Zurich University of Applied Sciences
Department School of Engineering
Center for Artificial Intelligence

MASTERS THESIS

Pathological-Llama: an Explainable Medical Visual Question Answering System

---

Author:
Sydney Nguyen

Supervisor:
Dr. Jasmina Bogojeska

Submitted on
February 17, 2024

Study program:
Data Science, M.Sc.

Imprint

Project: Masters Thesis
Title: Pathological-Llama: an Explainable Medical Visual Question Answering System
Author: Sydney Nguyen
Date: February 17, 2024
Keywords: Deep Neural Networks, Large Language Models, Computer Vision, Multimodality, Explainability, Medical VQA
Copyright: Zurich University of Applied Sciences

Study program:
Data Science, M.Sc.
Zurich University of Applied Sciences

Supervisor: Dr. Jasmina Bogojeska

Abstract

This thesis introduces Pathological-Llama, an explainable medical visual question answering system that integrates computer vision and natural language processing to accurately interpret medical images through a generative task approach. By addressing the demand for both accuracy and transparency in medical diagnostics, Pathological-Llama represents an advancement in the application of artificial intelligence within healthcare. This research undertakes a generative task, distinguishing it from traditional classification-based VQA systems, enabling the model to generate detailed, contextually relevant answers to complex medical questions.

Pathological-Llama was developed and fine-tuned using the PathVQA dataset, with a focus on explainability. The system employs the Integrated Gradients method and leverages GPT-4 for in-depth analysis, enhancing the interpretability of its decision-making process. These methods ensure that the system not only achieves high accuracy but also provides transparent explanations for its diagnostics. The effectiveness of Pathological-Llama is demonstrated through testing on previously unseen datasets, achieved impressive metrics with a BERT score of 0.591 and an F1 score of 0.419. These results confirm the model's robust generalization capabilities and highlight the role of optimizing hyperparameters and visual prefix lengths in enhancing the performance of generative VQA tasks.

Contributing to medical AI and VQA, Pathological-Llama demonstrates how generative models can deliver precise, explainable medical solutions. It establishes a benchmark for reliable, transparent AI in healthcare, aiming to enhance patient care and diagnostics. The evaluation of Pathological-Llama, highlighted by its performance metrics, underscores the value of generative tasks in medical VQA and paves the way for future advancements.

Zusammenfassung

Diese Dissertation führt Pathological-Llama ein, ein erklärbares medizinisches System zur Beantwortung visueller Fragen (VQA), das Computer Vision und natürliche Sprachverarbeitung integriert, um medizinische Bilder durch einen generativen Aufgabenansatz genau zu interpretieren. Indem es sowohl die Anforderungen an Genauigkeit als auch an Transparenz in der medizinischen Diagnostik adressiert, stellt Pathological-Llama einen Fortschritt in der Anwendung künstlicher Intelligenz (KI) im Gesundheitswesen dar. Diese Forschung unternimmt eine generative Aufgabe, die sie von traditionellen klassifikationsbasierten VQA-Systemen unterscheidet und es dem Modell ermöglicht, detaillierte, kontextuell relevante Antworten auf komplexe medizinische Fragen zu generieren.

Pathological-Llama wurde mit dem PathVQA-Datensatz entwickelt und feinabgestimmt, wobei der Schwerpunkt auf Erklärbarkeit liegt. Das System verwendet die Methode der integrierten Gradienten und nutzt GPT-4 für eine tiefe Analyse, was die Interpretierbarkeit seines Entscheidungsfindungsprozesses verbessert. Diese Methoden stellen sicher, dass das System nicht nur eine hohe Genauigkeit erreicht, sondern auch transparente Erklärungen für seine Diagnosen bietet. Die Wirksamkeit von Pathological-Llama wird durch Tests an zuvor ungesehenen Datensätzen demonstriert und erreicht beeindruckende Metriken mit einem BERT-Score von 0.591 und einem F1-Score von 0.419. Diese Ergebnisse bestätigen die robusten Generalisierungsfähigkeiten des Modells und heben die Rolle der Optimierung von Hyperparametern und der Länge des visuellen Präfixes zur Verbesserung der Leistung generativer VQA-Aufgaben hervor.

Mit einem Beitrag zur medizinischen KI und VQA zeigt Pathological-Llama, wie generative Modelle präzise, erklärbare medizinische Lösungen liefern können. Es setzt einen Maßstab für zuverlässige, transparente KI im Gesundheitswesen mit dem Ziel, die Patientenversorgung und Diagnostik zu verbessern. Die Bewertung von Pathological-Llama, hervorgehoben durch seine Leistungsmetriken, unterstreicht den Wert generativer Aufgaben in der medizinischen VQA und ebnet den Weg für zukünftige Fortschritte.

Acknowledgements

I extend my gratitude to Dr. Bogojeska, whose invaluable feedback and expert guidance have been pivotal throughout the various stages of this thesis. Her willingness to look into specific subjects related to my research and her adept handling of the challenges I faced have been highly appreciated. Her expertise and support have been a cornerstone of my academic journey.

I am also grateful to the Center for Artificial Intelligence for providing the necessary infrastructure, including both hardware and the collective knowledge of its members. Their support played an essential role in the realization of this research.

My appreciation goes to my family, whose love and support created an environment that allowed me to concentrate solely on my Master's.

A special thanks to my colleagues for their peer reviews and constructive feedback, which were instrumental in refining this thesis. Their contributions have enriched my work.

Lastly, while humorously intended, I acknowledge the hard work, late nights, and perseverance that I invested in this journey. It has been a process of learning, growth, and self-discovery.

Contents

Abstract III

Zusammenfassung IV

Acknowledgements V

List of Abbreviations VIII

1 Introduction 1
- 1.1 Motivation 1
- 1.2 Objectives and Contribution 1
- 1.3 Scope 2

2 Fundamentals 3
- 2.1 Transformers 3
  - 2.1.1 Vanilla Transformer 3
  - 2.1.2 Multimodal Transformers 4
- 2.2 Multimodal Learning 4
  - 2.2.1 Medical Multimodal Learning 5
  - 2.2.2 Supervised Multimodal Learning 5
  - 2.2.3 Self-Supervised Multimodal Learning 6
- 2.3 Explainable AI 6
  - 2.3.1 Explainability Methods for Multimodal Systems 6
  - 2.3.2 Integrated Gradients 7
  - 2.3.3 Importance of XAI in Healthcare 9
- 2.4 Medical VQA 10
  - 2.4.1 Comparison with Non-Medical VQA Systems 10
  - 2.4.2 Challenges and Opportunities 10

3 Related Work 12

4 Methodology of Pathological-Llama 14
- 4.1 Problem Statement 14
- 4.2 PathVQA Dataset Description 14
  - 4.2.1 Qualitative Analysis 14
  - 4.2.2 Distribution 15
  - 4.2.3 Dataset Version 15
- 4.3 Model Architecture 16
  - 4.3.1 Vision Model 17
  - 4.3.2 Causal Language Model 17
  - 4.3.3 Low-Rank Adaptation as a Parameter-Efficient Strategy 18
  - 4.3.4 Multimodal Prompt Construction 18

4.3.5 Masking Strategy 19
4.3.6 Answer Generation with Beam Search 19

5 Experimental Setup **21**
5.1 Dataset Split 21
5.2 Evaluation Metrics 21
5.2.1 BERT Score 22
5.2.2 F1 Score 23
5.3 Implementation Details 23

6 Results **25**
6.1 Findings 25
6.1.1 Benefit of Chat-Based Pre-Trained Models 25
6.1.2 Benefit of Greater Model Capacity 26
6.1.3 Benefit of Increased Visual Prefix Length 28
6.1.4 Benefit of Generative Task for Close-Ended Question Types 29
6.2 Evaluation 30

7 Explainability **31**
7.1 Implementation 31
7.2 Visualization and Analysis of Examples 32
7.2.1 Example 1: Cardiovacular 32
7.2.2 Example 2: Endocrine System 33
7.2.3 Example 3: Cut Surface 35

8 Discussion **37**
8.1 Discussion of Findings 37
8.2 Comparison to existing State-of-the-Art Models 38

9 Conclusion and Future Research **39**
9.1 Key Findings 39
9.2 Implications for Future Research 39
9.3 Conclusion 40

A Extended Experiments **41**

B Extended Explainability Examples **43**
B.1 Integrated Gradients Applied to Pathological-Llama with Llama2 43
B.2 Chatbot Generated Responses 46

Bibliography **51**

List of Abbreviations

AI	**A**rtificial **I**ntelligence
DNN	**D**eep **N**eural **N**etwork
FFN	**F**ully-connected **F**eed-forward **N**etwork
IG	**I**ntegrated **G**radients
LLM	**L**arge **L**anguage **M**odels
MHSA	**M**ulti-**H**ead **S**elf-**A**ttention
MLP	**M**ulti**L**ayer **P**erceptron
MLP	**Lo**w-**R**ank **A**daptation
QA	**Q**uestion-**A**nswer
SA	**S**elf **A**ttention
VQA	**V**isual **Q**uestion **A**nswering
ViT	**Vi**sion **T**ransformer
XAI	**EX**plainable **A**rtificial **I**ntelligence

Chapter 1

Introduction

Visual Question Answering (VQA) systems, developed from the combination of computer vision and natural language processing, offer a promising tool for interpreting visual data through questions and answers. These technologies have particular relevance in the medical field, where they can assist healthcare professionals by quickly providing information based on medical images. However, the requirements of medical decision-making extend beyond accuracy to include transparency in the systems' reasoning processes. This thesis focuses on developing "Pathological-Llama", an explainable medical VQA system that integrates the capabilities of VQA with the need for explainability in medical diagnostics.

1.1 Motivation

The potential of VQA to transform medical diagnostics, patient care, and treatment planning is immense, given its capacity to apply artificial intelligence for interpreting complex medical images. Despite the promise, creating effective VQA systems for medical use presents challenges, notably in delivering precise predictions and transparent explanations for those predictions. This research centers on constructing a multimodal deep neural network model trained on the PathVQA dataset [1], selected for its comprehensive pathological data. The objective is to craft a model that not only navigates the intricacies of medical data but also elevates the precision of medical diagnostics. Addressing the demand for explainability in medical AI is central; decisions in healthcare directly affect patient health, necessitating AI systems that are both accurate and transparent in their decision-making processes. Accordingly, this thesis attempts to enhance AI's utility in healthcare through medical VQA, confronting the dual challenges of accuracy and explainability to forward the application of AI in enhancing healthcare outcomes and efficiency.

1.2 Objectives and Contribution

This thesis makes key contributions to the domain of medical AI and VQA by:

1. **Developing Pathological-Llama:** An explainable multimodal VQA system tailored for the medical field, integrating advanced AI modeling techniques to interpret medical images and answer queries with high accuracy and transparency.

2. **Focusing on Explainability:** Incorporating explainability mechanisms, such as the Integrated Gradients method [2] and GPT-4 analysis [3], to provide insights into the model's reasoning, addressing a critical gap in current medical AI applications.

3. **Utilizing the PathVQA Dataset:** Training and evaluating the model with a dataset specifically selected for its extensive pathological data, underscoring the research's targeted application in pathology and its specificity in medical image interpretation.

4. **Advancing Medical VQA Technology:** By prioritizing precision and explainability, this work lays the groundwork for future research in medical VQA, aiming to bridge the gap between complex data analysis and practical medical usage.

1.3 Scope

This thesis is structured into nine chapters. Following this introduction, chapter 2, "Fundamentals" introduces technical concepts necessary for understanding the core topics discussed throughout the thesis. Chapter 3, "Related Work," provides a literature review on medical VQA and evaluation metrics, highlighting existing research in the field. Chapter 4, "Methodology," details the model architecture and the strategies employed in the development of the model, offering insight into the technical approaches taken. Chapter 5, "Experimental Setup," describes the baseline conditions for the experiments, outlining the configurations and preparations needed for evaluating the model's performance. Chapter 6, "Results" presents the outcomes of the conducted experiments, demonstrating the model's performance and effectiveness in the context of medical VQA. Chapter 7, "Explainability" focuses on explaining the decision-making process of the model, aiming to provide clarity on how it generates its conclusions. The thesis concludes with chapter 8, "Discussion" and chapter 9, "Conclusion and Outlook". These sections reflect on the research findings and explore their implications for the field of medical AI. The conclusion also identifies potential directions for future research.

Chapter 2

Fundamentals

This chapter lays the foundation to understand the methods throughout the thesis. Several concept are detailed: Transformers, with their advanced self-attention mechanism; multimodal learning, which combines diverse data types for enhanced AI performance; explainable AI (XAI), emphasizing transparency in AI decisions; and medical Visual Question Answering (VQA), an application bridging computer vision and language processing for medical image interpretation. It highlights the importance of these technologies in advancing AI research and applications, particularly their transformative potential in healthcare, focusing on improved diagnostics and decision-making processes.

2.1 Transformers

Transformers, especially the Vanilla Transformer, can be understood from a geometrically topological perspective. The self-attention mechanism allows modelling tokenized inputs as fully-connected graphs in the topological geometry space. This flexibility sets Transformers apart from other deep networks like CNNs [4], which are limited to aligned grid spaces.

2.1.1 Vanilla Transformer

The Vanilla Transformer serves as the foundation for Transformer-based research, employing an encoder-decoder structure. It takes tokenized inputs and uses Transformer layers/blocks for both encoding and decoding. Each block contains two sub-layers: multi-head self-attention (MHSA) and position-wise fully-connected feed-forward network (FFN). Residual connections with normalization layers aid gradient backpropagation. The output of the MHSA and FFN sub-layers can be represented as $\mathrm{Z} \leftarrow \mathrm{N}(\mathrm{sublayer}(\mathrm{Z}) + \mathrm{Z})$, where $\mathrm{sublayer}(\cdot)$ is the sub-layer's mapping and $\mathrm{N}(\cdot)$ denotes normalization [5].

Vanilla Transformer utilizes tokenized sequences as input, treating each token as a node in a graph. Tokenization offers advantages such as geometrically topological flexibility, flexible information organization, compatibility with task-specific tokens, and inherent support for multimodal data processing. Vanilla Transformer employs sine and cosine functions for position embedding. Position embeddings provide temporal or spatial information to the Transformer. Their necessity depends on the input type, and they can be seen as a form of additional information.

Self-Attention (SA) is a core component of the Vanilla Transformer, enabling each element of an input sequence to attend to all other elements. SA models the input as a fully-connected graph, providing a global perception similar to Non-Local Networks. Masked Self-Attention (MSA) modifies SA to incorporate contextual dependencies and prevent attending to future positions. MHSA stacks multiple SA sub-layers in parallel, allowing the model to jointly attend to information from multiple representation subspaces. MHSA acts as an ensemble mechanism, enhancing the Transformer's ability to process diverse information.

The output of the multi-head attention sub-layer in the Transformer passes through a position-wise FFN. The FFN consists of successive linear layers with non-linear activation. For example, a two-layer FFN can be represented as:

$$FFN(Z) = \sigma(ZW1 + b1)W2 + b2,$$

Here, W1, b1, W2, and b2 denote the weights and biases of the linear transformations, while $\sigma(\cdot)$ represents a non-linear activation function such as ReLU(·) [6] or GELU(·) [7].

2.1.2 Multimodal Transformers

The Transformer architecture can process each input as a fully-connected graph through self-attention. This allows Transformers to work with various modalities by treating the embedding of each token as a graph node. Users only need to tokenize the input and select an embedding space before inputting the data into Transformers. Tokenization and embedding approaches are highly flexible, offering alternatives such as using ROIs and CNN features, patches and linear projection, or object detection and graph features [5]. From a geometric topology perspective, each modality can be seen as a graph. For example, an RGB image represents a neat grid graph, while video and audio are clip/segment-based graphs with temporal and semantic patterns. Both uni-modal and multimodal Transformers utilize special/customized tokens as placeholders in token sequences. Common special tokens, such as [CLS] for classification and [SEP] as a separator, are defined to add semantic meaning to the token sequences. Token embedding fusion is a technique used in Transformers to combine multiple embeddings for each token position, which allows for early fusion of embeddings. Token-wise summing is a common method of fusion, providing flexibility in various Transformer models, including multimodal surveillance AI. In multimodal Transformer applications, this approach combines different embeddings using token-wise operators like addition. Examples include VisualBERT [8], Unicoder-VL [9], VL-BERT [10], InterBERT [11], and ImageBERT [12], which leverage token embedding fusion for improved performance in multimodal tasks.

In multimodal Transformers, self-attention and its variants are used for processing cross-modal interactions such as fusion and alignment [5].

2.2 Multimodal Learning

Multimodality in AI involves the integration of various types of data (e.g., text, images, audio) to enhance predictions and identifications [13]. This process, known as multimodal fusion, can be performed through early, late, or hybrid fusion methods. Early fusion integrates data before analysis, either by removing correlations or combining data at its lower-dimensional latent subspace. However, it can be challenging to synchronize data sources with variable sampling rates and convert them into a fixed representation [14]. Early fusion is depicted in Figure 2.1. Late fusion as seen in Figure 2.2 uses individual modality sources for fusion during decision-making, resembling human cognitive abilities, and can be integrated to generate a single common decision. Hybrid fusion, uses deep neural networks for intermediate fusion, changing input data to a higher-level abstraction and learning a joint representation of different modalities [14]. This is illustrated in Figure 2.3.

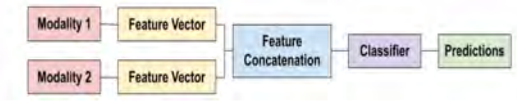


FIGURE 2.1: Early fusion.

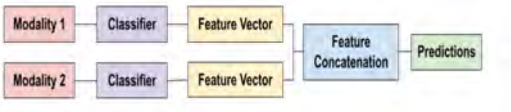


FIGURE 2.2: Late fusion.

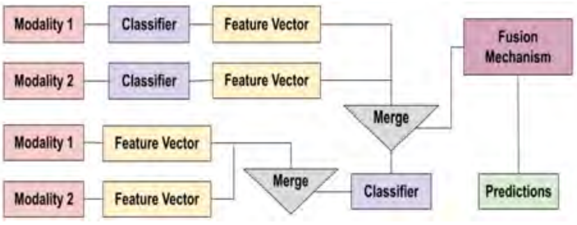


FIGURE 2.3: Hybrid Fusion

2.2.1 Medical Multimodal Learning

Medical data, which can be represented in various forms or modalities offers a rich source of information for analysis and diagnosis [15]. Multimodal learning in the medical field leverages these diverse data types to gain an understanding of medical conditions. The integration of different modalities can reveal insights that might not be apparent when examining data from a single source. For instance, in Alzheimer's diagnosis, combining FDG-PET scans and structural brain MRI provides a more complete picture than relying on either modality alone [16]. FDG-PET scans are effective in identifying hypermetabolism patterns, while structural MRI is used for detecting hippocampal atrophy. Together, these modalities offer a more accurate and robust diagnosis than either could alone.

In the context of machine learning, a unimodal labeled dataset is defined as $D_u = \{(X_{i1}, y_{i1})\}_{i=1}^{n}$, where X represents the input, y is the ground-truth label, and n is the number of samples. In contrast, a multimodal dataset containing k different modalities is denoted as $D_m = \{(X_{i1}, \ldots, X_{ik}, y_{i1})\}_{i=1}^{n}$. This definition allows for the incorporation of various types of data into a unified learning framework. There are scenarios where a multimodal dataset might lack labels for some or all samples. Such datasets are represented as $(X_{i1}, \ldots, X_{ik})$ without the corresponding y labels. This situation is particularly common in medical contexts, where obtaining labeling for all modalities can be challenging due to privacy concerns, the need for expert analysis, or incomplete data collection. Furthermore, it is not uncommon to encounter unpaired data within these datasets. This means that some instances X may not include all k modalities, presenting additional challenges for multimodal learning. The absence of certain modalities in some data points requires sophisticated approaches to handle missing data effectively while still extracting useful information from the available modalities. These complexities in medical multimodal learning emphasize the need for advanced machine learning techniques capable of handling diverse, unpaired, and sometimes unlabeled data to extract meaningful insights for medical diagnosis and treatment planning.

2.2.2 Supervised Multimodal Learning

Supervised multimodal learning [15] involves learning a predictive model M_θ parameterized by θ using ground truth labels y. The objective is to find θ^* which minimizes the supervised loss L_{sl} across the multimodal dataset D_m:

$$\theta^* = \arg\min_{\theta} \sum_{(X_k^i, y_k^i) \in D_m} L_{sl}(M_\theta(X_1^i, \ldots, X_k^i), y^i)$$

Supervised multimodal learning faces challenges in the medical domain due to the need for extensive labeled data. The labeling process, often requiring domain experts, is compounded by privacy concerns associated with making medical data publicly available.

2.2.3 Self-Supervised Multimodal Learning

Self-supervised learning typically involves two stages: pretraining and downstream task learning [15]. In the pretraining stage, the self-supervised method does not use ground-truth labels. Instead, it employs pseudo-labels $\hat{y}$, generated from the input modalities, to train the predictive model M_θ. The model parameters are optimized as follows:

$$\theta^* = \arg\min_{\theta} \sum_{(X_k^i, y^i)_k \in D_m} L_{ssl}(M_\theta(X_1^i, \ldots, X_k^i), \hat{y}^i)$$

Post-pretraining, the model M_θ is applied to specific downstream tasks involving ground-truth labels y. The pre-trained model can serve as a feature extractor or be fine-tuned for specific tasks. Pretraining with large-scale multimodal data enables the model to capture extensive general knowledge, facilitating task-specific learning. This approach, especially effective in the medical domain, reduces reliance on labeled data and often outperforms fully supervised models. In medical applications, self-supervised learning is highly effective due to the challenge of obtaining labeled data. Medical datasets often include images and detailed textual reports, which are valuable for self-supervised pre-training.

2.3 Explainable AI

Explainable AI (XAI) involves understanding the working mechanism and decision-making process of AI systems. It aims to answer questions like why the system made a particular prediction (interpretability) and how the system came to a specific decision (explainability) [17]. Model-specific explanations apply to a specific model, while model-agnostic methods are independent and irrespective of the model. Feature attribution-based methods highlight image regions that are contributors to decision-making. Distillation methods build an approximate local model or a surrogate model on top of the original model for interpretation. Intrinsic methods are explainable and self-explain using models' attention mechanisms to focus on important visual and textual regions. This category includes joint training approaches that combine predictions and explanations. Figure 2.4 shows the taxonomy of various deep explainability methods for both unimodal and multimodal scenarios [14].

2.3.1 Explainability Methods for Multimodal Systems

The following methods [14] aim to provide interpretable explanations for tasks like visual question answering, image captioning, and common sense reasoning.

Attention-based methods assign more weight and importance to specific factors in multimodal data. They are commonly used in tasks like visual captioning and visual question answering, where attention mechanisms align and fuse information from different modalities. These approaches generate explanations based on attention features and improve interpretability. However, challenges exist in evaluating and ensuring consistent explanations.

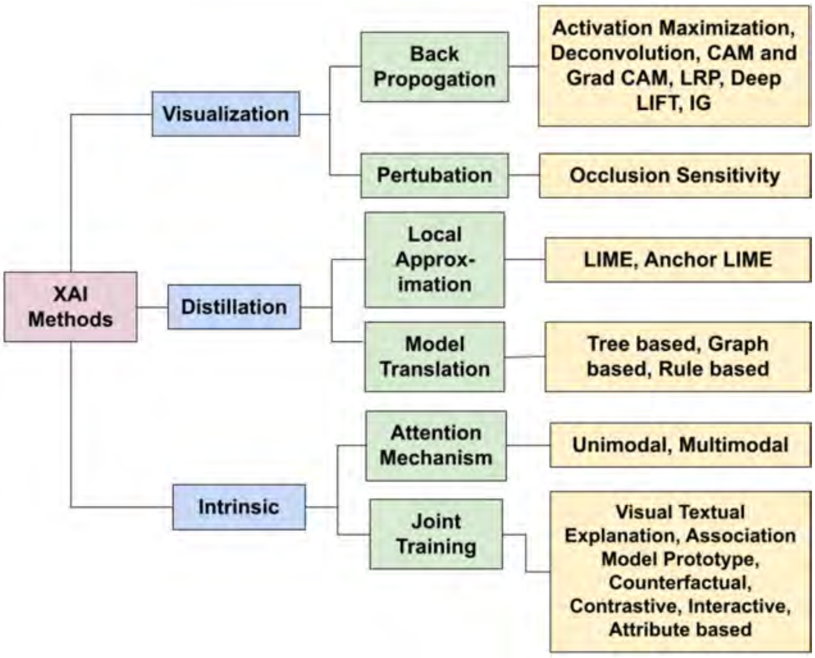

XAI Methods
Visualization
Distillation
Intrinsic
Back Propogation
Pertubation
Local Approx-imation
Model Translation
Attention Mechanism
Joint Training
Activation Maximization, Deconvolution, CAM and Grad CAM, LRP, Deep LIFT, IG
Occlusion Sensitivity
LIME, Anchor LIME
Tree based, Graph based, Rule based
Unimodal, Multimodal
Visual Textual Explanation, Association Model Prototype, Counterfactual, Contrastive, Interactive, Attribute based

FIGURE 2.4: XAI methods.

Counterfactual explanations focus on contrasting decisions and causal understanding. They recommend actionable insights and minimal changes to achieve desired outcomes. Counterfactual approaches have been applied in visual question answering, visual captioning, and image description tasks to analyze model behaviour and improve predictions.

Interactive explanations involve user feedback by combining model explanations, user annotations, and active learning to rectify incorrect predictions and enhance user trust. They have been applied to tasks like VQA and customer relationship management.

Graph-based methods leverage scene graphs and knowledge graphs to improve explanation quality. Scene graphs represent relationships between objects in an image, while knowledge graphs incorporate semantic information. These approaches enhance interpretability in tasks like visual question answering and neuro-symbolic AI.

Attribute-based methods focus on the importance of attributes in generating explanations. They associate visual features with attribute information to provide class-discriminative and concept-explaining explanations. Attribute maps, counter attributes, and spatiotemporal attention mechanisms are utilized to improve interpretability.

2.3.2 Integrated Gradients

Integrated Gradients (IG) is an attribute-based model interpretability technique that assigns importance scores to input features. It does so by approximating the integral of gradients of the model's output concerning the inputs along a straight-line path from given baselines to inputs. IG relies on two fundamental axioms, namely Sensitivity and Implementation Invariance, as stated in Definition 1 and Definition 2 of the paper "Axiomatic Attribution for Deep Networks" [2]. These axioms hold are believed to be essential traits of all attribution methods.

Definition 1 Axiom Sensitivity states that an attribution method satisfies Sensitivity if, when presented with two inputs and baselines that differ in only one feature, resulting in different predictions, the differing feature is attributed a non-zero value. In cases where the deep network's mathematical implementation does not rely on a particular variable, the attribution to that variable should always be zero.

Definition 2 Axiom Implementation Invariance defines functional equivalence between two networks, indicating that their outputs are identical for all inputs despite varying implementations. An attribution method satisfying Implementation Invariance would ensure that attributions remain the same for functionally equivalent networks.

The sensitivity axiom requires a baseline which is defined as an absence of a feature in an input. It can be understood as an "input from the input space that produces a neutral prediction." By treating the baseline as an input, counterfactual explanations can be generated, exploring how the model behaves while transitioning from the baseline to the original image. The authors contend that gradient-based methods violate Sensitivity (Def. 1). To illustrate this, they present a simple function, $f(x) = 1 - ReLU(1-x)$, as shown in Figure 2.5. When attempting to generate attribution for $x = 2$, the function's output changes from 0 to 1, but after $x = 1$, it becomes flat, resulting in a gradient of zero. Although x contributes to the result, the flatness of the function at the input being tested leads to invalid attribution, breaking Sensitivity. Breaking Sensitivity causes gradients to focus on irrelevant features [2].

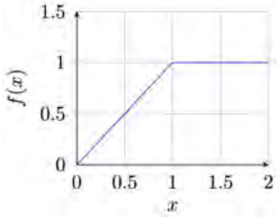


FIGURE 2.5: $f(x) = 1 - \mathrm{ReLU}(1-x), \quad x \in [0,2]$

Computing Integrated Gradients In the definition of IG, function F represents the model, where the input $x \in R^n$ (with n denoting the number of dimensions) and the baseline $x' \in R^n$. The method involves computing gradients along a straight-line path between x and x'. The integrated gradient along the ith dimension is formally defined as shown in equation 2.1.

$$IG_i(x) := (x_i - x_i') * \int_{\alpha=0}^{1} \frac{\delta F(x' + \alpha * (x - x'))}{\delta x_i} d\alpha \tag{2.1}$$

However, since the original definition involves an integral, it is infeasible to calculate directly. Therefore, the practical implementation of IG utilizes an approximation by replacing the integral with a summation, as in equation 2.2.

$$IG_i^{approx}(x) := (x_i - x_i') * \sum_{k=1}^{m} \frac{\delta F(x' + \frac{k}{m} * (x - x'))}{\delta x_i} * \frac{1}{m} \tag{2.2}$$

To obtain the approximated calculation (equation 2.2), parameter m is used to define the number of interpolation steps. For example, when visualizing the interpolations with m equals five (see Figure 2.6), the process can be better understood. In practice, the number of interpolation steps typically ranges from 20 to 300, with the most common value being 50. The results of applying IG can be observed in Figure 2.7.

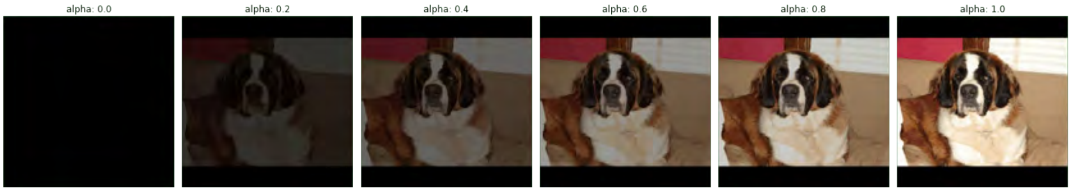


FIGURE 2.6: Five-step interpolation between the baseline x' and the input image x. The first image on the left (alpha:0.0) is not a part of the interpolation process [18].

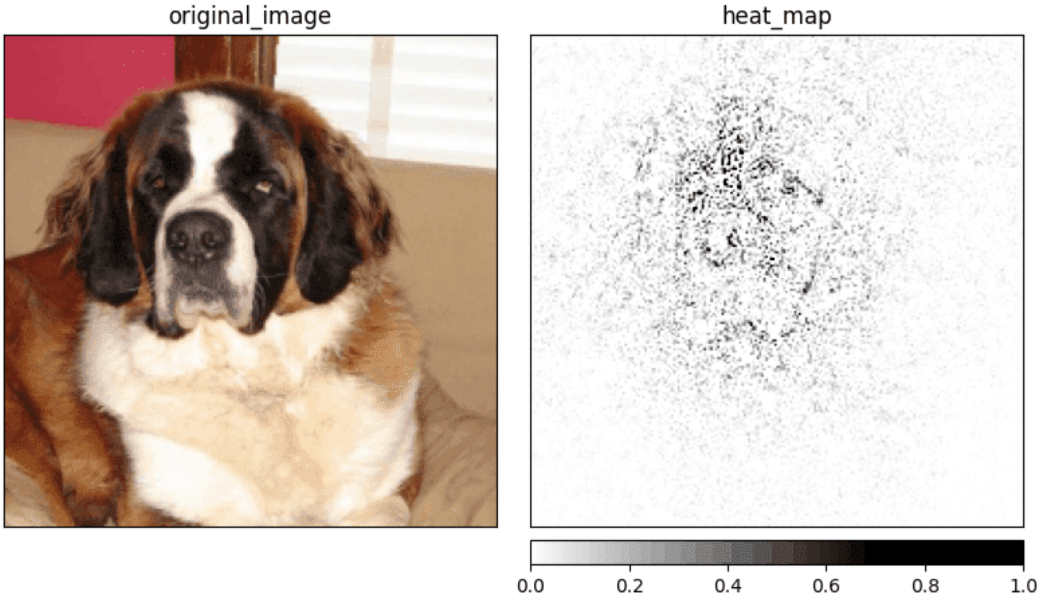


FIGURE 2.7: Visualization of the saliency map by the IG generated for the class saint-bernard. The result is averaged over 50 interpolation steps [18].

2.3.3 Importance of XAI in Healthcare

Rising healthcare costs globally underscore the need for AI to enhance healthcare efficiency. AI, especially as clinical decision support systems, offers advancements in disease diagnosis and treatment [19]. These AI systems, leveraging patient data-driven models, face challenges like technical complexities and legal, medical, and patient concerns, requiring a multidisciplinary approach. Explainability in healthcare is an important yet complex aspect of medical AI. It involves making AI's logic transparent, impacting legal, ethical, and practical facets of medical applications. Technologically, explainability varies in AI models; it is more straightforward in traditional models but complex in "black-box" models like DNNs, creating a trade-off between accuracy and clarity. Legally, the evolving healthcare landscape is increasingly prioritizing explainability in AI solutions, affecting informed consent, medical device certification, and liability. Medically, explainability enhances physicians' comprehension of AI, essential for clinical validation and application, aiding informed decision-making and trust-building in AI systems. For patients, explainability supports patient-centered care, facilitating shared decision-making and deeper patient engagement. Ethically, it intersects with autonomy, beneficence, non-maleficence, and justice, ensuring informed consent, aiding decision-making, addressing biases, and promoting equitable AI access in healthcare.

2.4 Medical VQA

Medical VQA merges computer vision and natural language processing to interpret medical images and answer related questions, aiding clinical decision-making and patient engagement [20]. Unlike typical medical AI, medical VQA addresses a broad range of questions with user-friendly answers. Medical VQA systems cover tasks, such as assisting radiologists in managing high workloads and aiding pathologists in diagnoses. They also serve as assistants to clinicians by providing second opinions and directly interacting with patients, especially in automated health exams or online information searches. The field faces challenges, such as creating datasets requiring expert annotation and designing models for specific, fine-grained medical analysis. Since its inception in 2018, medical VQA has seen growing interest, with developments in datasets and methodologies, marking it as an increasingly important area.

2.4.1 Comparison with Non-Medical VQA Systems

Medical VQA and general VQA integrate visual and language processing for answering questions, however, differ in applications, objectives, datasets, methodologies, and evaluation. Medical VQA focuses on clinical support and patient interaction, targeting medical diagnosis and decision-making with datasets that are smaller, less diverse, and require expert annotation [20]. Its methods emphasize image encoder pre-training due to complex datasets. In contrast, general VQA caters to a broader range of applications such as information retrieval and human-computer interaction, with larger, more diverse datasets and a focus on multi-modal fusion and reasoning. While medical VQA requires deep domain knowledge for understanding specialized medical terminology and imaging data, general VQA involves varied strategies for scene and object recognition. Evaluation in medical VQA prioritizes precision and medical relevance, addressing dataset imbalances, whereas general VQA seeks a balance in plausible answers, emphasizing overall correctness.

2.4.2 Challenges and Opportunities

Medical VQA faces distinct challenges compared to general VQA [20], offering research opportunities:

- Question Diversity: Enhancing question diversity is relevant, moving beyond basic abnormality identification to include more specific and varied queries. This requires sourcing questions from diverse materials, including textbooks and real-world interactions, and improving reasoning and medical domain knowledge within VQA systems.

- Integrating Extra Medical Information: Challenges include effectively incorporating Electronic Health Records and multiple images into VQA systems, necessitating new dataset designs and model adaptations.

- Interpretability and Reliability: Ensuring the interpretability of medical VQA systems is essential for reliable outputs. This involves tackling unimodal biases, integrating external knowledge, conducting evidence verification, and establishing interpretability benchmarks.

- Generalizability: Addressing generalizability issues is essential, focusing on training with medical-specific datasets and validating across different data domains to ensure robust model performance.

- Large Language Models (LLMs): Integrating LLMs presents opportunities for enhanced question-answering capabilities, though it raises concerns about domain-specific accuracy, inherent biases, and the need for expert medical knowledge.

- Integration in Medical Workflow: Effective integration of medical VQA into clinical workflows demands an understanding of clinicians' preferences and cognitive styles, aiming to improve efficiency and service quality.

Chapter 3

Related Work

Medical VQA is a task that combines image and text analysis to answer clinical questions based on medical images. The field faces challenges due to the scarcity of annotated medical datasets, leading researchers to utilize pre-trained models from non-medical datasets [21]. However, this introduces domain shift issues, as models such as VGG16 [22], ResNet [23], LSTM [24], and BERT [25] are not initially trained on medical data. To mitigate domain shifts, recent efforts focus on pre-training visual encoders specifically with medical images. Notable approaches include meta-learning strategies for visual representation [26], and the adoption of self-supervised pre-training objectives. The PubMed-CLIP framework [27] illustrates the effectiveness of aligning image-text features through pre-training, enhancing visual encoder performance for medical VQA tasks. Discriminative methods, treating VQA as a classification problem, have shown promising performance though are limited to predefined answers. In contrast, generative methods allow for open-ended answers, offering a more flexible approach suited for clinical questions. Multimodal medical VQA has greatly benefited from transformer-based models and self-supervised learning, with techniques like CLIP's contrastive pretraining improving zero-shot capabilities by effectively combining visual and textual data.

The PathVQA dataset [1] serves as a benchmark for evaluating VQA models, with methods employing a mix of GRU [28] and Faster R-CNN networks [29] (Method 1), CNN [4] and LSTM networks with multimodal pooling (Method 2), and stacked attention networks (Method 3). The evaluation underscores Method 1's effectiveness in localizing image regions and Method 3's superiority in multi-layer attention processing. The dataset's open-ended question challenge and small size highlight the need for innovative approaches in medical VQA. BiomedGPT [30], a pioneering model in the biomedical domain, introduces a generative pre-trained Transformer architecture tailored for multimodal inputs. The model's pretraining phase contains masked image modeling, image infilling, and object detection for visual inputs, along with masked language modeling for textual data. The fine-tuning stage extends to tasks such as image captioning, visual question answering, image classification, text summarization, and natural language inference, emphasizing the model's application across a broad spectrum of biomedical modalities. The authors of [31] present an approach to open-ended VQA by leveraging pre-trained language models to interpret visual features extracted from medical images. This model demonstrates a generative method's advantages over traditional classification-based VQA models, particularly in handling class imbalances and enhancing the generalization of generated responses. The methodology underscores the importance of prompt structures in improving model performance, emphasizing the critical role of textual inputs in answer generation. LaVIN [32] introduces the Mixture-of-Modality Adaptation technique for vision-language instruction tuning of large language models. By employing lightweight adapters and a routing algorithm, LaVIN achieves joint optimization for vision and language tasks without extensive pre-training. However, LaVIN is unable

to identify fine-grained visual content, such as text characters. LLaVA-Med [33] enhances a learning method to fine-tune a vision-language conversational assistant, utilizing a large-scale dataset from PubMed Central [34] and instruction generation via GPT-4 [3]. However, LLaVA-Med faces challenges with hallucinations and lacks deep reasoning. Suggestions from the authors include enhancing the model's quality and reliability. M2I2 [21] presents a self-supervised method for pretraining on medical image caption datasets using a combination of masked image modeling, masked language modeling, image text matching, and contrastive learning. The results of M2I2 indicate the potential of self-supervised learning methods in enhancing model interpretability and performance in the medical domain.
Motivated by these findings, this study aims to explore self-supervised Transformer training, integrating effective strategies from existing models to advance medical VQA.

The assessment of generative medical tasks is necessary for determining the effectiveness of AI systems, especially as applications such as ChatGPT become increasingly utilized, even for medical topics. Despite the absence of universal standards for evaluating generative tasks in healthcare, this section outlines prevalent methods in current research, including both human-based and automated approaches.
Automated evaluations, preferred for their quick and cost-effective nature, include corpus-based and trainable metrics. Corpus-based metrics, such as BLEU ([35] and ROUGE [36] scores, rely on n-gram comparisons, while model-driven metrics such as BLEURT [37] predict human judgments using BERT [25] which is trained on large datasets.
The metrics are categorized into assessments of text quality and medical accuracy [38]. Textual metrics assess word overlap (BLEU, ROUGE) and semantic congruence (BERT score), alongside QuestEval [39] for evaluating the completeness of information. Despite their utility, these automated metrics do not sufficiently reflect human judgments of system performance.
Regarding the Medical correctness evaluation, evaluations use metrics such as Medical Concept Coverage [40] and Negation Correctness [41], applying domain-specific extraction methods and Named Entity Recognition [42] to ensure alignment with medical lexicons such as the Unified Medical Language Systems [43]. The F1 score [44], blending precision and recall, frequently features in medical VQA evaluations to measure concept correctness.
Informed by these findings, this research incorporates a selection of automated evaluation metrics, encompassing corpus-based approaches, assessments of text quality, and measures of concept correctness.

While the selected metrics effectively evaluate the quality of medical VQA systems, enhancing their explainability is essential for building trust through deeper understanding of the system's decision-making processes. The methods for achieving explainability are detailed in 2.3, with this thesis specifically employing an attribute-based approach to analyze how decisions are made.

Chapter 4

Methodology of Pathological-Llama

The methodology presents the development process of the model Pathological-Llama which generates answers from input images and natural language questions. It introduces the PathVQA dataset, tailored for medical VQA, and emphasizes the specialized knowledge required for interpreting pathological images. The chapter outlines the model's encoder-decoder architecture, incorporating a vision Transformer and a causal language model to process inputs and generate answers. Key innovations such as Low-Rank Adaptation for efficient fine-tuning, multimodal prompt construction for integrating visual and textual inputs, and a strategic masking strategy to focus attention on relevant tokens are discussed. This section lays the technical groundwork for understanding how the Pathological-Llama model tackles the complexities of medical VQA.

4.1 Problem Statement

The objective is to devise a method that, upon receiving an input image I and a corresponding question Q phrased in natural language, sequentially constructs an answer A consisting of N tokens $A = \{A_0, A_1, ..., A_N\}$. This construction process is conditioned on both the input image and the question. The goal from a model development standpoint is to determine the ideal parameters θ^* for the model. This is achieved by optimizing the model to maximize the conditional log-likelihood, formulated as:

$$\theta^* = \text{argmax}_\theta \sum_{i=1}^{N} \log p_\theta(A_i | Q, I, A_{i-1}). \tag{4.1}$$

4.2 PathVQA Dataset Description

The PathVQA dataset [1] is a multimodal dataset, consisting of images with textual data, specifically tailored for medical VQA. This dataset integrates pathological images and corresponding text-based question-answer pairs (QA). In the following, each combination of an image, and its associated question and answer is referred to as a 'datapoint'.

4.2.1 Qualitative Analysis

Pathological images require specialized expertise for accurate interpretation. Professionals in this field, known as pathologists, must complete a certification exam from the American Board of Pathology in the United States to be qualified. The PathVQA dataset is centered on VQA for pathology, utilizing images and accompanying captions sourced from two publicly-available pathology textbooks "Textbook of Pathology" [45] and "Basic Pathology" [46], and a publicly-available digital library: "Pathology Education Informational Resource" [47]. A semi-automated method is employed to convert these captions into QA pairs, which are then reviewed and refined manually by experts. The construction of questions within the PathVQA dataset mirrors the format and complexity of the American Board of Pathology certification examination, serving as a measure of an AI system's proficiency in medical decision support. The design of PathVQA is intended to train AI systems to match the analytical abilities of professional pathologists. This dataset presents a challenge as it demands the AI to not only discern the details within the images but also to comprehend and respond to related text-based questions.

4.2.2 Distribution

The initial PathVQA [1] dataset is comprised of 32 799 QA pairs, which are associated with 4 998 pathology images. On average, there are 6.6 questions per image, with the number of questions per image ranging from a minimum of 1 to a maximum of 14. The average word count is 9.5 for questions and 2.5 for answers.

The dataset categorizes questions into seven types: what, where, when, whose, how, how much/how many, and yes/no. 4.1 details the count and percentage of each question type. Of these, the first six categories are open-ended, totaling 16 465 questions and forming 50.2% of the dataset. The remaining is close-ended “yes/no” questions, with a balanced distribution of “yes” and “no” answers, numbering 8 145 and 8 189 respectively. Questions span a range of visual aspects such as color, location, appearance, and shape.

Question Type	**Total Number and Percentage**
Yes/No	16 334 (49.8%)
What	13 402 (40.9%)
Where	1 268 (4.0%)
How	1 014 (3.0%)
How much/How many	294 (0.9%)
When	285 (0.9%)
Whose	202 (0.6%)

TABLE 4.1: Distribution of Question Types in the PathVQA Dataset.

In the PathVQA dataset, "yes/no" questions are the most prevalent, making up 49.8% of the dataset, followed by "what" questions at 40.9%. The remaining categories, including "where", "how", "how much/how many", "when", and "whose", collectively account for less than 10% of the questions. This distribution underlines the focus on diagnostic and descriptive queries typical in pathology. The frequency of answers exhibits a long-tail distribution, with certain answers appearing frequently while most are less common.

4.2.3 Dataset Version

In the context of this project, the PathVQA dataset version was updated by the dataset authors [1] on February 15th, 2023. This particular version contains a collection of 5 004 images and a total of 32 795 datapoints with a balanced distribution of 16 332 close-ended and 16 463 open-ended QA pairs. 4 289 images are directly linked to at least one QA pair, while the remaining 715 images do not have any associated questions or answers.

Upon a detailed examination of the dataset, it is observed that certain datapoints are repeated across all subsets. To maintain the integrity and uniqueness of the dataset, these duplicate datapoints are removed in a preprocessing step. This refinement process results in a final count of 32 632, consisting of 16 394 open-ended and 16 238 close-ended unique QA pairs, associated with 4 289 images.

Despite this minor difference in the number of QA pairs, the overall composition and distribution of the dataset remain largely unchanged. Therefore, the statistics presented in 4.2.2 accurately reflect the dataset's structure and are deemed appropriate for the purposes of this project. This ensures that the dataset's representation and its role in the experimental analysis remain consistent and reliable.

4.3 Model Architecture

The model is structured using an encoder-decoder architecture. It incorporates a dual-stream encoder and employs a language model, functioning as a decoder, as depicted in 4.1. The encoder is tasked with processing two different input modalities: the image I and the natural language question Q. In contrast, the decoder, a causal language Transformer, generates the answer A autoregressively. This approach aligns with the prefix tuning technique, prompting the language model to produce outputs in a specific format, here an answer generated from a question-image pair.

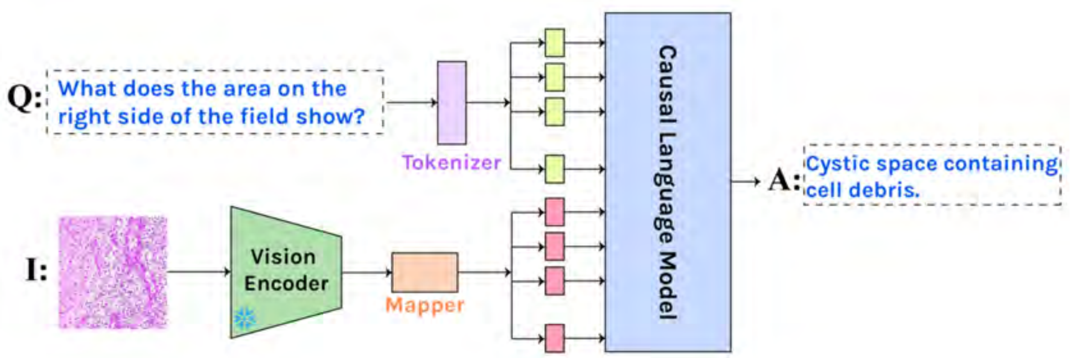

Q:
What does the area on the right side of the field show?
Tokenizer
I:
Vision Encoder
Mapper
Causal Language Model
A:
Cystic space containing cell debris.

FIGURE 4.1: Model Architecture [31].

For image encoding, a pre-trained vision encoder extracts visual features represented as $\{x_1, x_2, ..., x_{l_v}\}$. These features are then transformed into a visual prefix $\{v_1, v_2, ..., v_{l_v}\} \in R^{l_v \times e}$ compatible with the language model's latent space, where e is the embedding size and l_v is the visual prefix length. The transformation is achieved using a mapping network f_M, implemented as a three-layer Multilayer Perceptron (MLP).

For the textual input, first the question Q and the answer A are tokenized into a sequence of tokens. These tokens are then embedded using the embedding function of a pre-trained language model such that $Q = \{q_1, q_2, ..., q_{l_q}\} \in R^{l_q \times e}$ and $A = \{a_1, a_2, ..., a_{l_a}\} \in R^{l_a \times e}$. The language model treats VQA as a conditional text generation task. During training, it optimizes the standard maximum likelihood objective. The model takes a prompt sequence p as input and sequentially generates the answer A. At each step i, the model outputs logits that parameterize a categorical distribution $p_\theta(A)$ over the vocabulary, formulated as:

$$\log p_\theta(A) = \sum_{l_a} \log p_\theta(a_i | q_1, ..., q_{l_q}, v_1, ..., v_x, a_1, ..., a_{i-1}). \tag{4.2}$$

The language model's parameters are initialized from a pre-trained model, which has been extensively trained on large web-collected datasets. Details are described in 5.

4.3.1 Vision Model

The visual processing begins with a pre-trained vision Transformer model (ViT), specifically the ViT-B/32 variant [48]. The model processes input images through a series of self-attention and feed-forward layers, encoding the raw pixel data into a high-dimensional representation. This process is formally expressed as:

$$I \rightarrow \text{ViT-B/32} \rightarrow \{x_1, x_2, ..., x_{l_v}\}, \tag{4.3}$$

where I denotes the input image, and $\{x_1, x_2, ..., x_{l_v}\}$ represents the extracted visual features.

The subsequent stage involves translating these visual features into a visual prefix, ensuring compatibility with the latent space of the language model. This transformation bridges the gap between the visual and textual modalities, allowing the model to treat the visual information as if it were part of the natural language sequence. This transformation is achieved through a mapping network f_M, implemented as a MLP. The MLP is designed with dimensions tailored to project the visual features into the language model's embedding space. The process is formulated as:

$$\{x_1, x_2, ..., x_{l_v}\} \xrightarrow{f_M} \{v_1, v_2, ..., v_{l_v}\} \in R^{l_v \times e}, \tag{4.4}$$

where $\{v_1, v_2, ..., v_{l_v}\}$ denotes the visual prefix, and e signifies the embedding size of the language model.

The choice of the visual prefix length determines how much visual information, in terms of the number of embeddings, is incorporated into the language model's input sequence. It directly influences the balance between textual and visual information in the model's decision-making process. It can affect the model's performance, as it needs to be sufficient to convey the visual context without overwhelming the textual content.

4.3.2 Causal Language Model

The authors of [31] present a comparison between language models pre-trained on general text corpora and those trained specifically on medical texts. Models like GPT2 [49], which are trained on diverse, general text corpora, demonstrate superior overall performance compared to medically-trained models such as BioGPT [50] or BioMedLM [51]. One of the issues with medically-trained models like BioGPT and BioMedLM is their potential overoptimization to medical text corpora. This over-specialization often results in a lack of generalization capabilities when applied to different downstream domains. This limitation is particularly problematic in contexts where flexibility and adaptability across various types of data are essential, as in this study. In contrast, GPT2 benefits from exposure to a wide array of data during its pre-training phase, which includes medically oriented text. This broad spectrum of training data allows GPT2 models to easily generalize to other domains, an attribute that is highly valuable for diverse VQA datasets.

Motivated by these findings, Llama2 [52] and Tiny-Llama [53] were selected for their architectural strengths and diverse training backgrounds. The choice of Llama2 is particularly motivated by its training on a more extensive parameter set than GPT2-xl. This is expected to enhance its generalization in downstream tasks.

4.3.3 Low-Rank Adaptation as a Parameter-Efficient Strategy

Applying Low-Rank Adaptation (LoRA) [54] to the Pathological-Llama model presents a strategic solution to the challenges associated with the extensive parameter set of Transformer models, particularly relevant in the computationally demanding domain of medical VQA.

Fine-tuning a Transformer model traditionally involves a sequence of steps: in the forward pass, the input data is processed through the network, generating predictions. During loss calculation, the model's predictions are compared with the actual data to quantify the loss. Followed by the backward pass, the gradient of the loss with respect to the model's parameters is computed. Finally, to update the weights, the model's parameters are adjusted to minimize the loss. In this process, weight adjustment, specifically for a single linear projection, is typically denoted as $W' = W + \Delta W$, where W is the weight matrix, ΔW is the change-in-weights matrix, and W' are the updated weights.

LoRA introduces a modification to this traditional approach by keeping the weight matrix W and the change-in-weights matrix ΔW separate throughout the fine-tuning process. This approach is especially beneficial for this project due to its efficiency in updates. The forward pass under LoRA is represented as:

$$h = W_0 x + \Delta W x$$

where W_0 is the initial weight matrix, kept static during fine-tuning, ΔW is the adaptable change-in-weights matrix, and x is the input vector. The updates are specifically applied to ΔW, denoted as W_ϕ, and the updated change-in-weight matrix is expressed as $W'_\phi = W_\phi + \Delta W_\phi$.

The essence of LoRA lies in its underlying principle that W_ϕ possesses a low intrinsic rank and can therefore be efficiently approximated by the product of two lower-rank matrices, B and A, thus reducing redundancy and achieving a more compact representation. This is articulated as:

$$h = W_0 x + W_\phi x = W_0 x + BAx$$

The dimension r, a hyperparameter, defines the compression level; a smaller r means a more compact representation and fewer trainable parameters in BA compared to W_ϕ.

The strength of LoRA is further underscored in its learnable parameters within matrices B and A, which are fine-tuned during the process. Once fine-tuned, the product BA can be seamlessly integrated with the static weights W_0, thereby effectively incorporating the learned adaptations.

4.3.4 Multimodal Prompt Construction

In the training phase, the model integrates the visual prefix with the textual prompt, establishing a context for the language model to generate accurate answers. This integration is done by injecting the visual prefix into the embedding sequence, positioning it right after the question tokens. For a refined structure of the prompt sequences, the model adopts a descriptive labeling strategy. Descriptors such as "question:", "context:", and "answer:" are prepended to the respective question, image, and answer tokens. This strategy aims to enhance the model's focus and comprehension. A typical structured prompt embodies the format:

$$p = [\text{"question:"}\{q_1, ..., q_{l_q}\}, \text{"context:"}\{v_1, ..., v_{l_v}\}, \text{"answer:"}],$$

where $\{q_1,...,q_{l_q}\}$ represents the tokenized and embedded question, and $\{v_1,v_2,...,v_{l_v}\}$ denotes the visual prefix. This structured prompt, a fusion of textual and visual input, is then fed into the language model.

The study of [31] reveals that reversing the order of question embeddings and visual data leads to a performance degradation. This indicates that the model's processing and prioritization of information are influenced by the arrangement of the prompt components. Specifically, placing the visual information before the question often results in the model undervaluing or even ignoring the visual context, leading to responses that lack visual grounding. Furthermore, the authors of [31] emphasize the importance of the question component in the prompt structure. Experiments demonstrate a marked performance decline when the question element is omitted, unlike the removal of visual information.

4.3.5 Masking Strategy

The masking strategy employs a mask $M = \{m_1,m_2,...,m_{l_p}\} \in R^{l_p \times e}$, where l_p represents the length of the tokenized input prompt. Each element m_i in the sequence, for $i \in \{l_p\}$, is designated as either 0 or 1. The value is set to 1 for each prompt token and 0 for padding, which is appended to the right of the input sequence. This masking approach ensures that the model's attention is directed towards tokens of contextual importance.

During the training phase, the prompt is concatenated with the padded target answer A, and the mask is assigned a value of 0 for each a_i with $i \in \{l_a\}$. This setup allows for the computation of loss during training, aligning the output size with the target size. This configuration enables the model to concentrate exclusively on the question as textual input and the context with prefixes as visual input. In contrast, during the testing phase, A is excluded as the focus is shifted to token generation, which does not necessitate a predefined output length. This deliberate masking mechanism ensures that the model, uninformed by the actual target during training, retains its capacity to generate precise answers in the testing phase.

4.3.6 Answer Generation with Beam Search

The answer generation process utilizes a beam search algorithm [55] to sequentially construct accurate and contextually relevant answers. Beam search, a heuristic search algorithm, maintains a set of k most promising sequences (beams) at each generation step. The beam size, k, is a hyperparameter that influences the breadth of search and the diversity of generated answers.

Given an input prompt p, the goal is to generate an answer $A = \{a_1,a_2,...,a_n\}$ where each a_i represents a token in the answer sequence. The beam search algorithm iteratively predicts the next tokens based on the current state and updates the beams based on these predictions. For initialization, let $B_0 = \{\emptyset\}$, where B_0 is the initial beam set. For each step i, the algorithm predicts the next set of tokens T_i for each beam in B_{i-1}, using the model's current state and input data to generate output logits, subsequently applying temperature scaling and softmax to derive a probability distribution over the next possible tokens.

$$T_i = \text{softmax}\left(\frac{\text{logits}(B_{i-1},Q,I)}{\tau}\right), \tag{4.5}$$

where τ is the temperature parameter controlling the diversity of predictions. The beams are updated based on these predictions, selecting the top k sequences with the highest cumulative scores:

$$B_i = \text{topk}\left(\bigcup_{b \in B_{i-1}} \text{expand}(b, T_i)\right). \tag{4.6}$$

The process repeats until each beam in B_i reaches the predefined stop token or the maximum sequence length n. The final set of beams is then decoded into text, and the sequences are scored to identify the most accurate answers:

$$A^* = \arg\max_{a \in B_n} \text{score}(a) \tag{4.7}$$

where A^* represents the highest-scoring answer sequences.

This strategy, similarly to the one implemented by the authors of [31], differentiates in that it halts upon reaching a stop token, outputting only the sequence up to that point.

Chapter 5

Experimental Setup

This chapter describes the framework for evaluating the Pathological-Llama model, detailing the structured dataset split for training, validation, and testing, and the set of evaluation metrics used, including Cross-Entropy Loss, BLEU, BERT score, and F1 score. It emphasizes the use of the BERT score for semantic evaluation and the F1 score for accuracy assessment. The chapter also outlines the technical setup, mentioning implementation details and computational resources.

5.1 Dataset Split

The dataset, as detailed in 4.2.3, is organized into three subsets: the training set, the validation set and the testing set. This structure follows the official split of the PathVQA dataset and is designed to facilitate standardized performance comparisons. To ensure the integrity and meaningfulness of these subsets, a preprocessing step is additionally implemented to eliminate any duplicated datapoints. The distribution of datapoints across these distinct subsets is 0.6 for training, 0.2 for validation, and 0.2 for testing. This allocation strategy ensures a balanced and representative dataset for robust model evaluation and testing. The specifics of this dataset split are outlined in Table 5.1.

Dataset Split	**# Images**	**# Datapoints**	**# Open-ended**	**# Close-ended**
Training set	2 499	19 654	9 903	9 751
Validation set	1 499	6 719	3 357	3 362
Testing set	1 000	6 259	3 134	3 125

TABLE 5.1: Statistics of Images and QA Pairs in Dataset Splits.

The split and distribution of the dataset are visually represented in 5.2 and 5.1.

5.2 Evaluation Metrics

To evaluate the proposed model's performance, a range of evaluation metrics are chosen for the training, validation, and testing phases, with consideration of textual integrity and content correctness.

During the training phase, Cross-Entropy Loss [56] is utilized to refine the model's performance by minimizing the discrepancies between generated answers and the actual answers. The validation phase incorporates Cross-Entropy Loss alongside BLEU, BERT score, and F1 score to evaluate linguistic and semantic quality. BLEU measures textual similarity to the ground truth using n-grams [35], while BERT score evaluates semantic similarity [57],

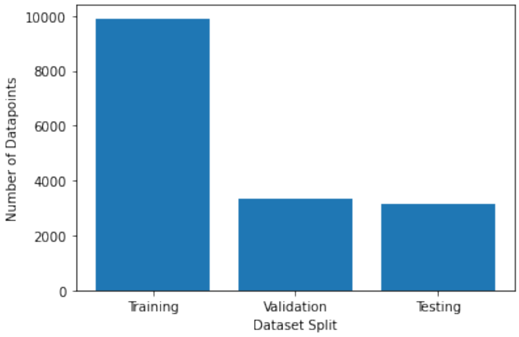


FIGURE 5.1: Open-ended Dataset Split.

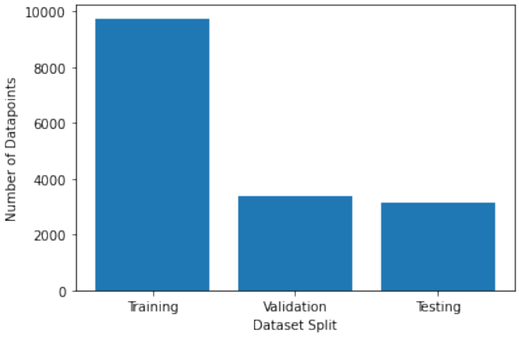


FIGURE 5.2: Close-ended Dataset Split.

emphasizing the model's capability to produce answers that are contextually aligned with the ground truth, thereby measuring relevance and coherence. The F1 score serves as a balanced metric for precision and recall [44], offering insight into the model's overall accuracy. Testing focuses on semantic precision via BERT score and contextual relevance, examining the model's capacity to deliver context-relevant responses. This phase emphasizes real-world applicability, with the F1 score providing an accuracy assessment.

5.2.1 BERT Score

BERT score provides an alternative for text generation evaluation, addressing the shortcomings of n-gram metrics such as BLEU score. Unlike n-gram approaches, which often misjudge paraphrases and struggle with capturing semantic reordering and long-range dependencies, BERT score uses contextual embeddings from models, which is BERT [25] in this study. This approach allows for a deeper semantic comparison between generated text and reference text based on cosine similarity.

The process involves converting both reference $x = \langle x_1, \ldots, x_k \rangle$ and candidate $\hat{x} = \langle \hat{x}_1, \ldots, \hat{x}_m \rangle$ sentences into sequences of vectors. These vectors are obtained by tokenizing the sentences into subwords and encoding them with a Transformer encoder, which applies self-attention and nonlinear transformations to capture contextual information. Evaluation is performed by greedily matching tokens from the candidate sentence to the reference sentence to maximize similarity, calculating recall and precision. These metrics are then combined to compute an F1 score:

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j$$

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j$$

$$F_{\text{BERT}} = 2 \cdot \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}$$

To enhance readability and interpretation, BERT score is rescaled based on an empirical lower bound b, calculated using Common Crawl data. This rescaling adjusts the score to a more intuitive range:

$$\hat{R}_{\text{BERT}} = \frac{R_{\text{BERT}} - b}{1 - b}$$

This method ensures scores are typically between 0 and 1, improving readability without compromising the score's ability to rank text generation quality or its correlation with human judgment.

5.2.2 F1 Score

The F1 score evaluates the balance between precision and recall. It measures the overlap between candidate and reference tokens, providing a single metric for model accuracy. In the context of the project's task, it identified common tokens between the candidate and reference texts, calculating the number of shared tokens by counting the minimum occurrence of each common token across both texts. Precision is then determined as the ratio of shared tokens to the total in the candidate text, and recall is the ratio of shared tokens to the total in the reference text. The F1 score is calculated as:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

This ensures high scores only when both precision and recall are high, emphasizing the model's effectiveness in generating relevant text. If there are no shared tokens or if any token set is empty, the F1 score defaults to 0. This metric is essential for models where accurate and complete text generation is aimed for.

5.3 Implementation Details

For the experiments, a pre-trained CLIP [48] model with a ViT backbone is employed to extract visual features. This model is selected for its effectiveness in processing a wide range of visual content. The output of the CLIP model is a 512-dimensional feature vector, providing a comprehensive representation of the visual data. The mapping network f_M transforms visual features into a format compatible with language models. This network consists of MLP layers. The layer sizes are defined as 512, $\frac{l_v * e}{2}, l_v * e$, with l_v being the variable visual prefix length. The prefix length l_v is adjusted throughout the experiments for optimal performance. Sequence lengths l_q and l_a, corresponding to questions and answers, are determined based on dataset characteristics. Specifically, these lengths are set to the mean number of tokens in the training set plus three times its standard deviation, ensuring coverage of most dataset instances while avoiding excessive padding for shorter sequences. To accommodate sequences of varying lengths in batch-wise learning, padding with End-Of-Sequence (EOS) token is applied. To configure LoRA, the rank (r) is set to 8, and the $lora_{\alpha}$ parameter is established at 32. Additionally, $lora_{dropout}$ is fixed at 0.1. The Adam optimizer [58] is used for training, with 600 warmup steps. These values were adopted based on findings of [31]. The experimental framework involves the use of two language models interchangeably:

- Llama2 [52] spans models from 7 billion to 70 billion parameters, aimed at text-based tasks using an auto-regressive language model with an optimized transformer architecture. This study uses the Llama-2-7b-chat-hf model from Meta AI, which has 7 billion parameters and is fine-tuned for chat applications. It uses a transformer encoder-decoder architecture which enables contextual understanding and text generation, making it well-suited for dialogue tasks. Its fine-tuning on dialogue datasets allows the model to produce conversational responses that are coherent and engaging. The model employs SentencePiece tokenization [59], handling a vocabulary of 32 000 tokens.

- Tiny-Llama [53] adopts the same architecture and tokenizer as its counterpart, Llama2, but is scaled down to 1.1 billion parameters, making it more accessible for projects

> constrained by computational and memory resources. It is presented as an intermediate model, having been trained up to 50,000 steps on a dataset comprising 105 billion tokens. For this research, the specific pretrained version employed is TinyLlama-1.1B-Chat-v0.1, which is a chatbot variant of the Tiny-Llama model, fine-tuned on the openassistant-guananco [60] dialogue dataset. An intermediate checkpoint is employed, as the model is still undergoing training as of this study. Despite the existence of a newer chat-based version, TinyLlama-1.1B-Chat-v0.4, TinyLlama-1.1B-Chat-v0.1 is chosen for its superior performance, as detailed in appendix A.

The baseline configuration for the experiments is determined by several parameters, mindful of computational resource constraints. A batch size of 6 is selected as it represents the upper limit that Llama2 can accommodate, allowing for the maximum visual prefix length achievable within the computational budget. Further details on these experiments are provided in appendix A. A learning rate of 1e-4 is adopted, close to the 5e-3 rate used in prior studies [31]. However, a rate of 5e-3 and 1e-3 resulted in model divergence, leading to the selection of 1e-4 as the starting point for experiments. The epoch count is set at 60 to balance training time with achieving desirable loss levels. This decision is supported by the observed trends in training and validation loss, detailed in each subsection. Additionally, a visual prefix length of 6 is chosen as the starting parameter, being the maximum feasible option for Llama2 even with a batch size of 1, due to computational resource limitations.

The experiments were conducted on a shared server cluster featuring Tesla V100-SXM2-32GB GPUs, 240 CPU cores, 1536 GiB of RAM, and 10.5 TiB of high-speed local SSD storage.

Chapter 6

Results

This chapter details the outcomes from testing various model configurations, focusing on model architecture, capacity, and visual prefix length. It aims to highlight configurations that enhance VQA performance through analysis across training (Train), validation (Val), and testing (Test) phases. Key findings include the impact of chat-based pre-training, the positive correlation between model capacity and performance, the role of visual prefix length in integrating visual-textual information, and the advantages of generative models for close-ended questions. The chapter concludes with an evaluation of the best-performing models, based on BERT and F1 scores.

6.1 Findings

The results presented in this chapter are derived from a series of experiments, as detailed in the corresponding subsections. Multiple models are trained and evaluated, with their performance metrics compared. The key aim is to identify the most effective model configuration in regard of the evaluation metrics, informed by insights gained from preceding experiments, to enhance overall performance, with a particular emphasis on evaluating Val BERT for model selection.

6.1.1 Benefit of Chat-Based Pre-Trained Models

This section examines how model architecture, particularly task-specific features like chat-based interaction, influences performance in this project's VQA task. The hypothesis is that architectural design tailored to specific tasks enhances a model's effectiveness in VQA scenarios. To test this, Tiny-Llama and Tiny-Llama-Chat were compared. Both models have 1.1 billion parameters but differ in their orientation towards chat-based tasks. They were fine-tuned on PathVQA with a learning rate of 1e-4 for 60 epochs, using visual prefix size of 6. Performance metrics at epoch 60 are summarized in table 6.1.

Model	Train Loss	Val Loss	Val BLEU	Val BERT	Val F1
Tiny-Llama	0.1259	0.7275	0.5056	0.521	0.322
Tiny-Llama-Chat	0.0626	0.7747	0.5071	0.550	0.363

TABLE 6.1: Performance Comparison of Tiny-Llama and Tiny-Llama-Chat.

The comparison clearly shows Tiny-Llama-Chat's improved performance by across most evaluated metrics compared to Tiny-Llama. This includes a decrease in training loss and improvements in BLEU, BERT, and F1 scores, as illustrated in 6.1 and 6.2. Such enhancements underscore the chat-specific features' effectiveness in VQA tasks, where answering questions often involves interpreting complex visual and textual data in a conversational manner.

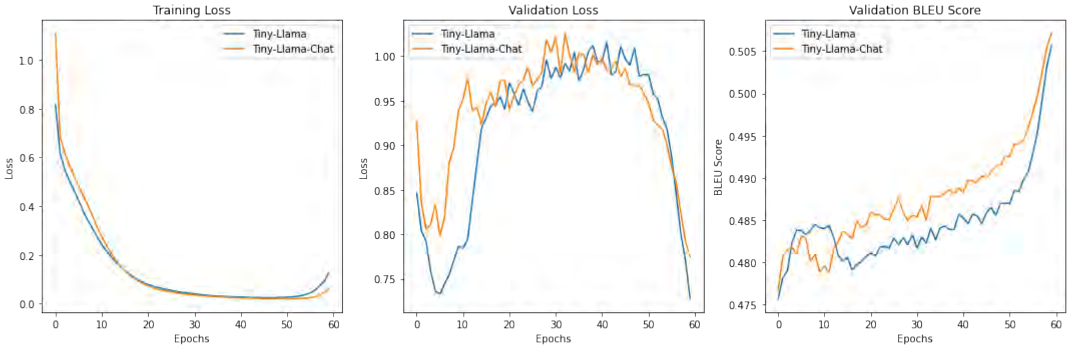


FIGURE 6.1: Train and Validation Performance of Tiny-Llama and Tiny-Llama-Chat over 60 Epochs.

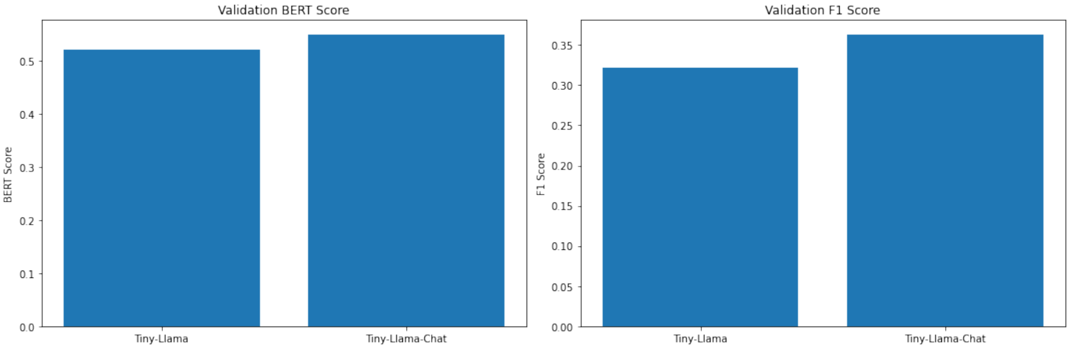


FIGURE 6.2: Validation BERT and F1 score at Epoch 60.

Tiny-Llama-Chat's advanced performance underlines the importance of task-aligned architectural design. The next sections will continue with only chat-based models, exploring in greater detail how these specialized architectures address the challenges and leverage the opportunities within the VQA task.

6.1.2 Benefit of Greater Model Capacity

This experiment investigates the connection between model capacity, indicated by the parameter count, and performance. The guiding hypothesis suggests that models with more parameters should exhibit superior performance, attributed to their enhanced capability to discern and assimilate complex data patterns. To test this, a comparative study was conducted between Tiny-Llama, with 1.1 billion parameters, and Llama2, boasting 7 billion parameters, both configured for chat-based functionalities. These models were assessed under uniform conditions, employing a visual prefix size of 6 across 60 epochs with a learning rate (lr) of 1e-4 and 1e-5. Table 6.2 presents the performance metrics for both models at different learning rates.

The data illustrates the relationship between model capacity, learning rate, and performance metrics. Notably, Llama2, with its higher parameter count, demonstrates variable performance based on the learning rate. At a lower learning rate of 1e-5, Llama2 outperforms Tiny-Llama across all metrics, showcasing the advantage of larger model capacity in capturing intricate patterns This is evidenced by the highest validation BERT score and F1 score at

Model	Lr	Train Loss	Val Loss	Val BLEU	Val BERT	F1
Tiny-Llama	1e-4	0.0626	0.7747	0.5071	0.550	0.363
Tiny-Llama	1e-5	0.0698	0.8208	0.5032	0.540	0.351
Llama2	1e-4	0.0340	0.8713	0.5054	0.528	0.330
Llama2	1e-5	0.0178	0.8475	0.5091	0.585	0.404

TABLE 6.2: Comparison of Model Performance with Different Learning Rates.

this learning rate, highlighting the model's improved comprehension and contextual processing abilities.

Conversely, with a learning rate of 1e-4, the performance advantage of Llama2 in terms of Val Loss and Val BERT scores diminishes, suggesting that the benefits of increased model capacity can be contingent upon the optimal tuning of learning rates. Tiny-Llama, despite its smaller size, maintains competitive performance, especially notable in its Val BERT and Val F1 scores at 1e-4, indicating that smaller models can still effectively process and generate conversational content when optimally configured. This is visualized in 6.3 and 6.4.

The comparison underscores the complexity of the relationship between model capacity and performance. While larger models like Llama2 have the potential to achieve superior performance, this capability is highly dependent on the careful tuning of learning rates and other hyperparameters. This highlights the importance of not only considering model size in the development of VQA systems but also the optimization of training configurations to fully leverage the models' capabilities.

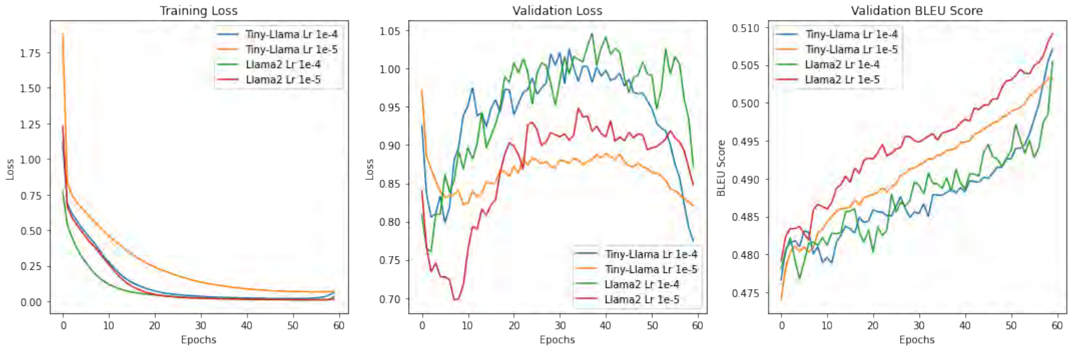


FIGURE 6.3: Train and Validation Performance over 60 Epochs.

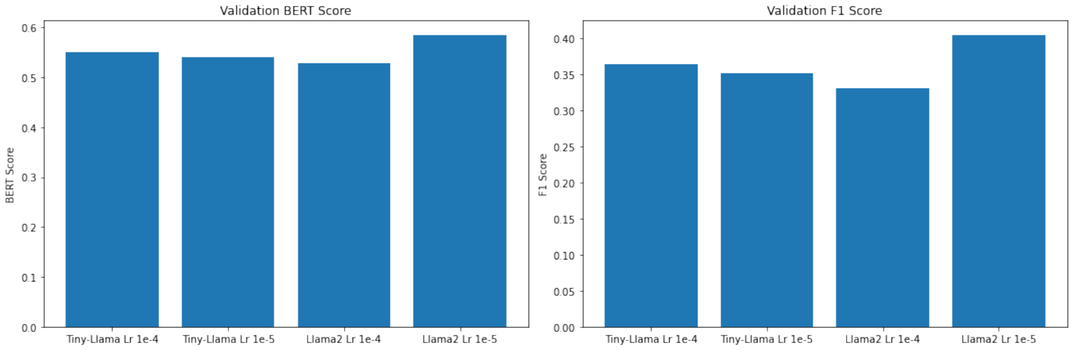


FIGURE 6.4: Validation BERT and F1 score at Epoch 60.

The results confirm the hypothesis with emphasizes on the importance of thorough tuning. Tiny-Llama demonstrates optimal performance at a learning rate of 1e-4, whereas Llama2 achieves its best results using a learning rate of 1e-5. Consequently, the subsequent analysis of results will focus on these specific learning rate configurations for each model.

6.1.3 Benefit of Increased Visual Prefix Length

The visual prefix length impacts how models integrate visual data into their decision-making, hypothesizing an optimal balance that best combines visual and textual inputs. Too short a visual prefix may lack necessary context, while an excessively long one could overwhelm the textual information.

This study compared Tiny-Llama and Llama2, both chat-adapted models, across different visual prefix lengths. For Tiny-Llama, a learning rate of 1e-4 was used, exploring prefix lengths of 2, 6, and 12, the latter being the maximum feasible given computational constraints. Llama2, at a 1e-5 learning rate, was tested with prefix lengths of 2, 4, and 6, which is the maximum possible for this configuration. The performance data is summarized in table 6.3.

Model	Prefix	Train Loss	Val Loss	Val BLEU	Val BERT	Val F1
Tiny-Llama	2	0.1366	0.7347	0.5524	0.518	0.327
Tiny-Llama	6	0.0626	0.7747	0.5071	0.550	0.363
Tiny-Llama	12	0.0668	0.7843	0.4500	0.570	0.402
Llama2	2	0.0775	0.7879	0.5525	0.516	0.316
Llama2	4	0.0300	0.8217	0.5309	0.581	0.405
Llama2	6	0.01780	0.8475	0.5091	0.585	0.404

TABLE 6.3: Impact of Visual Prefix Length on Model Performance

The results reveal that an increase in visual prefix length generally improves validation BERT and F1 scores across both Tiny-Llama and Llama2 models, suggesting a better capability for contextual understanding and processing with more visual data. However, this comes with a corresponding rise in validation loss, indicating the challenge of balancing visual and textual inputs effectively.

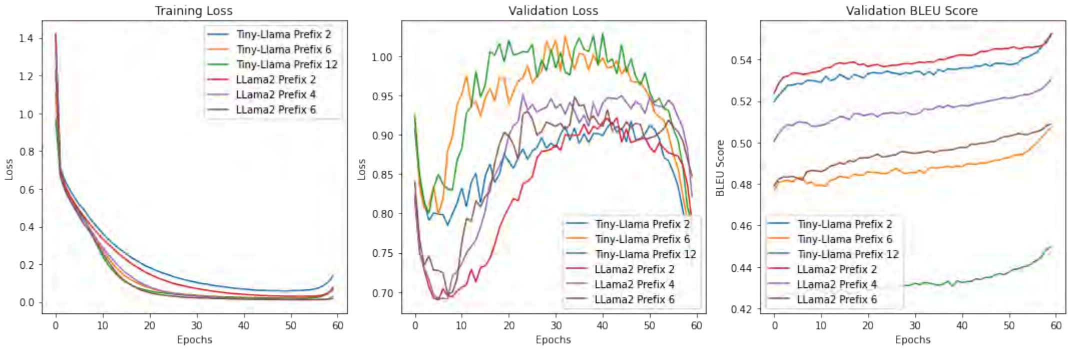


FIGURE 6.5: Train and Validation Performance over 60 Epochs.

Tiny-Llama reaches peak performance with a visual prefix length of 12, achieving the highest validation BERT and F1 scores despite increased validation loss, indicating an enhanced ability to generate contextually relevant responses. Llama2, attaining its best performance at a visual prefix length of 6, demonstrates superior contextual processing and accuracy, evidenced by its high validation BERT and F1 scores, though at the cost of higher validation loss.

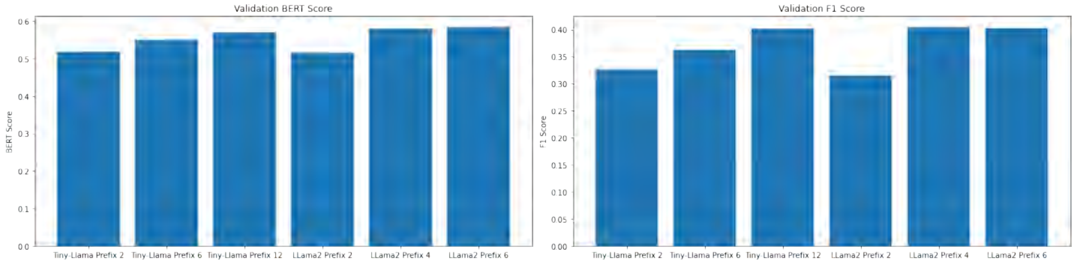


FIGURE 6.6: Validation BERT and F1 score at Epoch 60.

These patterns suggest that while extending the visual prefix length can improve a model's response accuracy and relevance in a VQA context, finding the optimal length is needed to maximizing both visual understanding and textual coherence without compromising model performance.

6.1.4 Benefit of Generative Task for Close-Ended Question Types

This section investigates the effectiveness of using a generative approach over a classification approach for answering close-ended questions. This experiment is based on the hypothesis that generative models are more suited for this task than traditional classification models, considering the inherent nature and capabilities of the LLaMA-2 model in generative tasks.

The experiment involves the comparison between the Llama2 model configured for a generative task and the same model adapted for a classification task. Both models were trained with a learning rate of 1e-5, over 30 epochs and a visual prefix size of 6, specifically trained and evaluated on close-ended questions. The classification model was structured similarly to the generative model with the addition of a fully connected layer after the model's output logits, designed to output two classes: 'Yes' or 'No'.

Model	Val Loss	Test Accuracy	Test F1
Llama2 (Classification)	0.6967	0.5011	0.5161
Llama2 (Generative)	0.0697	0.875	0.937

TABLE 6.4: Performance of Llama2 Models for Classification and Generative Tasks.

The results as in 6.4 indicate that the Llama2 model configured for the generative task outperforms its classification counterpart. Notably, the generative model achieves higher accuracy and F1 score compared to the classification model. This difference in performance can be attributed to the intrinsic design and strengths of the Llama2 model, which is suited for generative tasks. The generative approach allows the model to leverage its full potential in understanding and generating language, leading to a more nuanced and accurate interpretation of close-ended questions. In contrast, while the classification model is designed to provide direct 'Yes' or 'No' answers, it may not fully encapsulate the contextual nuances and depth of understanding that the generative model possesses. This could result in a more superficial analysis of the questions, as reflected in the lower performance metrics. These findings underscore the importance of aligning the model architecture and task type, highlighting that Llama2's capabilities are more effectively used in generative tasks, especially for handling close-ended questions in a more context-aware manner.

6.2 Evaluation

The testing phase employs a previously unseen dataset to evaluate the models' performance, focusing on key metrics such as BERT and F1 scores. The selected best models for this phase are Tiny-Llama with a learning rate of 1e-4 and a visual prefix length of 12 and Llama2 with a learning rate of 1e-5 and a visual prefix length of 6, both chat pre-trained.

These configurations were identified as optimal based on their performance in the validation phase, specifically targeting the enhancement of the models' comprehension and response generation in the context of medical VQA. The test phase results are summarized in the table 6.5:

Model	Learning Rate	Prefix Length	Test BERT	Test F1
Tiny-Llama	1e-4	12	0.573	0.403
Llama2	1e-5	6	0.591	0.419

TABLE 6.5: Test Phase Performance of Optimized Models

The test results reinforce the findings from the validation phase, with both models demonstrating robust generalization capabilities. Llama2, in particular, shows a slight improvement in both Test BERT (0.591) and Test F1 scores (0.419) compared to its validation performance, suggesting a strong ability to understand and process visual and textual data in VQA tasks.

Interestingly, despite Llama2's more than six times larger parameter count, Tiny-Llama's performance in the testing phase closely rivals that of Llama2, with a BERT score of 0.573 and an F1 score of 0.403. This comparison underlines Tiny-Llama's exceptional efficiency and efficacy as a model, particularly in scenarios where computational resources might be constrained.

These outcomes highlight the effectiveness of the selected configurations for each model, indicating the importance of optimizing learning rates and visual prefix lengths to enhance model performance. The slight improvements and stability in test scores compared to validation metrics suggest that both models are well-tuned to their respective task demands, capable of delivering consistent and accurate responses across different datasets.

Chapter 7

Explainability

Pathological-Llama model's reasoning is explained using the Integrated Gradients method, highlighting how the textual and visual inputs influence predictions through visualization techniques. This method provides insight into the model's logic across various medical scenarios, such as cardiovascular and endocrine systems. To address the expertise required for interpreting pathological images, GPT-4 for is employed for additional analysis, comparing these AI-generated explanations with the model's decisions. By integrating advanced AI for explainability, this chapter aims to enhance transparency and trust in the model's healthcare applications, demonstrating an approach to understanding AI-driven diagnostic processes.

7.1 Implementation

The explainability analysis employs the Integrated Gradients (IG) method on Pathological-Llama with Tiny-Llama as the causal language model with the configuration as in 6.2. The choice of the smaller model for this analysis is due to computational efficiency and its close performance alignment with its counterpart, Llama2 as causal language model. This method would also work on the latter, with a reduction in some parameters. The interpretability results for this are recorded in the appendix B.1

This interpretability algorithm assigns importance scores to input features by estimating the gradients of the model's output concerning its inputs along a straight-line path from selected baselines to the actual inputs. For this analysis, the critical inputs include both the visual and textual components of the model. Baselines are defined as a black image for visual inputs and a sequence of embedded padding tokens for textual inputs, representing the absence of signal from which the integral calculation starts. The comparison involves attributing importance to model inputs by contrasting them with their baselines. For visual attributions, the raw pixel values are compared against the baseline and the original image, necessitating that they maintain the same dimensions. Textual attributions require the embeddings of the text, as IG attributes importance to embeddings rather than indices, given the method's reliance on gradient-based attribution which cannot be computed with integer values. Initially, pixel values are converted into embeddings, or visual prefixes, during preprocessing. However, IG necessitates the use of raw pixel values. A forward hook is implemented to take pixel values and apply the same transformations as during training to accommodate this. This includes generating embeddings, inputting the visual embedding into the vision model to create visual prefixes, and then projecting these visual prefixes into the text embedding space. These processed inputs are then fed into the language model to generate outputs, which are subsequently analyzed using IG. The attribution method uses 30 steps for approximation and the Right Riemann Sum [61] method for integral calculation. The Right Riemann Sum approach estimates integrals by summing areas of rectangles aligned with the right side of each interval, offering an effective means to quantify input feature contributions to model predictions.

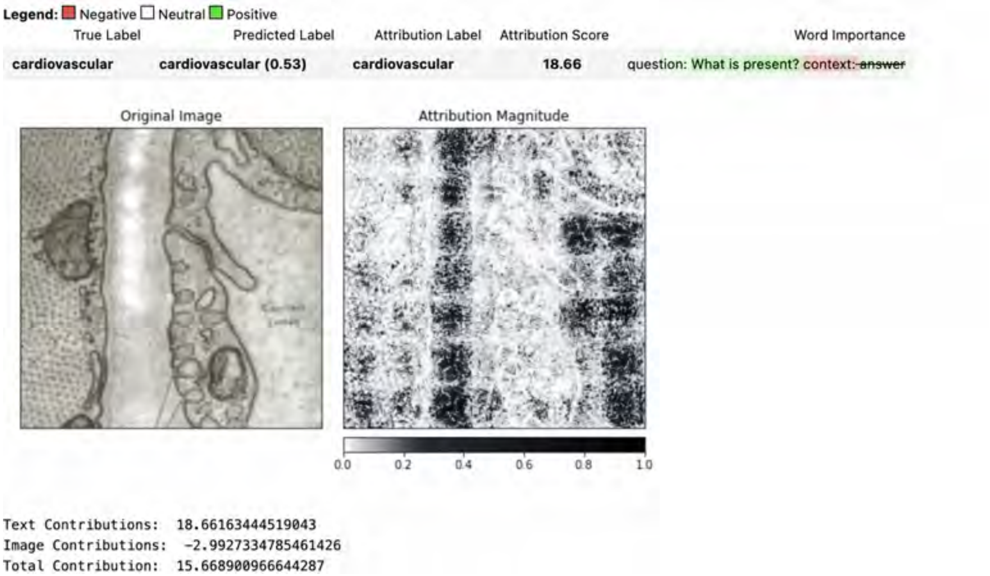
7.2 Visualization and Analysis of Examples

In the visualizations, red denotes negative, green positive, and white neutral attributions for text. "True Label" is the actual label of the data, while "Predicted Label" shows the model's prediction along with a probability score. The IG method, adapted from [2] using PyTorch [62] and the Captum [63] library, originally suits classification tasks with single-word labels. This study extends it to generative tasks, where "Predicted Label" becomes the first token's probability in the generated sequence. The "Attribution Label" specifies the label for which the attribution is calculated. Attribution scores reveal how each input feature influences the prediction: positive scores indicate a feature's positive contribution, negative scores the opposite, and zero means no contribution. For text, word importance visualizations display attributions without visual prefixes, highlighting how each token impacts the model's output. Image attributions are represented through an attribution map, where black pixels denote absolute values, emphasizing the importance of each pixel in the image analysis. The darkness of a pixel correlates with the magnitude of its integrated gradient, with colorless pixels indicating no gradient presence. Both text and image attributions are combined to illustrate their overall effect on the model's prediction.

To analyze visual attributions, expertise in pathological imagery is required, which falls outside the scope of this study. Consequently, OpenAI's Chatbot, GPT-4 [3], is utilized to generate responses to various prompts. For each example, a new conversation with GPT-4 is initiated to ensure unbiased responses by preventing the chatbot from recalling previous interactions. This approach serves the purpose of allowing a comparison between GPT-4's responses and Pathological-Llama's predictions and offers potential explanations for Pathological-Llama's decision-making process. The GPT-4 generated responses are shortened in a subsequent prompt, ensuring that the information is still contained but in a more concise form for better comparison for this analysis. This analysis aims to extend the study with additional insights and does not substitute for expert evaluation. Similarly, Google's Chatbot Gemini [64] is employed following the same methodology, with the outputs detailed in the appendix B.2, alongside additional examples. The responses from the chatbots are shortened by the chatbots themselves, ensuring that the essential information is preserved in a more concise format to enable a clearer comparison for this analysis.

7.2.1 Example 1: Cardiovacular

Figure 7.1 positive attributions are assigned to "what" and "present," indicating their importance in forming the question's diagnostic intent. The term "context:" is given a negative attribution, reflecting its minimal impact in the presence of keywords directing towards a diagnostic query. Visual attributions center on the image's hollow region, aligning with the "cardiovascular" prediction for this sample. The textual input, with a contribution score of 18.66 overshadows the visual input's minor contribution of -2.99. The preference for textual information over visual ones was elaborated by the authors of [31], whose model was similarly constructed as in this study, highlighting the model's dependency on textual information for accurate diagnostics.



Legend: 🟥 Negative ☐ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
cardiovascular	**cardiovascular (0.53)**	**cardiovascular**	**18.66**	question: What is present? context: ~~answer~~

```
Text Contributions:  18.66163444519043
Image Contributions:  -2.9927334785461426
Total Contribution:  15.668900966644287
```

FIGURE 7.1: Example 1: Cardiovascular from Test Set.

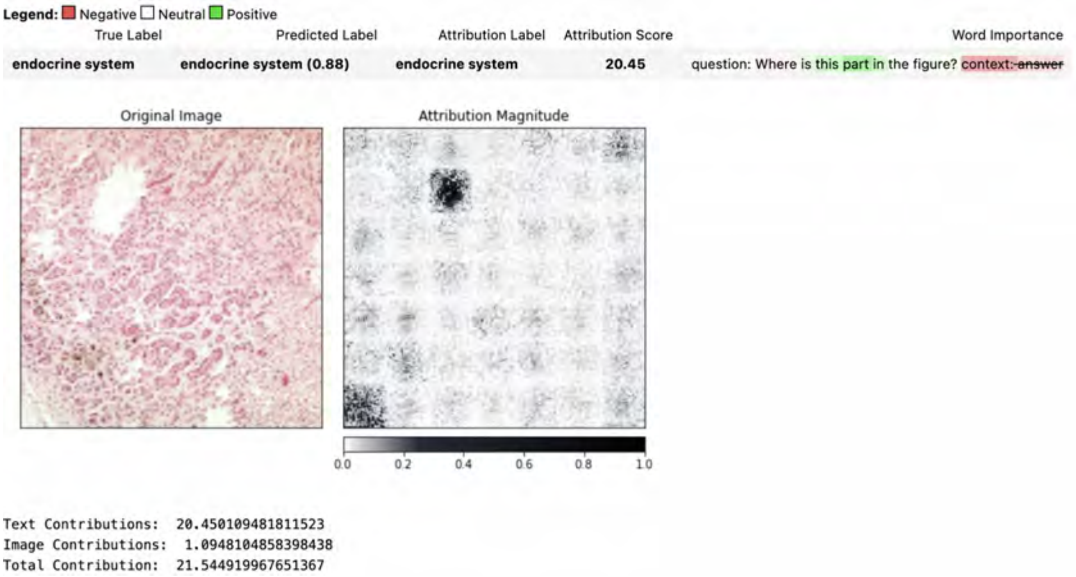
Table 7.1 presents prompts and responses from GPT-4, offering insights into Pathological-Llama's attribution process. Initially, GPT-4's analysis, based solely on the original image and prompt, identifies a capillary and its lumen, key cardiovascular features, to which Pathological-Llama's visual attributions also point. Further interaction, incorporating the correct "cardiovascular" label and visual focus on the image's hollow areas, leads GPT-4 to affirm the interpretability method's rationale applied in this example.

Prompt: What is present?
GPT-4: The image you've uploaded appears to be an electron micrograph, which is a type of high-resolution photograph taken with an electron microscope. The details in the image are consistent with a cross-sectional view of biological tissues at the microscopic level. We can see what looks like a capillary (a small blood vessel) with its lumen (the hollow part through which blood flows) labeled. Surrounding the capillary are various cells and extracellular structures, which could be part of the tissue that supports the vessel, such as connective tissue.
Prompt: My model correctly predicts the answer "cardiovascular". It focuses on the hollow passage in the middle of the figure and on the hollow part on the right of the figure. Give a possible explanation.
GPT-4: Your model likely predicts "cardiovascular" because it recognizes the hollow passage as a capillary lumen, a characteristic feature of the cardiovascular system. It also may identify the surrounding tissue and its arrangement as typical of this system, possibly using visual patterns it has learned during training to make this association. If there's text within the image, like "Capillary Lumen," the model could be using that as a contextual clue to bolster its prediction.

TABLE 7.1: Chatbot Generated Response to Example 1: Cardiovascular.

7.2.2 Example 2: Endocrine System

In figure 7.2, the model highlights the terms "is", "this", "part" and "in" with positive attributions, indicating that the question concerns the location. This underscores the identification of the question type as pertaining to location. Visual attributions are directed towards the bright spots in the image, pinpointing areas of interest relevant to the prediction. Textual elements impact the model's prediction, with a contribution score of 20.45, which greatly exceeds the visual contribution of 1.09. The negative attribution assigned to the word "context" and the greater emphasis on textual attributions follow the pattern observed in the previous example, reflecting the model's consistent approach to prioritizing textual information over visual cues.



Legend: 🟥 Negative ☐ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
endocrine system	**endocrine system (0.88)**	**endocrine system**	**20.45**	question: Where is this part in the figure? context:~~answer~~

```
Text Contributions:  20.450109481811523
Image Contributions:  1.0948104858398438
Total Contribution:  21.544919967651367
```

FIGURE 7.2: Example 2: Endocrine from Test Set.

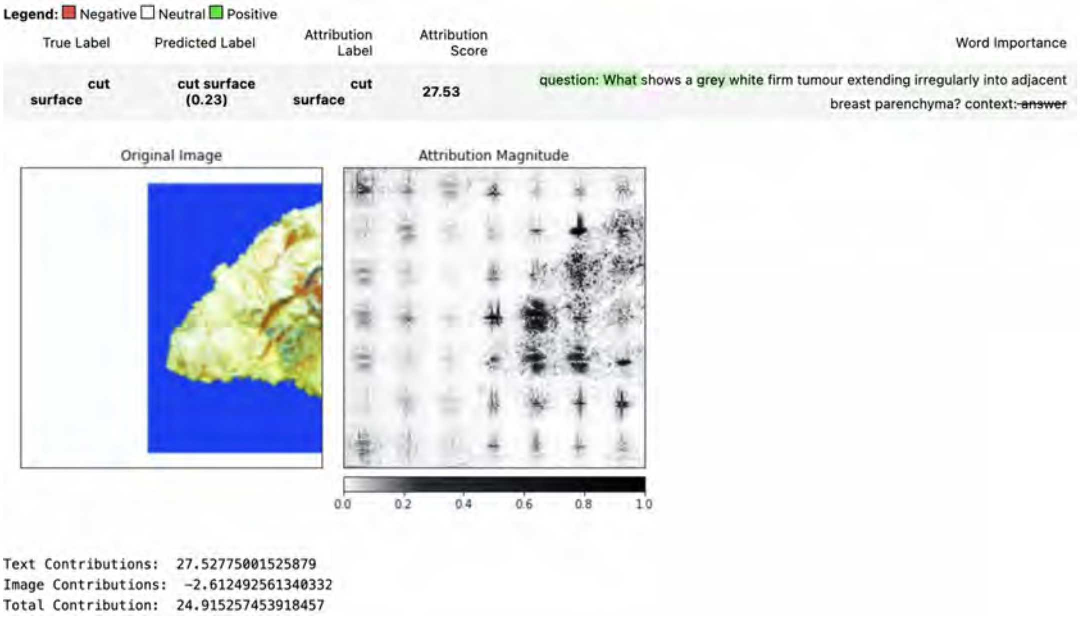
Table 7.2 presents responses generated by GPT-4 to specific prompts. Initially, GPT-4's inability to provide an answer from only the image and question underscores the challenges inherent in pathology-focused VQA. Subsequently, incorporating the correct "endocrine system" label and focusing on the bright spots gives GPT-4 sufficient context to offer a possible explanation.

Prompt: Where is this part in the figure?
GPT-4: The image you've provided appears to be a microscopic view of a tissue sample, likely from a histological examination. However, without additional context or information regarding the specific type of tissue and staining technique used, I cannot provide a precise location or diagnosis for the tissue in the figure. Histological images like this one are typically used to study the microscopic anatomy of biological tissues. The pattern of staining, the morphology of the cells, and the structure of the tissue all provide clues to its identity. This could be a section from any number of organs or tissues within the body, as many tissues can appear similar at this magnification and without specific markers.
Prompt: My model correctly predicts the answer "endocrine system". It focuses on the bright spots on the figure. Give a possible explanation.
GPT-4: The bright spots in the image likely represent hormone-secreting cells of endocrine tissue. These cells often appear lighter on stained histological slides because they contain secretory granules or vesicles that don't absorb the stain as much. This is common in endocrine glands like the pancreas, thyroid, and others, where such cells stand out against the darker tissue background. If your model is trained to recognize these patterns, it may use the contrast between the bright and dark areas to identify the tissue as part of the endocrine system.

TABLE 7.2: Chatbot Generated Response to Example 2: Endocrine System.

7.2.3 Example 3: Cut Surface

In Figure 7.3, the model focuses on the terms "What" and "grey" prominently, while "white," "irregularly," and "parenchyma" receive lighter positive attributions. This indicates that the model is pinpointing specific keywords relevant to diagnosing. Visual attributions target the white areas within the object on the right against the blue background, identifying these spots as important for diagnosis. The textual input greatly influences the model's prediction, with a contribution score of 27.52.



Legend: 🟥 Negative ☐ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
cut surface	cut surface (0.23)	cut surface	27.53	question: What shows a grey white firm tumour extending irregularly into adjacent breast parenchyma? context:~~answer~~

```
Text Contributions:  27.52775001525879
Image Contributions:  -2.612492561340332
Total Contribution:  24.915257453918457
```

FIGURE 7.3: Example 3: Cut Surface from Test Set.

Table 7.3 showcases the responses generated by GPT-4 to various prompts. Initially, without the target answer, GPT-4's response to the image and question highlights "dissected tissue," aligning closely with the "cut surface" descriptor. Upon integrating the target answer "cut surface" and noting Pathological-Llama's visual focus, GPT-4 elaborates on the rationale behind such a prediction, suggesting the model's adeptness at identifying the reflective qualities typical of freshly cut tissues.

Prompt: What shows a grey white firm tumour extending irregularly into adjacent breast parenchyma?
GPT-4: The image you've provided appears to depict a pathological specimen, likely from a medical scan or a photograph of a dissected tissue, which shows a grey-white firm tumor with irregular extensions into the surrounding breast parenchyma. Such a description is typical of certain types of breast tumors, possibly malignant ones like invasive ductal carcinoma, which is known to have these characteristics. In a clinical setting, the diagnosis would be confirmed through a combination of imaging studies, like mammograms or MRIs, and a biopsy with subsequent histopathological examination. The firmness and irregularity of the growth, as well as its infiltration into adjacent tissues, are hallmarks of malignant lesions, but definitive diagnosis always requires a pathological assessment.
Prompt: My model correctly predicts the answer "cut surface". It focuses on the brightest part of the tumor on the right of the figure. Give a possible explanation.
GPT-4: Your model's prediction of "cut surface" when focusing on the brightest part of the tumor could be due to its training on recognizing the reflective properties of tissues in pathological specimens. The brightness often indicates a freshly cut surface that reflects more light. The model might combine techniques like contrast detection, texture and color segmentation, and edge detection to differentiate the cut surface from the rest of the tissue. Additionally, the patterns recognized by the model are likely informed by its training data, which may have included many examples where brighter areas correlate with cut surfaces of tumors. The accuracy of such predictions depends on the model's ability to integrate these features effectively, avoiding reliance on brightness alone, which could lead to misinterpretation in varied scenarios.

TABLE 7.3: Chatbot Generated Response to Example 3: Cut Surface.

Chapter 8

Discussion

8.1 Discussion of Findings

This chapter discusses the findings from the experimentation on Pathological-Llama, particularly emphasizing how chat-based configurations, model capacity, visual prefix length, and generative approaches influence medical VQA systems' performance.

The comparison between Tiny-Llama and Tiny-Llama-Chat models as causal language models for Pathological-Llama highlighted the importance of task-specific architectural features. Tiny-Llama-Chat's performance, with a 5.6% increase in validation BERT score (from 0.521 to 0.550) and a 12.7% improvement in validation F1 score (from 0.322 to 0.363), confirms the benefits of integrating chat-based features for VQA tasks. This concrete evidence supports the hypothesis that models designed with task-specific considerations can enhance VQA performance.

Further examination of the performance differences between Tiny-Llama and Llama2 highlighted that Llama2, with its substantially higher parameter count, demonstrated visible performance improvements. In detail, Llama2 achieved a validation BERT score that was 6.5% higher, reaching 0.585, and experienced a 10.6% increase in its validation F1 score, achieving 0.404, all at a learning rate of 1e-5. This underlines the efficacy of larger models in deciphering complex data patterns when hyperparameters are optimized to leverage the model's extensive capacity. However, it is noteworthy that despite Llama2 having over six times more parameters than Tiny-Llama, the optimally configured Tiny-Llama approaches Llama2's performance closely, with a BERT score of 0.573 and an F1 score of 0.403 in the test set. This marginal difference of only 3% and 4% respectively in favor of Llama2 showcases Tiny-Llama's exceptional efficiency and effectiveness, making it a commendable alternative for tasks demanding high computational efficiency alongside robust performance.

Investigating visual prefix length revealed that longer prefixes generally enhance model performance. For Tiny-Llama, extending the prefix to 12 led to a 14.5% increase in validation BERT score and a 20.2% improvement in validation F1 score in comparison to its counterpart setup of prefix 2, indicating the value of providing the model with an extended visual context for improved data integration and interpretation. Similarly, Llama2 exhibited optimal performance with a prefix length of 6, pointing to the necessity of finding an ideal balance for visual prefix length that maximizes performance within computational constraints.

The study's exploration into the use of generative approaches for close-ended questions showed effectiveness over classification models. The generative Llama2 model outperformed its classification counterpart in testing accuracy and F1 scores. This distinction illustrates the potential of generative models to provide a richer, more nuanced understanding of medical VQA tasks.

The application of the Integrated Gradients method, alongside GPT-4's analysis, offered profound insights into the Pathological-Llama model's reasoning process. This innovative approach to explainability not only validated the model's predictions but also established a novel paradigm for interpreting AI decisions in complex scenarios, thereby enhancing model transparency and trustworthiness.

A notable challenge identified in this study is the lack of standardized evaluation metrics for generative text in the medical domain. These findings advocate for a multifaceted evaluation approach that include semantic similarity and medical correctness. Addressing hardware limitations also influenced many experimental decisions and highlighting the importance of developing lightweight models that balance efficiency with computational feasibility.

8.2 Comparison to existing State-of-the-Art Models

In the PathVQA task, comparing Pathological-Llama to other state-of-the-art models presents a challenge, primarily due to inconsistencies in evaluation metrics, dataset splits, and dataset versions across related work. Notably, some models employ different metrics, or when similar metrics are used, the dataset split varies (e.g., some use a 0.5, 0.3, 0.2 split [21]) or rely on earlier versions of the dataset. These discrepancies underscore the necessity for standardized benchmarks in tasks of this nature to enable fair and meaningful comparisons.

From the original papers on PathVQA, various models report diverse performance metrics:

- PathVQA Method 1 [1] reported an F1 score of 0.24,
- GPT2-xl [31] boasted an F1 score of 0.58 and a BERT score of 0.78,
- LLaVA-Med [33] documented a Recall of 39.60.

However, an attempt to replicate the results reported by GPT2-xl during this study encountered difficulties. Despite utilizing the code and documented parameters provided by the authors, with missing parameters supplemented from the default values in their codebase, the results did not align with those initially reported. The replicated results yielded a BERT score of 0.563 and an F1 score of 0.223 at epoch 60 on the test set split used in this study.

Despite inconsistencies in evaluation standards making direct comparisons difficult, Pathological-Llama's performance, with a BERT score of 0.591 and an F1 score of 0.419, positions it competitively among current medical VQA models. This underscores its potential as a valuable diagnostic tool. The challenges in replicating results from other studies underscore the need for transparent documentation and uniform reporting standards to improve reproducibility and facilitate progress in the field. Establishing such standards would not only validate the advancements made by models like Pathological-Llama but also push forward the capabilities of medical VQA technologies.

Chapter 9

Conclusion and Future Research

This thesis conducted an in-depth investigation into the Pathological-Llama model, examining its performance across diverse configurations in the field of medical VQA.

9.1 Key Findings

This study's key findings are:

- **Task-Specific Architecture:** Incorporating causal language models that are constructed for chat/dialogue resulted in performance improvements, highlighting the importance of aligning model architecture with the specific demands of VQA tasks.

- **Model Capacity:** Larger models, exemplified by Llama2, demonstrated superior performance when optimally tuned with appropriate hyperparameters, showcasing their enhanced capability to process complex data patterns.

- **Efficiency vs. Capacity:** Despite Llama2 possessing over six times the parameters of Tiny-Llama, the latter demonstrated remarkably close performance. This highlights Tiny-Llama's efficiency, positioning it as an effective model for scenarios where computational resources are limited.

- **Visual Prefix Length:** Extending the visual prefix length generally led to better model performance, indicating that a more comprehensive visual context leads to improved data integration and interpretation.

- **Generative Approaches:** Generative models proved more effective than classification models for close-ended questions, offering a nuanced and contextually aware approach to VQA.

- **Explainability:** The application of the Integrated Gradients method, complemented by GPT-4's analysis, provided valuable insights into the model's reasoning process, enhancing transparency and trust in AI-driven diagnostics.

9.2 Implications for Future Research

The study's implications extend beyond the immediate findings, suggesting several avenues for future research:

- Further exploration of task-specific architectures and model capacities could uncover additional enhancements to VQA performance.
- Investigating the impact of even longer visual prefixes, given computational resources allow, could provide deeper insights into optimal data integration strategies.
- Expanding the application of generative models to other types of questions and datasets may reveal broader benefits and limitations within medical VQA systems.
- Developing a standardized set of evaluation metrics for generative text in the medical domain could facilitate more refined assessments of model performance.
- Addressing computational constraints through the development of more efficient model architectures could broaden the applicability of findings and facilitate the adoption of VQA systems across various medical settings.

9.3 Conclusion

This thesis contributes to the understanding and optimization of medical VQA systems, laying the groundwork for future advancements in AI-driven healthcare diagnostics. By navigating the complexities of model design and functionality within computational constraints, this study has highlighted the potential of advanced AI applications in improving medical diagnostics. The insights concluded from this study not only enhance the academic discourse on medical AI but also hold promise for practical applications, aiming to improve patient care and outcomes through the integration of AI technologies.

Appendix A

Extended Experiments

The comparison is between TinyLlama-1.1B-Chat-v0.1 (V1), the initial chat-based model, and its latest iteration, TinyLlama-1.1B-Chat-v0.4 (V4). Both versions were fine-tuned on the PathVQA dataset, applying a learning rate of 1e-4 across 60 epochs and a visual prefix size of 6. Performance metrics at epoch 60 are summarized in table A.1. The comparative evaluation reveals that TinyLlama V1 exhibits superior performance across several key metrics relative to TinyLlama V4. Notable improvements include reduced training loss and enhanced scores in BLEU, BERT, and F1 metrics, as visualized in A.1 and A.2. This enhanced performance of TinyLlama V1 may stem from its more effective generalization to the PathVQA dataset, suggesting an advantage in fine-tuning on the PathVQA task with a less extensively pre-trained model on a broader dataset.

Model	**Train Loss**	**Val Loss**	**Val BLEU**	**Val BERT**	**Val F1**
Tiny-Llama V1	0.0626	0.7747	0.5071	0.550	0.363
Tiny-Llama V4	0.1015	0.7414	0.5047	0.539	0.349

TABLE A.1: Performance Comparison of Tiny-Llama V1 and Tiny-Llama V4.

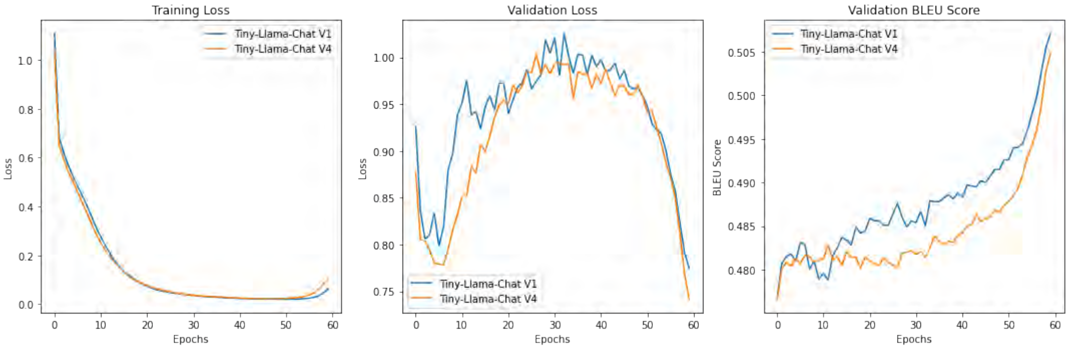


FIGURE A.1: Train and Validation Performance of Tiny-Llama V1 and Tiny-Llama V4 over 60 Epochs.

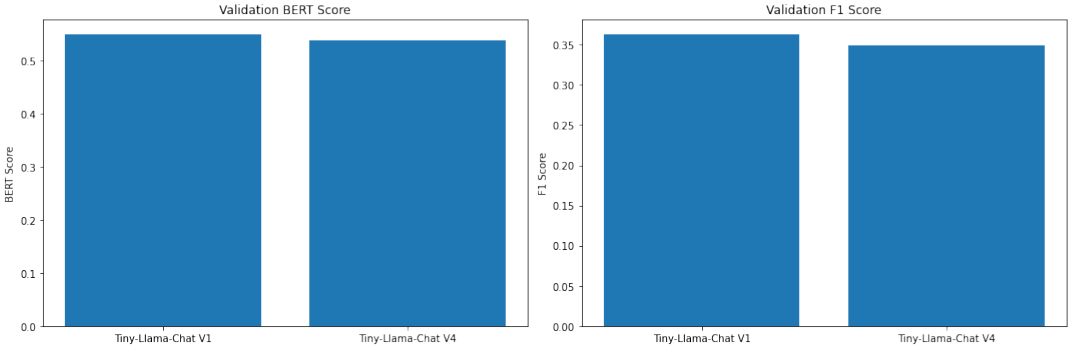


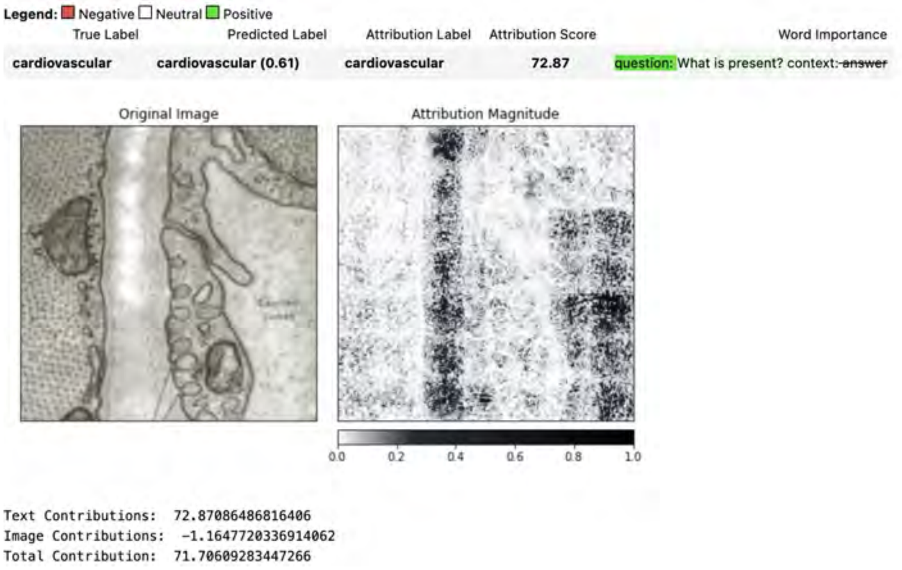
FIGURE A.2: Validation BERT and F1 score at Epoch 60.

Appendix B

Extended Explainability Examples

B.1 Integrated Gradients Applied to Pathological-Llama with Llama2

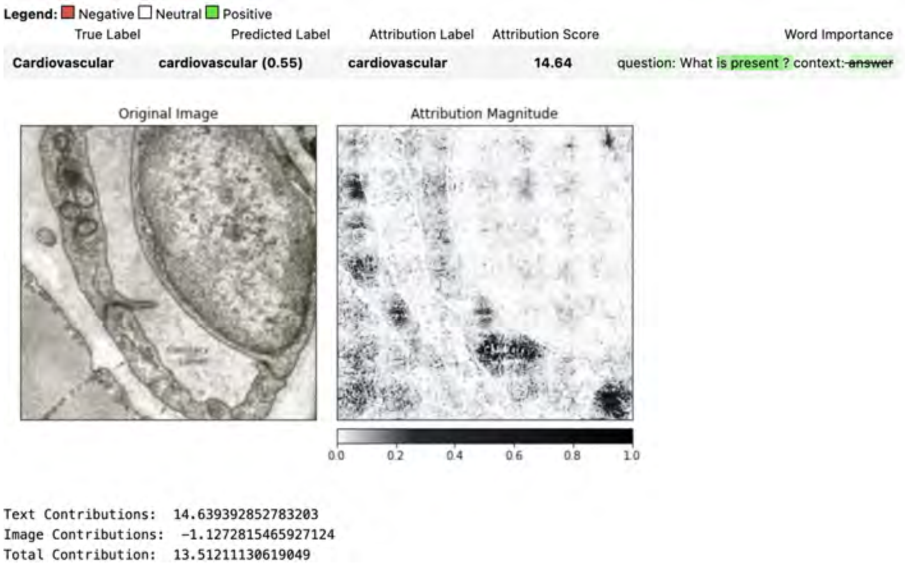
The figures presented below illustrate the application of Integrated Gradients (IG) to the Pathological-Llama model, which incorporates Llama2 as the causal language model, configured according to the specifications outlined in 6.2. The implementation follows the details described in 7, with the exception that the attribution method employs 10 steps.



True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
cardiovascular	**cardiovascular (0.61)**	**cardiovascular**	**72.87**	question: What is present? context:~~answer~~

```
Text Contributions:  72.87086486816406
Image Contributions:  -1.1647720336914062
Total Contribution:  71.70609283447266
```

FIGURE B.1: Example 1: Cardiovascular from Test Set.

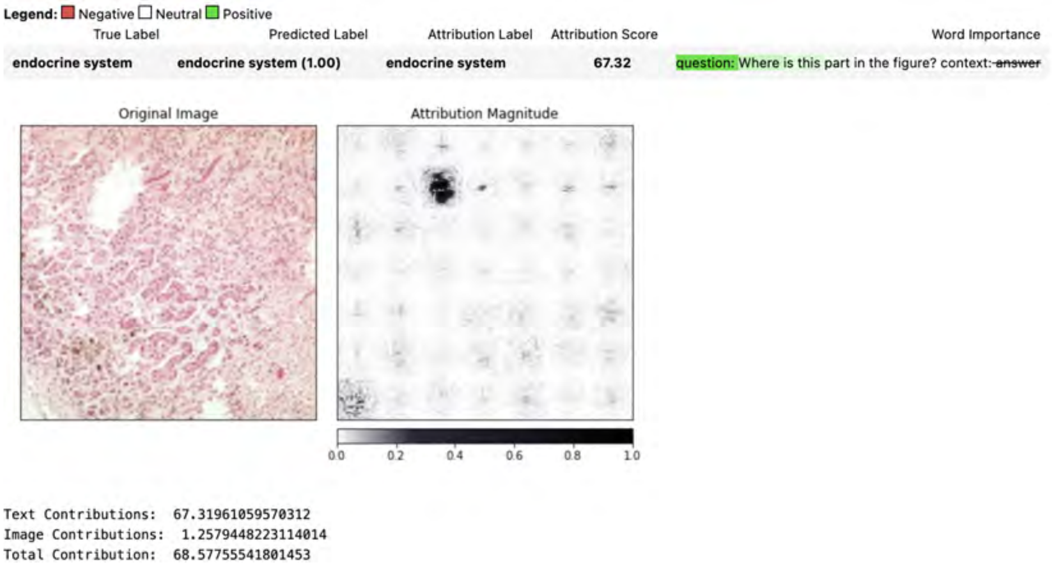


Legend: 🟥 Negative ☐ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
Cardiovascular	**cardiovascular (0.55)**	**cardiovascular**	**14.64**	question: What is present ? context: ~~answer~~

```
Text Contributions:  14.639392852783203
Image Contributions:  -1.1272815465927124
Total Contribution:  13.51211130619049
```

FIGURE B.2: Example 2: Cardiovascular from Test Set.

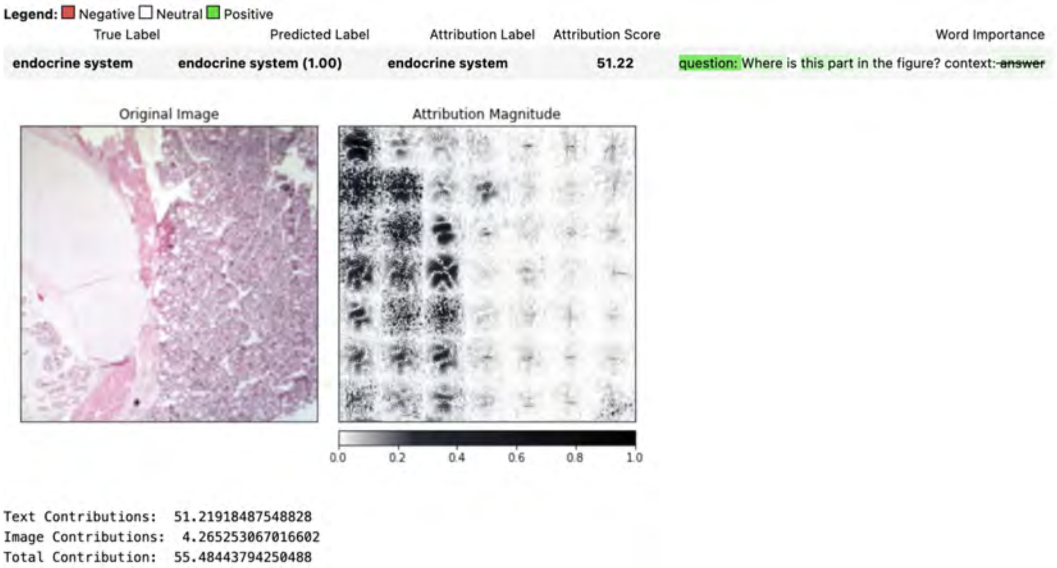


Legend: 🟥 Negative ☐ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
endocrine system	**endocrine system (1.00)**	**endocrine system**	**67.32**	question: Where is this part in the figure? context: ~~answer~~

```
Text Contributions:  67.31961059570312
Image Contributions:  1.2579448223114014
Total Contribution:  68.57755541801453
```

FIGURE B.3: Example 3: Endocrine from Test Set.

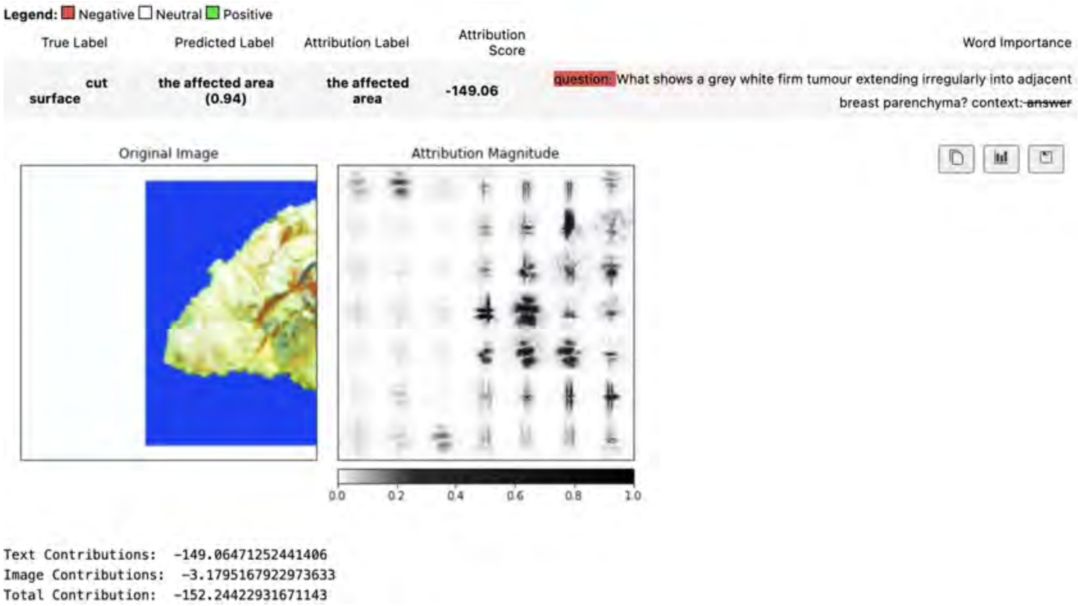


Legend: 🟥 Negative ☐ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
endocrine system	**endocrine system (1.00)**	**endocrine system**	**51.22**	question: Where is this part in the figure? context: ~~answer~~

```
Text Contributions:  51.21918487548828
Image Contributions:  4.265253067016602
Total Contribution:  55.48443794250488
```

FIGURE B.4: Example 4: Endocrine from Test Set.



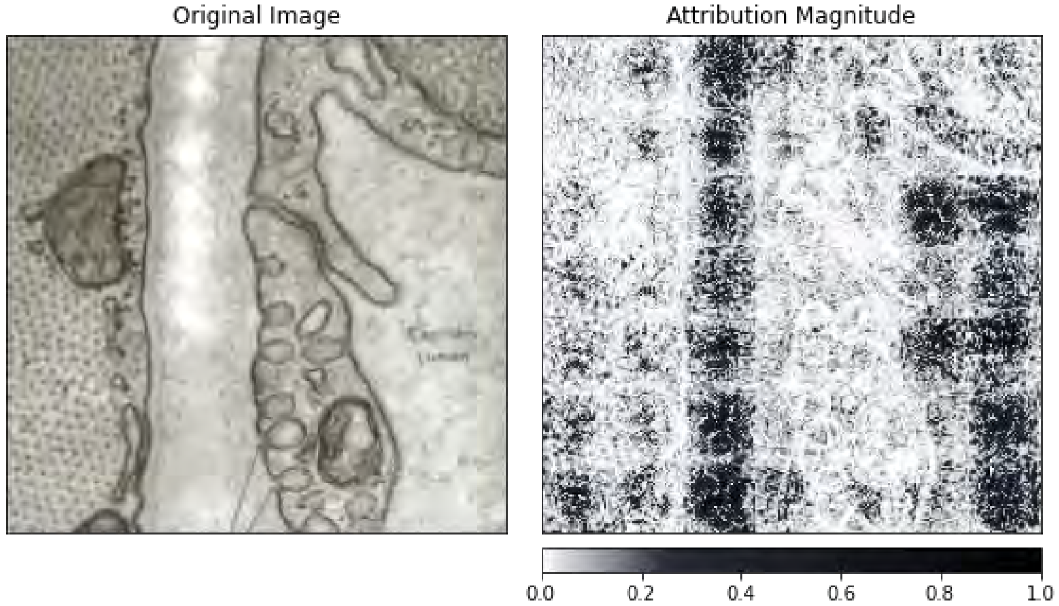
Legend: 🟥 Negative ☐ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
cut surface	**the affected area (0.94)**	**the affected area**	**-149.06**	question: What shows a grey white firm tumour extending irregularly into adjacent breast parenchyma? context: ~~answer~~

```
Text Contributions:  -149.06471252441406
Image Contributions:  -3.1795167922973633
Total Contribution:  -152.24422931671143
```

FIGURE B.5: Example 5: Cut Surface from Test Set.

B.2 Chatbot Generated Responses



Prompt: What is present?

GPT-4: The image you’ve uploaded appears to be an electron micrograph, which is a type of high-resolution photograph taken with an electron microscope. The details in the image are consistent with a cross-sectional view of biological tissues at the microscopic level. We can see what looks like a capillary (a small blood vessel) with its lumen (the hollow part through which blood flows) labeled. Surrounding the capillary are various cells and extracellular structures, which could be part of the tissue that supports the vessel, such as connective tissue.

Gemini: The image you sent is a black and white drawing of a biological cell. The specific type of cell is not labeled, but it has a nucleus, endoplasmic reticulum, Golgi apparatus, mitochondria, and lysosomes. These are all organelles, which are membrane-bound structures that perform specific functions within the cell. The text on the image says "Capillary" and "Lumen," which suggests that the cell is part of a capillary, which is a small blood vessel. The lumen is the hollow space inside the capillary.

Prompt: My model correctly predicts the answer "cardiovascular". It focuses on the hollow passage in the middle of the figure and on the hollow part on the right of the figure. Give a possible explanation.

GPT-4: Your model likely predicts "cardiovascular" because it recognizes the hollow passage as a capillary lumen, a characteristic feature of the cardiovascular system. It also may identify the surrounding tissue and its arrangement as typical of this system, possibly using visual patterns it has learned during training to make this association. If there’s text within the image, like "Capillary Lumen," the model could be using that as a contextual clue to bolster its prediction.

Gemini: Your model’s focus on hollow spaces in the image could explain its "cardiovascular" prediction. It might see them as blood vessel features or infer their role in blood flow, aligning with the cardiovascular system’s function.

TABLE B.1: Chatbot Generated Response Example Cardiovascular 1.

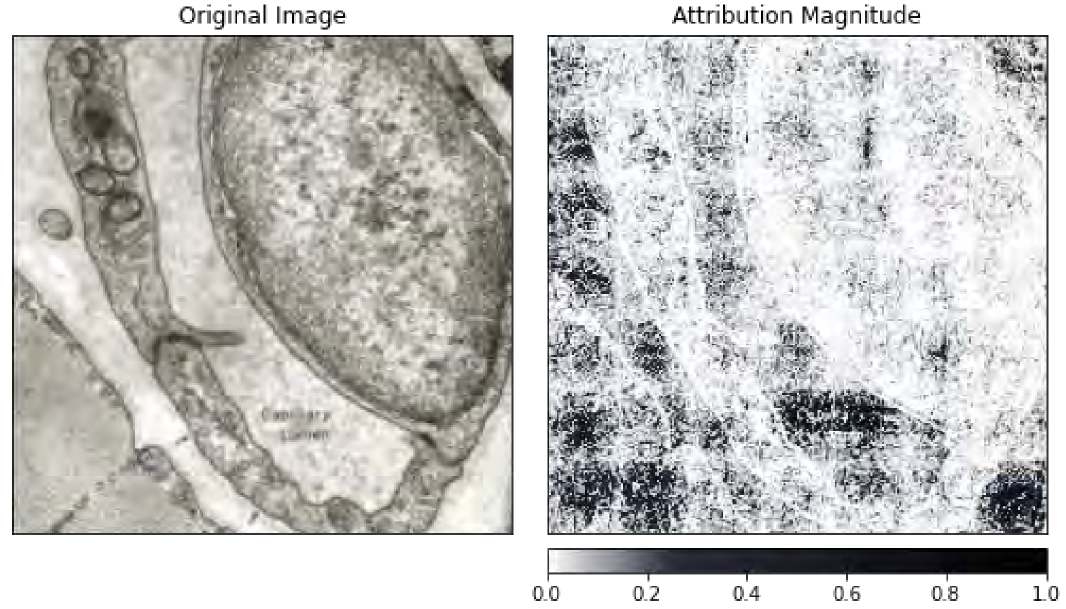
 Original Image Capillary Lumen Attribution Magnitude 0.0 0.2 0.4 0.6 0.8 1.0
Prompt: What is present ?
GPT-4: The image you've uploaded appears to be a transmission electron microscope (TEM) image of a biological sample. The large, oval structure labeled "Capillary Lumen" indicates that this is likely a cross-section of a capillary, a small blood vessel. Inside the capillary lumen, we typically expect to see blood cells and plasma. The surrounding cellular details are not clearly visible, but it's common to see endothelial cells lining the capillary, which play a crucial role in the exchange of substances between the blood and surrounding tissues. The other structures in the image are not labeled, but they might be other cell types or extracellular components depending on the tissue this sample was taken from.
Gemini: The image you sent appears to be a black and white micrograph of a neuron, a nerve cell. It has a large, irregular cell body that contains the nucleus. The nucleus is the round, light colored structure in the center of the cell body. It contains the cell's genetic material. Extending from the cell body are thin fibers called dendrites and a longer fiber called an axon. Dendrites receive signals from other neurons, while the axon transmits signals to other neurons. The text on the image says "Capillary," but this is incorrect. Capillaries are tiny blood vessels that deliver oxygen and nutrients to cells and remove waste products. The structure in the image is a neuron, not a capillary.
Prompt: My model correctly predicts the answer "cardiovascular". It focuses on the hollow passage on the left of the figure and on the part where "Capiliary Lumen" is written. Give a possible explanation.
GPT-4: Your model's correct prediction of "cardiovascular" could be due to its recognition of the labeled "Capillary Lumen," a term specific to the cardiovascular system, and its identification of the hollow structure typical of blood vessels. Additionally, if it was trained on similar images, it may have learned to associate certain morphological features with the cardiovascular system.
Gemini: The model predicted "cardiovascular" because the neuron's structure resembled a blood vessel, and the misleading "Capillary Lumen" label influenced its thinking. These instances help us improve models by recognizing misleading clues and avoiding similar misinterpretations.

TABLE B.2: Chatbot Generated Response Example Cardiovascular 2.

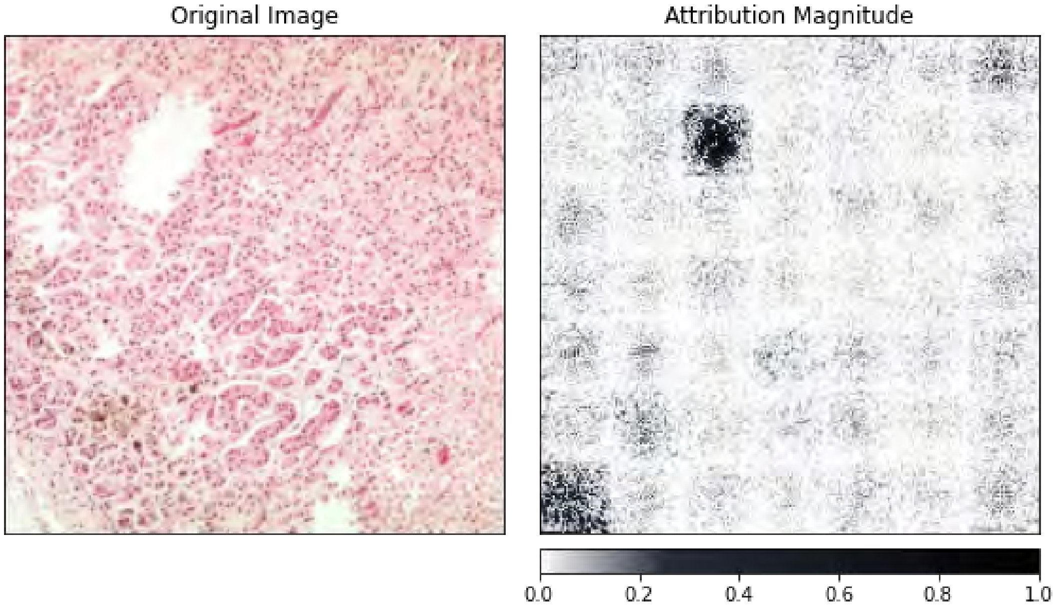
 Original Image Attribution Magnitude 0.0 0.2 0.4 0.6 0.8 1.0
Prompt: Where is this part in the figure?
GPT-4: The image you've provided appears to be a microscopic view of a tissue sample, likely from a histological examination. However, without additional context or information regarding the specific type of tissue and staining technique used, I cannot provide a precise location or diagnosis for the tissue in the figure. Histological images like this one are typically used to study the microscopic anatomy of biological tissues. The pattern of staining, the morphology of the cells, and the structure of the tissue all provide clues to its identity. This could be a section from any number of organs or tissues within the body, as many tissues can appear similar at this magnification and without specific markers.
Gemini: I can't tell you where the tissue sample in the image came from.
Prompt: My model correctly predicts the answer "endocrine system". It focuses on the bright spots on the figure. Give a possible explanation.
GPT-4: The bright spots in the image likely represent hormone-secreting cells of endocrine tissue. These cells often appear lighter on stained histological slides because they contain secretory granules or vesicles that don't absorb the stain as much. This is common in endocrine glands like the pancreas, thyroid, and others, where such cells stand out against the darker tissue background. If your model is trained to recognize these patterns, it may use the contrast between the bright and dark areas to identify the tissue as part of the endocrine system.
Gemini: The pink/red regions in your histological image, likely due to H&E staining, might be triggering your model's "endocrine system" prediction. These vibrant areas could reflect the cell characteristics of endocrine glands or highlight their distinct tissue architecture. Alternatively, your model's training data might have ingrained an association between brightness and the endocrine system.

TABLE B.3: Chatbot Generated Response Example Endocrine 1.

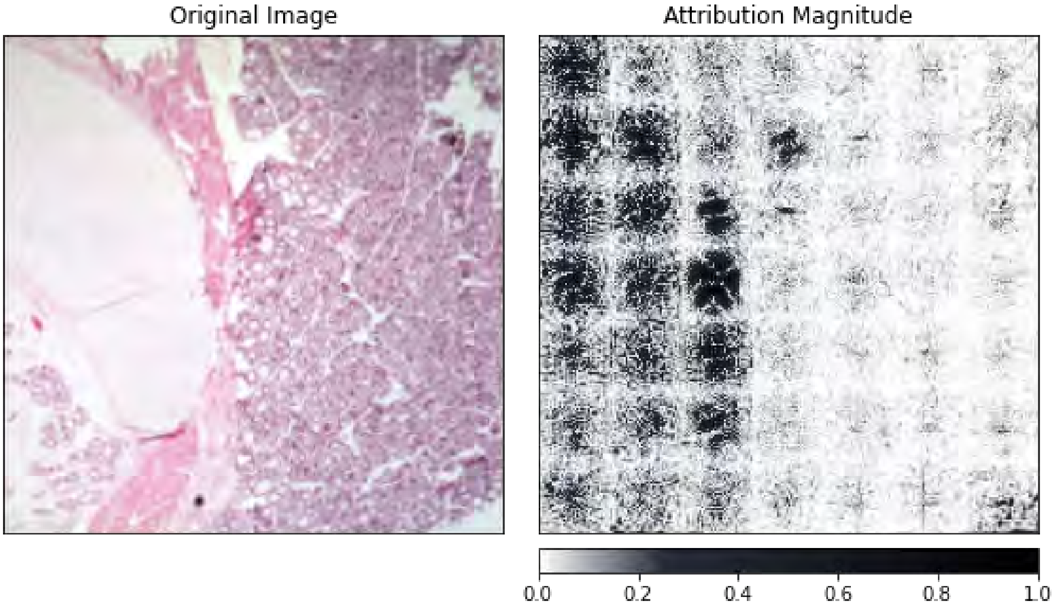
 Original Image Attribution Magnitude 0.0 0.2 0.4 0.6 0.8 1.0
Prompt: Where is this part in the figure?
GPT-4: The image you have provided appears to be a photomicrograph from a histological examination, most likely a section of tissue stained and viewed under a microscope. Unfortunately, without additional context or information, I cannot definitively determine the exact origin of the tissue. It could potentially be a section from an organ or a specific type of tissue from the human body or from an animal. Typically, the features that can be observed such as the arrangement of cells, the presence of any specialized structures, and the staining pattern can provide clues to its identity. For example, certain tissues are characterized by the presence of specific cell types (like adipose tissue with its large fat cells), unique structures (like the villi in the small intestine), or particular arrangements of cells (like the layered structure in skin).
Gemini: The part of the figure you sent is a tissue sample mounted on a microscope slide. It is difficult to say definitively what tissue it is, but it looks like it could be from the thyroid gland. The thyroid gland is located in the front of the neck, just below the Adam's apple. It is a butterfly-shaped gland that produces hormones that regulate metabolism. Microscope slides are used to hold samples of tissue or other materials for examination under a microscope. The tissue sample is typically mounted on the slide and then stained with dyes to make it easier to see the different cells and structures. The slide is then placed on the stage of the microscope and viewed through the lens.
Prompt: My model correctly predicts the answer "endocrine system". It focuses on the bright part on the left of the figure. Give a possible explanation.
GPT-4: If your model identifies the tissue as part of the endocrine system, it might be focusing on the bright area to the left because it resembles the features of endocrine glands, which are usually organized into distinct structures like cords or follicles and are highly vascularized for hormone secretion into the bloodstream. These glands don't have ducts, and the contrast in the image could be highlighting the dense cellular regions that are typical of endocrine tissues, next to lighter areas that could be connective tissue or spaces around blood vessels. The precise identification would still depend on the stain used, magnification, and the sample's origin.
Gemini: The bright part on the left likely represents thyroid tissue due to the presence of characteristic features like colloid-filled follicles, follicular epithelium, and parafollicular cells. The distinct pink color of the colloid might have attracted the model's attention.

TABLE B.4: Chatbot Generated Response Example Endocrine 2.

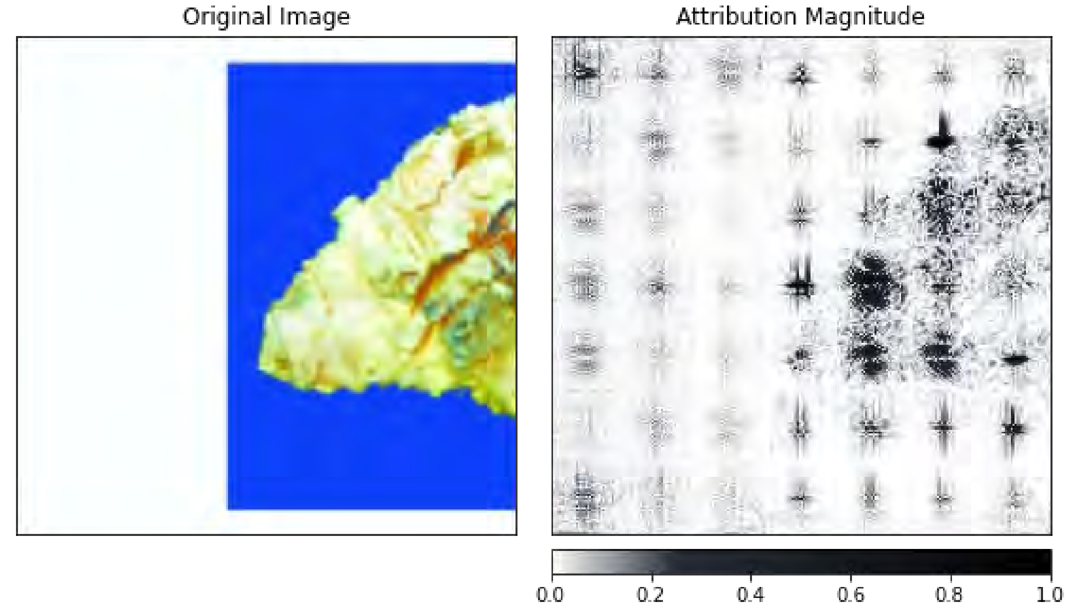
 Original Image Attribution Magnitude 0.0 0.2 0.4 0.6 0.8 1.0
Prompt: What shows a grey white firm tumour extending irregularly into adjacent breast parenchyma?
GPT-4: The image you've provided appears to depict a pathological specimen, likely from a medical scan or a photograph of a dissected tissue, which shows a grey-white firm tumor with irregular extensions into the surrounding breast parenchyma. Such a description is typical of certain types of breast tumors, possibly malignant ones like invasive ductal carcinoma, which is known to have these characteristics. In a clinical setting, the diagnosis would be confirmed through a combination of imaging studies, like mammograms or MRIs, and a biopsy with subsequent histopathological examination. The firmness and irregularity of the growth, as well as its infiltration into adjacent tissues, are hallmarks of malignant lesions, but definitive diagnosis always requires a pathological assessment.
Gemini: I am unable to diagnose medical conditions or provide medical advice.
Prompt: My model correctly predicts the answer "cut surface". It focuses on the brightest part of the tumor on the right of the figure. Give a possible explanation.
GPT-4: Your model's prediction of "cut surface" when focusing on the brightest part of the tumor could be due to its training on recognizing the reflective properties of tissues in pathological specimens. The brightness often indicates a freshly cut surface that reflects more light. The model might combine techniques like contrast detection, texture and color segmentation, and edge detection to differentiate the cut surface from the rest of the tissue. Additionally, the patterns recognized by the model are likely informed by its training data, which may have included many examples where brighter areas correlate with cut surfaces of tumors. The accuracy of such predictions depends on the model's ability to integrate these features effectively, avoiding reliance on brightness alone, which could lead to misinterpretation in varied scenarios.
Gemini: Your model's prediction of a "cut surface" for the bright tumor region is reasonable for several reasons. Brighter areas in various imaging techniques often indicate cut surfaces due to differing light reflection/absorption. The location on the edge, resembling a knife cut shape, further supports this.

TABLE B.5: Chatbot Generated Response Example Tumor.

Bibliography

[1] Xuehai He et al. “PathVQA: 30000+ Questions for Medical Visual Question Answering”. In: *CoRR* abs/2003.10286 (2020). arXiv: `2003.10286`. URL: `https://arxiv.org/abs/2003.10286`.

[2] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. *Axiomatic Attribution for Deep Networks*. 2017. arXiv: `1703.01365 [cs.LG]`.

[3] OpenAI et al. *GPT-4 Technical Report*. 2023. arXiv: `2303.08774 [cs.CL]`.

[4] Keiron O’Shea and Ryan Nash. *An Introduction to Convolutional Neural Networks*. 2015. arXiv: `1511.08458 [cs.NE]`.

[5] Peng Xu, Xiatian Zhu, and David A. Clifton. *Multimodal Learning with Transformers: A Survey*. 2023. arXiv: `2206.06488 [cs.CV]`.

[6] Abien Fred Agarap. *Deep Learning using Rectified Linear Units (ReLU)*. 2019. arXiv: `1803.08375 [cs.NE]`.

[7] Dan Hendrycks and Kevin Gimpel. *Gaussian Error Linear Units (GELUs)*. 2023. arXiv: `1606.08415 [cs.LG]`.

[8] Liunian Harold Li et al. *VisualBERT: A Simple and Performant Baseline for Vision and Language*. 2019. arXiv: `1908.03557 [cs.CV]`.

[9] Gen Li et al. *Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training*. 2019. arXiv: `1908.06066 [cs.CV]`.

[10] Weijie Su et al. *VL-BERT: Pre-training of Generic Visual-Linguistic Representations*. 2020. arXiv: `1908.08530 [cs.CV]`.

[11] Junyang Lin et al. *InterBERT: Vision-and-Language Interaction for Multi-modal Pre-training*. 2021. arXiv: `2003.13198 [cs.CL]`.

[12] Di Qi et al. *ImageBERT: Cross-modal Pre-training with Large-scale Weak-supervised Image-Text Data*. 2020. arXiv: `2001.07966 [cs.CV]`.

[13] Peng Xu, Xiatian Zhu, and David A. Clifton. *Multimodal Learning with Transformers: A Survey*. 2023. arXiv: `2206.06488 [cs.CV]`.

[14] Zhihong Lin et al. “Medical Visual Question Answering: A Survey”. In: *CoRR* abs/2111.10056 (2021). arXiv: `2111.10056`. URL: `https://arxiv.org/abs/2111.10056`.

[15] Prashant Shrestha et al. *Medical Vision Language Pretraining: A survey*. 2023. arXiv: `2312.06224 [cs.CV]`.

[16] Gaël Chételat. “Multimodal neuroimaging in Alzheimer’s disease: Early diagnosis, physiopathological mechanisms, and impact of lifestyle”. In: *J. Alzheimers. Dis.* 64.s1 (June 2018), S199–S211.

[17] *XxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*. Vienna, Austria: Springer-Verlag, 2020. ISBN: 978-3-031-04082-5.

[18] Aditya Khosla et al. “Novel Dataset for Fine-Grained Image Categorization”. In: *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*. Colorado Springs, CO, June 2011.

[19] Jürgen Amann, Alessandro Blasimme, Effy Vayena, et al. “Explainability for artificial intelligence in healthcare: a multidisciplinary perspective”. In: *BMC Medical Informatics and Decision Making* 20.1 (2020), p. 310. DOI: `10.1186/s12911-020-01332-6`. URL: `https://doi.org/10.1186/s12911-020-01332-6`.

[20] Zhihong Lin et al. “Medical visual question answering: A survey”. In: *Artificial Intelligence in Medicine* 143 (Sept. 2023), p. 102611. ISSN: 0933-3657. DOI: `10.1016/j.artmed.2023.102611`. URL: `http://dx.doi.org/10.1016/j.artmed.2023.102611`.

[21] Pengfei Li et al. *Self-supervised vision-language pretraining for Medical visual question answering*. 2022. arXiv: `2211.13594 [cs.CV]`.

[22] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: `1409.1556 [cs.CV]`.

[23] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *CoRR* abs/1512.03385 (2015). arXiv: `1512.03385`. URL: `http://arxiv.org/abs/1512.03385`.

[24] Ralf C. Staudemeyer and Eric Rothstein Morris. “Understanding LSTM - a tutorial into Long Short-Term Memory Recurrent Neural Networks”. In: *CoRR* abs/1909.09586 (2019). arXiv: `1909.09586`. URL: `http://arxiv.org/abs/1909.09586`.

[25] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: `1810.04805 [cs.CL]`.

[26] Binh D. Nguyen et al. *Overcoming Data Limitation in Medical Visual Question Answering*. 2019. arXiv: `1909.11867 [cs.CV]`.

[27] Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. *Does CLIP Benefit Visual Question Answering in the Medical Domain as Much as it Does in the General Domain?* 2021. arXiv: `2112.13906 [cs.CV]`.

[28] Junyoung Chung et al. *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling*. 2014. arXiv: `1412.3555 [cs.NE]`.

[29] Shaoqing Ren et al. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. 2016. arXiv: `1506.01497 [cs.CV]`.

[30] Kai Zhang et al. *BiomedGPT: A Unified and Generalist Biomedical Generative Pre-trained Transformer for Vision, Language, and Multimodal Tasks*. 2024. arXiv: `2305.17100 [cs.CL]`.

[31] Tom van Sonsbeek et al. *Open-Ended Medical Visual Question Answering Through Prefix Tuning of Language Models*. 2023. arXiv: `2303.05977 [cs.CV]`.

[32] Gen Luo et al. *Cheap and Quick: Efficient Vision-Language Instruction Tuning for Large Language Models*. 2023. arXiv: `2305.15023 [cs.CV]`.

[33] Chunyuan Li et al. *LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day*. 2023. arXiv: `2306.00890 [cs.CV]`.

[34] *Home — ncbi.nlm.nih.gov*. `https://www.ncbi.nlm.nih.gov/pmc/`. [Accessed 15-02-2024].

[35] Kishore Papineni et al. “Bleu: a Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Ed. by Pierre Isabelle, Eugene Charniak, and Dekang Lin. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318. DOI: `10.3115/1073083.1073135`. URL: `https://aclanthology.org/P02-1040`.

[36] Kavita Ganesan. *ROUGE 2.0: Updated and Improved Measures for Evaluation of Summarization Tasks*. 2018. arXiv: `1803.01937 [cs.IR]`.

[37] Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. *BLEURT: Learning Robust Metrics for Text Generation*. 2020. arXiv: `2004.04696 [cs.CL]`.

[38] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. *Evaluation of Text Generation: A Survey*. 2021. arXiv: 2006.14799 [cs.CL].

[39] Thomas Scialom et al. *QuestEval: Summarization Asks for Fact-based Evaluation*. 2021. arXiv: 2103.12693 [cs.CL].

[40] Qiu-Yue Zhong et al. *Medical Concept Representation Learning from Claims Data and Application to Health Plan Payment Risk Adjustment*. 2019. arXiv: 1907.06600 [cs.LG].

[41] Gaetano Geck et al. *Parallel-Correctness and Containment for Conjunctive Queries with Union and Negation*. 2015. arXiv: 1512.06246 [cs.DB].

[42] Greg M. Silverman et al. *An Empirical Study of UMLS Concept Extraction from Clinical Notes using Boolean Combination Ensembles*. 2021. arXiv: 2108.02255 [cs.CL].

[43] Rui Yang et al. *Integrating UMLS Knowledge into Large Language Models for Medical Question Answering*. 2023. arXiv: 2310.02778 [cs.CL].

[44] D. M. W. Powers. "Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation". In: *Journal of Machine Learning Technologies* 2.1 (2011), pp. 37–63.

[45] H. Mohan. *Textbook of Pathology*. Jaypee Brothers, 2000. ISBN: 9788171797837. URL: https://books.google.ch/books?id=znoZnQAACAAJ.

[46] V. Kumar et al. *Robbins Basic Pathology*. Robbins Pathology. Elsevier Health Sciences, 2007. ISBN: 9781437700664. URL: https://books.google.ch/books?id=-keXQ6LaXVIC.

[47] Pathology Education Instructional Resource. *Main Page — Pathology Education Instructional Resource*. [Online; accessed 8-January-2024]. 2020. URL: %5Curl%7Bhttps://peir.path.uab.edu/index.php?title=Main_Page&oldid=3421%7D.

[48] Alec Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *CoRR* abs/2103.00020 (2021). arXiv: 2103.00020. URL: https://arxiv.org/abs/2103.00020.

[49] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: (2018). URL: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

[50] Renqian Luo et al. "BioGPT: generative pre-trained transformer for biomedical text generation and mining". In: *Briefings in Bioinformatics* 23.6 (Sept. 2022). ISSN: 1477-4054. DOI: 10.1093/bib/bbac409. URL: http://dx.doi.org/10.1093/bib/bbac409.

[51] Bernal Jiménez Gutiérrez, Huan Sun, and Yu Su. *Biomedical Language Models are Robust to Sub-optimal Tokenization*. 2023. arXiv: 2306.17649 [cs.CL].

[52] Hugo Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. 2023. arXiv: 2307.09288 [cs.CL].

[53] Peiyuan Zhang et al. *TinyLlama: An Open-Source Small Language Model*. 2024. arXiv: 2401.02385 [cs.CL].

[54] Edward J. Hu et al. "LoRA: Low-Rank Adaptation of Large Language Models". In: *CoRR* abs/2106.09685 (2021). arXiv: 2106.09685. URL: https://arxiv.org/abs/2106.09685.

[55] Yuu Jinnai, Tetsuro Morimura, and Ukyo Honda. *On the Depth between Beam Search and Exhaustive Search for Text Generation*. 2023. arXiv: 2308.13696 [cs.CL].

[56] Anqi Mao, Mehryar Mohri, and Yutao Zhong. *Cross-Entropy Loss Functions: Theoretical Analysis and Applications*. 2023. arXiv: 2304.07288 [cs.LG].

[57] Tianyi Zhang et al. “BERTScore: Evaluating Text Generation with BERT”. In: *CoRR* abs/1904.09675 (2019). arXiv: `1904.09675`. URL: `http://arxiv.org/abs/1904.09675`.

[58] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: `1412.6980 [cs.LG]`.

[59] Taku Kudo and John Richardson. *SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing*. 2018. arXiv: `1808.06226 [cs.CL]`.

[60] Aibek Bekbayev et al. *The Poison of Alignment*. 2023. arXiv: `2308.13449 [cs.CL]`.

[61] Ludovick Bouthat. *On the monotonicity of left and right Riemann sums*. 2023. arXiv: `2311.01208 [math.CA]`.

[62] Adam Paszke et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. 2019. arXiv: `1912.01703 [cs.LG]`.

[63] Narine Kokhlikyan et al. “Captum: A unified and generic model interpretability library for PyTorch”. In: *CoRR* abs/2009.07896 (2020). arXiv: `2009.07896`. URL: `https://arxiv.org/abs/2009.07896`.

[64] Gemini Team et al. *Gemini: A Family of Highly Capable Multimodal Models*. 2023. arXiv: `2312.11805 [cs.CL]`.