A QUANTITATIVE STUDY ON IMPROVING RADIOLOGY REPORT DESCRIPTIVENESS USING LARGE LANGUAGE MODELS

RSNA 2024

Building Intelligent Connections

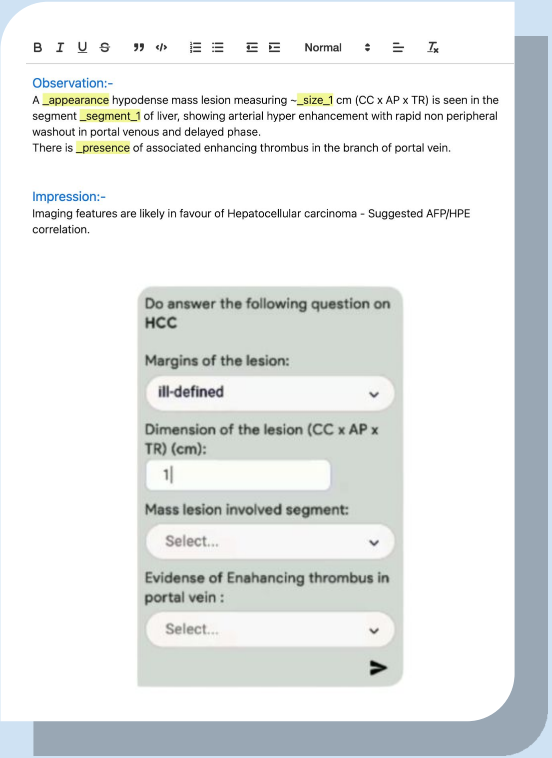
Kalyan Sivasailam

Co-Founder & CEO

5C Network

"**Consistency is a hallmark of quality**. What draws us to the burgers at Five Guys or the desserts at the Cheesecake Factory is the consistency of the product across time and location. Such consistency will be the watchword for the **radiology report in 2025.**"

Curtis P. Langlotz, *The Radiology Report: A Guide to Thoughtful Communication for Radiologists and Other Medical Professionals*, Chapter 12



Challenges in Radiology Reporting

Addressing Variability, Clarity, and Resource Limitations for Improved Patient Outcomes

Ambiguous Radiology Reports

Reports could lack in **clarity**, making interpretation difficult clinicians.

Lack of Detailed Descriptions

The absence of standardized reporting practices results in **insufficient information** for clinical decision-making.

Sole Dependence on Radiologist's Expertise

A shortage of radiologists exacerbates the challenge, placing a premium on individual expertise for **maintaining report quality**.

Time Constraints Impacting Report Quality

High caseloads and time pressures can limit the depth and clarity of reports, increasing the likelihood of errors and inconsistencies in radiology reporting.

Building AI to help Radiology become more reliable

5C

Establishing Baseline Descriptive Scores

Framework and Interventions for Improving Radiology Report Quality

Descriptive Score

The descriptors were systematically classified according to their relevance to over **400 pathologies in CT Abdomen & Pelvis Studies,** including incidental findings. A model was developed to evaluate each report at the category level, assigning a score out of 10.

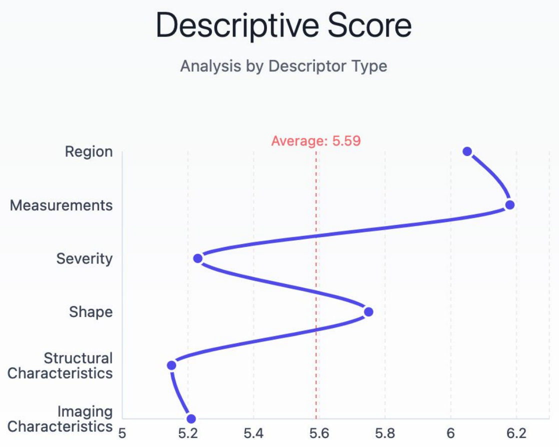
Method

Used an Instruction Fine Tuned Model to rate 109,639 reports for descriptive score based on 6 categories we defined for all abdomen pathologies.

Findings

Radiologist report without AI assistance, our **descriptive score was 5.59**

Limitations: There are no clear or accepted ways to test and validate these models yet. We have worked with expert radiologists in India to define the metrics that we are measuring.



Study (Modality): **CT Abdomen & Pelvis**

Total # reports read: **109,639**

Total Descriptor categories: **6**

Baseline assessment Descriptor Score: **5.59**

Establishing Baseline Error rate and Turn Around Time (TAT)

Framework and Interventions for Improving Radiology Report Quality

Error Rate & TAT

TAT and error rate were meticulously tracked to assess reporting efficiency and accuracy. TAT measured the time from initial report generation to final review, while the error rate quantified the frequency of diagnostic discrepancies.

Method

The RADPEER Scoring Analysis was applied to categorize error grades (1, 2A, 2B, 3A, 3B) across a dataset of **51,238 radiology reports**, providing a systematic evaluation of diagnostic accuracy and consistency.

Findings

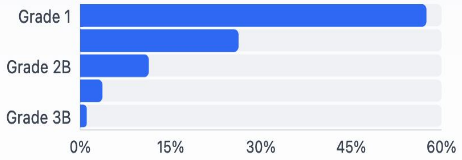
An overall **error rate of 8.37%** was observed, with **83.8%** of these errors categorized as grade 1 or 2A, representing lower-severity discrepancies and TAT was **17 Mins** for reports without AI assistance.

Error Rates Analysis

RADPEER Scoring in Radiology Reports

8.37%
Average Error Rate

51,238
Reports Analyzed



Total # reports read: **51,238**

Baseline assessment Error Rate : **8.37%**

Baseline assessment TAT: **17 Mins**

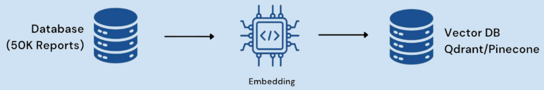
Leveraging LLMs: Experiment

Evaluating Model Enhancements from Baseline Metrics to Optimal Quality

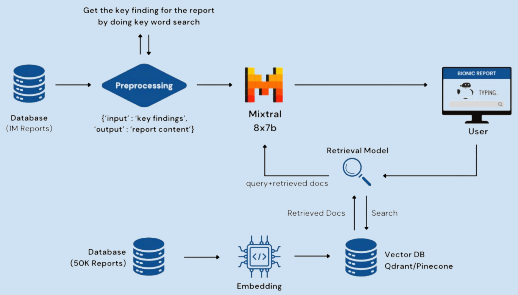
This study assessed the impact of AI assistance on radiology reporting by comparing baseline metrics from 15 radiologists to enhanced results over 12 weeks, focusing on descriptor scores and error rates.

RAG

Phase 1: (1-4 Weeks)
Introduced RAG model to support reporting, improving descriptiveness but limited in handling complex pathologies.



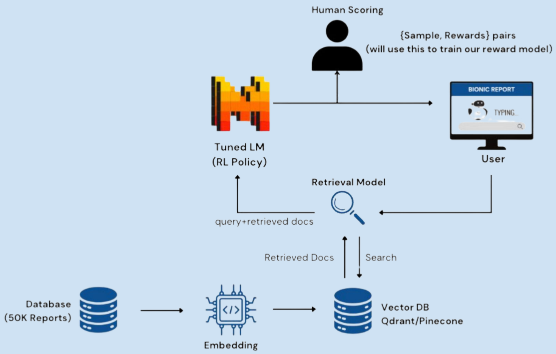
Instruction Fine-tuning + RAG



Phase 2: (4-8 Weeks)
Instructional Fine-tuning + RAG
Enhanced model precision with specific instructions, achieving greater clarity, though occasional deviations from human preferences persisted.

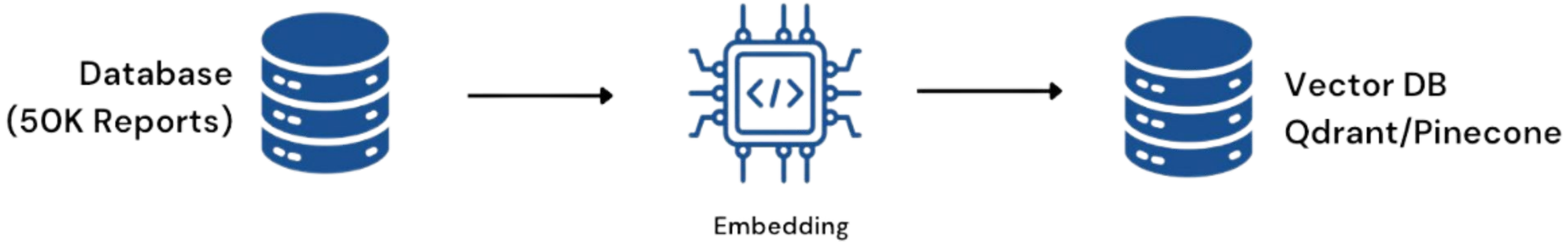
Reinforcement Learning from Human Feedback + RAG

Phase 3: (8-12 Weeks)
Reinforcement Learning from Human Feedback, significantly increasing descriptive accuracy and minimizing errors for reliable, actionable reports.



Retrieval Augmented Generation

Phase 1



Limitations: Despite improvements in report format and style mimicry, RAG struggles to fully capture the nuanced understanding of radiologists in pathology explanations. Resulted in increased reporting times

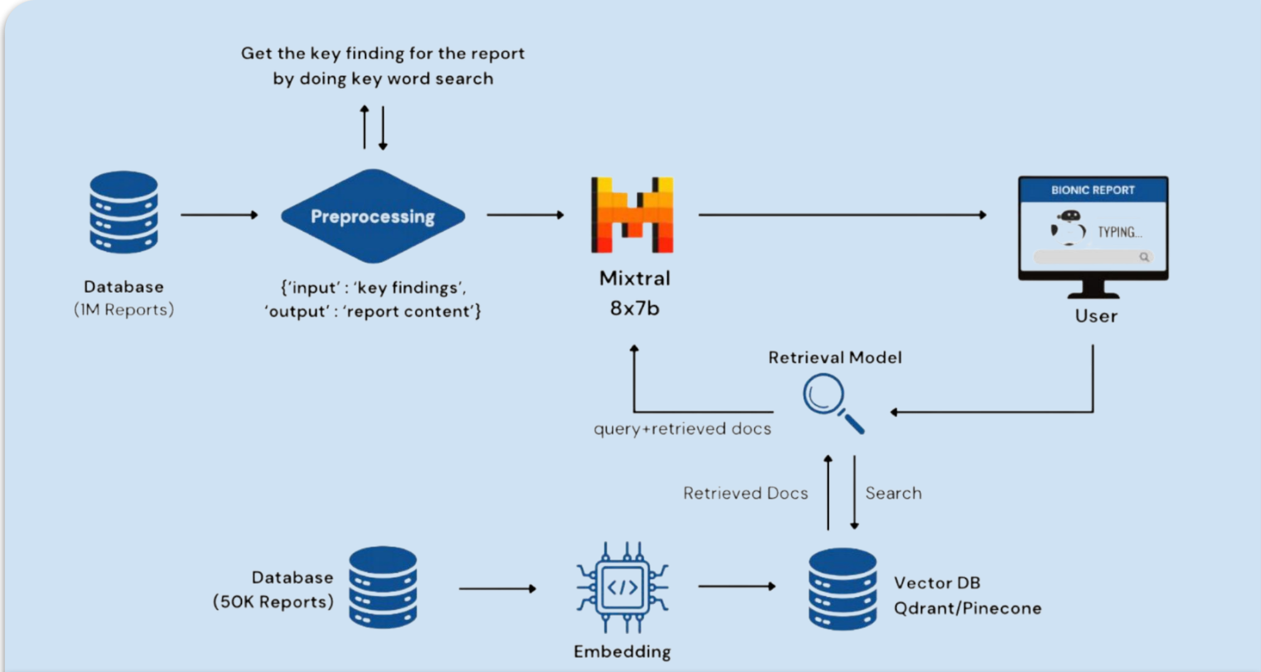
7.45%
Average Error Rate

6.1
Descriptive Score

21
Mins, Average TAT

Instruction Fine-tuning + RAG

Phase 2



Limitations: Despite the notable increase in accuracy, there were occasional divergence from human preferences, necessitating further human assessment for refinement.

4.5%
Average Error Rate

7.6
Descriptive Score

15
Mins, Average TAT

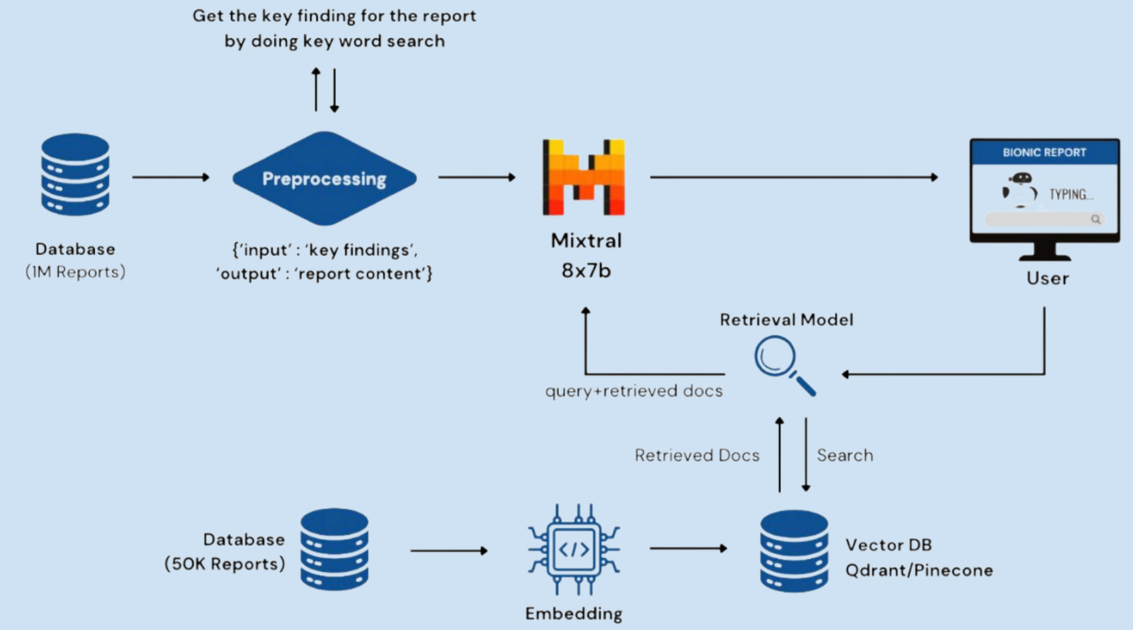
Reinforcement Learning from Human Feedback + RAG

Phase 3

8.75 Descriptive Score

1.2% Average Error Rate

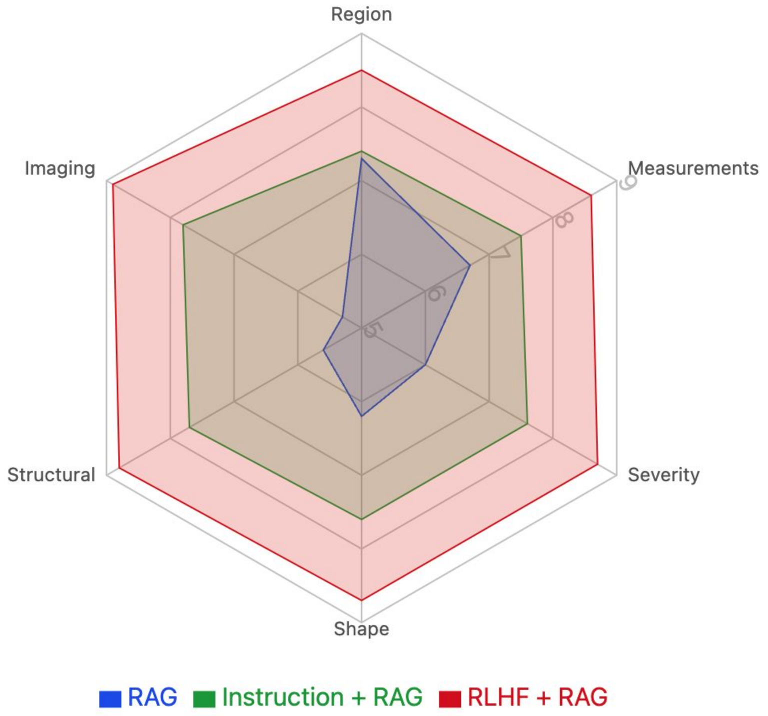
14 Mins, Average TAT



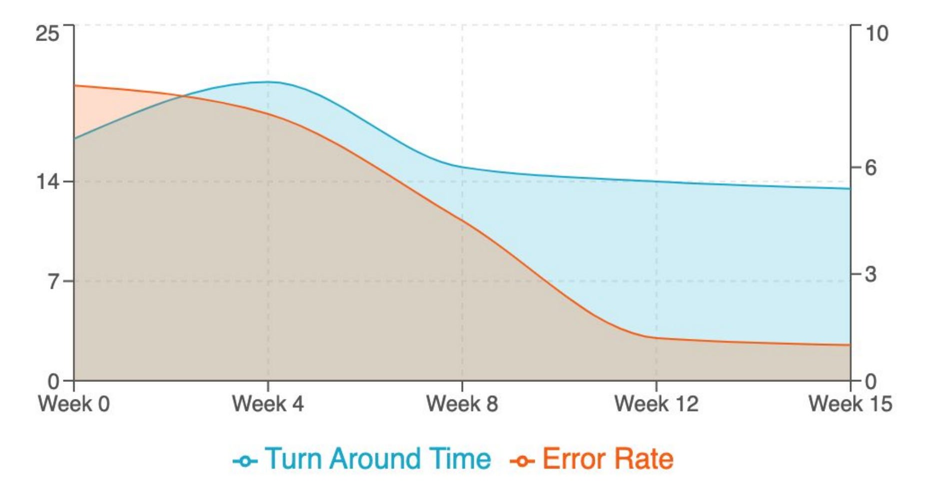
Limitations: Despite the notable increase in accuracy, there were occasional divergence from human preferences, necessitating further human assessment for refinement.

Study Outcome

Mertics: Descriptive Score, TAT and Error Rate



Descriptive Score



TAT and Error Rate outcomes

Key Insights & Critical Considerations

Comprehensive Analysis of LLM Integration in Radiology Practice

Participant Demographics

150 Radiologists	37.2 years Average Age
27-58 Age Range	35 years Median Age

- For "small language" usecases, **RLHF + IF + RAG > IF + RAG > RAG**
- LLM-powered reporting could significantly improve actionability of Radiology reports
- Making radiologists focus systematically could reduce error rates

Study Limitations

- Pathology Identification: Limited impact on reducing Grade 3 errors due to reliance on radiologist expertise
- Descriptive Score Methodology: Potential for improved performance with more nuanced weighting of descriptor categories
- Radiologist Memory: Same scan read thrice in 12-week period could lead to familiarity bias

Get in Touch

Author: Kalyan Sivasailam, CEO
5C Network
ceo@5cnetwork.com

RSNA® 2024
Building Intelligent Connections