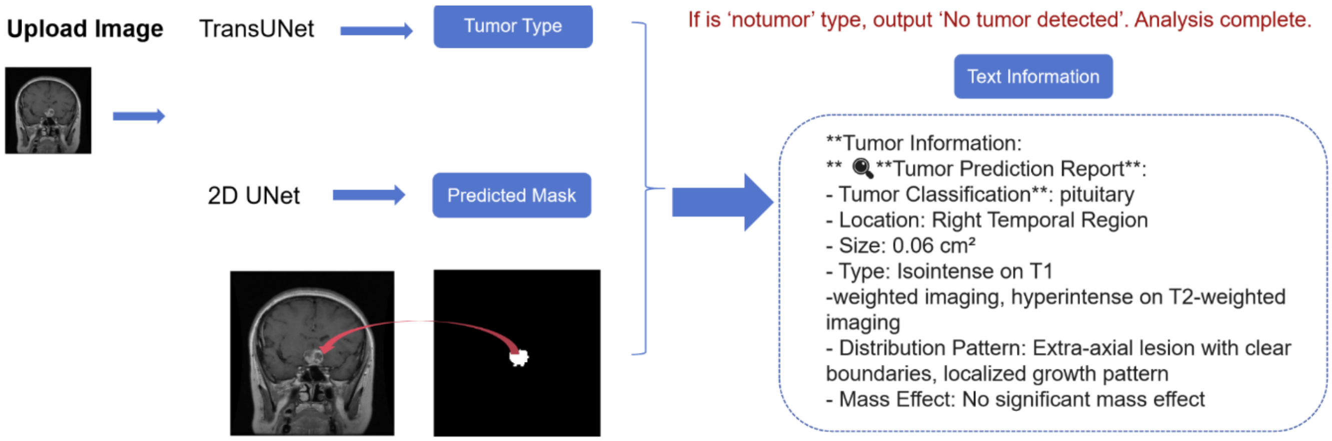
Each entry included key attributes such as:

- **Title:** Title of the MRI-related publication.
- **Authors:** List of contributing authors.
- **Journal:** Name of the publishing journal.
- **Publication Date:** Standardized to the format `"DD MM YYYY"`.
- **PMC URL:** Link to the original PMC article.
- **MRI Findings:** Diagnostic observations related to MRI scans.
- **Image URL:** Link to MRI images (if available).

Missing values were handled by assigning `"N/A"` placeholders to incomplete records to ensure data integrity. Additionally, text entries were lowercased, and redundant spaces were removed for consistency.

4.1.1 Data Summary and Impact.

The resulting dataset comprises **100,000 MRI-related records**, ensuring coverage across various tumor types and diagnostic patterns. This dataset was crucial in improving the model’s generalization capabilities, particularly in generating structured MRI diagnostic reports.


Upload Image
TransUNet
Tumor Type
2D UNet
Predicted Mask
If is 'notumor' type, output 'No tumor detected'. Analysis complete.
Text Information

**Tumor Information:
** 🔍**Tumor Prediction Report**:
- Tumor Classification**: pituitary
- Location: Right Temporal Region
- Size: 0.06 cm²
- Type: Isointense on T1
-weighted imaging, hyperintense on T2-weighted imaging
- Distribution Pattern: Extra-axial lesion with clear boundaries, localized growth pattern
- Mass Effect: No significant mass effect

Figure 3: Deep Learning Part Program Workflow

By leveraging medical literature directly from peer-reviewed sources, the dataset improved our model's ability to:

- Accurately recognize MRI-specific terminology.
- Generate coherent and structured diagnostic summaries.
- Produce medically relevant explanations aligned with clinical standards.

4.1.2 *Challenges*.

Several challenges emerged during data preparation:

- **Incomplete Metadata:** Some entries lacked author information or MRI findings. Placeholder values such as "N/A" were assigned to preserve these records.
- **Inconsistent Text Formats:** MRI findings varied significantly in length and structure. To address this, text cleaning techniques were applied to improve consistency in the final dataset.

Despite these challenges, our data preparation pipeline successfully produced a high-quality dataset that was pivotal in achieving strong model performance.

4.2 Model Design

4.2.1 *Fine-Tuned LLaMA with LoRA*.

For MRI report generation, we employed the **LLaMA (Large Language Model Meta AI)** architecture enhanced with **LoRA (Low-Rank Adaptation)** fine-tuning. This approach was chosen for its efficiency in adapting large-scale language models to specialized domains such as medical text generation.

4.2.2 *Why LoRA?*.

Fine-tuning large language models is often computationally expensive and memory-intensive. LoRA addresses this challenge by introducing low-rank matrices to efficiently update only a small subset of the model's parameters. This drastically reduces the number of trainable parameters while retaining strong performance. This selective fine-tuning allowed us to achieve strong performance with minimal computational overhead.

4.2.3 *Data Preparation for Training*.

To prepare data for model fine-tuning, we constructed a structured dataset of 100,000 MRI-related records. Each entry included:

- **Input Prompt:** "Write a radiology report for '*Title*'."
- **Output:** The corresponding MRI findings from the dataset.

The dataset was split into a **90% training set** and a **10% evaluation set** to facilitate robust model evaluation.

4.2.4 *Training Configuration*.

To optimize resource usage, we leveraged **4-bit quantization** via the BitsAndBytes library. This strategy significantly reduced the model's memory footprint while preserving precision. During training, the model checkpoints were saved every 500 steps to ensure recovery options during interruptions.

4.3 Model Evaluation

The fine-tuned LLaMA model underwent training for **18 epochs**, achieving stable convergence with no signs of overfitting. The training process and final evaluation results are summarized as follows.

4.3.1 *Training Process Overview*.

The training and validation loss curves are presented in Figure 4. The model's training loss decreased from **0.23** to **0.0076**, while the validation loss stabilized at **0.0072**, indicating effective convergence.

These metrics highlight the model's computational efficiency and demonstrate that LoRA fine-tuning effectively reduced the number of trainable parameters without compromising performance.

4.3.2 *Discussion*.

The stable convergence pattern and low final loss value confirm the model's ability to generalize effectively without overfitting. This outcome underscores the success of LoRA in efficiently adapting the LLaMA model to medical language tasks.

4.4 Model Performance Comparison

To assess the effectiveness of our fine-tuned LLaMA model, we conducted a comprehensive evaluation using key NLP metrics:

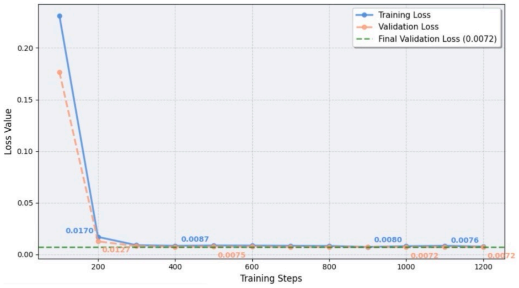


Figure 4: Enhanced Training Loss vs Validation Loss. Stable convergence suggests no overfitting.

BLEU, ROUGE, and BERTScore. Results were obtained by averaging outcomes from **100 independent runs** to ensure statistical robustness.

4.4.1 *Evaluation Metrics.*

The following metrics were employed to evaluate model performance:

- **BLEU Score:** Measures text fluency and word overlap with reference content. Higher scores indicate improved content accuracy.
- **ROUGE-1, ROUGE-2, and ROUGE-L:** Assess recall across unigrams, bigrams, and longest common subsequences to evaluate content coherence and structure.
- **BERTScore:** Measures semantic similarity between generated and reference texts using contextual embeddings. Higher scores indicate better understanding of complex language.

4.4.2 *Results and Analysis.*

The comparison of the base model and the fine-tuned model is visualized in Figure 5. Key observations include:

- The fine-tuned model achieved substantial improvements in **BLEU** (+100.7%) and **ROUGE-1** (+79.5%), highlighting its enhanced ability to produce fluent and accurate text.
- The fine-tuned model also demonstrated improved performance in **ROUGE-2** (+62.9%) and **ROUGE-L** (+51.7%), indicating better sentence coherence and structure.
- The **BERTScore** improved moderately (+5.3%), reflecting increased stability in semantic similarity performance.

4.4.3 *Discussion.*

The performance improvements observed across all metrics suggest that fine-tuning the LLaMA model with LoRA effectively enhanced its capability to generate accurate, fluent, and medically relevant MRI diagnostic reports. The most notable improvement occurred in the **BLEU Score**, which doubled in value, indicating the fine-tuned model's enhanced ability to produce coherent medical language.

These gains demonstrate that LoRA's efficient parameter adaptation effectively improved the model's ability to capture domain-specific terminology while minimizing computational overhead.

Future improvements may focus on:

- Incorporating larger MRI datasets to further improve model generalization.

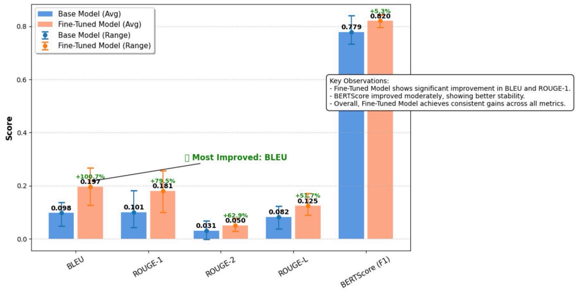


Figure 5: Performance Comparison between Base Model and Fine-Tuned Model.

- Implementing domain-specific prompts to better align model outputs with radiological standards.
- Exploring multi-modal learning strategies to combine MRI images with textual content for enhanced diagnostic precision.

4.5 Prompt Engineering

To enhance the performance of our fine-tuned LLaMA model in generating structured MRI diagnostic reports, we applied various **Prompt Engineering** techniques. These included **Few-shot Prompting**, **Instruction-based Prompting**, and **Role-based Prompting**. Each technique played a distinct role in improving the model's understanding of medical language and generating well-structured outputs.

4.5.1 *Few-shot Prompting.*

Few-shot prompting leverages multiple high-quality examples to guide the model's response. This method enhances the model's ability to generate MRI findings by providing structured exemplars that align with medical standards.

Example Prompt:

> **Imaging Description:** MRI scan shows a well-defined extra-axial mass in the left parietal region, measuring 3.0 cm. The lesion is isointense on T1-weighted imaging and hyperintense on T2-weighted imaging. The mass exhibits a clear boundary with localized growth extending along the dura. There is no evidence of surrounding edema or significant mass effect.
>
> **Example Output (Follow this format):** - Tumor Classification: Meningioma - Location: Left parietal region - Size/Extent: 3.0 cm - Type: Isointense on T1-weighted imaging, hyperintense on T2-weighted imaging - Distribution Pattern: Extra-axial lesion with clear boundaries, localized growth pattern - Mass Effect: No significant mass effect or surrounding edema

Impact: Few-shot prompting effectively improved the model's ability to generate coherent, structured medical text by mimicking standard MRI report templates.

4.5.2 *Instruction-based Prompting.*

Instruction-based prompting enhances model guidance by providing explicit step-by-step instructions. This technique was particularly effective in standardizing the 'Risk Assessment' section.

Example Prompt:

> **Example Output (Follow this format):** - Risk Score: [Total score out of 100] - Risk Level: [Low/Medium/High] - Justification: - Tumor Classification: [Explain contribution to risk score] - Location: [Explain contribution to risk score] - Size/Extent: [Explain contribution to risk score] - Distribution Pattern: [Explain contribution to risk score] - Mass Effect: [Explain contribution to risk score]

Impact: Instruction-based prompting provided precise control over output format, improving consistency across generated reports. The model consistently followed the required structure, enhancing readability for medical professionals.

4.5.3 *Role-based Prompting*.

Role-based prompting assigns the model a defined "role," prompting it to emulate the behavior of a radiologist or medical expert. This technique improved the model's ability to align generated content with professional language and clinical standards.

Example Prompt:

> **MRI Findings:** - Tumor Classification: Glioma - Location: Right frontal lobe - Size/Extent: 4.5 cm - Type: Hyperintense on T2-weighted imaging with irregular margins and heterogeneous enhancement - Distribution Pattern: Intracranial mass with irregular margins, causing significant midline shift - Mass Effect: Significant vasogenic edema present, with approximately 5 mm midline shift
> **Risk Assessment:** High Risk (Score: 70/100)
> **Output Format:** - Imaging Recommendations: [Recommended imaging follow-ups] - Follow-up Schedule: [Recommended MRI review timeline] - Treatment Suggestions: [Surgical options, radiation therapy, etc.]

Impact: Role-based prompting significantly improved the model's ability to produce medically accurate, structured treatment suggestions by simulating the decision-making patterns of a clinical expert.

4.5.4 *Prompt Engineering Impact*.

Combining these prompt engineering strategies improved model performance across key criteria:

- **Content Structure:** Few-shot prompting improved the organization and clarity of MRI findings.
- **Output Consistency:** Instruction-based prompting ensured standard formatting across reports.
- **Domain-Specific Precision:** Role-based prompting enabled the model to generate expert-level language aligned with clinical standards.

This strategic integration of prompt engineering enhanced the model's capability to generate coherent, structured, and medically accurate MRI reports, effectively aligning with professional diagnostic practices.

4.6 Conclusion

Our proposed system successfully integrates fine-tuned LLaMA with LoRA and strategic Prompt Engineering techniques to generate accurate and structured MRI diagnostic reports. The combination of model adaptation and prompt design proved effective in improving both content quality and performance metrics.

4.6.1 *Key Contributions*.

Our work offers the following key contributions:

- Developed a customized MRI report generation pipeline leveraging **LLaMA + LoRA**, reducing memory requirements while achieving effective convergence.
- Introduced strategic **Prompt Engineering** techniques, including Few-shot, Instruction-based, and Role-based prompting, to improve content structure, clarity, and medical language precision.
- Achieved substantial performance gains, with notable improvements in **BLEU (+100.7%)**, **ROUGE-1 (+79.5%)**, and **BERTScore (+5.3%)** over the baseline model.

4.6.2 *Limitations*.

Despite its strengths, our system presents some limitations:

- **Lack of Follow-up Recommendations:** While the system effectively generates MRI reports and treatment suggestions, it lacks the ability to dynamically recommend follow-up strategies based on patient response or clinical progression. This limitation stems from the model's reliance on static input data rather than continuous clinical updates. As a result, the system may overlook changes in the patient's condition that warrant adjusted follow-up timelines or treatment modifications.
- **Moderate BERTScore Gain:** While BLEU and ROUGE scores improved significantly, BERTScore gains were relatively modest, suggesting room for improvement in semantic understanding.

5 FUTURE WORK

5.1 Deep Learning Model

To enhance the current deep learning diagnosis system and push it closer to clinical deployment, we propose the following improvements:

- **Building a More Advanced Deep Learning Model:** Future iterations of our system will explore unified architectures capable of performing both tumor classification and segmentation within a single end-to-end model. Multi-task learning techniques will be applied to ensure consistency between tasks and reduce overall inference latency.
- **Creating a Modular AI Pipeline:** We aim to package the entire workflow—from preprocessing to report generation—into a modular and configurable pipeline. This design will allow easy integration into existing radiology systems and support future extension, such as integrating with hospital PACS systems or cloud-based inference engines.
- **Expanding Dataset Diversity:** We plan to increase the dataset scale by incorporating additional brain tumor MRI datasets, especially those covering rare subtypes and varying imaging conditions. This will improve the system's generalization capability and robustness in real-world scenarios.

- **Exploring Joint Training with Medical LLMs:** As a future extension, we will explore joint training strategies between the vision module and the LLM report generator, enabling the system to better understand context and uncertainty across modalities.

5.2 Medical LLM Model

To address limitations, we propose the following improvements:

- Expanding the dataset to include rare MRI cases and broader anatomical variations.
- Incorporating **adaptive prompting techniques** to enable dynamic prompt selection based on input complexity.
- **Developing a Follow-up Prediction Mechanism:** A key focus will be to integrate a follow-up recommendation system that dynamically suggests appropriate follow-up intervals based on MRI findings, risk scores, and potential treatment responses. We envision this mechanism using:
  - **Reinforcement Learning:** To adapt follow-up timelines by learning from past clinical data and optimizing decisions based on long-term outcomes.
  - **Dynamic Risk Score Systems:** To continuously update the patient's risk profile and recommend adjusted follow-up plans as new information becomes available.
- Exploring multi-modal learning by integrating MRI image data with textual content to enhance diagnostic precision.

By further refining these approaches, we aim to improve the system's robustness and reliability in real-world clinical environments.

REFERENCES

[1] Najjar, R. 2023. *Redefining radiology: a review of artificial intelligence integration in medical imaging*. Diagnostics, 13(17), 2760. https://doi.org/10.3390/diagnostics13172760

[2] Menze, B.H., Jakab, A., Bauer, S., et al. 2015. *The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS).* IEEE Transactions on Medical Imaging, 34(10), 1993–2024.

[3] Ronneberger, O., Fischer, P., and Brox, T. 2015. *U-Net: Convolutional Networks for Biomedical Image Segmentation.* In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 234–241.