Automated Medical Impression Generation: A Fine-Tuned Approach for Radiology Reports

Sabber Ahamed

September 21, 2025

1 Goal and Context

The goal of this project is to develop a comprehensive approach to automate medical impression generation from radiology findings using fine-tuned large language models. I developed a findings-focused data processing pipeline that filters for substantial clinical content, implemented quality-based segmentation strategies, and fine-tuned Microsoft's **MediPhi-Instruct** model using LoRA based techniques. The resulting model demonstrates significant performance improvements over the base model across multiple imaging modalities, achieving a 19.6% increase in ROUGE-1 scores and a 56.6% increase in ROUGE-2 scores.

2 Exploratory Data Analysis

2.1 Dataset Overview

I began with 30,135 de-identified radiology reports spanning 6 clinics with diverse imaging modalities. Key findings from my EDA include:

- **Modality Distribution**: MR imaging dominates (59.9%, 18,046 reports), followed by CT (17.7%, 5,322), CR (7.3%, 2,205), US (7.1%, 2,139), XR (2.9%, 879), and NM/Other (1.1%, 143)
- **Clinical Content Quality**: Only 46.8% (14,091 reports) contain both substantial findings ($\geq$100 characters) and complete impressions
- **Clinic Variation**: Significant heterogeneity in reporting styles, with some clinics favoring structured numbered lists while others use paragraph format

I incorporated all of these insights into the prompt design and segmentation strategy.

2.2 Data Quality Assessment

Critical filtering revealed substantial data quality challenges:

- 1,401 reports (4.6%) missing basic metadata (clinic, modality, or structured content)
- 14,372 reports (47.7%) with insufficient findings content (<100 characters)
- 1,387 reports (4.6%) with inadequate impression sections

2.3 Key Insights

1. **Findings-Rich Focus**: Reports with substantial findings ($\geq$100 characters) demonstrate significantly higher clinical value and impression quality
2. **Modality-Specific Patterns**: Different imaging modalities exhibit distinct reporting conventions and complexity levels
3. **Clinic Style Diversity**: Clear institutional preferences for formatting, terminology, and level of clinical detail

3 Segmentation Strategy and Justification

3.1 Evolution from Clinic-Based to Quality-Based Segmentation

Initially, I pursued clinic-based segmentation to capture institutional reporting styles. However, discussing with Mark, I pivoted to quality-focused segmentation that better serves the core objective of generating medically accurate impressions.

3.2 Implemented Segmentation Approach

3.2.1 Findings-Rich Filtering

After discussing with Mark, I implemented a quality filter that prioritizes clinical substance. I used the filtered data for training:

- **Minimum Findings Length**: 100 characters (ensures substantial clinical observations)
- **Impression Completeness**: 20-1000 characters
- **Essential Fields Only**: Focus on findings and impressions, excluding auxiliary metadata

On top of this filtering, I used modality and clinic id as segmenting data to fine tune and evaluate performance variations (See more details in evaluation section).

3.3 Justification

This quality-based filtering and modality-based approach offers several advantages:

- **Clinical Relevance**: Emphasizes both medically meaningful and institutional preferences
- **Scalability**: Generalizes across institutions without overfitting to specific clinics
- **Efficiency**: Maximizes training value from high-quality examples
- **Performance**: Focuses model learning on clinically substantial cases

4 Fine-Tuning Approach and Results

4.1 Model Selection

I selected **Microsoft MediPhi-Instruct** as the base model due to its:

- Medical domain pre-training and instruction-following capabilities

- Proven performance on clinical text generation tasks
- Efficient fine-tuning characteristics suitable for budget constraints

4.2 Technical Implementation

The fine-tuning approach leveraged efficient techniques to stay within the less than $100 budget. I used runpod to manage the training on an RTX 4090 GPU instance.

4.2.1 LoRA Configuration

```
lora_config = LoraConfig(
    r=8,                    # Low-rank dimension
    lora_alpha=32,          # Scaling parameter
    target_modules=['o_proj', 'qkv_proj',
                    'gate_up_proj', 'down_proj'],
    task_type="CAUSAL_LM"
)
```

4.2.2 Training Parameters

I have used the following training parameters. For more details of the code, please refer to `02_model_fine_finetuning.ipynb` notebook.

- **Dataset**: 8,865 training samples, 1,901 validation, 1,915 test
- **Batch Size**: 2 per device with 16 gradient accumulation steps
- **Learning Rate**: 2e-4 with cosine scheduler
- **Quantization**: 4-bit NF4 for memory efficiency
- **Max Length**: 1024 tokens with sequence packing

4.3 Training Results

The fine-tuning process completed successfully over 1 epoch with progressive improvement across 150 training steps. Final training metrics are presented in Table 1.

The training demonstrated strong convergence, with training loss decreasing from 1.112 at step 30 to 0.4465 at step 150, while validation loss improved from 0.640 to 0.430, indicating effective learning without overfitting. ROUGE scores show excellent content overlap, with ROUGE-1 reaching 0.86 and ROUGE-L achieving 0.82, demonstrating strong semantic alignment with reference impressions.

4.3.1 Style Metrics Analysis

The style metrics (structured and bullet format ratios) currently show 0.0 values, indicating that the model is not generating impressions in the expected structured formats. This occurred due to:

1. **Prompt Template Issues**: The model outputs include system messages rather than clean impressions. Due to time constraints, I could not refine the prompt templates further.

Table 1: Fine-Tuning Performance Metrics (Final Step: 150)

Metric	Value
Trainable Parameters	12.6M (0.33% of total)
Training Loss (Final)	0.4465
Validation Loss (Final)	0.4301
ROUGE-1	0.8579
ROUGE-2	0.7296
ROUGE-L	0.8243
Structured Format Ratio	0.0000
Bullet Format Ratio	0.0000
Validation Samples	614

2. **Tokenization Challenges**: The evaluation pipeline captures full model responses instead of extracted impressions.

3. **Format Detection Limitations**: Current regex patterns may not capture the variety of clinical formatting styles

4.3.2 Next Steps for Improvement

- **Output Parsing**: Implement robust impression extraction from model responses
- **Template Refinement**: Optimize chat templates to ensure clean impression-only outputs
- **Style Pattern Enhancement**: Expand format detection to include clinical paragraph styles
- **Evaluation Framework**: Develop comprehensive evaluation pipeline with proper text processing

4.4 Model Deployment

The fine-tuned adapter was successfully deployed to Hugging Face Hub:

`https://huggingface.co/sabber/medphi-radiology-summary-adapter`

5 Evaluation and Results

5.1 Systematic Modality-Based Evaluation

I implemented a comprehensive evaluation framework that systematically tests both the base Microsoft MediPhi-Instruct model and my fine-tuned adapter across all imaging modalities. The evaluation uses 20 randomly sampled cases per modality from the test set (138 total samples).

5.1.1 Evaluation Framework

- **Automated Metrics**: ROUGE-1, ROUGE-2, and ROUGE-L scores measuring content overlap with reference impressions
- **Systematic Sampling**: 20 samples per modality with random seed for reproducibility

- **Baseline Comparison**: Direct comparison between base model and fine-tuned adapter
- **Modality-Specific Analysis**: Performance assessment across 7 imaging modalities

5.2 Comparative Evaluation Results

5.2.1 Overall Performance Improvement

The fine-tuned model demonstrates significant improvement over the base model across all metrics:

Table 2: Overall Model Comparison (138 test samples)

Metric	Base Model	Fine-Tuned
ROUGE-1	0.3465	0.4146 (+19.6%)
ROUGE-2	0.1800	0.2818 (+56.6%)
ROUGE-L	0.2727	0.3720 (+36.4%)

5.2.2 Modality-Specific Performance

Performance varies significantly across imaging modalities, as shown in Table 3:

Table 3: ROUGE-1 Performance by Modality

Modality	Base Model	Fine-Tuned	Improvement
MR	0.4642	0.6274	+0.1632 (+35.1%)
Unspecified	0.4186	0.5655	+0.1469 (+35.1%)
CR	0.3283	0.3970	+0.0687 (+20.9%)
XR	0.2859	0.3812	+0.0953 (+33.3%)
NM	0.3440	0.2872	-0.0568 (-16.5%)
US	0.3073	0.3394	+0.0321 (+10.4%)
CT	0.2836	0.2978	+0.0142 (+5.0%)
OTHER	0.2745	0.2276	-0.0469 (-17.1%)

5.2.3 Key Evaluation Insights

1. **Strongest Improvements**: MR imaging shows the largest absolute improvement (+0.1632 ROUGE-1), likely due to its prevalence in the training data (59.9% of dataset)
2. **Consistent Gains**: Six out of eight modalities show performance improvements, with four modalities achieving >20% relative improvement
3. **Challenging Modalities**: NM and OTHER modalities show performance decreases, possibly due to limited training examples and high variability
4. **Clinical Relevance**: Major modalities (MR, CT, CR, XR) all demonstrate meaningful improvements, covering >85% of clinical cases

6 Strategic Scaling Recommendations

I had discussed couple of hypothesis during my interview about scaling. Of course with the growth of the business, we can not fine tune multiple model targeting different clinics or modalities. Here are some of the recommendations to scale this approach:

- **Option-1: One unified model**: Train a single model that takes modality as an additional input feature, allowing it to adapt its generation style based on the specified imaging type. This reduces the need for multiple models while still capturing modality-specific nuances.
- **Option-2: Cluster-based style adaptation fine tuning**: In this option we can cluster clinics and radiologists reporting styles and fine-tune a model for each cluster. This balances the need for customization with scalability, as fewer models are required compared to clinic-specific fine-tuning.
- **Option-3: Data augmentation techniques**: Use data augmentation techniques to expand the training dataset, especially for underrepresented modalities. Synthetic data generation can help balance the dataset and improve model robustness.

7 Conclusion

This technical assessment demonstrates a practical, cost-effective approach to automated medical impression generation that achieves significant performance improvements over baseline models. By focusing on findings-rich content and implementing quality-based segmentation, I developed a robust fine-tuning pipeline that produces clinically relevant impressions while maintaining scalability and efficiency. The fine-tuned model shows strong gains across multiple imaging modalities, particularly in MR and CR, which are critical for clinical practice. While some challenges remain in output formatting and style consistency, the overall results validate the effectiveness of the approach.