R2GEN2

Fine-tuning Multimodal AI for Radiology

Aayush Verma Ashmita Mukherjee Rishab Mohan Leo Li

Meet The Team

Rishab Mohan

Leo Li

Aayush Verma

Ashmita Mukherjee

01

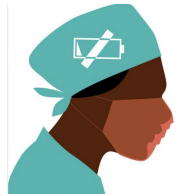
PROBLEM STATEMENT

Background and Research Problem

Meet Dr. Cooper

Dr Cooper is a radiologist at IU Health leading a team of accomplished radiologists.

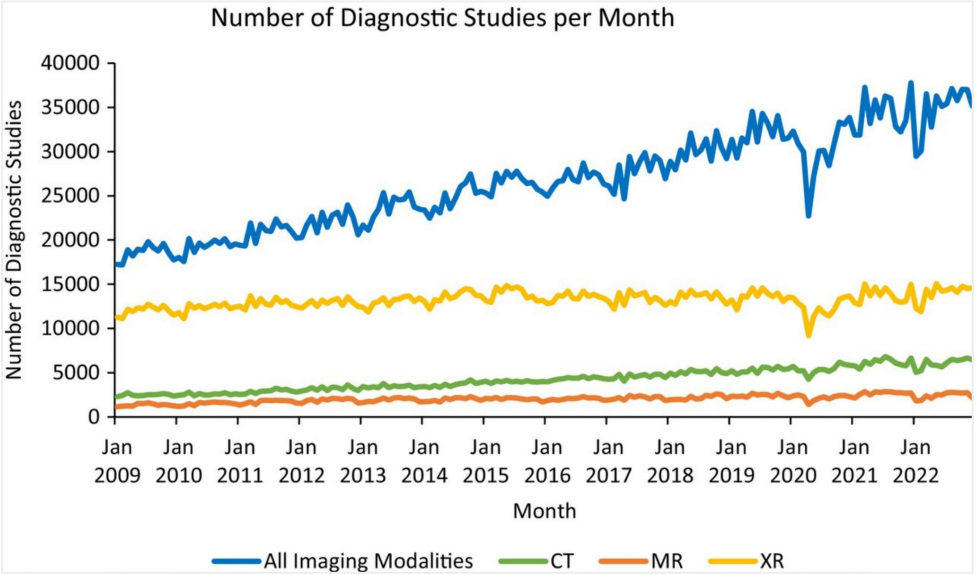
She and her team along with many other radiologists in the USA contend with extreme circumstances and challenges daily.



- ➔ Ever Increasing Imaging Volume
- ➔ Increasing Work Hours spilling into personal life
- ➔ Non Essential Administrative Tasks and Desk Work
- ➔ Pressure To Maintain Accuracy under Immense stress

We believe AI Integration is the key to address some of their burdens !

Let's talk Numbers

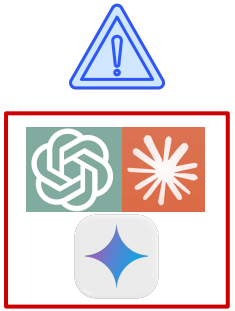


- Dr. Cooper and her team work with **~35000 studies** a Month. This leads to her and her team working more **than 35 hours per day total**.
- **~49%** of their valuable time is **wasted on documentation and desk work*.**
- **In fact, 54% of all Radiologists** report experiencing **"long-term, unresolved, job-related stress leading to exhaustion, cynicism, and lacking a sense of personal accomplishment."**

Sinsky, C., Colligan, L., Li, L., Prgomet, M., Reynolds, S., Goeders, L., Westbrook, J., Tutty, M., & Blike, G. (2016). Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties. *Annals of Internal Medicine*, *165*(11), 753. https://doi.org/10.7326/m16-0961
https://appliedradiology.com/Articles/playing-with-fire-burnout-among-radiologists-a-growing-concern https://pubmed.ncbi.nlm.nih.gov/7551783/#:~:text=Data%20were%20collected%20on%202345,systematic%20variation%20in%20reporting%20time.

Challenges for **enterprise VLMs** to deliver **personalized** solutions.

- ➔ Radiologists develop formulaic easy to reproduce report writing styles unique to themselves.
- ➔ HIPAA Compliance Laws restrict the flow of data slowing AI Innovation.
- ➔ Accuracy and Factual correctness thresholds.

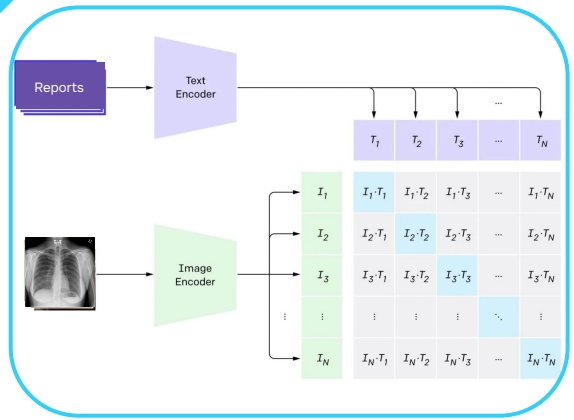


Enterprise Solutions are not the Answer

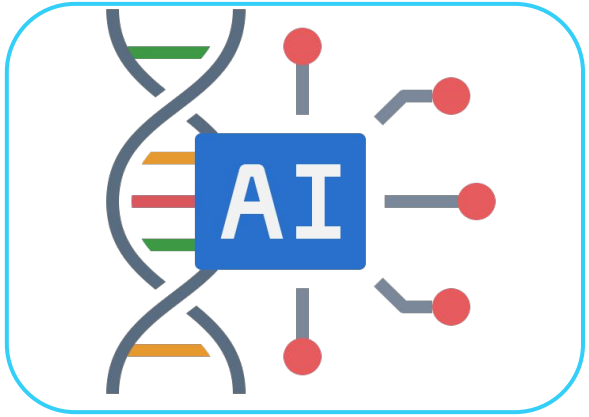
- ➔ Lack of standardization in reporting styles hinders the adoption of generalized AI
- ➔ Ensuring data security and compliance with regulations like HIPAA remains a critical barrier to implementation.
- ➔ Not holistically trained in medical data leading to factual correctness errors.

Literature review focused on significant advancements in AI: **Advancement of vision-language models** and **automated systems for healthcare applications.**

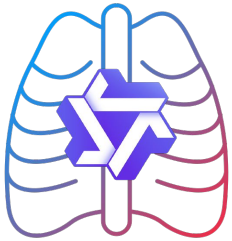
Vision Language Models



AI/ LLMs in Healthcare



VLMs in Healthcare



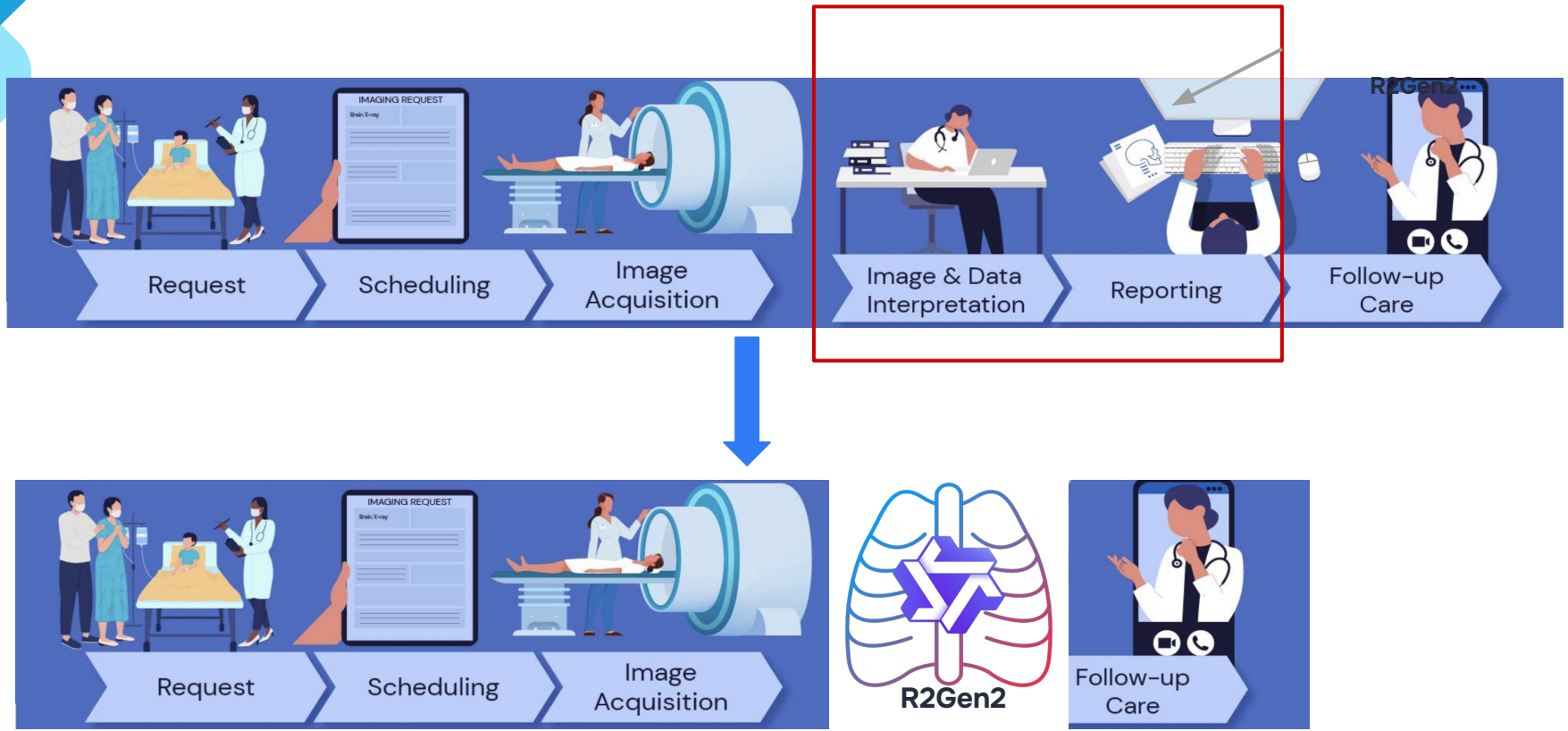
R2Gen2

- Brings Transformer Architecture to Vision Models (**GPT with Eyes !**)
- We surveyed Current Work Involving **Feature Extraction from Images in a structured format.**

- Current work consists of **converting Labels to Findings and Impressions**
- Some authors also leveraged AI to **create customized Patient Plans**

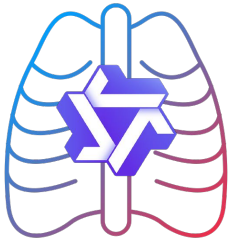
- We **combine** the intents of previous work to propose R2Gen2
- Harnessing Vision language Models we synthesize **Reports and Features directly from Images!**

R2Gen2 in the Radiology Life Cycle



R2Gen2 for Advanced Medical Feature Extraction and Personalized Radiology Findings Reporting


```
"anatomical entity":null,
"location_descriptor":null,
"procedure":[null,]
"clinical findings":
[
    {
        "clinical finding":"retrocardiac opacity",
        "existence":"pos_dx",
        "descriptive_term":"retrocardiac",
        "observation":"opacity"
    },
]
```

R2Gen2

Generated Findings :

The heart, pulmonary vasculature and mediastinum are within normal limits. There is no pleural effusion or pneumothorax. There is no focal air space opacity to suggest a pneumonia. There is an interim upper cervical spinal fusion partly evaluated.

Feature Extraction Model

Automatically extracts relevant features.

- *fracture lines, soft tissue swelling, or lung opacities*

JSON output easily integrates with radiology reporting systems or EMRs

- ”fracture: true”, ”location: distal radius”, ”severity: displaced”

Radiology Findings Reporting Model

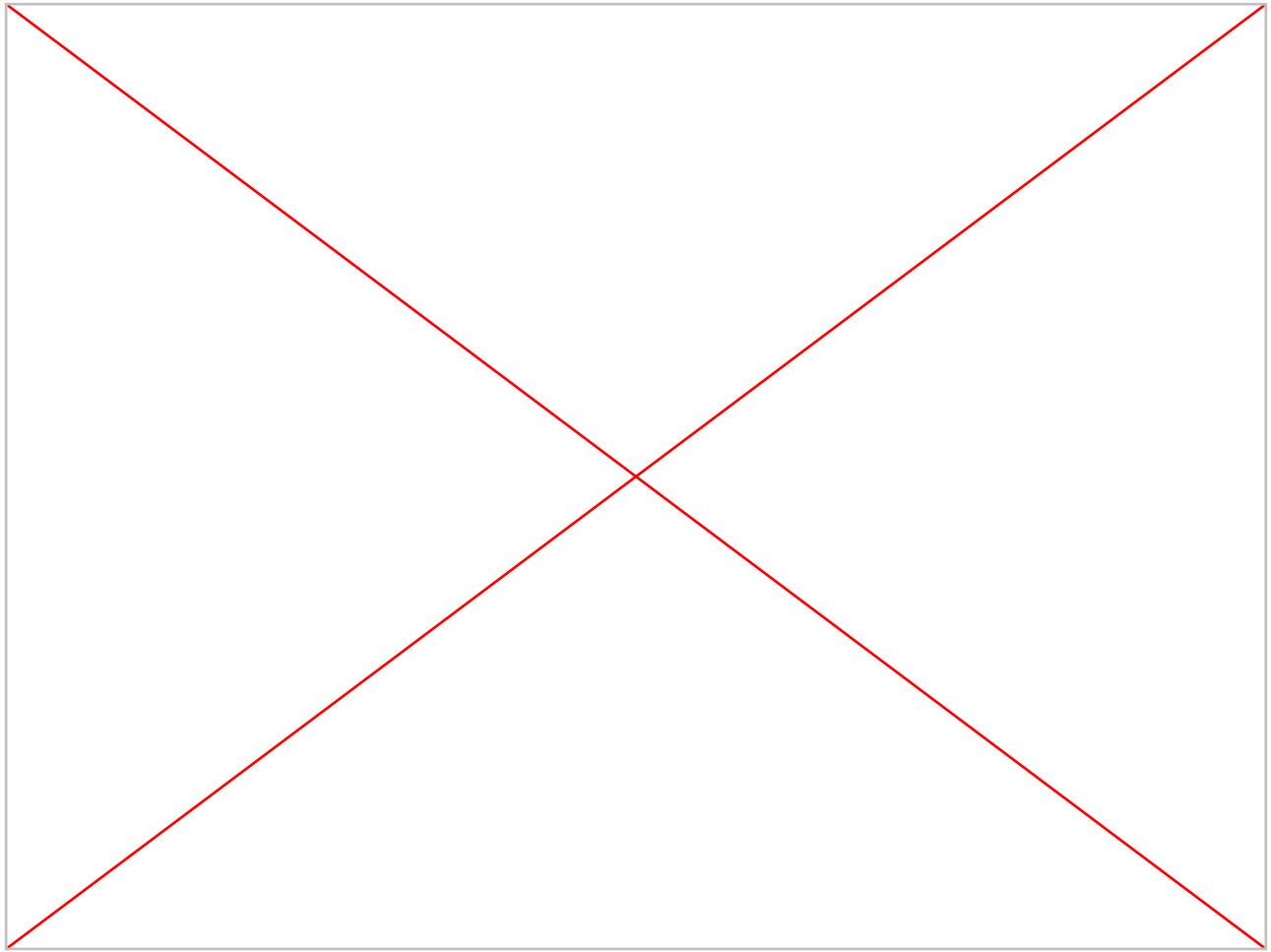
Adapt to each radiologist’s distinct reporting style.

- *Ensures consistent use of terminology and formats across all reports*

Provide preliminary findings to streamline workflow.

- *Reduces the time spent manually transcribing findings*

R2Gen2 in Action!



02

DATA OVERVIEW

Exploratory Analysis and Preprocessing

Example of a real radiology report

Exam Number: 12345678 Report Status: Final
Type: Chest 2 Views
Date/Time: 01/01/2014 10:30
Exam Code: XRCH2
Ordering Provider: Wayne, John Michael MD

HISTORY:
- Cough and Fever

REPORT Frontal and lateral views of the chest.

COMPARISON: None

FINDINGS:
Lines/tubes: None.

Lungs: The lungs are well inflated and clear. There is no evidence of pneumonia or pulmonary edema.

Pleura: There is no pleural effusion or pneumothorax.

Heart and mediastinum: The cardiomediastinal silhouette is normal.

Bones: The visualized skeleton is normal.

IMPRESSION:
Clear lungs without evidence of pneumonia.

RECOMMENDATION:
None.

PROVIDERS:
Doe, Jane Lynn MD

SIGNATURES:
Doe, Jane Lynn MD

If you have questions or concerns regarding this report, feel free to contact us by phone at 555-555-5555, or by e-mail at contact@aplusradiology.com

→ **R2Gen2's Target Area !**

A dataset of **Chest X-ray images** and **reports** sourced from the **Indiana Network for Patient Care** was utilized.

```
Label="INDICATION">Positive TB test
 Label="FINDINGS">The cardiac silhouette and
 mediastinum size are within normal limits. There
 is no pulmonary edema. There is no focal
 consolidation. There are no XXXX of a pleural
 effusion. There is no evidence of pneumothorax.
 Label="IMPRESSION">Normal chest x-XXXX.
<Affiliation>Indiana University</Affiliation>
  <LastName>Kohli</LastName>
  <ForeName>Marc</ForeName>
  <Initials>MD</Initials>
 <Author Valid Y/N ="Y">
  <LastName>Rosenman</LastName>
  <ForeName>Marc</ForeName>
  <Initials>M</Initials>
 <PublicationType>RadiologyReport</PublicationType>
```

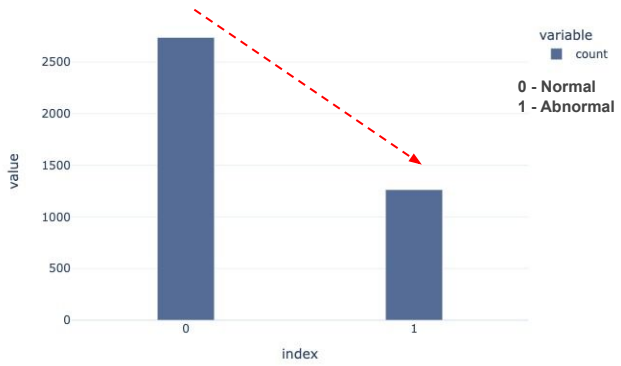
Our research is focused on accurate generation of findings

Indication | Findings | Impression | Affiliation | Radiologist | Technician | Report Type

Source Link : https://openi.nlm.nih.gov/imgs/collections/NLMCXR_reports.tgz

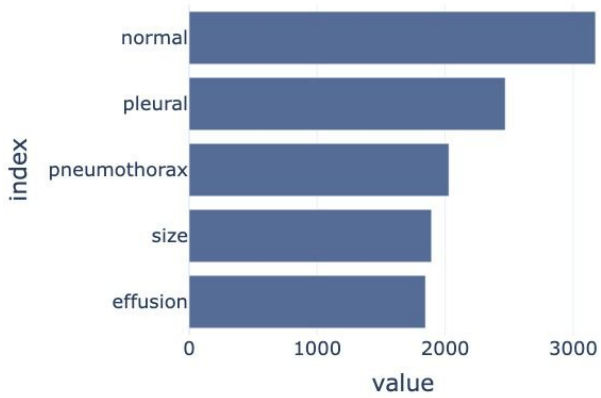
textual exploratory data analysis unlocked a deeper understanding of the dataset.

Normal v/s Abnormal Report Count



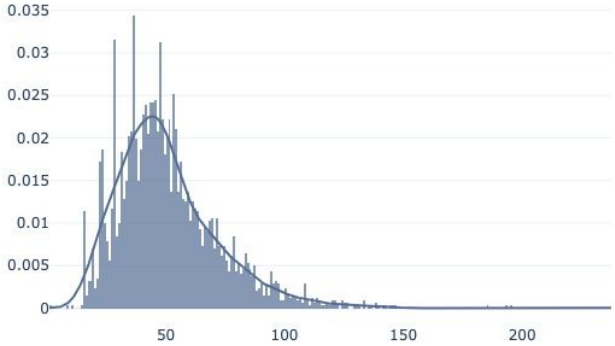
- 67.56% of the reports are categorized as normal.
- 32.43% are classified as abnormal.

Frequently used terms in Reports



- The most frequent terms in the dataset include "pleural", "normal", "focal", "lungs".
- This aligns with the typical language and findings found in chest X-ray reports.

Distribution of Report Lengths



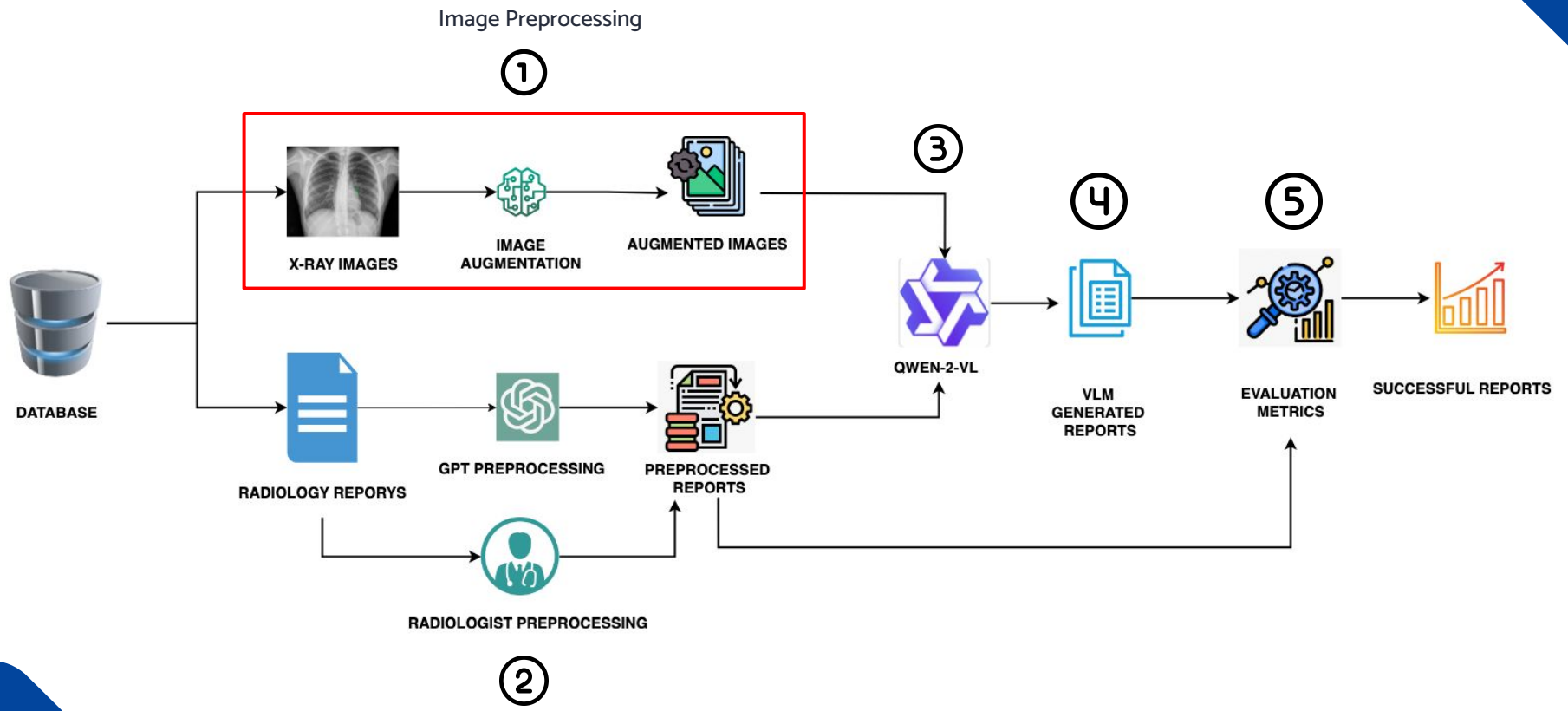
- The radiology reports are typically ~50 words in length.
- This enables the report generation process for the VLM, to produce accurate and efficient outputs.

03

METHODOLOGY

Workflow and Model Fine Tuning

R2Gen2 Under the Microscope

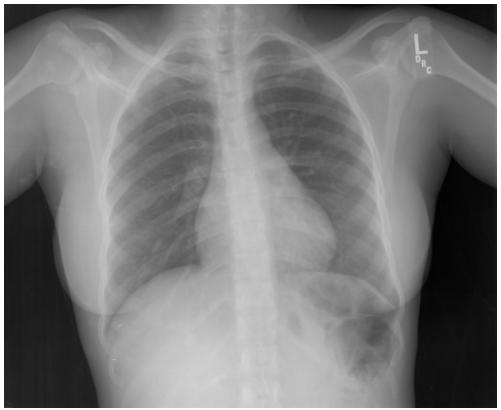


Given the limited size of the dataset, it was necessary to **generate additional samples** to **mitigate** the risk of **overfitting**.

Transformers are particularly prone to overfitting when trained on small datasets.

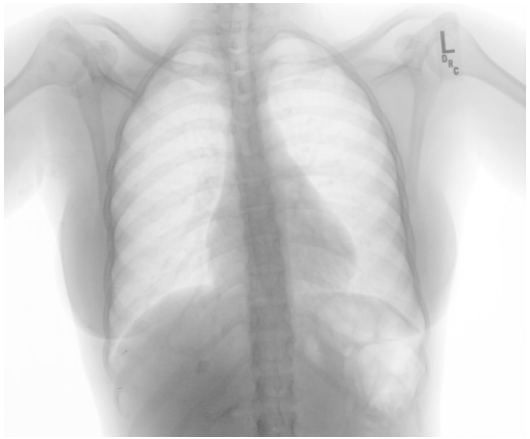
This approach boosts robustness and generalizability.

Original Image

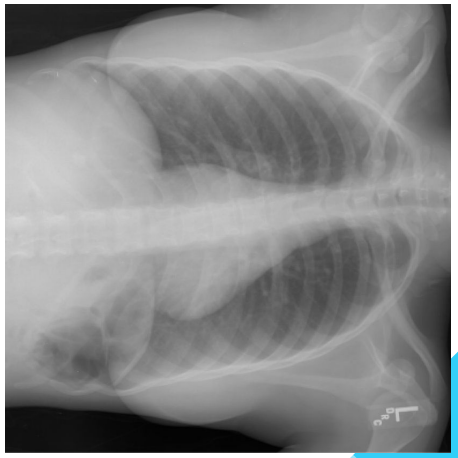


AUGMENTATION

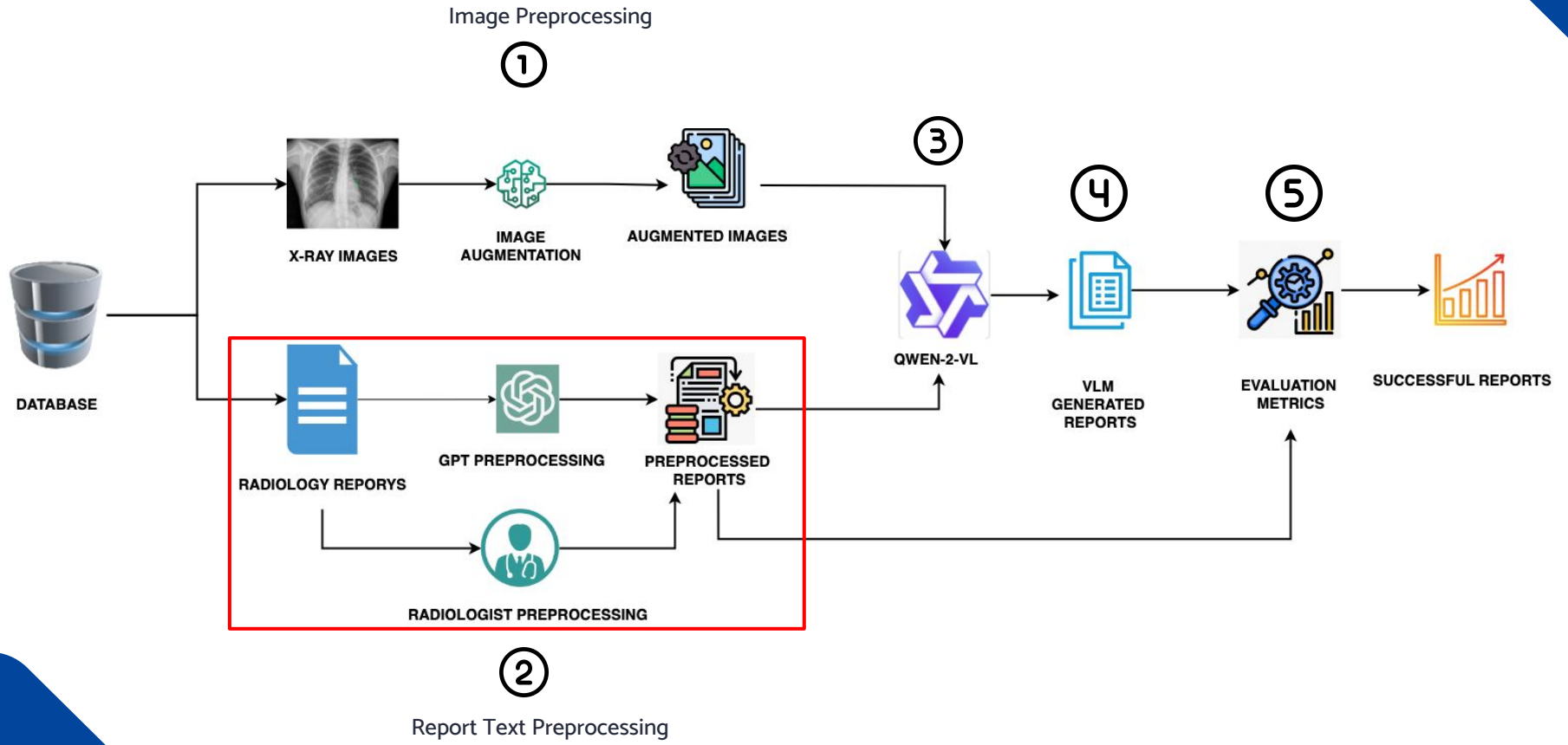
Color Inversion



Rotation



R2Gen2 Under the Microscope



Imputation of **anonymized medical findings from 37%** of the reports compromising the dataset's utility for research.

Example of a Report with missing information

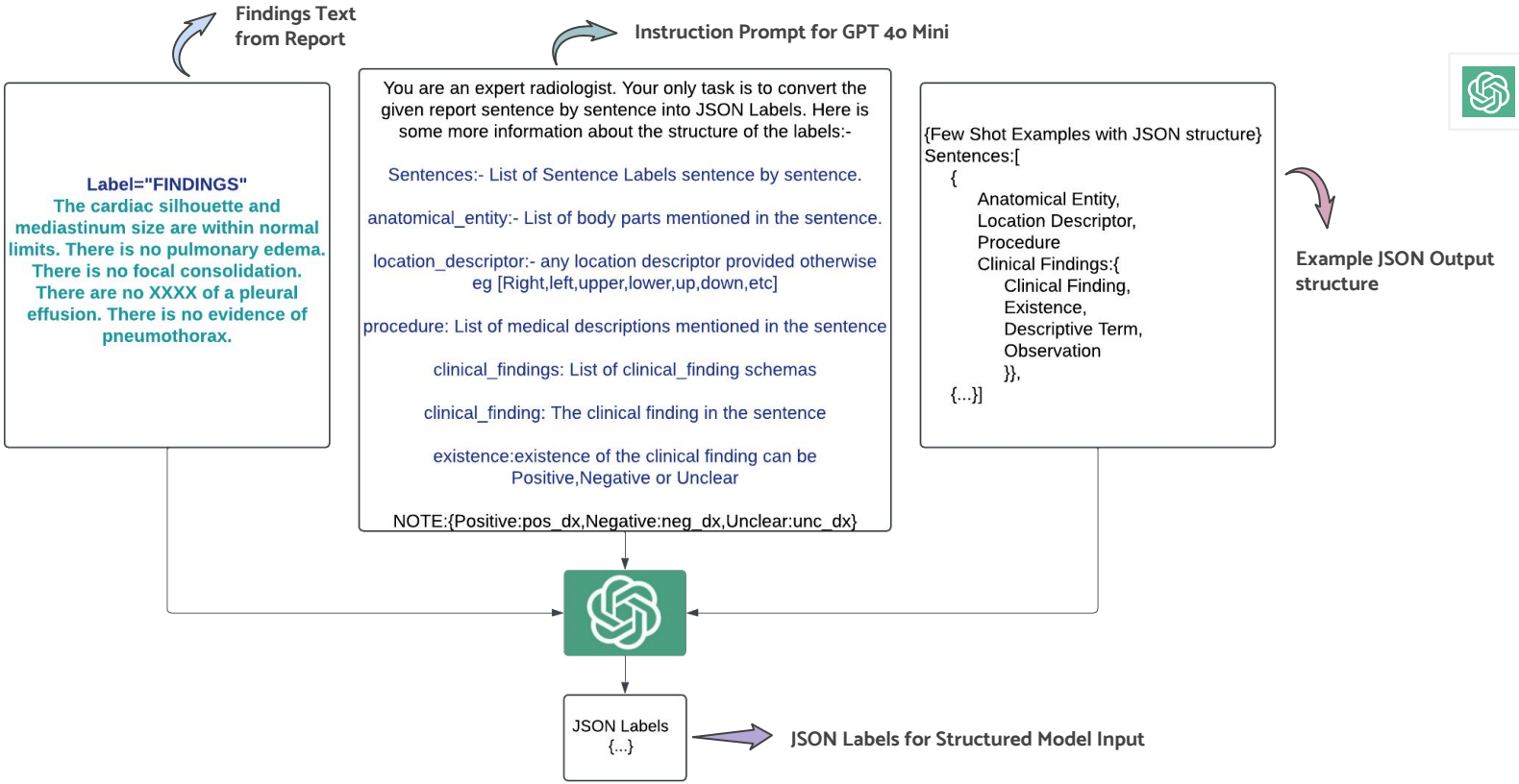
Redacted Text

Findings:
The cardiac silhouette and mediastinum size are within XXXX limits. There is no pulmonary edema. There is no focal consolidation. There are no XXXX of a pleural effusion. There is no XXXX of pneumothorax.

To reconstruct the de-identified dataset, we implemented the following strategies

- **Collaborated with a certified radiologist. Reconstructed findings maintaining overall privacy.**
- **The radiologist cross-referenced original images with redacted data.**
- **Restored critical observations while ensuring the clinical meaning and context of the data were intact**

Reports are transformed into **structured JSON labels** for the Structured Fine tuned Model



Example of the **original report** plus **newly generated labels** by **GPT-4o Mini**.

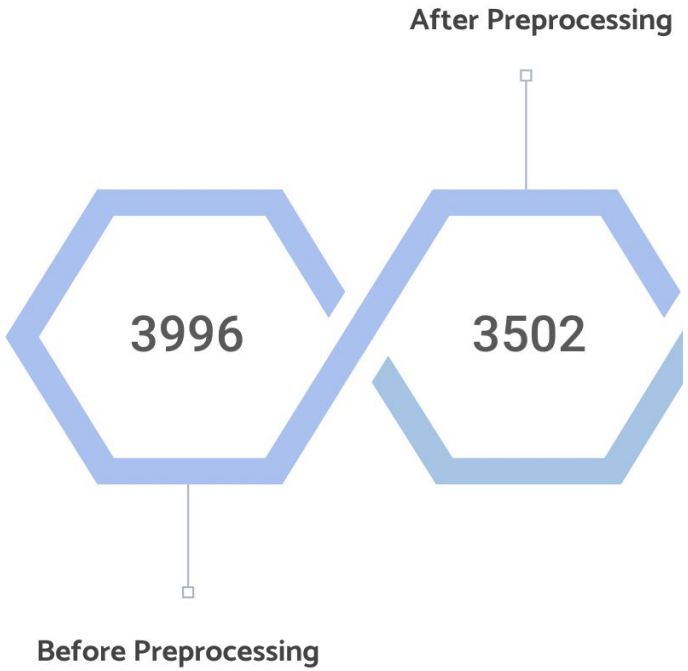
Each sentence → structured JSON, detailing the anatomical entity, the procedure performed, and the specific clinical findings associated with that entity

Retrocardiac opacity which may represent atelectasis and or or small effusion is stable.

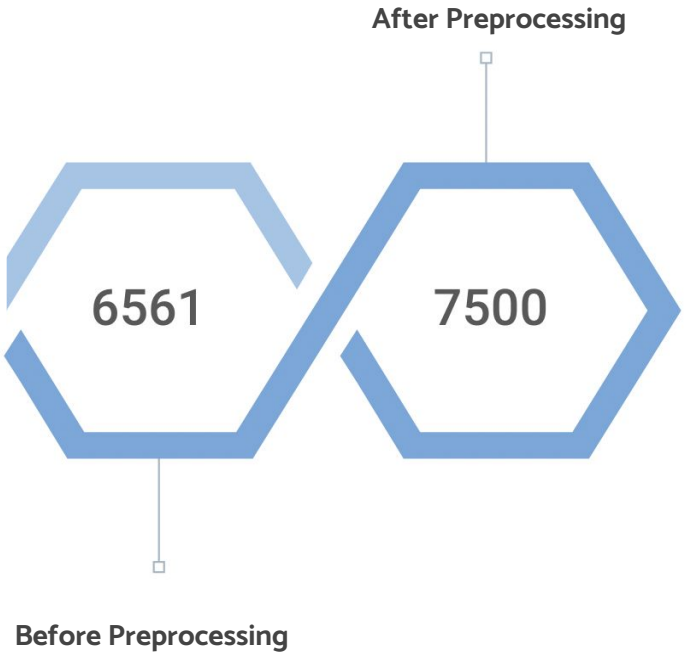

```
{
    "anatomical entity":null,
    "location_descriptor":null,
    "procedure":[null,]
    "clinical findings":
[
        {
            "clinical finding":"retrocardiac opacity",
            "existence":"pos_dx",
            "descriptive_term":"retrocardiac",
            "observation":"opacity"
        },
        {
            "clinical finding":"atelectasis",
            "existence":"unc_dx",
            "observation":"atelectasis"
        },
        {
            "clinical finding":"or small effusion",
            "existence":"unc_dx",
            "descriptive_term":"or small",
            "observation":"effusion",
        }
]

}
```

Preprocessing resulted in a **12% reduction in X-ray reports**, and **augmentation increased the number of images by 14%.**

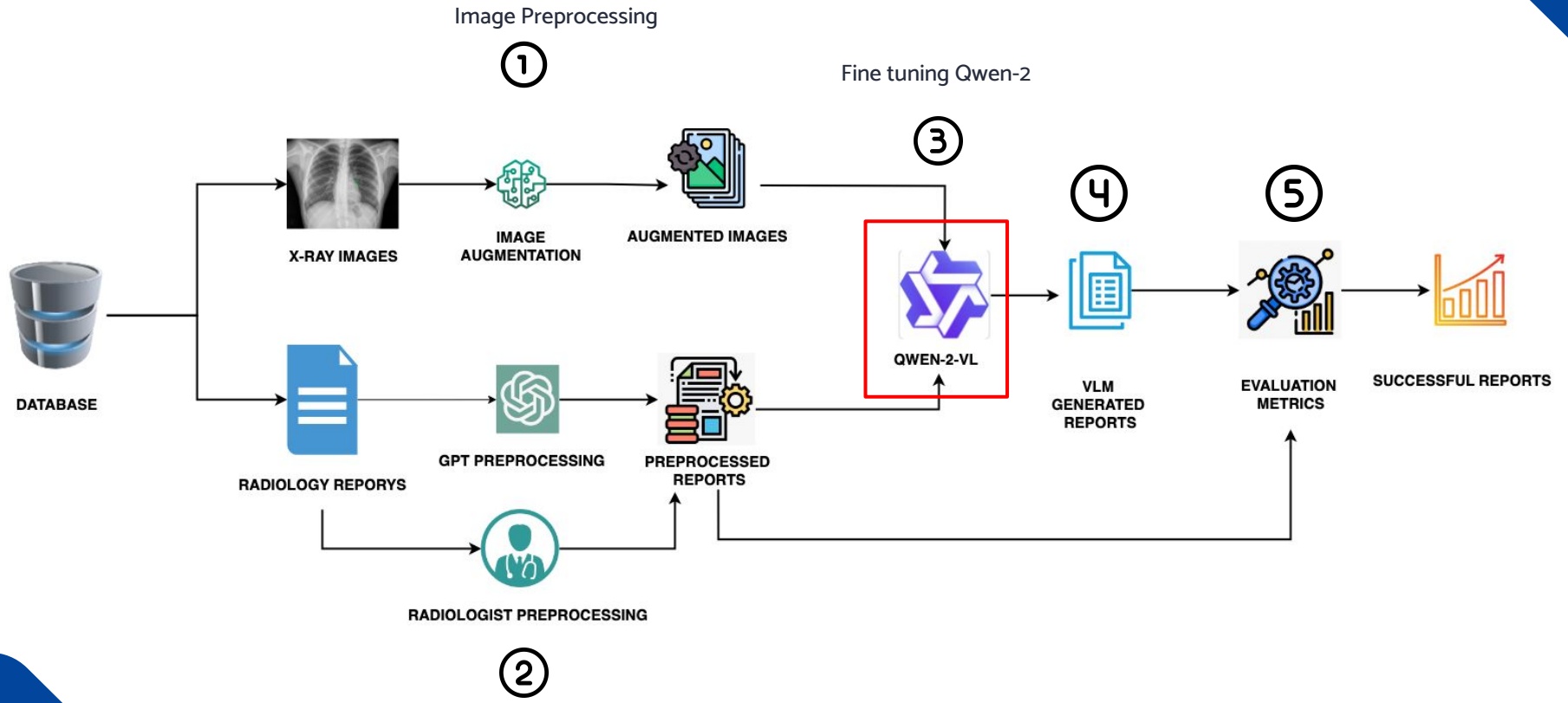
Number of Chest X-Ray Reports



Number of Chest X-Ray Images

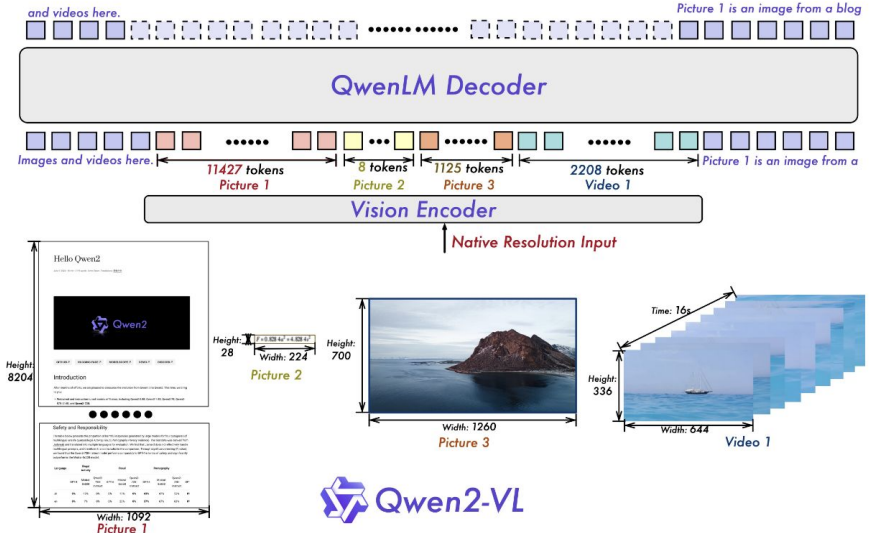


R2Gen2 Under the Microscope



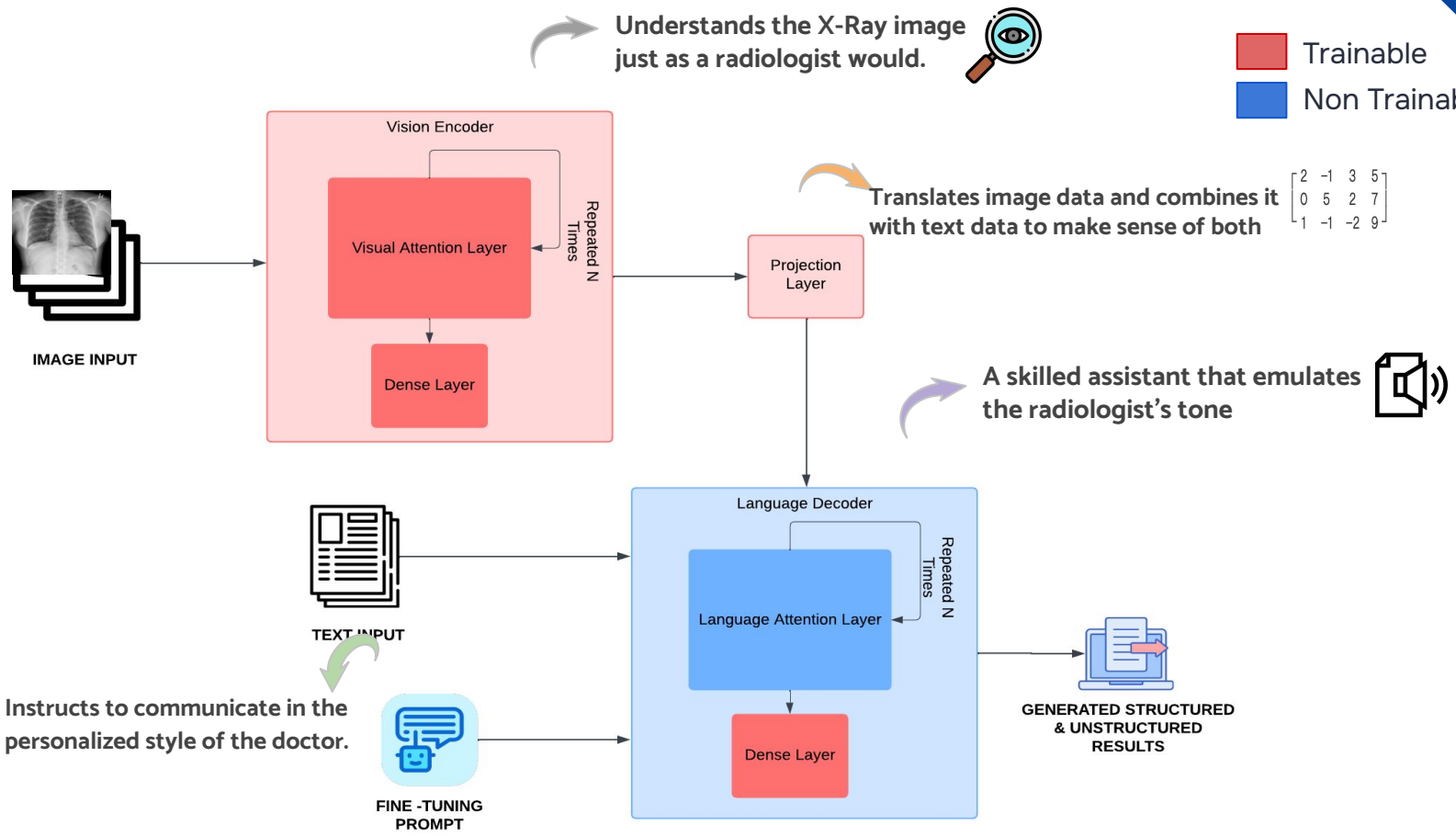
Our **Literature Review** of vision models revealed that **Qwen 2** demonstrated **superior performance** on Image comprehension and Document semantics benchmarks, **surpassing** LLaMA and GPT.

- Advanced multimodal architecture
- Handles Arbitrary Image sizes and resolutions
- Positional understanding across text, visual, and video domains

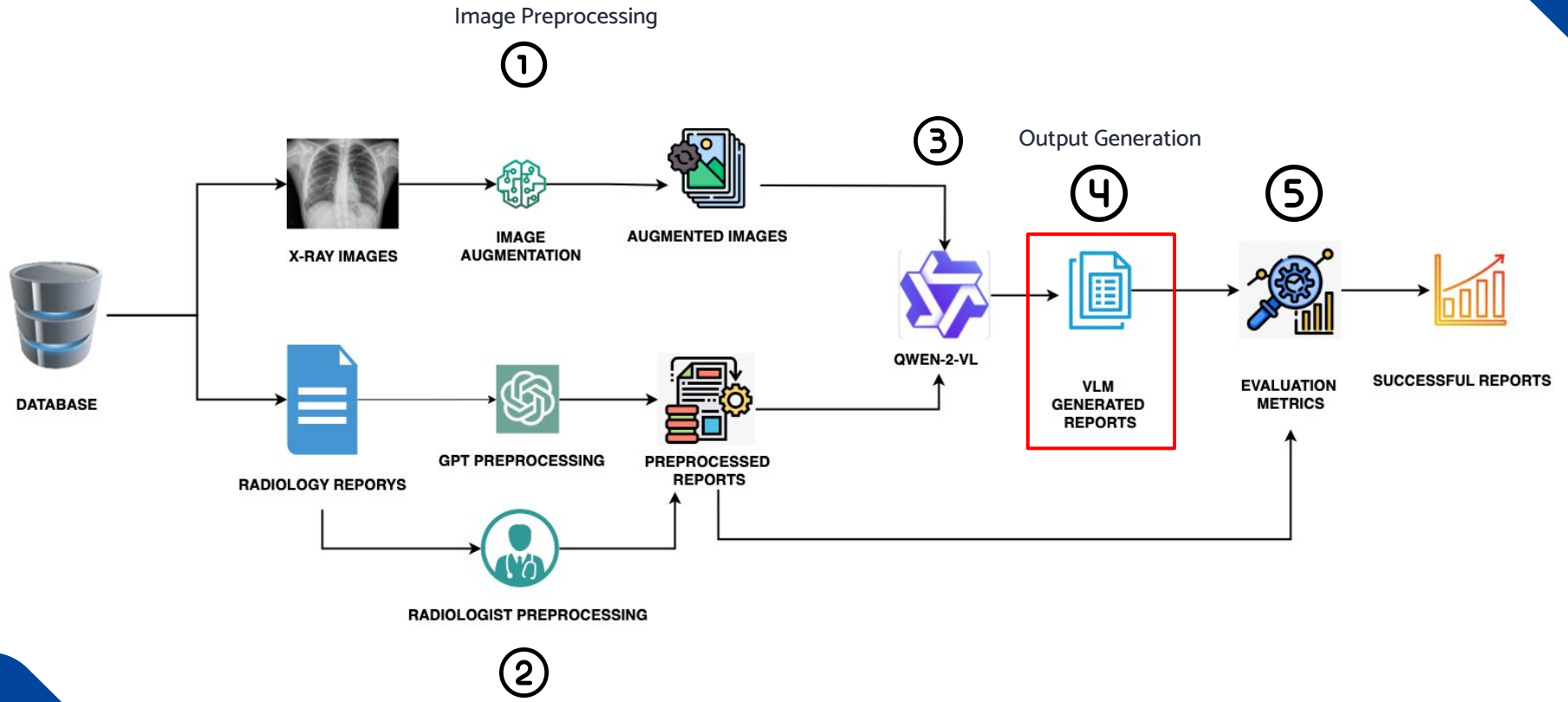


Benchmarks \ Models	Qwen-2-VL-7B	Llama-3.2-11B	GPT-4o-Mini
DocVQA	94.5	91.6	-
InfoVQA	84.3	80.1	81.7
MMBench1.1	80.7	79.4	76.0

How to train your VLM

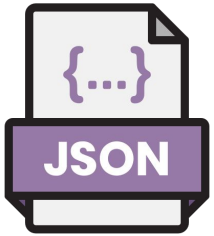


R2Gen2 Under the Microscope



We have **two specialized models** to streamline the workflow: **one focused on extracting relevant features** and **second dedicated to generating detailed reports in the radiologist's personalized tone boosting customization.**

Structured Fine-tuned Model



Purpose: Extract and distill clinical findings from X-rays into structured JSON labels.

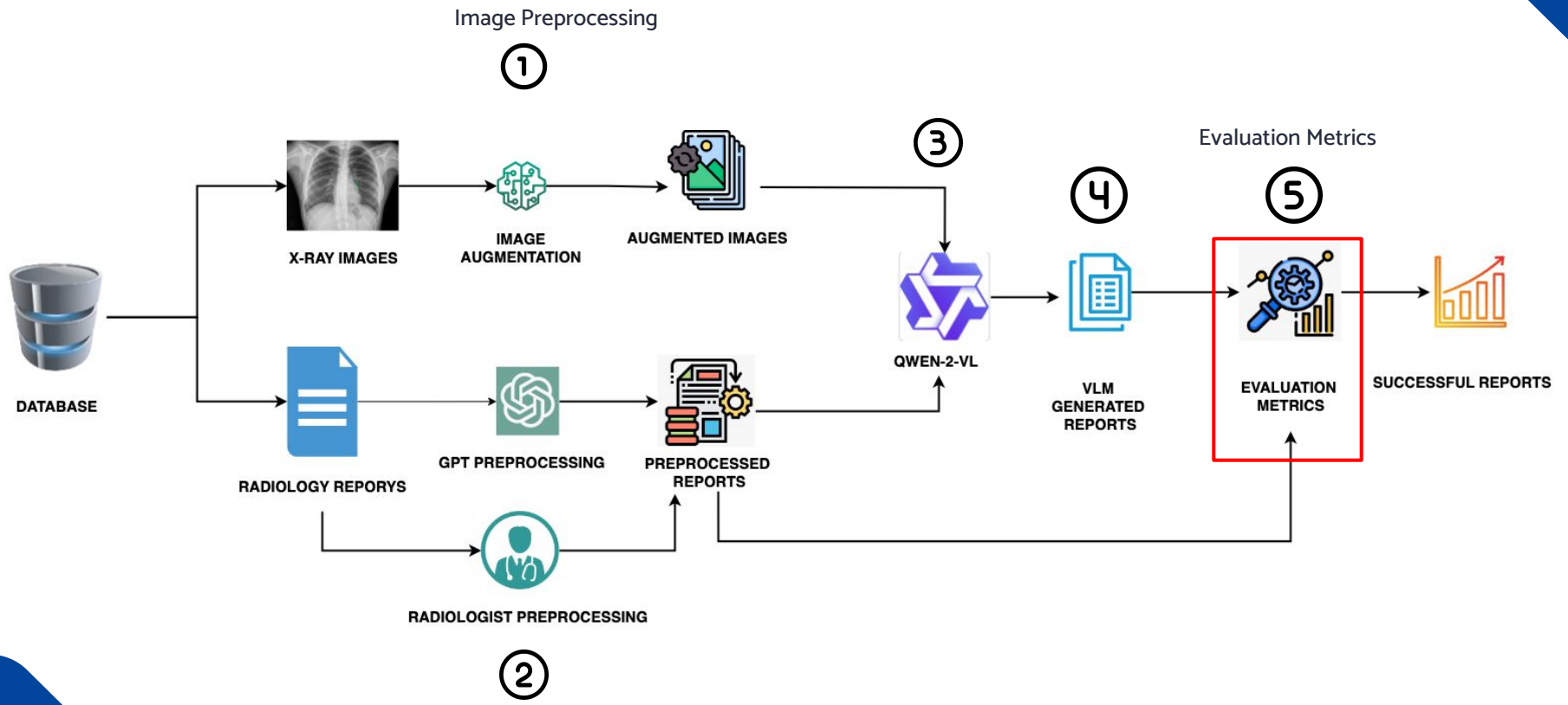
Use: Metadata filtering. Data Analysis.

Unstructured Fine-tuned Model

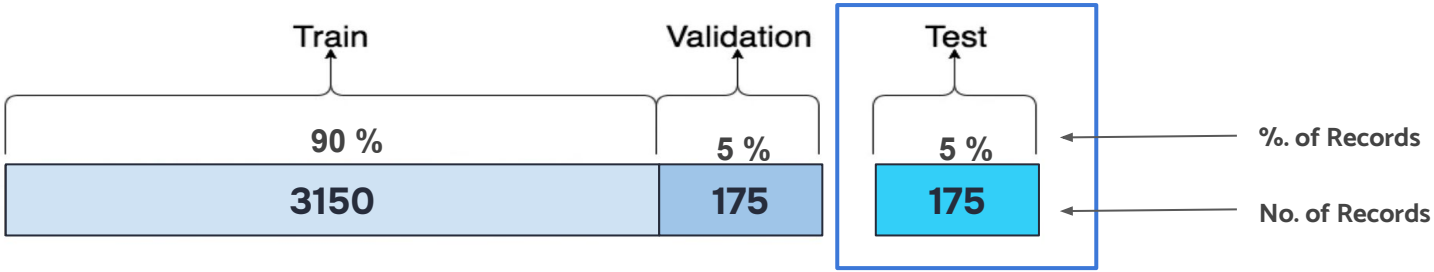
Purpose: Designed for customization, translating X-rays into a comprehensive radiology findings report tailored to the reporting style of the radiologist

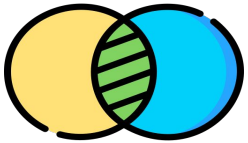
Use: Rapid prototyping. Drafting of reports during time-sensitive scenarios.

R2Gen2 Under the Microscope



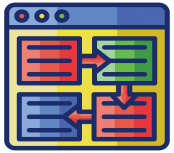
Test data and **evaluation methods** used to assess the performance of both models.





[Text Similarity]

- GLUE
- ROUGE
- Custom JSON Overlap



[Semantic Understanding]

- BERT Precision
- BERT Recall

[Clinical Evaluation]

- Radiologist Review

04

RESULTS

Novelties, Model Performance and Next Steps

Novelties we developed during our Journey

Personalized Report Generation Assistant

- ➔ VLM customized to mimic the specific reporting styles of individual radiologists and perform data analysis via feature extraction

Open Source Dataset Contribution

- ➔ Collaborated with a certified radiologist to decode dataset.
- ➔ Created a new, curated open source dataset.

Customized Evaluation Metric

- ➔ Development of our own token overlap metric for structured output generation (JSON).

GPT 4o V/s R2Gen2

The right lung appears darker, suggesting a potential pneumothorax (collapsed lung).The left lung has visible opacities, which might suggest consolidation or mass effect.The heart size and borders seem slightly shifted, likely secondary to the findings in the lung.The right hemidiaphragm seems elevated, which is often associated with pneumothorax or loss of lung volume.

The heart size is enlarged. There is left lower lobe airspace disease identified. There is moderate left pleural effusion. No visualized pneumothorax. Surgical suture material noted.

- ➔ Using Public Large Language Models does not guarantee security and HIPAA compliance.
- ➔ Writing style is generic and not specific to hospital/team
- ➔ Possess limited medical knowledge.

- ➔ R2Gen2 is Privately Hosted and does not leave Hospital System IT Architecture.
- ➔ Writing tailored to IU standards.
- ➔ Has learned the meanings of medical terms and providing accurate reports.

Incorrect | Partially Correct | Correct

Structured and unstructured model results

Structured Model Results

Customized JSON evaluation metric score → **0.7632**

Unstructured Model Results

Metrics	GLUE	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSum	BERT Precision	BERT Recall
Report Generation Model	0.8302	0.8132	0.7616	0.8139	0.8144	0.9134	0.8675

Next Steps for R2Gen2


Model **Overfitting**

MIMIC-CXR dataset, with **< 200,000** studies


The **model's size** leads to **slower inference times.**

Implement **model pruning** techniques .


Findings extraction and **customized generation** model

An **aftercare suggestion framework** using **Retrieval Augmented Generation**

Special Thanks

- Dr. Anjalie Verma, MBBS, DNB (Radiology)
- A Certified Radiologist In India For Helping with Data Imputation and Output Evaluation

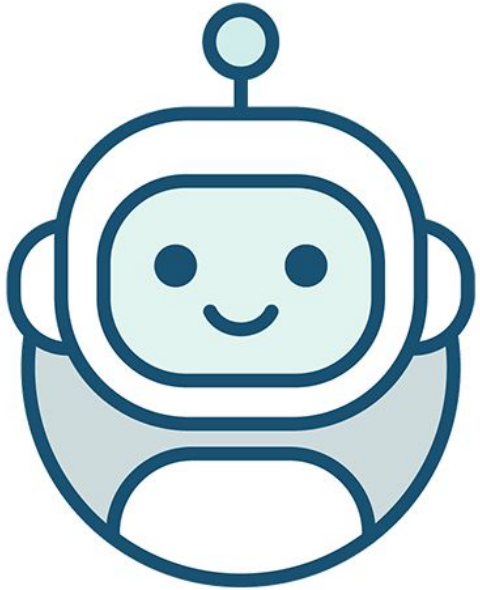
- Dr. Utku Pamuksuz, PHD
- Our Supervisor, Mentor for helping us overcome challenges and get to the finish line.

- We would also like to thank our Instructional Assistants for advice with every small hiccup and round the clock availability.
  - Tegan Keigher, MS
  - Mary Erikson, MS
  - Andrew Alvarez, MS

Questions?

Please redirect any further questions to :-

aayushv2001@uchicago.edu
ashmitam@uchicago.edu
rmohan22@uchicago.edu
leoli0907@uchicago.edu



APPENDIX

54% of Radiologists report experiencing “long-term, unresolved, job-related stress leading to exhaustion, cynicism, detachment from job responsibilities, and lacking a sense of personal accomplishment.”

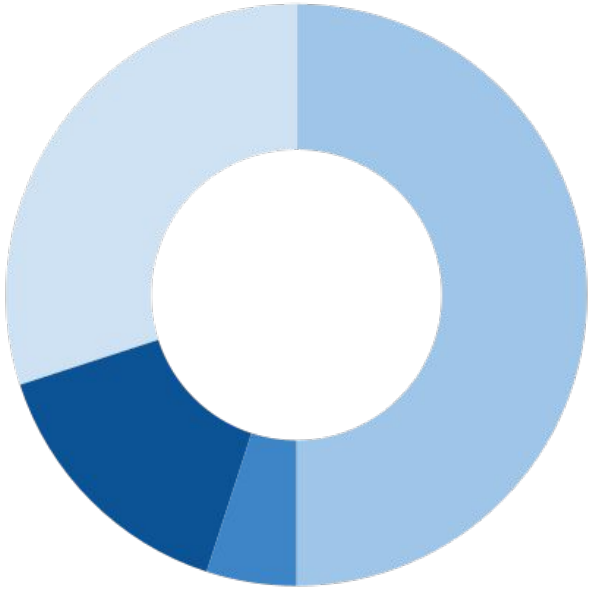
Meet Dr. Cooper

- Dr Cooper Is a radiologist at the IU Health hospital.
- She and her team work with upwards of **13000 X-Rays a Month.** This leads to her and her team working more than **35 hours per day altogether.**
- **50% of this time is wasted in documentation work*.**
- Because Radiologists like Dr. Cooper have such heavy load
  - they have developed a f**ormulaic reporting style** which is **unique to every team and hospital**.
  - This **hampers AI Adoption** due to it being generalized.
- Hospital Systems like IU Health **need to have a large amount of quality data for analysis** to spur on improvement.
  - Current Techniques focus on creating unstructured data **leaving a large quantity of data uncaptured**.
  - There is a focus on **data security** to adhere to **HIPAA**

Sinsky, C., Colligan, L., Li, L., Prgomet, M., Reynolds, S., Goeders, L., Westbrook, J., Tutty, M., & Blike, G. (2016). Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties. *Annals of Internal Medicine*, *165*(11), 753.

BACKGROUND

Radiology is among the top medical specialties experiencing burnout, with an average of 47.7 work hours per week and pressures from both clinical responsibilities and the need to keep up with advancements in imaging technology.



Time with Patients	**30%**
Documentation Work	**50%**
Administrative Tasks	**5%**
Miscellaneous Tasks	**15%**

***Example** of the **original report alongside the newly generated labels by GPT-4o Mini** illustrates the transformation:*
Each sentence is now represented as a structured JSON object, detailing the anatomical entity, the procedure performed, and the specific clinical findings associated with that entity

Each **sentence** from the report is converted into a **structured JSON format**

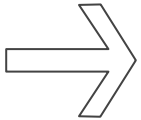
Original Report

Retrocardiac opacity which may represent atelectasis and or or small effusion is stable.

Right Lung is otherwise clear.

No pneumothorax.

NG tube tip below the diaphragm.



[{"anatomical entity":null,
"location_descriptor":null,
"procedure":[null,]
"clinical findings":[{
"clinical finding":"retrocardiac opacity",
"existence":"pos_dx",
"descriptive_term":"retrocardiac",
"observation":"opacity"
},
{
"clinical finding":"atelectasis",
"existence":"unc_dx",
"observation":"atelectasis"
},
{
"clinical finding":"or small effusion",
"existence":"unc_dx",
"descriptive_term":"or small",
"observation":"effusion",
}
]},

{"anatomical entity":"lungs"
"location_descriptor":Right,
"procedure":[null]
"clinical findings":[{
"clinical finding":"clear",
"existence":"neg_dx",
"observation":"clear"
}]},
{"anatomical entity":null,
"location_descriptor":null,
"procedure":[null]
"clinical findings":[{
"clinical finding":"pneumothorax",
"existence":"neg_dx",
"observation":"pneumothorax"
}]},
{ "anatomical entity":"diaphragm",
"location_descriptor":null,
"procedure":[null]
"clinical findings":[null]
}]}"""}

LITERATURE REVIEW

	Large Language Models	Vision Language Models
Other Healthcare	❑ Generalization in Healthcare AI: Evaluation of a Clinical Large Language Model ❑ InMD-X: Large Language Models for Internal Medicine Doctors	❑ GP-VLS: A general-purpose vision language model for surgery ❑ PA-LLaVA: A human pathology specialized large language vision assistant
Radiology	❑ Radiology-Llama2: Best-in-Class Large Language Model for Radiology ❑ R2GenGPT: Radiology Report Generation with Frozen LLMs	❑ RaDialog: A Large Vision-Language Model for Radiology Report Generation and Conversational Assistance ❑ Vision Transformer and Language Model Based Radiology Report Generation ❑ Visual Prompt Engineering for Medical Vision Language Models in Radiology

Our Research Area

*The Data comprising **3996 radiology reports and 6561 associated images of Chest X-Rays** is sourced from the **Indiana Network for Patient Care**.*

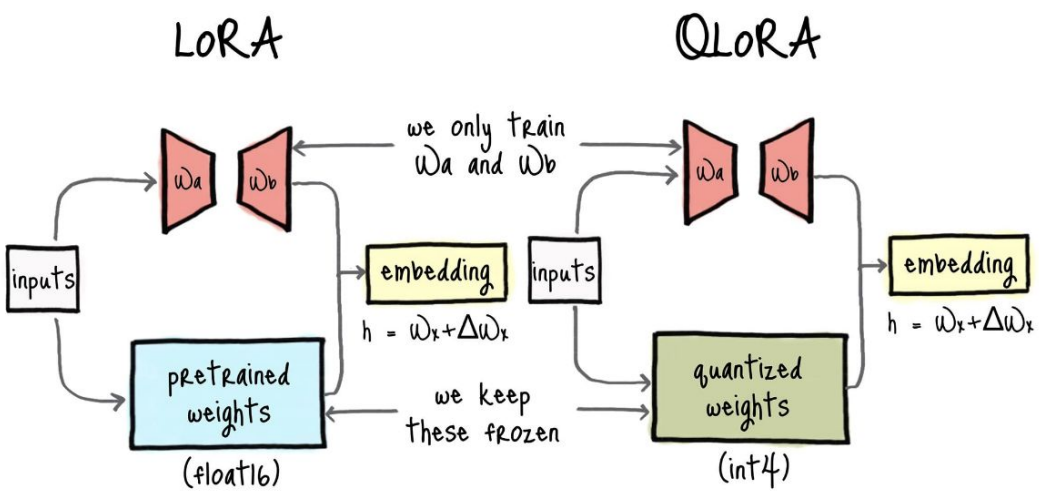
Data Source : Indiana Network for Patient Care

Data Description: 3996 radiology reports and 6561 associated images of Chest X-Rays.

Source Link : *https://openi.nlm.nih.gov/imgs/collections/NLMCXR_reports.tgz*

1486 Reports with XXXX, 3735 sentences with XXXX

QLoRA :QUANTIZED LOW-RANK ADAPTATION



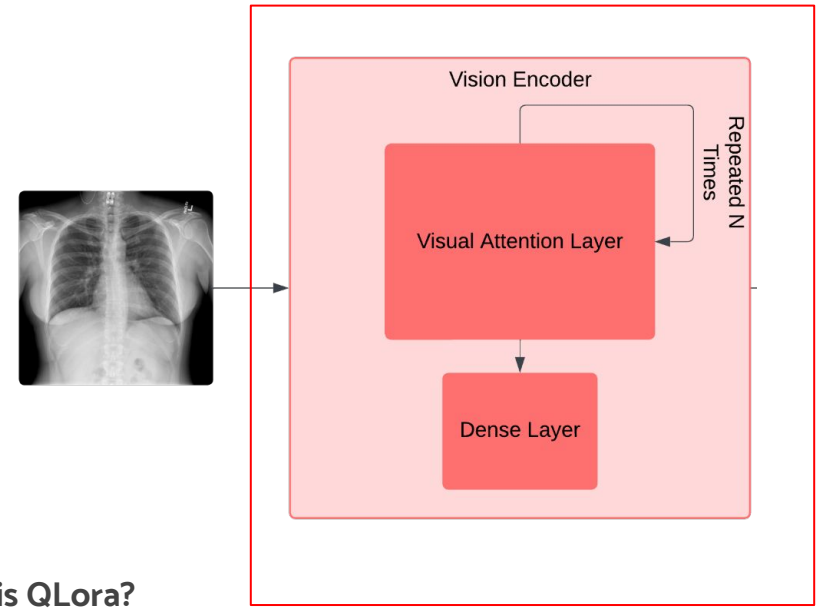
- QLoRA is a technique used for fine-tuning deep neural networks, making **training faster** and more efficient **in situations with limited computational resources**.
- QLoRA can be explained by breaking it down into two parts:-
  - **Quantization**:- This technique decreases the model's precision to decrease the size it holds in the memory and the speed of processing
  - **Low Ranking Adapters**:- This technique decreases the computation size by decreasing the order of the matrices which are used to issue updates to the model weights.

The fine-tuning process of the VLM involves several steps, beginning with training the Vision Encoder to interpret images through its complex architecture to perform object detection and and identifying anatomical structures

VISION ENCODER

- The VLM was fine-tuned using QLoRA, enabling us to selectively fine-tune sections, **optimizing resource utilization** while maintaining overall performance.
- We chose to tune the following components:
  - **Visual Attention Layer**: This Layer conducts attention over the image and creates understanding of the Input
  - **Dense Layer**: This Layer distills the gained knowledge into embeddings of set length.

Trained to interpret the X-Ray just as a radiologist would.



What is QLora?

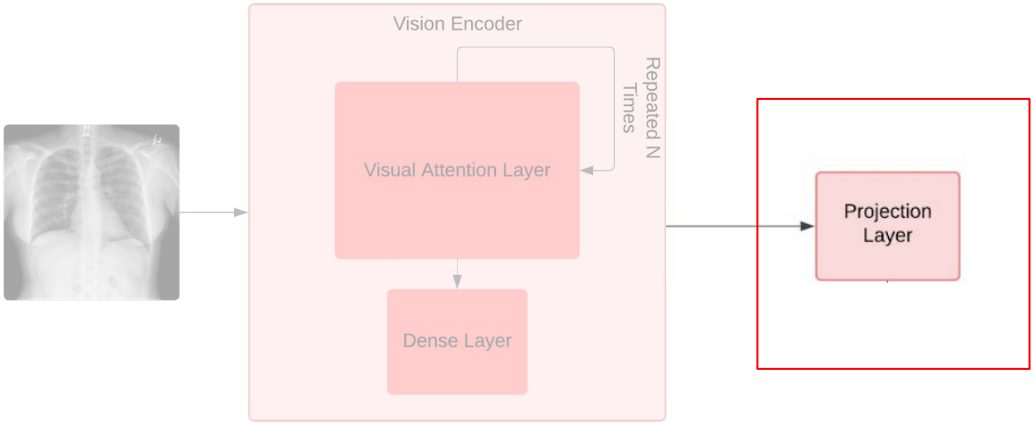
Method that simplifies training VLMs by reducing the size of the model using advanced compression techniques, while keeping its performance intact.

The next step is to take the information from the vision encoder and ***turn it into numbers that the model can understand and process.***

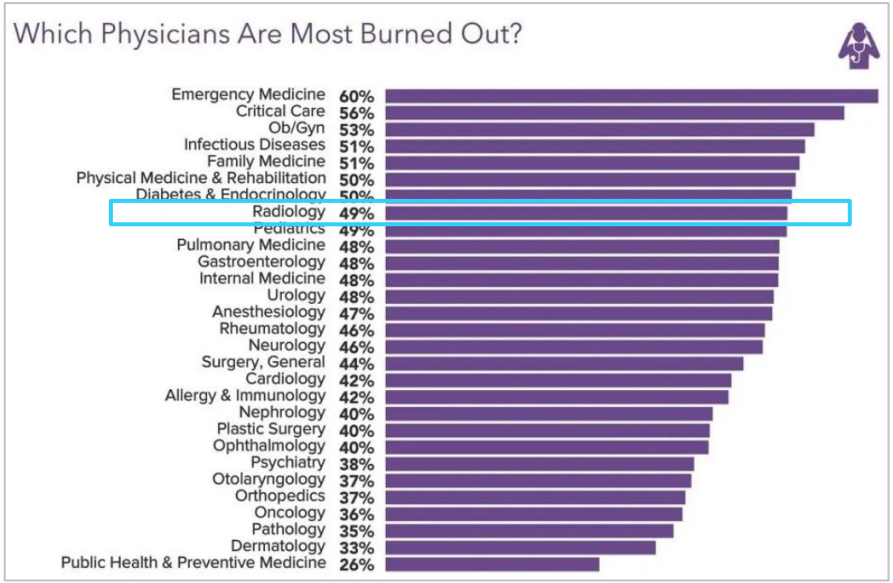
STEP 2 : PROJECTION LAYER

Acts like a **translator.**

This step ensures that the model can work with different types of data to make sense of both visuals and words.



54% of Radiologists report experiencing “long-term, unresolved, job-related stress leading to exhaustion, cynicism, detachment from job responsibilities, and lacking a sense of personal accomplishment.”



- ➔ High Imaging Volume
- ➔ Long Work Hours
- ➔ Administrative Tasks
- ➔ Pressure for Accuracy

*We have developed a highly adaptable model with **personalized customization** and **individual user specifications and preferences**.*

- ➔ Fine-tune the Vision Language Model overcoming GPU memory constraints.
- ➔ Address computational demands with quantization techniques.

- ➔ Adapt to each radiologist's distinct reporting style.
- ➔ Enable personalized analysis tailored to individual preferences.

- ➔ Design a supportive tool specifically for radiologists
- ➔ Provide preliminary findings to streamline workflow.
- ➔ Reduce time spent on routine assessments.

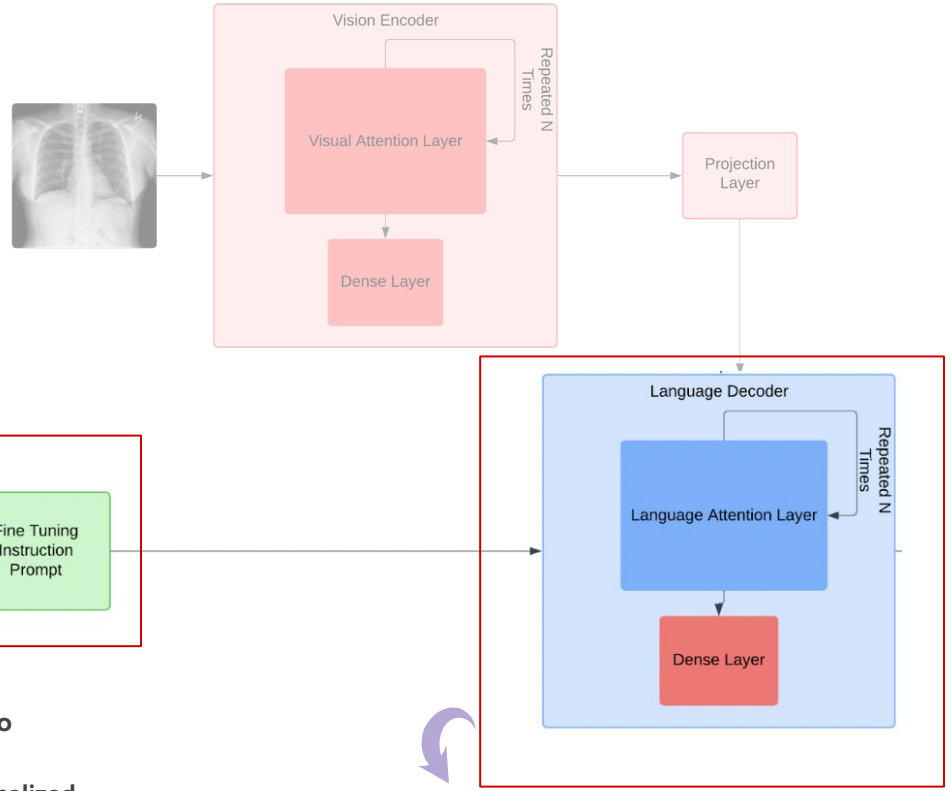
FINE TUNING PROCESS

PROMPT

The Fine Tuning Instruction Prompt refers to the **Instruction given to the model** informing it of its task.

LANGUAGE DECODER

- In the decoder block we have kept **everything frozen** to maintain the model's understanding of language.
- The Final Dense Layer was **unfrozen to change the most probable token** at each step for our use case.



Instructs the model to communicate in the distinctive and personalized style of the doctor.

Conveys the results in the personalized tone of the radiologist utilizing the model.

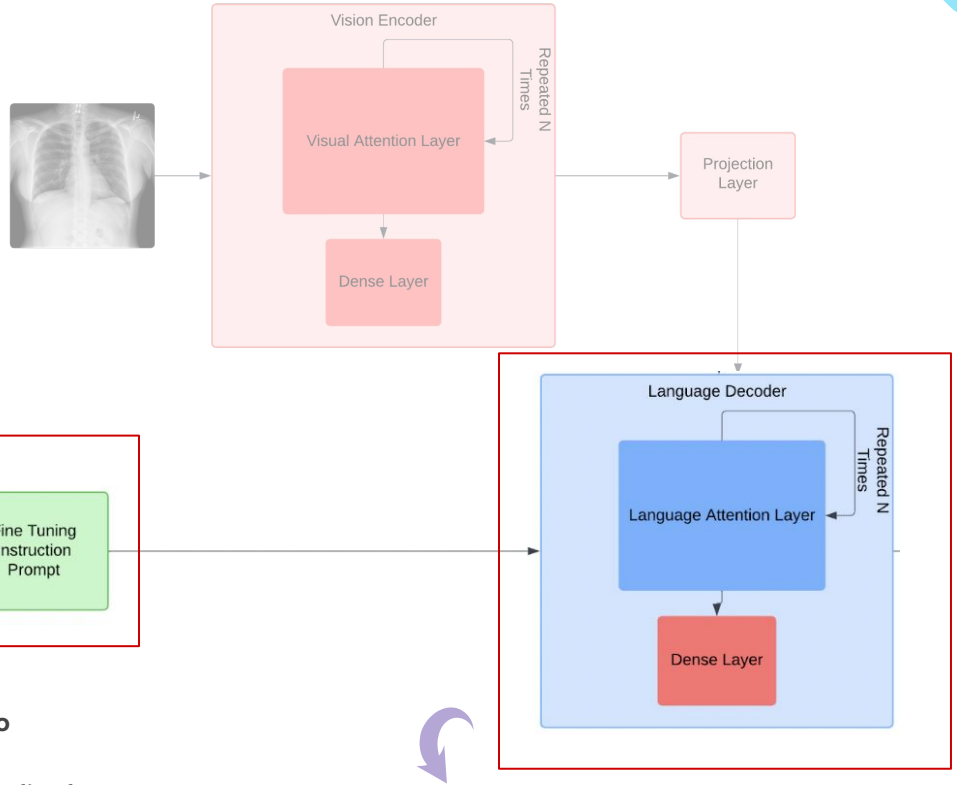
*The fine-tuning process of the VLM progresses by **prompting the model** and **training the Language Decoder to recognize** and adapt to the nuances of the **radiologist's particular reporting style.***

STEP 3 : PROMPT

The Fine Tuning Instruction Prompt refers to the **Instruction given to the model** informing it of its task.

STEP 4 : LANGUAGE DECODER

A skilled assistant that learns how a radiologist writes and speaks.
It takes the findings and translates them into reports using the same tone, style, and language the radiologist would use, making the results personalized.

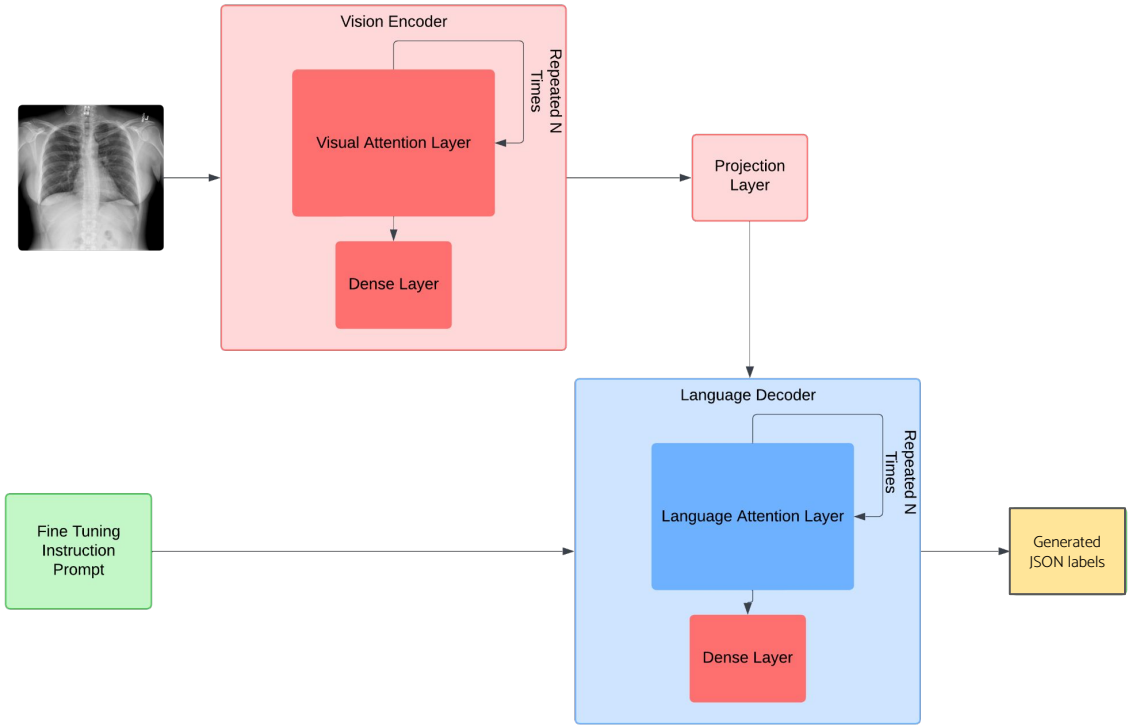


Instructs the model to communicate in the distinctive and personalized style of the doctor.

Conveys the results in the personalized tone of the radiologist utilizing the model.

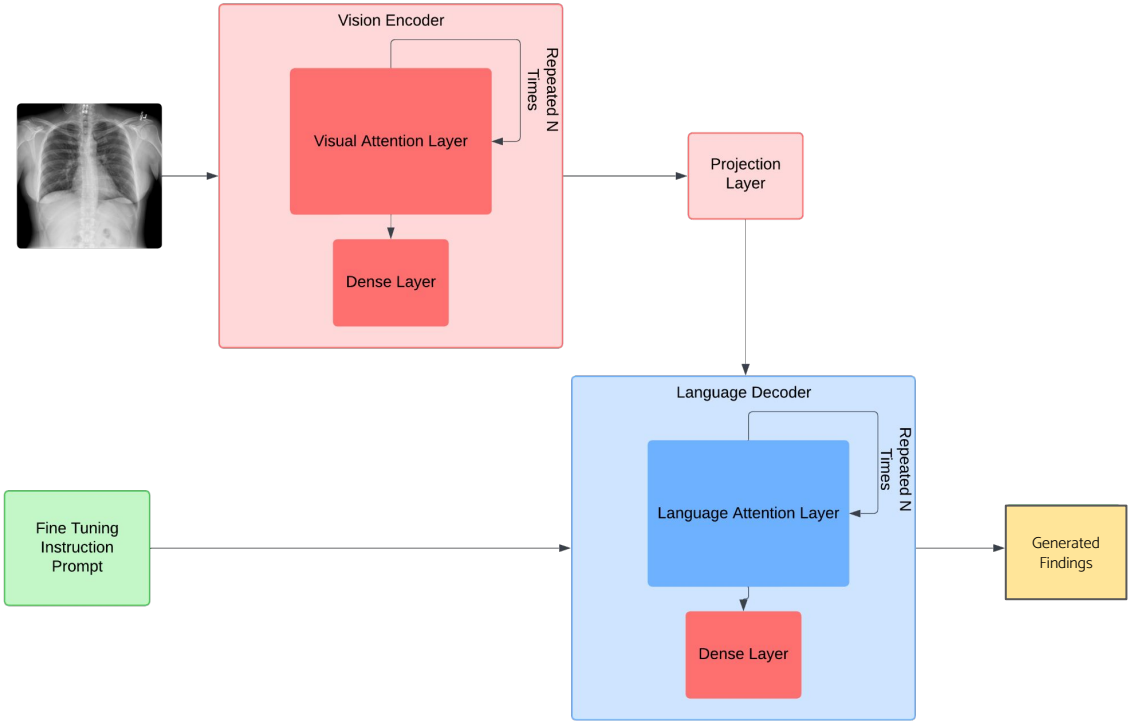
FINE TUNING PROCESS : Model I

- The first model we created was the **Image to JSON scheme model.**.
- This model was focused on **extracting** and distilling the symptoms and **clinical findings** found in the XRAY into **representative JSON Labels**.
- This is **useful** for radiologists as **metadata for filtering and quick summarization.**



FINE TUNING PROCESS : Model II

- The second model we created was the **Image to text reports scheme model**.
- This model was focused on understanding and **capturing the symptoms and clinical findings** found in the XRAY into a full radiology **findings write-up of a radiology report.**
- It serves as a valuable tool for radiologists, **enabling rapid prototyping and drafting of reports during time-sensitive scenarios**.



*A deep dive into the **VLM architecture**.*

Qwen2.5

Image Input

Extracts high-level visual features from X-Ray images

Vision Encoder

Maps encoded visual features into a compatible latent space

Projection Layer

Generated JSON labels/Report Findings

Attention

Heart
Normal,Mediastinum
Normal...

You are an Expert Radiologist. Your only Task is to generate succint and accurate clinical finidings from the given image.

Instruction Prompt

Text Decoder

Focuses on the most relevant parts of the input to enhance feature understanding and output accuracy.

Generates descriptive or diagnostic text based on encoded visual and/or contextual input.

LIMITATIONS

- ❏ The model currently exhibits overfitting, primarily due to the limited number of examples in the dataset, despite the use of augmentations to mitigate this issue.
- ❏ Additionally, the imbalance in the dataset, with a higher proportion of normal reports compared to abnormal ones, has led the model to favor generating normal reports, reducing its accuracy in handling abnormal cases.
- ❏ The model does not generate impressions, diagnoses, or treatment plans, focusing solely on findings extraction and draft preparation.

NEXT STEPS

- ❏ **Expand Dataset:** Request access to the **MIMIC-CXR dataset**, which includes over **200,000 studies** (both normal and abnormal), to enhance the diversity and size of the training dataset.
- ❏ **Optimize Model Performance**: Implement **model pruning technique**s to create smaller, more efficient models while maintaining performance.
- ❏ **Integrate Aftercare Framework**: Once the model achieves sufficient accuracy and reliability, develop an **aftercare suggestion framework using Retrieval-Augmented Generation (RAG)** to fully automate the radiology pipeline.

NEXT STEPS

- **Expand Dataset:** Request access to the **MIMIC-CXR dataset**, which includes over **200,000 studies** (both normal and abnormal), to enhance the diversity and size of the training dataset.
- **Optimize Model Performance**: Implement **model pruning technique**s to create smaller, more efficient models while maintaining performance.
- **Integrate Aftercare Framework**: Once the model achieves sufficient accuracy and reliability, develop an **aftercare suggestion framework using Retrieval-Augmented Generation (RAG)** to fully automate the radiology pipeline.

Custom Metric

- Text Evaluation has had many breakthroughs with many evaluation suites coming up in recent times.
- Structured Generation is more of new focus for the community hence there is not a lot of work done in this arena.
- We have created our own token overlap metric for evaluating the JSON Labels generated from our model compared to the findings in the text report.
- The metric works by checking if the values in the JSON key-value pairs are present in the given report acting as a check to see if we have captured all the clinical findings and symptoms in the report.

The three evaluation methods we use to assess the performance of both models.

Text Similarity Metrics

GLUE (General Language Understanding Evaluation): Measures the overlap between the model-generated text and reference reports, useful for **short structured outputs**.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation): Evaluates recall-based overlaps, often used for **summarization tasks**.

Customized JSON Evaluation:

Semantic Understanding Metrics

BERT Precision: Measures the **relevance of the retrieved tokens** or segments by a BERT-based model **compared to the ground truth**.

BERT Recall: Evaluates **how well** a BERT-based **model retrieves all relevant tokens** or segments from the ground truth.

Human Evaluation Metrics

Expert Review Scores: **Radiologists assess** the quality of the generated reports based on relevance, accuracy, and comprehensiveness.