REPUBLIC OF TURKEY
YILDIZ TECHNICAL UNIVERSITY
DEPARTMENT OF COMPUTER ENGINEERING

VISUAL QUESTION ANSWERING FOR MEDICAL IMAGES

21011013 — ARINÇ AYDEMİR
21011902 — SALİH DEMİROZ

SENIOR PROJECT

Advisor
Prof. Dr. Mine Elif KARSLIGİL

November, 2025

ACKNOWLEDGEMENTS

We are grateful to our advisor, Prof. Dr. Mine Elif KARSLIGİL, for her oversight and technical advice throughout the development phase. Furthermore, we appreciate the facilities and support provided by Yıldız Technical University.

ARINÇ AYDEMİR
SALİH DEMİROZ

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

3-D	Three-Dimensional
AI	Artificial Intelligence
ANN	Artificial Neural Network
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy (metric for text similarity)
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DICOM	Digital Imaging and Communications in Medicine
ECE	Expected Calibration Error (model confidence metric)
EM	Exact Match (metric for string-level accuracy)
FN	False Negative
FP	False Positive
GPU	Graphics Processing Unit
G.T	Ground-Truth
ImageCLEF	Image Cross Language Evaluation Forum (benchmark and competition platform)
IoU	Intersection over Union
LLaVA	Large Language and Vision Assistant (multimodal architecture)
LLM	Large Language Model
LoRA	Low-Rank Adaptation (parameter-efficient fine-tuning method)
LSTM	Long Short-Term Memory
MedBLIP	Medical Bootstrapping Language-Image Pretraining (model name)
MedVInT	Medical Visual Instruction Tuning (vision-language model)

METEOR	Metric for Evaluation of Translation with Explicit ORdering (evaluation metric for text generation)
ML	Machine Learning
MRI	Magnetic Resonance Imaging
NEM	Normalized Exact Match (metric ignoring punctuation and case)
NLP	Natural Language Processing
NumPy	Numerical Python (Python library for numerical computation)
PyTorch	Deep learning framework for machine learning and AI research
QA	Question and Answer
RAM	Random Access Memory
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic (curve metric for confidence analysis)
ROUGE	Recall-Oriented Understudy for Gisting Evaluation (evaluation metric for text generation)
SGD	Stochastic Gradient Descent
TensorFlow	Deep learning framework for machine learning and AI applications
TL	Turkish Lira (currency unit)
TN	True Negative
TP	True Positive
UI	User Interface
VLM	Vision–Language Model (multimodal architecture)
VQA	Visual Question Answering
VQA-Med	Visual Question Answering – Medical (ImageCLEF benchmark dataset)

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

VISUAL QUESTION ANSWERING FOR MEDICAL IMAGES

ARINÇ AYDEMİR
SALİH DEMİROZ

Department of Computer Engineering
Senior Project

Advisor: Prof. Dr. Mine Elif KARSLIGİL

Medical imaging technologies have been developing rapidly in recent years. Because of that, the amount of healthcare data is becoming huge. This creates a lot of pressure on radiologists and doctors. Reading these images take long time. It requires expert knowledge to do it right. So developing AI systems is important to help them and solve this problem. It can speed up the process and stop errors.

In this project, our goal is to build a "Medical Visual Question Answering" (Med-VQA) system. Basically, the system looks at clinical images and answers questions about them. We used the VQA-Med 2019 dataset for our experiments. We tried to compare how different Vision-Language Models (VLM) work in the medical field. We tested 4 models. These are: LLaVA-1.5, InstructBLIP, PaliGemma and Qwen2.5-VL. We used Low-Rank Adaptation (LoRA) and quantization methods for fine-tuning. Because we wanted to run them on our hardware efficiently. Results showed that performance depends on question type. For example, asking about organ system or abnormality. PaliGemma model got the best results among them. It achieved the highest accuracy and BLEU scores. This is because its image encoder (SigLIP) catches small details better. LLaVA-1.5 was stable during training. For result, we saw that multimodal models can be used in medicine. But image encoder is a very important factor.

Keywords: Medical Visual Question Answering, Multimodal AI, LLaVA, PaliGemma, InstructBLIP, Deep Learning, Radiology.

ÖZET

ARINÇ AYDEMİR
SALİH DEMİROZ

Bilgisayar Mühendisliği Bölümü
Bitirme Projesi

Danışman: Prof. Dr. Mine Elif KARSLIGİL

Son yıllarda tıbbi görüntüleme teknolojilerindeki hızlı gelişmeler, sağlık verilerinin çok büyük hacimlere ulaşmasına neden olmuştur. Bu durum, radyologlar ve doktorlar üzerinde ciddi bir iş yükü ve zaman baskısı oluşturmaktadır. Tıbbi görüntülerin yorumlanması hem zaman alıcıdır hem de yüksek düzeyde uzmanlık gerektirir. Bu nedenle, yapay zekâ tabanlı sistemlerin geliştirilmesi, sağlık profesyonellerine destek olmak açısından büyük önem taşımaktadır. Bu sistemler, tanı sürecini hızlandırabilir ve insan kaynaklı hataların azaltılmasına katkı sağlayabilir.

Bu projede amacımız, "Tıbbi Görsel Soru Cevaplama" (Med-VQA) sistemi oluşturmaktır. Temel olarak, sistem klinik görüntülere bakar ve bunlar hakkında soruları cevaplar. Deneylerimiz için VQA-Med 2019 veri setini kullandık. Farklı Görsel-Dil Modellerinin (VLM) tıp alanında nasıl çalıştığını karşılaştırmaya çalıştık. 4 model test ettik: LLaVA-1.5, InstructBLIP, PaliGemma ve Qwen2.5-VL. Donanımımızda verimli bir şekilde çalıştırabilmek için ince ayar için Düşük Dereceli Adaptasyon (LoRA) ve niceleme yöntemlerini kullandık . Sonuçlar, performansın soru türüne bağlı olduğunu gösterdi. Örneğin, organ sistemi veya anormallik hakkında soru sorulduğunda, PaliGemma modeli bunlar arasında en iyi sonuçları aldı. En yüksek doğruluk ve BLEU puanlarına ulaştı. Bunun nedeni, görüntü kodlayıcısının (SigLIP) küçük ayrıntıları daha iyi yakalamasıdır. LLaVA-1.5 eğitim sırasında istikrarlıydı. Sonuç olarak, çok modlu modellerin tıpta kullanılabileceğini gördük. Ancak görüntü kodlayıcı çok önemli bir faktördür.

Anahtar Kelimeler: Tıbbi Görsel Soru Cevaplama, Çok Modlu Yapay Zeka, LLaVA, PaliGemma, InstructBLIP, Derin Öğrenme, Radyoloji.

1
INTRODUCTION

In recent years, medical artificial intelligence got a lot of attention, particularly in areas related to image based diagnosis. Medical images such as CT, MRI, and X-ray scans contain a large amount of diagnostic information. However, interpreting these images correctly often requires expert knowledge and considerable amount of time. This situation places a serious workload on radiologists and clinicians, particularly as the volume of medical data continues to grows significantly every passing year.

Visual Question Answering (VQA) aims to combine visual understanding with natural language processing to support clinical education, clinical decision-making, and patient education. Applied to medicine, this concept can help clinicians by allowing them to obtain rapid and accurate information directly from medical images. Instead of manually examining every detail in an image, clinicians can just ask their questions to the VQA models

In this project, titled Visual Question Answering for Medical Images, we are studying the application of VQA techniques to clinical images. Our main focus is to evaluate how modern vision language models behave when they are adapted to medical data. We are aiming to fine tune couple of visual language models using VQA Med dataset, and we want to analyze model's performance across different question types. By doing this we are aiming to be better at understanding the advantages and disadvantages of current multimodal architectures in the medical imaging domain and to assess their potential as supportive tools rather than standalone diagnostic systems.

1.1 Objective

Our main objective is designing a functioning medical VQA system. It should interpret images and provide answers to the questions asked by clinicians.

Overall, the project aims to advance the integration of visual and textual intelligence. By doing so, it seek to establish a basis for clinical systems.

1.2 Preliminary Review

In this section we are defining boundaries within which study has been conducted. Also, the section specifies data sources to be utilized and presents core components involved in the analysis process. We are reviewing the necessities and limits of our work.

1.2.1 Project Need

Medical data is exploding, so the doctors need to make decisions fast. We think this creates a bottleneck in the system. Radiologists have too much work on their own and which is could be leading to delays or even errors.

So, as we can see, there is a need for automated systems. Automated systems that can look at image and answer questions and provide assistance to doctors. We believe Medical VQA is a great option here. If we build a solid model that interprets images correctly, it acts as a second eye for experts. It essentially makes healthcare accessible.

1.2.2 Project Scope

This project is focusing on evaluating pretrained vision language models on the VQA-Med dataset. The system is not designed for real clinical use yet and we are working with a limited dataset and limited capabilities. We are trying to learn and analyze how models are working. We want to compare results across different image and question types and identify the strengths and weaknesses of current models in medical VQA.

Our work is mainly focused on understanding how vision language models working and training, and how different models behaving and generating results.

2
LITERATURE REVIEW

2.1 Similar Studies

In this section, we are reviewing similar research and paperwork to our project.

At first, we started our review before transformer based projects even started which are the projects before year 2020. Early studies and papers mainly focused on small datasets. Earlier projects had around 4000 images and 15000 QA pairs and used neural networks or deep learning for question answering. Deep learning models combined with modern transformers have increased and made it easier to do VQA projects.[1].

After that, researchers began to build larger datasets. Pathology resources expanded question space to 30000 pairs [2]. In parallel, large multimodal archives were extracted from scientific literature. Now, from our research, we found out the latest projects can have up to one million images and QA pairs [3]. Pretrained encoders improved zero shot performance. This show value of big data.

Later surveys pointed out that medical images pose unique challenges, such as the fact that abnormalities are most of the time are difficult to detect. [4]. Researchers found that attention mechanisms help networks and it aligns the image regions with text.

Recent development in LLMs and VLMs made VQA a generative problem. From our reviews we can see standard fine-tuning of a language model could hurt its generalization capability, especially if a dataset used for fine-tuning has a small size and is domain-specific as in our case.[5].

2.2 General Approaches

In this section our goal is explaining general approach to VQA and medical VQA projects.

From our research we see that Med-VQA projects are grouped into three categories: discriminative, attention based, and generative. Before the advent of transformers VQA studies at general used discriminative approach and in which visual features extracted by CNNs and RNNs handled question encoding [1]. We see that from early approaches combined image and question features to predict answers from fixed sets [1]. When technology started to develop more attention mechanism were introduced to the world and it was ground breaking for VQA projects. It massively helped to link the visual and textual information [4].

Generative and instruction based models have become more important recently. These models treat QA as text generation task [5]. Vision language models generate free form responnses. Instruction tuning with domain specific prompts proven effective [6].

Building on these developments, for our project we had comparative approach on four different models. LLaVA is selected for its capability [6]. InstructBLIP is included for Q-Former architecture [7]. Furthermore, PaliGemma is incorporated. We also trained on Qwen because it is a large LLM, so we can see if our training is dependent on base LLM performance or not.

3
SYSTEM ANALYSIS AND FEASIBILITY

In this chapter, we will talk about our project's plan and it's feasibility based on couple of factors listed. We will give short preview of the project and early steps of the project. And after that we will analyse it's feasability.

3.1 System Analysis

In this section we are presenting workflow of the project step by step. Our goal is showing aim of the project and how it's functionality going to work below.

3.1.1 Workflow and Development Phases

System development process defined with the steps below:

- **Dataset Selection:** We are using publicly available VQA-Med 2019 dataset. The dataset contains anonymized radiology images paired with QA pairs in .txt files.
- **Data Preprocessing:** In this phase we are cleaning the dataset and formatting it for model training.
- **Model Selection and Fine-Tuning:** This part is the most important part of our project. We searched on sites like Github and Huggingface to find best suited VLM's for the project. After careful research four models selected: LLAVA 1.5, InstructBLIP, PaliGemma and Qwen. We fine tuned all models and trained them in VQA2019 dataset. We also used parameter efficient techniques like (LoRA). LoRa technique made fine tuning possible for university level thesis since training the entire model would be unfeasible and require substantial resources.
- **User Interface (UI):** After we finished training of VLMs, we are planning to design a user friendly interface. With this interface we can input a medical

image and see generated answer from one of our models. Our interface should feature all of the models for answer generating

- **Performance Benchmarking and Evaluation:** In this bit we are analysing results of our training. Classical metrics like BLEU and accuracy will probably used for assessment

3.2 Feasibility Analysis

Feasibility analysis determines if the Med-VQA system can be realistically developed within available technical and economic constraints. For this we are evaluating it with four main parts. Technical, hardware, legal, and economic feasibility.

3.2.1 Technical Feasibility

In this section, we are analysing technical feasibility of the project.Technical feasibility of the project is high at first glance. Since we already know how GPU intensive deep learning and transformer based LLMs. For this project we are using Python as programming language and PyTorch and Tensorflow for deep learning frameworks. Based on our studies and researches, we could use DeepSpeed library for GPU and RAM optimization.

Thanks to techniques like LoRa we don't need to fine tune large models from scratch. After our careful researches we can say this project is technically feasible for university thesis because of cloud services and optimization methods.

3.2.1.1 Hardware Feasibility

As students, we don't have the hardware for the Medical VQA project and LLM training in our home. But thanks to our university's workstation and Google Colab's cloud based GPUs project is feasible.

Workstation provides us with NVIDIA RTX A5000 with 24GB VRAM and Colab provides us with both T4 and A100 GPUs

Both of us can use personal laptops, and our laptops have NVIDIA RTX 3060, which can't run training in a feasible time, but can run testing.

Table 3.1 Hardware Specifications of Current Systems

Specification	**Laptop 1 (Monster Tulpar)**	**Laptop 2 (Lenovo Legion 5 Pro)**
Processor	Intel Core i7-12700H (12th Gen)	AMD Ryzen 7 5800H
GPU	NVIDIA RTX 3060 (6 GB)	NVIDIA RTX 3060 (6 GB)
RAM	32 GB	16 GB

3.2.2 Legal Feasibility

We access the dataset used from official VQA-Med 2019 Github. Dataset is publicly available for research and all medical images are fully anonymized. Images have no personal information so no doxxing ensured. Dataset is only used for research purposes and in the official Github repository states that the dataset is available for research use, provided that appropriate citation is given.

In terms of model selection, three of the models are open source and public. Only Paligemma requires Huggingface access token and after that we can use it as well

3.2.3 Economic Feasibility

In this section we are analysing expenses of the project. We are not using any priced LLMs and all of the models used in the project are free without token limit.

Project is carried by two researches and man-month table provided below.

Table 3.2 Economic Feasibility Summary (Values in Turkish Lira, 2025, 4-Month Project Period)

Cost Category	**Monthly Cost (TL)**	**Total Cost (TL)**
Software and Tools	0	0
Dataset (VQA-Med 2019)	0	0
GPU Workstation (Usage + Depreciation)	10,000–11,250	40,000–45,000
Electricity Consumption	100–125	400–500
Google Colab Pro Subscription	165	660
Cloud / Backup Storage (Optional)	125	500
Personal Expenses per Researcher (×2)	3,500	14,000
Total Estimated Cost	**14,000–15,000**	**55,000–60,000**

3.2.4 Labor and Time Feasibility

For this section we are providing detailed Gantt chart in Figure 3.1. To organize our project's workflow development tasks are shared evenly between two team members so we can ensure balanced workload

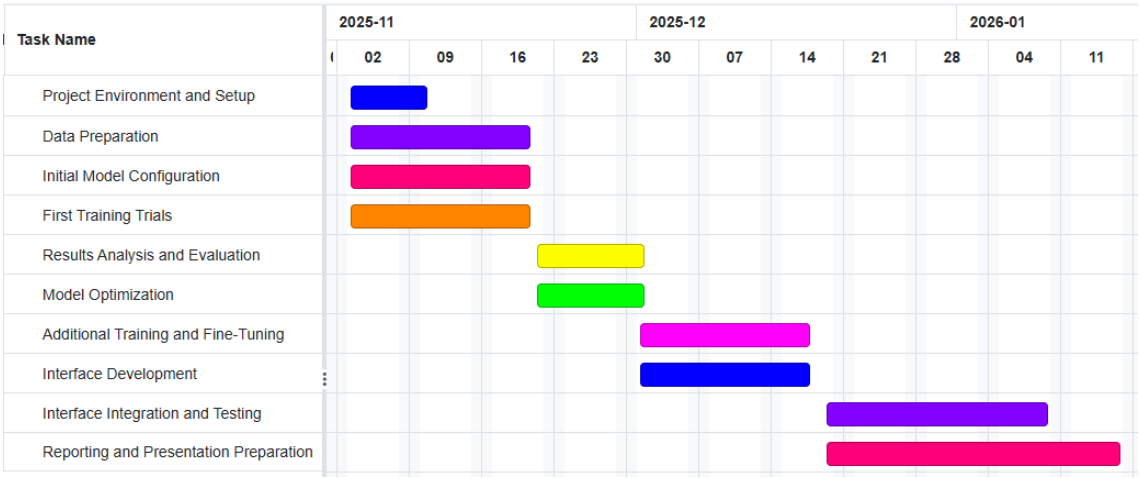


Figure 3.1 Gantt Chart of Timeline

4
SYSTEM DESIGN

4.1 Materials and Dataset

For this project we use the official ImageCLEF 2019 VQA-Med set. This dataset contains different types of medical images (e.g., CT, CTA, MRI, ultrasound, X-ray) and their QA pairs. Each QA pair is defined over a specific image and divided to four categories such as the imaging modality, anatomical region or the organ, imaging plane, or the abnormality.

The training split used in this work contains **12,792** QA pairs associated with **3,200** unique medical images. The validation split contains **2,000** QA pairs linked to **500** unique medical images.

In addition to these, the official set provides **test set** consisting of **clinical questions with reference answers**. This test split includes **500 images** paired with **500 QA pairs**, and is used evaluation of our training

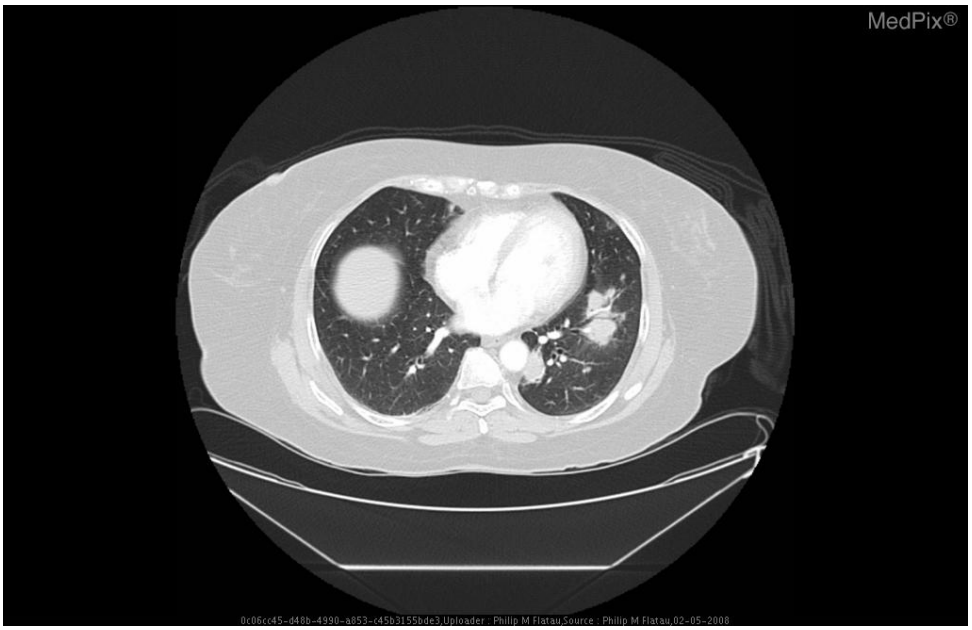
4.1.1 Question Categories

The VQA-Med questions have four categories listed below:

- **Modality**: identifies the imaging technique or MRI weighting used (e.g., CT, CTA, MRI T1, ultrasound).
- **Organ / organ system**: determines the anatomical region or system captured (e.g., lung, mediastinum, pleura).
- **Imaging plane**: specifies the geometric orientation of the slice (e.g., axial, coronal, longitudinal).
- **Abnormality**: identifies the visible disease, lesion, or clinical findings (e.g., arteriovenous malformation, ectopic pregnancy etc).

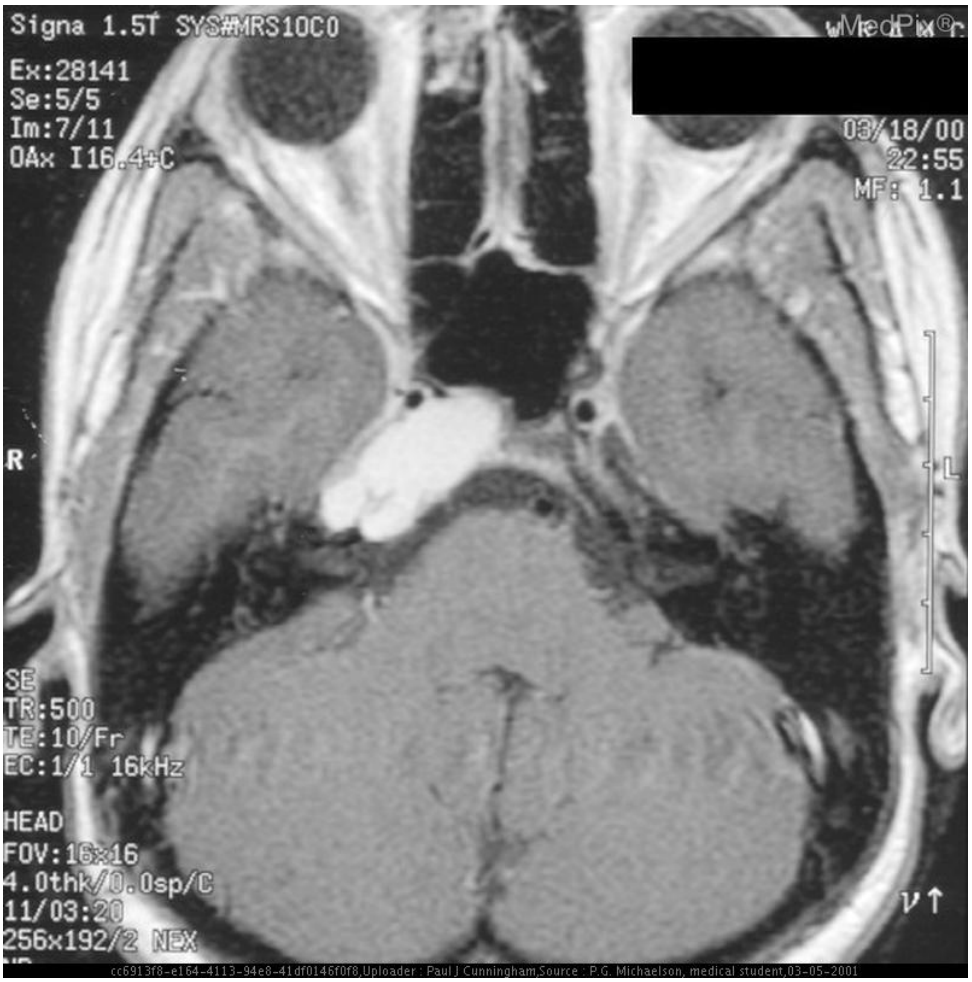
4.1.2 Representative Examples

Here below we are showing three examples of images and their QA pairs to explain how our dataset works and what type of images we are using for this project.



Question	**Answer**
what kind of image is this?	cta – ct angiography
which plane is this image taken?	axial
which organ is captured by this ct scan?	lung, mediastinum, pleura
what is abnormal in the ct scan?	cryptococcal pneumonia in an immunocompetent host

Figure 4.1 CTA chest image with QA pairs.



Question	Answer
what is the mr weighting in this image?	t1
what plane is seen?	axial
what organ system is imaged?	face, sinuses, and neck
what abnormality is seen in the image?	cholesterol granuloma of the petrous apex

Figure 4.2 Head MRI example with QA pairs.

Question	Answer
what type of imaging modality is used to acquire the image?	us – ultrasound
what plane is this ultrasound in?	longitudinal
which organ system is imaged?	genitourinary
what is the primary abnormality in this image?	ectopic pregnancy

Figure 4.3 Ultrasound image with QA pairs.

4.2 Methods

In this section we explain how framework of the project preprocessed, used and evaluated.

4.2.1 Data Preprocessing

The training and validation sets of the VQA-Med 2019 dataset contains medical images and QA pairs in .txt files. Each image have question about it's modality/plane/organ/abnormality and G.T of it

All images are loaded and normalized using a standard preprocessing pipeline [8], and each QA pair is transformed into a unified multimodal instruction format.

The model is trained to act as a medical assistant and produce short and image-grounded answers[6].

We are creating short answers to match the test set as a medical VQA project because we can't use semantic based metrics since we can score high scores by stating wrong illness in the same organ or area which would be bad in real life situation

The instruction prompt is:

"You are a medical visual question answering assistant. Answer briefly and only based on the image content."

This prompt design encourages the model to rely on the depicted anatomy and pathology rather than memorized textual patterns. All prompts follow the same template, ensuring uniformity across the training set [9].

4.2.2 Model and Fine-Tuning Procedure

We trained and used four models below here:

LLaVA 1.5 (7B) [6]: Utilizes a CLIP-ViT-L/336px vision encoder connected to a Vicuna LLM via a two-layer MLP projection. It is known for strong general-purpose visual reasoning [9].

InstructBLIP: employs a Query Transformer (Q-Former) to align visual features from the frozen image encoder with the frozen LLM. This architecture is specifically optimized for instruction-tuning tasks.

PaliGemma: A versatile open VLM by Google that combines a SigLIP vision encoder with the Gemma language model, designed for fine-grained image understanding. You need a hugging face account also access token to utilize Paligemma .

Qwen-VL: Integrates a vision encoder with the Qwen large language model through a unified multimodal framework. It supports high-resolution visual inputs and demonstrates strong performance in visual grounding, document understanding, and complex vision-language reasoning tasks.

Fine-tuning for all models is conducted in Google Colab using NVIDIA A100 GPUs. To reduce memory consumption while maintaining performance, Low-Rank Adaptation (LoRA) and quantization (4-bit/8-bit) are applied uniformly across all models [10].

We also save every epoch by creating a checkpoint folder so we do not lose our progress of training.

It is simply not possible for us to fine-tune entire model because it would take multiple

GPU's and larger RAM to do that.

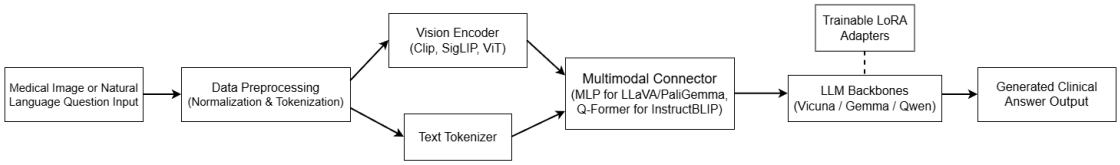


Figure 4.4 Workflow of the Med-VQA System.

Figure 4.4 Workflow of the Med-VQA system applied across all four architectures (LLaVA, InstructBLIP, PaliGemma, Qwen2.5-VL). The system processes medical images using specialized Vision Encoders (e.g., CLIP, SigLIP) and aligns them with text instructions. Instead of full-model training, Low-Rank Adaptation (LoRA) is utilized to fine-tune specific adapter layers within the Large Language Model (LLM), ensuring efficient adaptation to the medical domain while keeping the pre-trained backbone frozen. [6].

4.2.3 Evaluation Protocol

We normalize the data to make our final results easy to compare. To normalize answers we clean up the text by converting it to lowercase and removing punctuation.

Once we move to token level scoring, NLTK is used to filter out common English stopwords **a, an, the, and etc**. However, we keep *yes* and *no* in the answers . The remaining tokens are processed using Snowball stemming. [11].

All evaluations are performed on the official VQA-Med 2019 test set [1]. Each image is processed together with its corresponding question, while keeping the same prompt setup used during training. Preserving the original prompt is important to ensure a fair comparison.

Reference answers follow the official ImageCLEF VQA-Med rules [1]. If a question has multiple correct answers, they are separated using the # symbol like **face, sinuses, and neck # skull and contents** . Some answers may include additional details shown in parentheses, which are treated as optional and do not need to be matched exactly like **dermoid tumor (inclusion cyst) of cns**. After normalization, a prediction is considered correct if it matches the reference answer in test set.

4.2.3.1 Evaluation Metrics

The following normalized metrics are used:

- **Normalized Exact Match (NEM)**: exact string equality after normalization.
- **BLEU (unigram)**: BLEU-1 precision on normalized tokens with smoothing [12].
- **Fuzzy Partial Similarity**: partial string overlap for minor phrasing differences.
- **ROUGE-L**: longest common subsequence similarity [13].
- **METEOR**: alignment-based metric with stemming and synonym matching [14].
- **BERTScore F1**: contextual semantic similarity using pretrained transformer embeddings [15].

Yes/No questions are also calculated . Only exact matches to *yes* or *no* are accepted because we are running a medical model. Extended phrases are not considered correct.

Results are also reported by question category, including modality, organ system, imaging plane, and abnormality.

4.3 Graphical User Interface (GUI) Design

We put together a local GUI using Streamlit while working on the project. It ended up being useful for trying things and checking how the models behave. Everything runs on our the local machine because answering one question does not take a lot of time.

4.3.1 Sidebar Configuration

Interface is organized around a sidebar. From there, it is easy to switch between different ways of using the system . You can choose singe image reference for testing the model and Batch Metrics Dashboard for metric results .

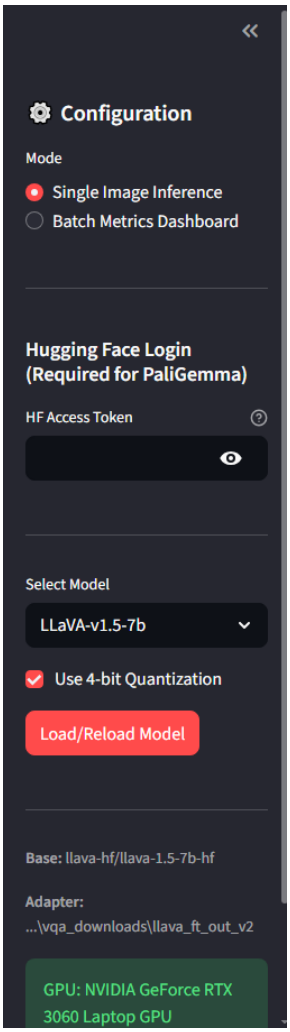


Figure 4.5 Sidebar mode selection in our GUI.

The GUI is set up so that we can easily move between different models while testing. Switching from one model to another is straightforward, whether it is LLaVA, InstructBLIP, PaliGemma, or Qwen2.5-VL. When memory becomes an issue, a 4-bit option can be turned on, and the pipeline can be refreshed with a single click.

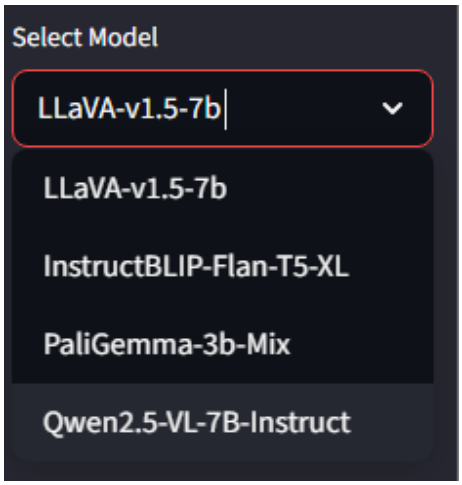


Figure 4.6 Model selection in our gui and image of it .

4.3.2 Single Image Inference Mode

In this part of the GUI we upload an image and display our medical image .Next to it you can choose pre determined question from dataset or custom question like whatever you want to ask to the model .

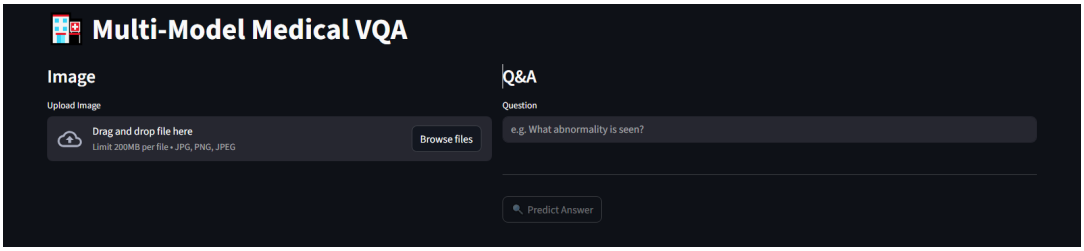


Figure 4.7 Single image inference display (image selection + Q&A panel).

When you select a pre determined question you can see the ground truth and generated answer. Below that we are showing in the GUI the BLEU scores and matching accuracy

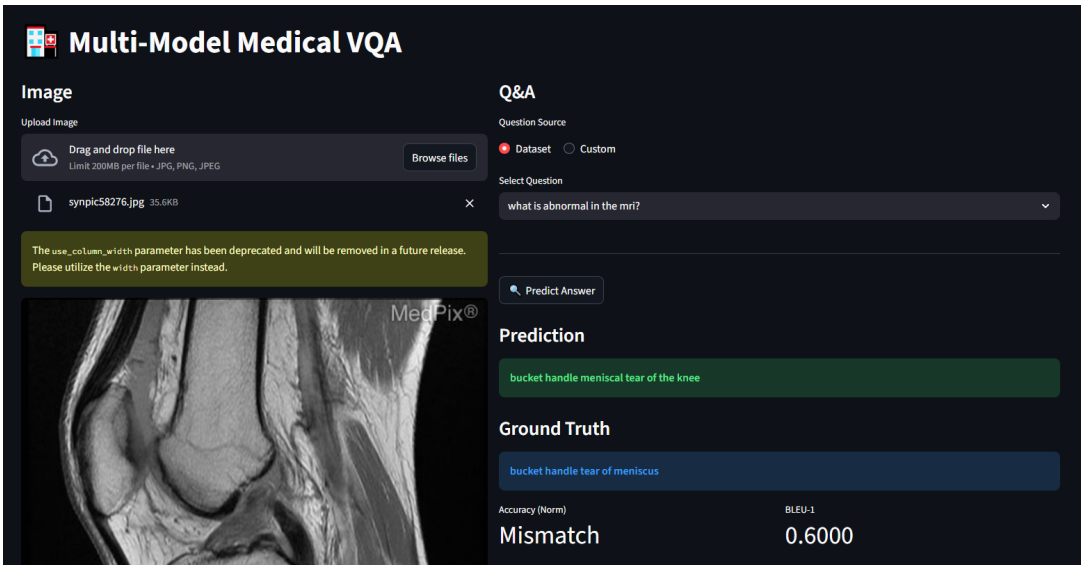


Figure 4.8 Dataset question selection and category display.

Continuation of the image before here we can see how normalizing is applied and what functions we used . After that we can see normalization results for G.T and generated answer

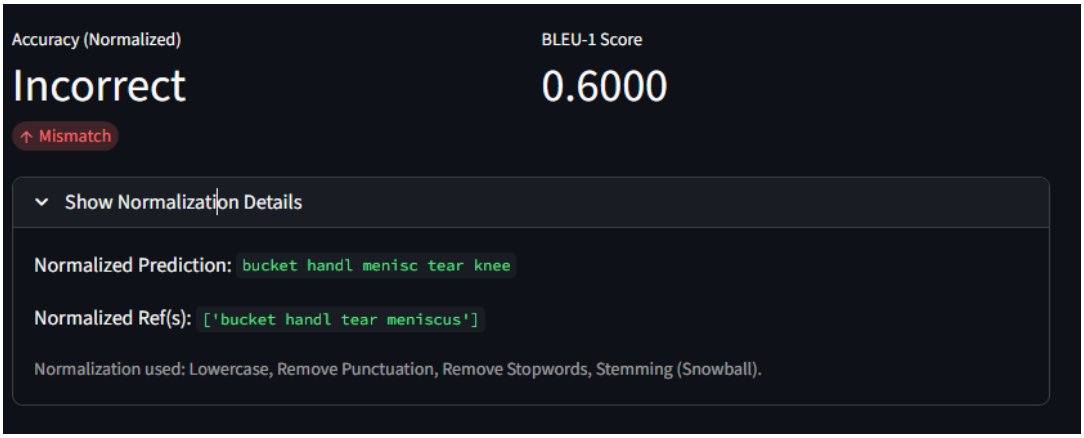


Figure 4.9 The generated answer, ground truth, and metrics

4.3.3 Batch Evaluation Dashboard

At this section we show the Batch Evaluation Dashboard while it is being used . For this section we need to upload tested model results and it's .json file . .json file includes image id, question, G.T, and predicted answer

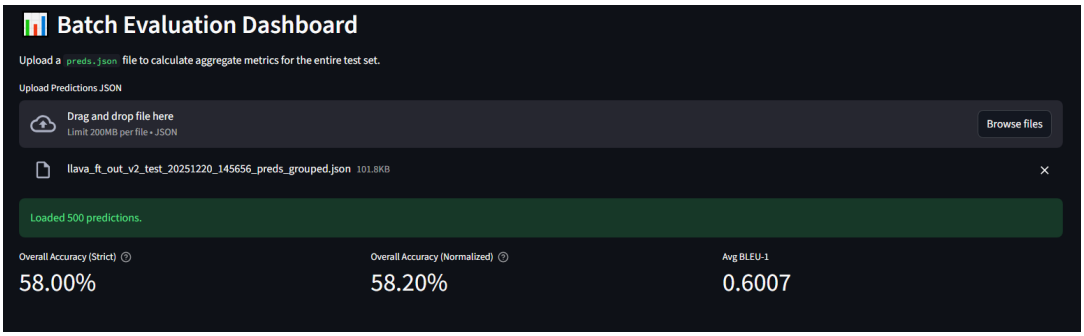


Figure 4.10 Batch evaluation dashboard: metric results

The GUI also shows results in by category wise. You can see the results below on a graph and also on a table.

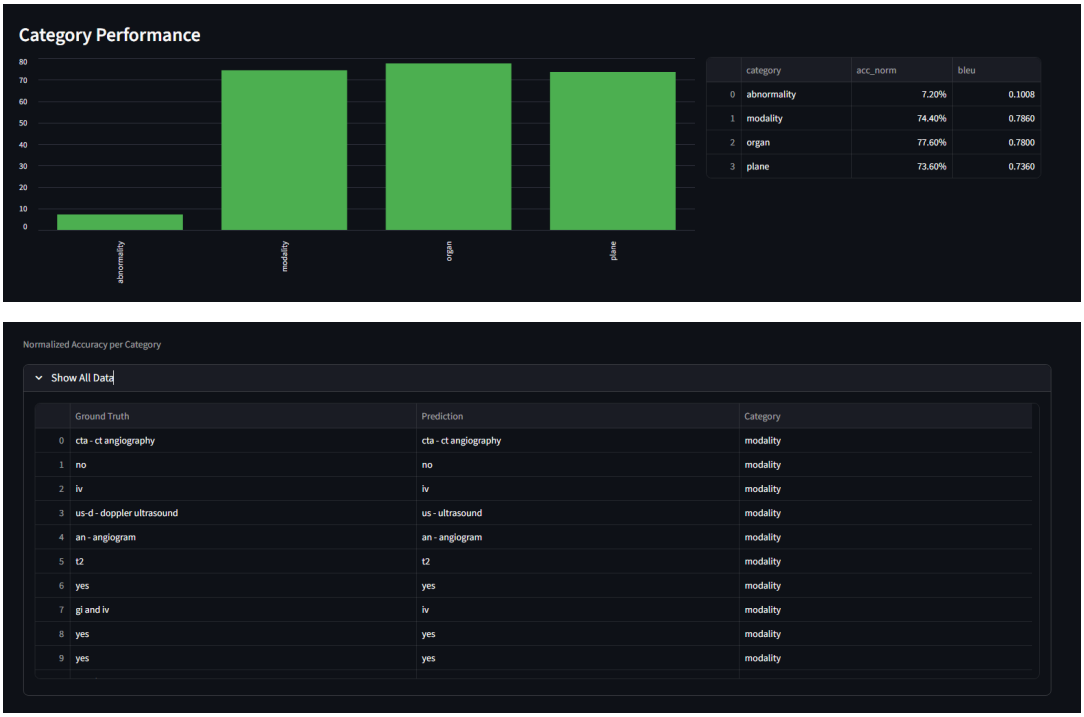


Figure 4.11 Category wise accuracy and BLEU results of the models

5
EXPERIMENTAL RESULTS

In this chapter, we are presenting a comparative evaluation of our VLM's used. As we explained in previous chapters, results are based on LLaVA-1.5, InstructBLIP, PaliGemma, and Qwen. We are using example based analysis and metric results of the models based on categories in the test set.

5.1 Performance Analysis

In this section we take closer look at four fine tuned models. We are comparing how are models performing as QA pairs and also how are they learning via loss graphs. And what kind of differences emerge between the models. We are looking at how our models are training and how they perform in our test set with different types of questions. This helps us understand what each model does well.

5.1.1 Training Logs and Results

The training graphic of the LLaVA-1.5 (7B) model.

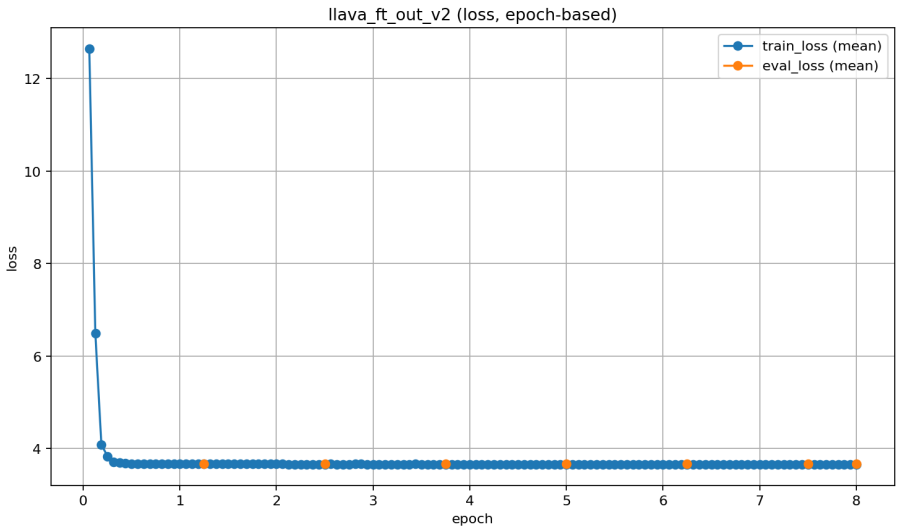


Figure 5.1 Epoch-based training and evaluation loss during fine-tuning.

We can analyse training loss decreases rapidly at the beginning. Model's training basically ends around first epoch.

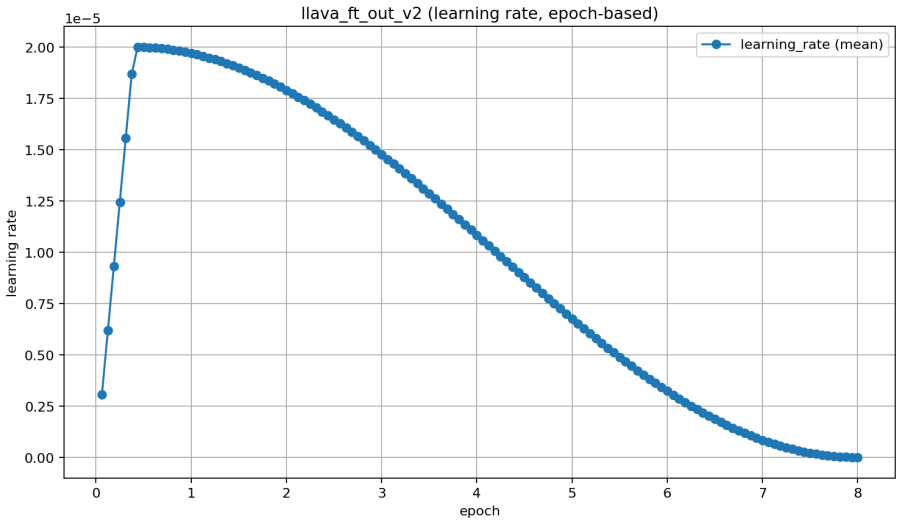


Figure 5.2 Epoch-based learning rate schedule used during fine-tuning.

The learning rate increases during early phases then it decreases slowly.

Training process for InstructBLIP was monitored over 18 epochs. Figure 5.3 shows training and validation loss curves.

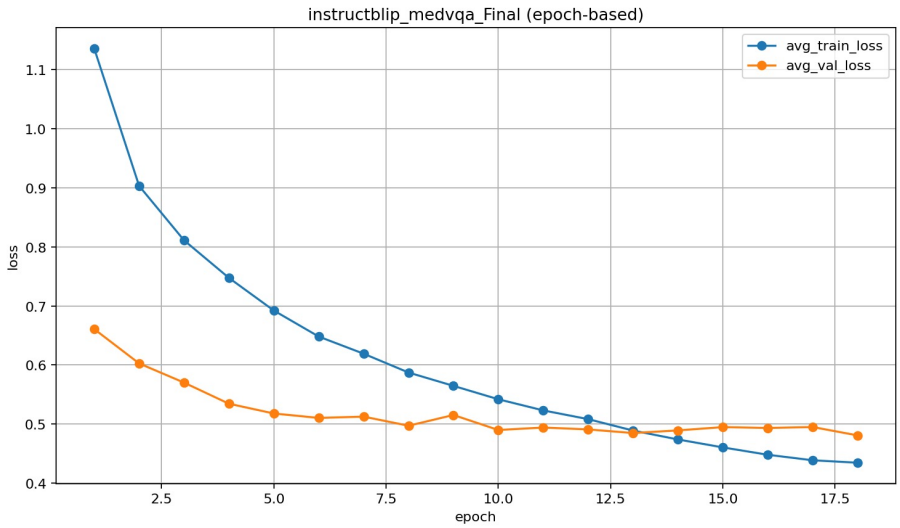


Figure 5.3 Epoch-based training and evaluation loss during InstructBLIP fine-tuning.

As we can see in Figure 5.3, training loss (blue line) decreases steadily. The validation loss (orange line) drops until 10th epoch. After this point, it stabilizes and flattens. This indicates that the model reaches its optimal performance near 10th epoch.

The training results of PaliGemma, shown in Figure 5.4,

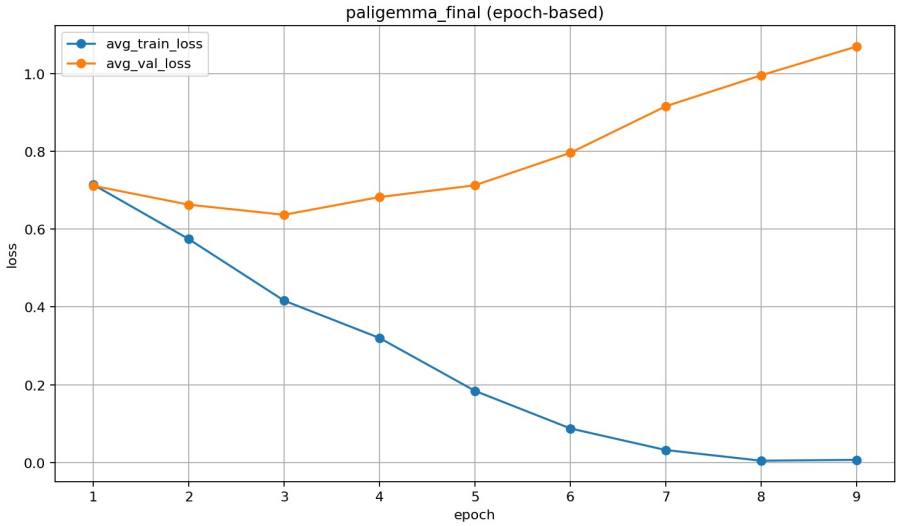


Figure 5.4 Epoch-based training and evaluation loss during PaliGemma fine-tuning.

The training loss decreases rapidly in the early epochs. As we can see model learns fast at the beginning of the training. Validation loss reaches its lowest point around epoch three. After that, it starts to increase. Later training brings little improvement. The model starts to memorize the data and further training just causes overfit.

The training dynamics of Qwen2.5-VL are shown in Figures 5.5 and 5.6.

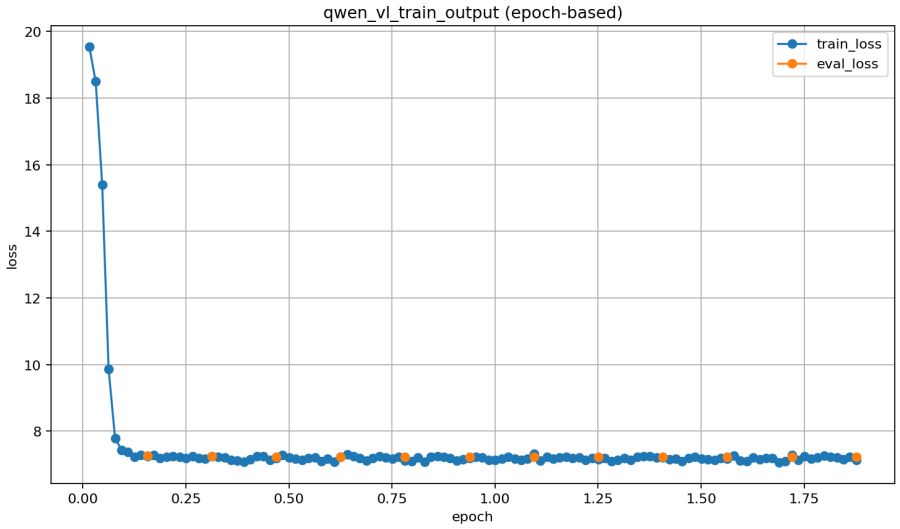


Figure 5.5 Epoch based training and evaluation loss graph of Qwen2.5-VL fine-tuning.

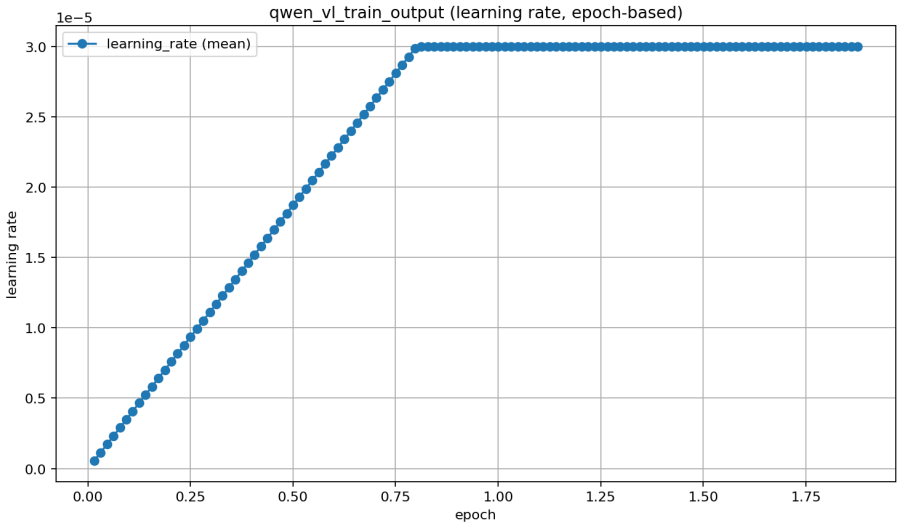


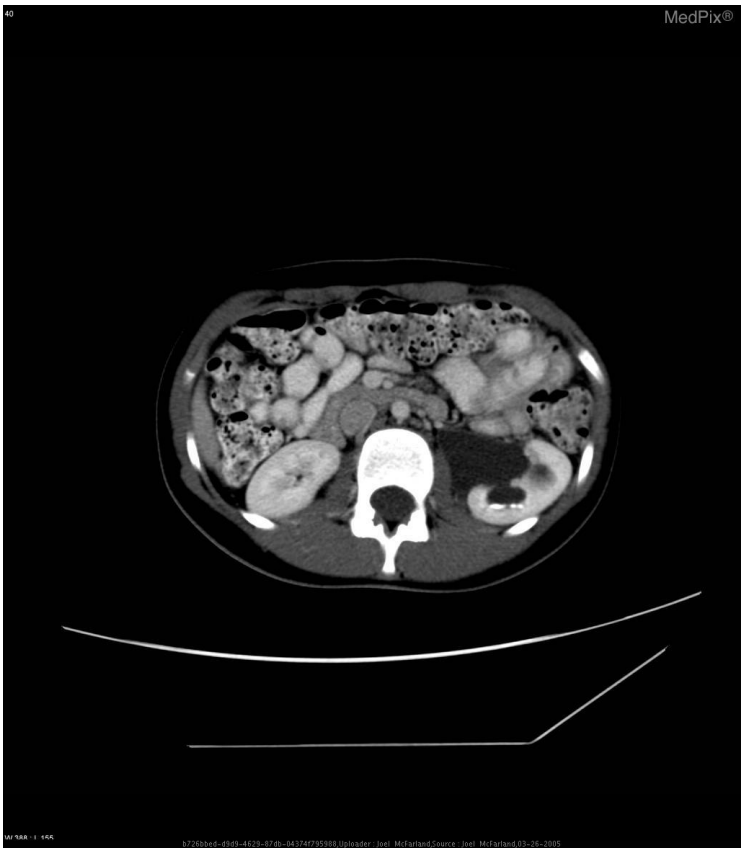
Figure 5.6 Epoch based learning rate graph of Qwen2.5-VL fine-tuning.

We can observe loss decreases rapidly during the early epochs. Both training and evaluation losses stabilize quickly.

Qwen2.5-VL is a large LLM and only limited training is possible with the available medical data.

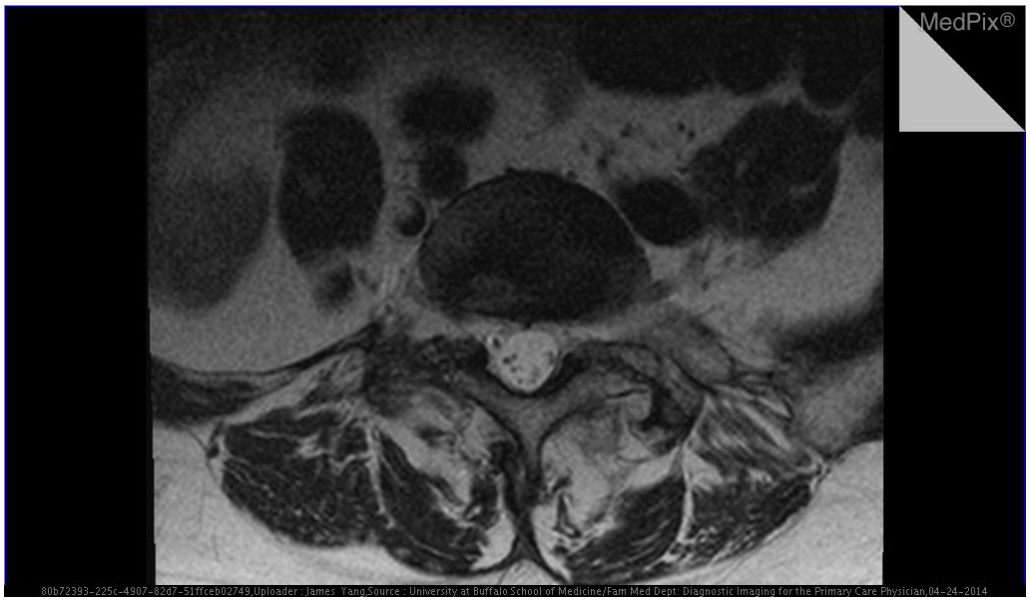
5.1.2 Example Results

To provide a better understanding of all models, we are showing a couple of examples from our test set. Each example consists of medical image and its QA and our model's generated answer



Category	Modality
Question	what imaging modality was used to take this image?
Ground Truth	ct with gi and iv contrast
Prediction (LLaVA)	ct with gi and iv contrast
Prediction (InstructBlip)	ct with iv contrast
Prediction (PaliGemma)	ct with iv contrast
Prediction (Qwen2.5-VL)	ct noncontrast

Figure 5.7 Modality example. LLaVA predicts the full contrast detail. InstructBLIP and PaliGemma are partially correct. Qwen predicts a different contrast setting.



Category	Modality
Question	is this image modality t1, t2, or flair?
Ground Truth	t2
Prediction (LLaVA)	t2
Prediction (InstructBlip)	t2
Prediction (PaliGemma)	t2
Prediction (Qwen2.5-VL)	t2

Figure 5.8 Modality example. All models predict the correct sequence (t2).

Category	Organ
Question	what organ system is shown?
Ground Truth	face, sinuses, and neck
Prediction (LLaVA)	face, sinuses, and neck
Prediction (InstructBlip)	face, sinuses, and neck
Prediction (PaliGemma)	spine and contents
Prediction (Qwen2.5-VL)	skull and contents

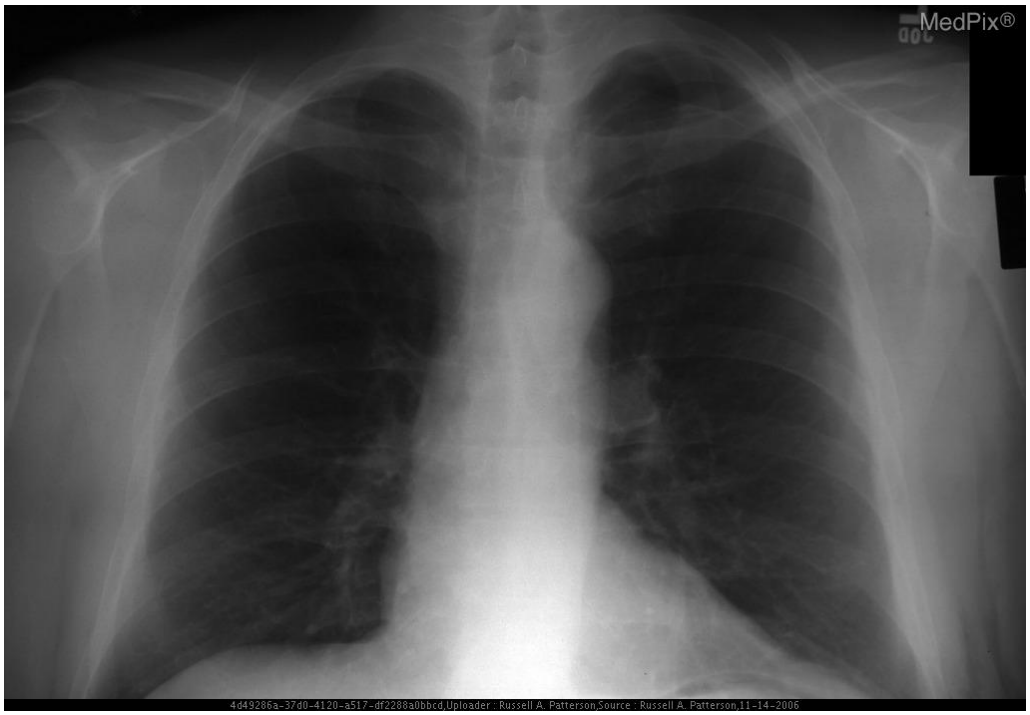
Figure 5.9 Organ system example. LLaVA and InstructBLIP match the ground truth. PaliGemma and Qwen drift to a different anatomy description.

Category	Organ
Question	what organ system is evaluated primarily?
Ground Truth	musculoskeletal
Prediction (LLaVA)	musculoskeletal
Prediction (InstructBlip)	musculoskeletal
Prediction (PaliGemma)	musculoskeletal
Prediction (Qwen2.5-VL)	musculoskeletal

Figure 5.10 Organ system example. All models predict the correct organ system label.

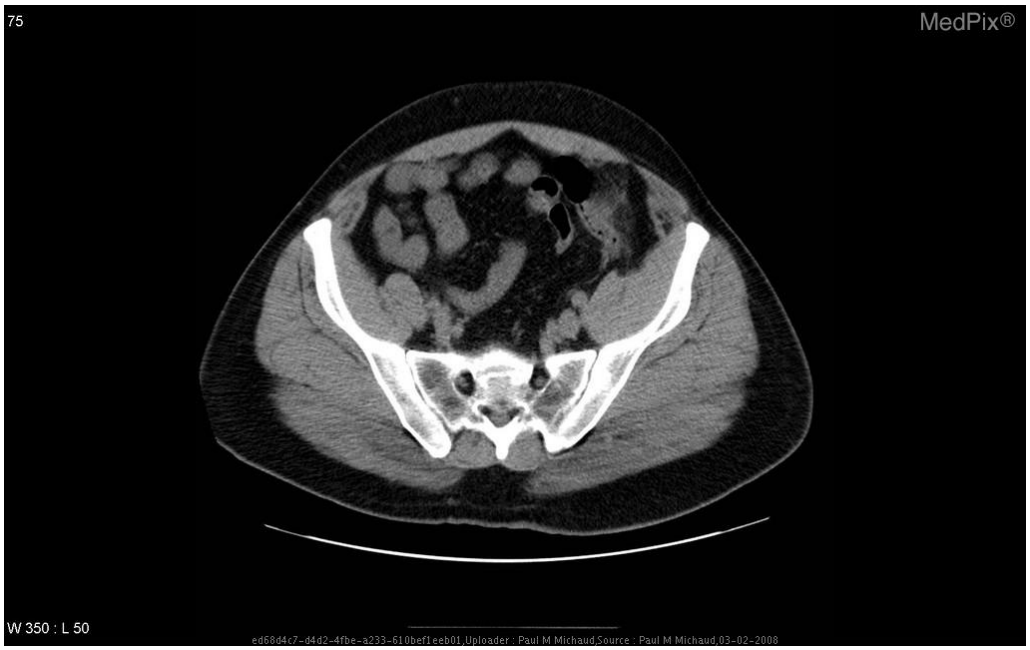
Category	Imaging Plane
Question	which plane is this image taken?
Ground Truth	axial
Prediction (LLaVA)	axial
Prediction (InstructBlip)	axial
Prediction (PaliGemma)	axial
Prediction (Qwen2.5-VL)	axial

Figure 5.11 Plane example. All four models correctly predict the plane (axial).



Category	Plane
Question	what is the plane of the x-ray?
Ground Truth	pa
Prediction (LLaVA)	pa
Prediction (InstructBlip)	pa
Prediction (PaliGemma)	pa
Prediction (Qwen2.5-VL)	ap

Figure 5.12 Plane example. LLaVA, InstructBLIP, and PaliGemma match the ground truth. Qwen confuses PA/AP in this case.

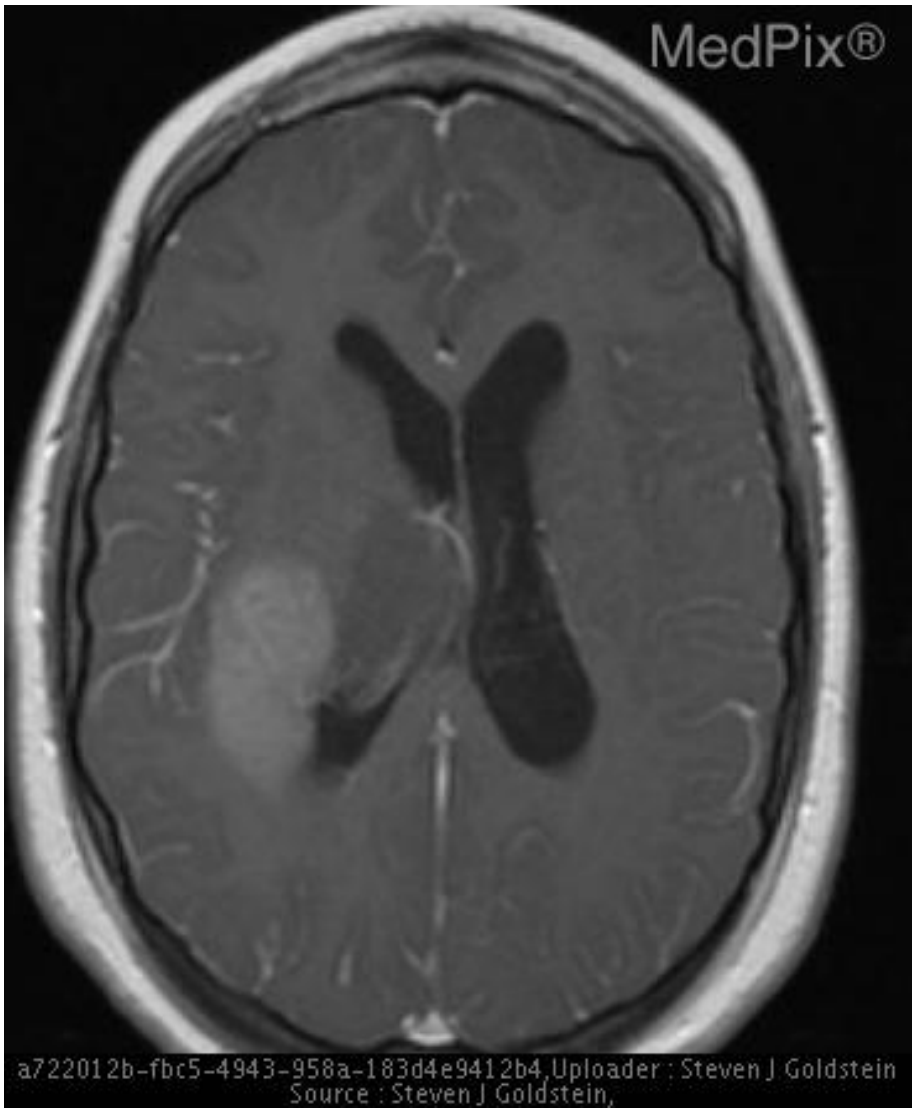


Category	Abnormality
Question	what is most alarming about this ct scan?
Ground Truth	epiploic appendagitis
Prediction (LLaVA)	acute appendicitis
Prediction (InstructBlip)	aortic dissection with aberrant right subclavian artery (arsa)
Prediction (PaliGemma)	rectovaginal fistula
Prediction (Qwen2.5-VL)	acute appendicitis

Figure 5.13 Abnormality example. All models miss the specific pathology. Predictions are incorrect.

Category	Abnormality
Question	what is abnormal in the mri?
Ground Truth	bucket handle tear of meniscus
Prediction (LLaVA)	bucket handle meniscal tear of the knee
Prediction (InstructBlip)	osteochondritis dissecans
Prediction (PaliGemma)	comminuted fracture of medial femoral condyle. small joint effusion.
Prediction (Qwen2.5-VL)	osteochondritis dissecans

Figure 5.14 Abnormality example. LLaVA finds out meaning of the image. InstructBLIP, PaliGemma, and Qwen predict false results.



Category	Abnormality
Question	what is the primary abnormality in this image?
Ground Truth	glioblastoma multiforme
Prediction (InstructBlip)	glioblastoma multiforme (gbm)
Prediction (PaliGemma)	glioblastoma multiforme (gbm)
Prediction (Qwen2.5-VL)	osteochondritis dissecans

Figure 5.15 Abnormality example. InstructBLIP, PaliGemma match to the ground truth. Qwen predict different conditions.

5.1.3 Overall Metric Results

Performance of the models are evaluated using the metrics we mentioned before in test set. Below are the results of the four models we used.

Table 5.1 Overall normalized performance of the LLaVA-1.5 (7B) model on the ImageCLEF 2019 VQA-Med test set

Metric	**LLaVA-1.5 (7B)**
Number of evaluated question–answer pairs	500
Exact Match Accuracy	58.00%
Fuzzy Partial Accuracy (threshold ≥ 80)	61.40%
Fuzzy Partial Accuracy (threshold ≥ 90)	60.80%
Average Fuzzy Similarity Score	77.43
BLEU-1 Score	0.6007
METEOR Score	0.4022
Yes/No Question Accuracy (strict)	81.25%

Table 5.2 Overall normalized performance of the InstructBLIP model on the ImageCLEF 2019 VQA-Med test set

Metric	**InstructBLIP**
Number of evaluated question–answer pairs	500
Exact Match Accuracy	55.60%
Fuzzy Partial Accuracy (threshold ≥ 80)	58.20%
Fuzzy Partial Accuracy (threshold ≥ 90)	57.80%
Average Fuzzy Similarity Score	75.10
BLEU-1 Score	0.5783
METEOR Score	0.3893
Yes/No Question Accuracy (strict)	78.12%

Table 5.3 Overall normalized performance of the PaliGemma model on the ImageCLEF 2019 VQA-Med test set

Metric	**PaliGemma**
Number of evaluated question–answer pairs	500
Exact Match Accuracy	64.00%
Fuzzy Partial Accuracy (threshold ≥ 80)	66.00%
Fuzzy Partial Accuracy (threshold ≥ 90)	65.60%
Average Fuzzy Similarity Score	80.32
BLEU-1 Score	0.6537
METEOR Score	0.4423
Yes/No Question Accuracy (strict)	92.19%

Table 5.4 Overall normalized performance of the Qwen2.5-VL model on the ImageCLEF 2019 VQA-Med test set

Metric	**Qwen2.5-VL**
Number of evaluated question–answer pairs	500
Exact Match Accuracy	55.20%
Fuzzy Partial Accuracy (threshold ≥ 80)	57.80%
Fuzzy Partial Accuracy (threshold ≥ 90)	57.00%
Average Fuzzy Similarity Score	76.18
BLEU Score	0.5659
METEOR Score	0.3803
Yes/No Question Accuracy (strict)	78.12%

5.1.4 Category-wise Results

Here we are evaluating results based on four categories of our test set.High scores are expected on modality/plane/organ type questions. Abnormality is hard to detect since there are a lot of them and we use limited training

Table 5.5 Category-wise normalized Exact Match Accuracy and BLEU-1 scores of the LLaVA-1.5 (7B) model

Question Category	**Exact Match Accuracy**	**BLEU-1 Score**
Modality / Imaging Sequence Questions	74.40%	0.7860
Organ / Organ System Questions	77.60%	0.7800
Imaging Plane / View Questions	73.60%	0.7360
Abnormality / Pathology Questions	6.40%	0.1008

Table 5.6 Category-wise normalized Exact Match Accuracy and BLEU-1 scores of the InstructBLIP model

Question Category	**Exact Match Accuracy**	**BLEU-1 Score**
Modality / Imaging Sequence Questions	71.20%	0.7615
Organ / Organ System Questions	72.80%	0.7280
Imaging Plane / View Questions	74.40%	0.7480
Abnormality / Pathology Questions	4.00%	0.0758

Table 5.7 Category-wise normalized Exact Match Accuracy and BLEU-1 scores of the PaliGemma model

Question Category	**Exact Match Accuracy**	**BLEU-1 Score**
Modality / Imaging Sequence Questions	88.00%	0.9015
Organ / Organ System Questions	76.80%	0.7720
Imaging Plane / View Questions	76.80%	0.7720
Abnormality / Pathology Questions	14.40%	0.1691

Table 5.8 Category-wise normalized Exact Match Accuracy and BLEU scores of the Qwen2.5-VL model

Question Category	**Exact Match Accuracy**	**BLEU Score**
Modality / Imaging Sequence Questions	71.20%	0.7437
Organ / Organ System Questions	72.00%	0.7280
Imaging Plane / View Questions	70.40%	0.7080
Abnormality / Pathology Questions	7.20%	0.0840

6
CONCLUSION AND DISCUSSION

In this chapter, we are summarizing the main outcomes of our (Med-VQA) project and discussing the results obtained during the development and evaluation stages.

We focused on medical visual question answering in this project. VQA tools are increasingly prevalent in healthcare. While we worked on this project, we also looked at how medical VQA has evolved and what new ideas researchers are following.

Our research consists of four models. It didn't take long to see how they were different. We didn't want to rely solely on final scores so we analysed training steps and logs also.

As we can see Paligemma performed best on test set. Paligemma has scored highest on accuracy and BLEU and also have much higher score in modality based questions.

LLAVA-1.5 was the first model we used. At first we saw very bad results on metrics results But after couple of parameter changes and better prompting model showed promising results and had increased accuracy

InstructBLIP was probably best model to train. As we can analyse from it's graph it continuously has validation and training loss drops until 10th epoch. And also model has decent results based on categories

When we analyze the results and graphs we can see Qwen2.5-VL behaves like Llava. Both are big LLMs. When we are looking at the charts we are noticing different results from the other models. Losses dropping sharp at first sight then stay flat after that. From this performance we can say it picks up patterns fast. But progress stops earlier than expected this just shows to us Qwen is already a big LLM so it does not need training or we can't change how it answers with limited resources.

References

[1] A. Ben Abacha, S. A. Hasan, V. V. Datla, J. Liu, D. Demner-Fushman, and H. Müller, "Vqa-med: Overview of the medical visual question answering task at imageclef 2019," in *Working Notes of CLEF 2019*, ser. CEUR Workshop Proceedings, vol. 2380, Lugano, Switzerland: CEUR-WS.org, Sep. 2019. https://ceur-ws.org/Vol-2380/paper_272.pdf.

[2] X. He, Y. Zhang, L. Mou, E. Xing, and P. Xie, *Pathvqa: 30000+ questions for medical visual question answering*, 2020. arXiv: 2003.10286 [cs.CL]. https://arxiv.org/abs/2003.10286.

[3] X. Zhang et al., *Pmc-vqa: Visual instruction tuning for medical visual question answering*, 2024. arXiv: 2305.10415 [cs.CV]. https://arxiv.org/abs/2305.10415.

[4] Z. Lin et al., "Medical visual question answering: A survey," *Artificial Intelligence in Medicine*, vol. 143, p. 102611, 2023, ISSN: 0933-3657. DOI: 10.1016/j.artmed.2023.102611 https://www.sciencedirect.com/science/article/pii/S0933365723001252.

[5] Q. Chen, X. Hu, Z. Wang, and Y. Hong, *Medblip: Bootstrapping language-image pre-training from 3d medical images and texts*, 2023. arXiv: 2305.10799 [cs.CV]. https://arxiv.org/abs/2305.10799.

[6] H. Liu, C. Li, Q. Wu, and Y. J. Lee, *Visual instruction tuning*, 2023. arXiv: 2304.08485 [cs.CV]. https://arxiv.org/abs/2304.08485.

[7] W. Dai et al., "Instructblip: Towards general-purpose vision-language models with instruction tuning," *arXiv preprint arXiv:2305.06500*, 2023.

[8] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019. https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee

[9] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 139, PMLR, 18–24 Jul 2021, pp. 8748–8763. https://proceedings.mlr.press/v139/radford21a.html.

[10] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, *Qlora: Efficient finetuning of quantized llms*, 2023. arXiv: 2305.14314 [cs.LG]. https://arxiv.org/abs/2305.14314.

[11] S. A. Hasan, Y. Ling, O. Farri, J. Liu, H. Müller, and M. Lungren, "Overview of the imageclef 2018 medical domain visual question answering task," in *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, 2018. https://ceur-ws.org/Vol-2125/paper_212.pdf.

[12] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P. Isabelle, E. Charniak, and D. Lin, Eds., Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. DOI: 10.3115/1073083.1073135 https://aclanthology.org/P02-1040/.

[13] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. https://aclanthology.org/W04-1013/.

[14] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, Eds., Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72. https://aclanthology.org/W05-0909/.

[15] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020. https://openreview.net/forum?id=SkeHuCVFDr.

Curriculum Vitae

FIRST MEMBER

Name-Surname: ARINÇ AYDEMİR
Birthdate and Place of Birth: 18.03.2002, İstanbul
E-mail: arinc.aydemir@std.yildiz.edu.tr
Phone: 0536 333 15 88
Practical Training:

SECOND MEMBER

Name-Surname: SALİH DEMİROZ
Birthdate and Place of Birth: 02.06.2002, İstanbul
E-mail: salih.demiroz@std.yildiz.edu.tr
Phone: 0543 304 77 09
Practical Training:

Project System Informations

System and Software: Windows Operating System, Python
Required RAM: 8GB
Required Disk: 30-40GB