UNIVERSITAT DE
BARCELONA

Bachelor's Thesis

BACHELOR IN COMPUTER ENGINEERING

Faculty of Mathematics and Computer Science

University of Barcelona

EXPLORING A MULTIMODAL FOUNDATION MODEL ON BREAST CANCER VISUAL QUESTION ANSWERING

José Javier Iglesias Murrieta

Supervisor: Dr. Oliver Díaz Montesdeoca

Conducted at: Department of Mathematics
and Computer Science

Barcelona, June 10, 2025

Abstract

Cancer remains a leading cause of mortality worldwide, with breast cancer being the most frequently diagnosed. Early and accurate detection is critical to improving patient outcomes, and recent advances in artificial intelligence (AI) have demonstrated significant potential in supporting this goal. Machine learning (ML) and deep learning (DL) techniques have been widely applied to medical imaging tasks enhancing diagnostic accuracy across modalities such as mammography, ultrasound, and magnetic resonance imaging (MRI). However, most models require task-specific training and large annotated datasets, limiting their scalability and generalizability.

In response to these limitations, foundation models (FMs) have emerged as a promising shift in AI research. These large scale models are pre-trained on diverse data and can be adapted to a wide range of downstream tasks, including multimodal medical applications. Their capacity for zero-shot and few-shot learning presents opportunities for improving diagnostic support in data constrained settings. This research explores the application of FMs in breast cancer analysis, specifically assessing their ability to perform visual question answering (VQA) on the BCDR-F01 and BreakHis breast imaging datasets.

The study involves selecting a suitable vision-language FM and evaluating zero-shot and fine-tuning strategies to breast imaging data. Results demonstrate that while FMs show promising zero-shot performance and flexibility, their effectiveness depends heavily on model scale, fine-tuning approach, and task formulation, especially in complex multimodal tasks such as VQA. Instruction tuning and multimodal alignment emerged as critical factors for improving clinical relevance. This research highlights the potential of FMs to serve as integrative tools for breast cancer analysis, leveraging multimodal data with minimal retraining. Nonetheless, challenges remain in optimizing performance for clinical deployment, particularly around interpretability, domain-specific adaptation, and computational cost.

Resumen

El cáncer sigue siendo una de las principales causas de mortalidad en todo el mundo, siendo el cáncer de mama el más frecuentemente diagnosticado. La detección temprana y precisa es fundamental para mejorar los resultados de los pacientes, y los recientes avances en inteligencia artificial (AI) han demostrado un potencial significativo para respaldar este objetivo. Las técnicas de aprendizaje automático (ML) y aprendizaje profundo (DL) se han aplicado ampliamente a las tareas de imagenología médica, mejorando la precisión diagnóstica en modalidades como la mamografía, la ecografía y la resonancia magnética. Sin embargo, la mayoría de los modelos requieren entrenamiento específico para cada tarea y grandes conjuntos de datos anotados, lo que limita su escalabilidad y generalización.

En respuesta a estas limitaciones, los modelos fundacionales (FM) han surgido como un avance prometedor en la investigación de la IA. Estos modelos a gran escala se entrenan previamente con diversos datos y pueden adaptarse a una amplia gama de tareas posteriores, incluyendo aplicaciones médicas multimodales. Su capacidad para el aprendizaje de disparo cero y de pocos disparos presenta oportunidades para mejorar el apoyo diagnóstico en entornos con datos limitados. Esta investigación explora la aplicación de los modelos básicos en el análisis del cáncer de mama, evaluando específicamente su capacidad para realizar preguntas y respuestas visuales (VQA) en los conjuntos de datos de imágenes mamarias BCDR-F01 y BreakHis.

El estudio implicó la selección de un modelo fundacional de visión-lenguaje adecuado y la evaluación de estrategias de disparo cero y ajuste fino para los datos de imágenes mamarias. Los resultados demuestran que, si bien los FM muestran un rendimiento y una flexibilidad prometedores en el disparo cero, su eficacia depende en gran medida de la escala del modelo, el enfoque de ajuste fino y la formulación de tareas, especialmente en tareas multimodales complejas como VQA. El ajuste de instrucciones y la alineación multimodal resultaron ser factores críticos para mejorar la relevancia clínica. Esta investigación destaca el potencial de los FM para servir como herramientas integradoras para el análisis del cáncer de mama, aprovechando datos multimodales con un reentrenamiento mínimo. No obstante, persisten desafíos para optimizar el rendimiento para la implementación clínica, en particular en cuanto a la interpretabilidad, la adaptación específica del dominio y el coste computacional.

Resum

El càncer és una de les principals causes de mortalitat a nivell mundial, sent el càncer de mama un dels més diagnosticats. La detecció precoç i precisa és fonamental per millorar els resultats dels pacients, i els avenços recents en intel·ligència artificial (IA) han constatat un potencial significatiu per donar suport a aquest objectiu. Les tècniques d'aprenentatge automàtic (ML) i aprenentatge profund (DL) s'han aplicat àmpliament a tasques d'imatge mèdica, millorant la precisió diagnòstica en modalitats com la mamografia, l'ecografia i la ressonància magnètica. Tanmateix, la majoria dels models requereixen d'un entrenament específic per a la tasca i grans conjunts de dades anotades, cosa que limita la seva escalabilitat i generalització.

En resposta a aquestes limitacions, els models fundacionals (FM) han sorgit com un canvi prometedor en la investigació sobre IA. Aquests models a gran escala estan pre-entrenats amb dades diverses i es poden adaptar a una àmplia gamma de tasques posteriors, incloses les aplicacions mèdiques multimodals. La seva capacitat d'aprenentatge *zero-shot* i *few-shot* presenta oportunitats per millorar el suport diagnòstic en entorns amb dades restringides. Aquesta investigació explora l'aplicació dels FM en l'anàlisi del càncer de mama, avaluant específicament la seva capacitat per realitzar respostes visuals a preguntes (VQA) als conjunts de dades d'imatge de mama BCDR-F01 i BreakHis.

L'estudi inclou la selecció d'un model fundacional de llenguatge-visió adequat i l'avaluació d'estratègies d'ajustament precís i de *zero shot* per a les dades d'imatges de mama. Els resultats demostren que, si bé els models de fonamentació mostren un rendiment i una flexibilitat prometedors de *zero shot*, la seva eficàcia depèn en gran mesura de l'escala del model, l'enfocament d'ajustament precís i la formulació de tasques, especialment en tasques multimodals complexes com l'anàlisi de la resposta visuals a preguntes (VQA). L'afinació de les instruccions i l'alineació multimodal van sorgir com a factors crítics per millorar la rellevància clínica. Aquesta investigació destaca el potencial dels FM per ser utilitzada com a eines integradores per a l'anàlisi del càncer de mama, aprofitant les dades multimodals amb un reentrenament mínim. No obstant, encara existeixen certs reptes per optimitzar el rendiment per al desplegament clínic, especialment pel que fa a la interpretabilitat, l'adaptació específica del domini i el cost computacional.

Acknowledgments

I dedicate this work to the ones who have faced cancer - those who continue to fight, those who have survived, and those we remember with love. Your resilience and bravery will forever inspire me.

Special thanks to Dr. Oliver Díaz Montesdeoca for his guidance, trust, and empathy. To all my professors, who gave me the skills and knowledge to tackle this endeavor. To the friendships that shared hardship and laughter with me.

Gracias

Pia, por enseñarme a soñar

Paco, por celebrar mi curiosidad

Rodrigo, por ser mi mentor

Paula, Javier, Angelina y Pepe por su amor, alegría, sabiduría, y esfuerzo

Cesia, por recordarme lo bonito en la vida

Table of Contents

Chapter 1

Introduction

1.1. Context

Cancer is a group of diseases characterized by the uncontrolled growth of abnormal cells, which can invade nearby tissues and spread to other organs. It is the second leading cause of death globally, accounting for an estimated 9.6 million deaths in 2018 (World Health Organization, 2025). Despite significant advancements in treatment, early detection remains crucial for improving patient outcomes. Ongoing advances in medical technology are enhancing diagnostic precision, while digital innovations are reshaping clinical approaches to cancer diagnosis and treatment.

Among these innovations, artificial intelligence (AI) has emerged as a powerful tool for medical applications. AI is a broad field encompassing various technologies and advancements, including machine learning (ML) and deep learning (DL). These have been increasingly used to support medical practitioners with their decision making. In oncology, AI shows promise in cancer detection and diagnosis (Karger & Kureljusic, 2023). Since overcoming early technological limitations in the 2000s, AI driven models now analyze complex algorithms and self-learn, enhancing accuracy and workflow efficiency in clinical practice (Kaul & Gross, 2020).

1.2. Motivation

Foundation models (FM), such as large language models and vision transformers, are AI architectures that have shown remarkable capabilities in various domains due to their ability to leverage vast amounts of data and transfer knowledge across tasks. This project seeks to explore how these powerful models can be adapted and applied to specific applications such as breast imaging tasks, potentially revolutionizing detection, diagnosis, and prognosis in breast cancer care.

FMs are large-scale, pre-trained models that can be fine-tuned for a wide range of downstream tasks. In the context of medical imaging, these models could potentially capture complex patterns and relationships in breast images that may not be apparent to human observers or traditional machine learning approaches.

1.3. Objectives

This research aims to demonstrate that applying multimodal foundation models in zero-shot and fine-tuning regimes to diverse breast imaging datasets can achieve performance comparable to or exceeding that of conventional ML/DL approaches in tasks such as lesion detection, tumor classification, and cancer subtype prediction. Moreover, it seeks to highlight the differentiating features of FMs such as their ability to handle complex multimodal and data constrained scenarios to enable more explainable, contextually grounded, and clinically relevant outcomes. A set of secondary objectives were defined as part of this goal.

I. To identify and evaluate suitable FMs that can be adapted for breast imaging tasks, such as vision transformers or multimodal models that can process both images and associated clinical data.

II. To develop methodologies for fine-tuning these FMs on breast imaging datasets, including mammograms, ultrasounds, and magnetic resonance imaging.

III. To assess the performance of fine-tuned FMs on various breast imaging tasks, such as lesion detection, classification of benign vs. malignant tumors, and prediction of cancer subtypes.

IV. To compare the performance of foundation model-based approaches with traditional machine learning and DL methods in breast imaging analysis.

V. To investigate the potential of these models for zero-shot or few-shot learning in rare breast cancer subtypes or uncommon imaging findings.

VI. To analyze the interpretability and explainability of foundation model decisions in the context of breast imaging, ensuring that their outputs can be understood and trusted by clinicians.

VII. To explore the potential of FMs in integrating multimodal data, including imaging, clinical, and genomic information, for comprehensive breast cancer analysis.

1.4 Planning

The work described here was planned to be performed in four months, following the time available in the spring semester of the academic year (**Figure 1**).

The research part is focused on the clinical context of breast cancer and the state of the art of foundation models. It also covered selecting the data and model that would be explored. On the other hand, the development stage is focused on preparing the data and environment for testing zero-shot and fine tuning on visual question answering (VQA).

PHASE	TASK TITLE	START DATE	DUE DATE	DURATION	Research								Development								Report			
					February				March				April				May				June			
					1	2	3	4	1	2	3	4	1	2	3	4	5	1	2	3	4	1	2	3	4
1	**Research**																								
1.1	Context	2/3/25	2/16/25	13																					
1.2	State of the art	2/16/25	3/2/25	16																					
1.3	Data selection	3/2/25	3/16/25	14																					
1.4	Model selection	3/9/25	3/23/25	14																					
2	**Development**																								
2.1	Data preparation	3/24/25	4/6/25	12																					
2.2	Environment setup	4/6/25	4/20/25	14																					
2.3	Fine tunning	4/13/25	5/18/25	35																					
2.4	Analysis	5/11/25	5/25/25	14																					
3	**Report**																								
3.1	Writing	5/25/25	6/3/25	8																					
3.2	Corrections	6/3/25	6/10/25	7																					
3.3	Turn in		6/10/25																						
3.4	Presentation	6/17/25	6/29/25	12																					

Figure 1. Gantt chart of initial planning of the project.

Chapter 2

Breast Cancer

2.1. Epidemiology

Breast cancer is the most commonly diagnosed cancer worldwide, with an estimated 2.3 million new cases in 2020. Women are specially affected, accounting for 685,000 deaths, a figure projected to reach 1 million by 2040 (Arnold, et. al, 2022). In Spain, it is the most common type of malignancy in women, representing 30% of total cancer cases. It is also the country's female leading cause of death (Contra el Cáncer España, 2024).

This disease imposes both social and economic burdens that are unequally distributed. It is estimated that between 2020 and 2050, cancers will cost the world economy $25.2 trillion, with 7.7% corresponding to breast cancer alone (Chen, et. al, 2023). Noticeably, transitioned countries have double the incidence rate, while transitioning countries have a 17% increased mortality rate (**Figure 2**) (Arnold, et. al, 2022).

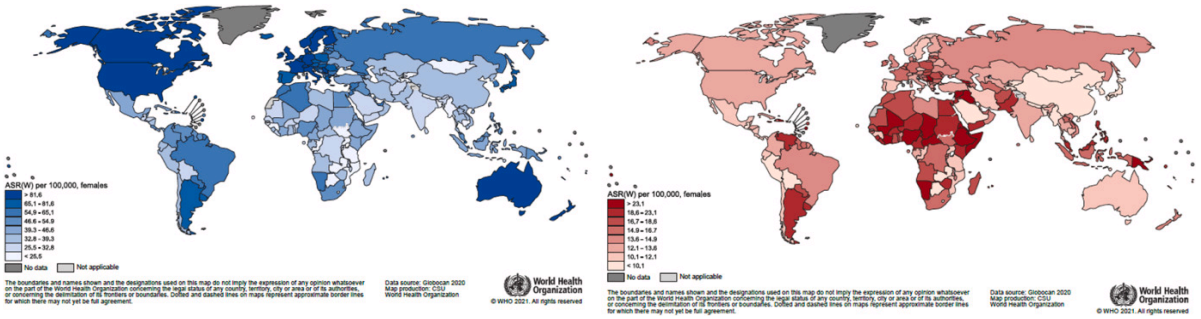


Figure 2. Age-standardized breast cancer incidence (blue) and mortality (red) rates per 100,000 females. Breast cancer cases and deaths by country (Arnold, et. al, 2022).

2.2. Subtypes

Molecular Classification

Breast cancer is not a single disease; it comprises multiple subtypes that differ in genetic, molecular, and histopathological features. New understanding of its molecular biology led to changes in how it is classified. The molecular classification of breast cancer uses biomarkers to identify each subtype, and guide diagnosis, treatment, and prognosis. However, the enormous heterogeneity and number of factors involved still make interpretation a challenging task.

Based on gene expression profiling, breast cancer is classified into four categories (**Figure 3**):

I. Luminal A Carcinomas
II. Luminal B Carcinomas
III. HER-2 Enriched Carcinomas
IV. Basal Carcinomas and Triple Negative Carcinomas

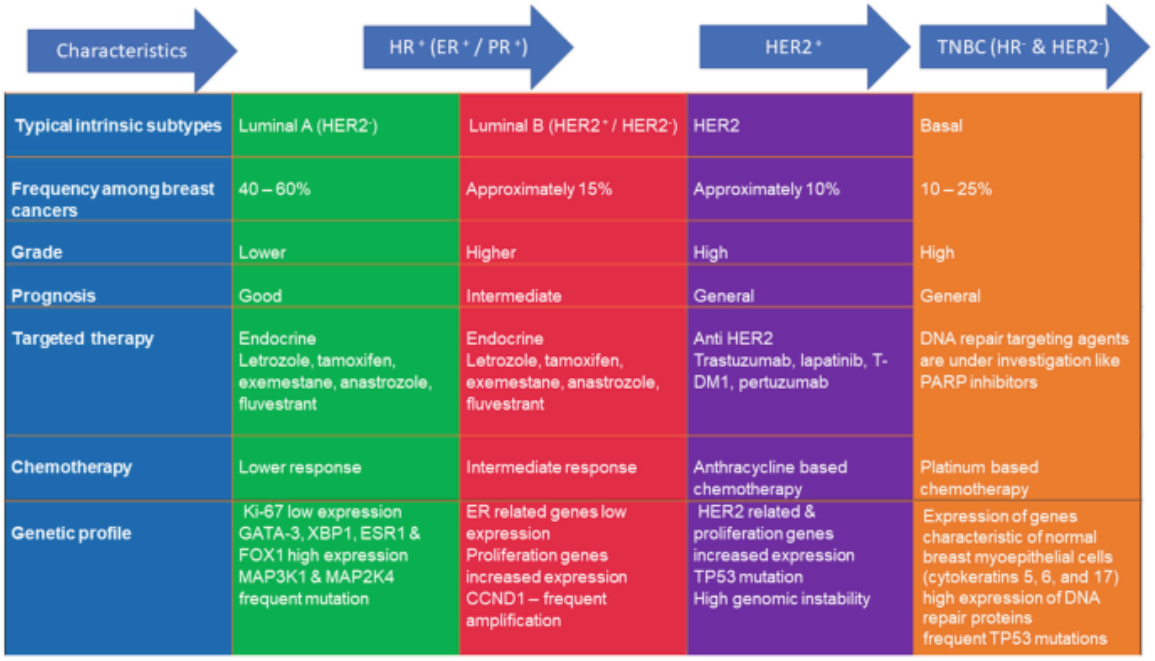


Figure 3. Breast cancer tumor's molecular subtypes (Malik, et al. 2020).

Alternatively, due to time and cost constraints, the standard practice uses the surrogate classification based on immunohistochemical assessment of biomarkers: estrogen (ER) receptor, progesterone (PR) receptor, HER2, and Ki-67 (Fernandes, 2022).

2.3. Medical imaging

There exist several imaging modalities for early breast cancer detection. The most common imaging techniques in clinical practice are: mammograms (MG), ultrasound (US), magnetic resonance imaging (MRI), and histopathology (HP). In fact, 50% of datasets are MGs, 20% US, 18% MRI and 8% HP. Each of these modalities can be further categorized into different subtypes (**Figure 4**).

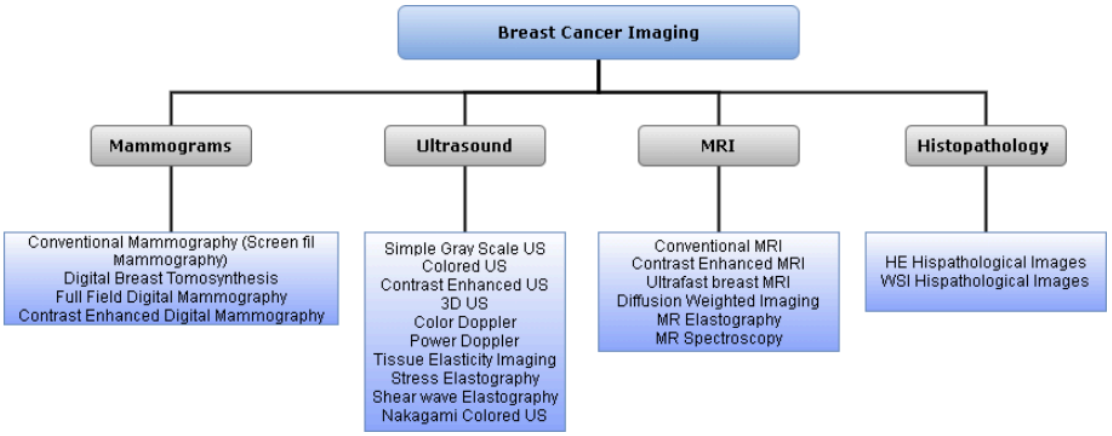


Figure 4. Most common imaging modalities and their subtypes used for breast cancer analysis (Shah, Khan, Arif, & Sajid, 2022)

2.3.1. Mammograms (MG)

MGs are low intensity X-ray images of human breast where glandular tissue, cancerous tumors and calcium deposits may appear brighter than surrounding tissue (e.g., adipose). As part of the standard protocol, two complementary views are captured for each breast: a craniocaudal view from above, and a mediolateral oblique view taken at an angle. These perspectives help provide a more complete assessment of lesions or abnormalities by reducing tissue overlap and improving localization. (**Figure 5**).

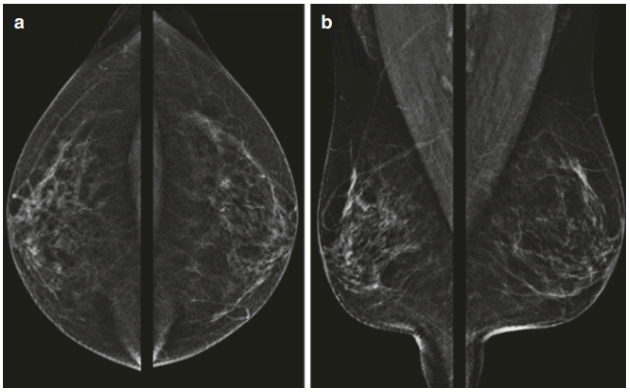


Figure 5. Standard mammography views: (a) craniocaudal; (b) mediolateral oblique (Morris & Kim, 2022).

MGs have widely been used for breast lesion detection and classification. While also used to detect breast cancer in early stages, it is not a preferred method due to reduced sensitivity in dense breast tissue and limited capabilities in capturing micro calcifications.

2.3.2. Ultrasound (US)

US are imaging techniques that use high frequency sound waves to create real time pictures of internal body structures. These are performed to detect the location of suspicious lesions in areas of interest in the breast (**Figure 6**). They come in 3 broad combinations: (1) 2D grayscale images, color images with (2) Shear Wave Elastography (SWE) added features, and (3) Nakagami. SWE enhances lesion differentiation by measuring stiffness, while Nakagami provides additional statistical parameters for localization.

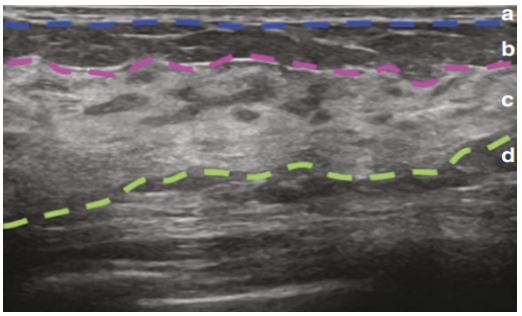


Figure 6. (a) Skin surface, (b) subcutaneous tissue, (c) mammary (d) retromammary zone (Morris & Kim, 2022).

However, Ultrasound suffers from two key problems that make it unreliable for general breast cancer screening, especially in asymptomatic women: the images are hard to interpret due to speckle noise, and the screening results have unacceptably high rates of both false positives and false negatives.

2.3.3. Magnetic Resonance Imaging (MRI)

MRIs capture multiple breast images at different angles to combine them together as a detailed view (**Figure 7**). Compared to previous techniques, they offer greater sensitivity in dense breasts and provide clearer soft tissue imaging.

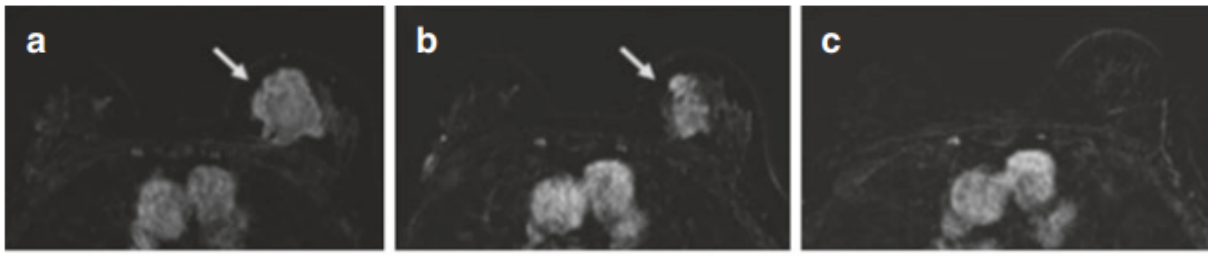


Figure 7. MRI scans of an invasive ductal carcinoma (arrow) before chemotherapy (a), after one cycle (b), and after eight cycles (c) (Morris & Kim, 2022).

Despite being an effective technique, due to its high cost and the possibility of missing some cancer tissue detectable by MGs, it is typically used as a secondary test to confirm a pathology or as a tool to follow-up during treatment.

2.3.4. Histopathology (HP)

Histopathology is the procedure of extracting a tissue sample from a suspicious human body region for microscopic examination and diagnosis.

Images are produced by fixing the sample glass stained with Haemotoxylin and Eosin, which create a colored visualization of the tissue (**Figure 8**). These images are available in 2 forms: (1) Whole Slide Images (WSI) and (2) Image patches extracted from WSI.

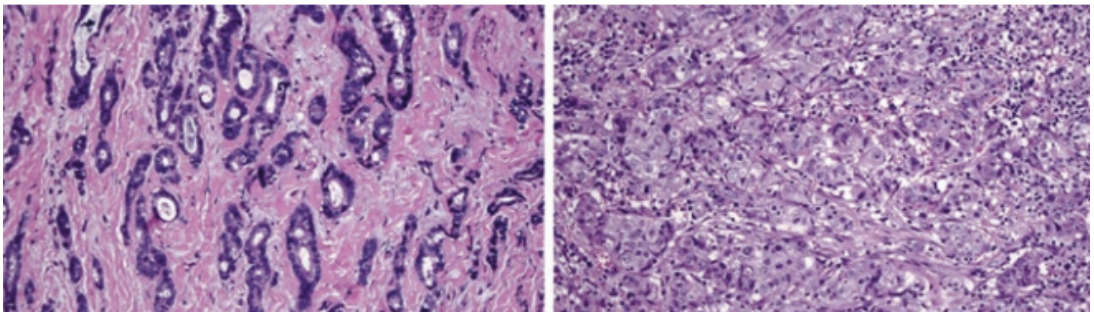


Figure 8. Histopathology of invasive carcinoma of no special type (Morris & Kim, 2022).

Patches with different zooming factors are used to diagnose multiple breast cancer types, which are impossible to diagnose with simple grayscale images. This tissue level examination has been successfully used for multi-class breast cancer classification (Shah, Khan, Arif, & Sajid, 2022).

Chapter 3

Artificial Intelligence

3.1. Definition

The field of AI aims to develop systems capable of performing tasks that typically require human intelligence, such as learning, reasoning, problem solving, and decision making. Its origins date back to 1950 when Alan Turing published *Computing Machinery and Intelligence*. In this paper, Turing raised the question "Can machines think?" and proposed an evaluation method for machine intelligence that later became known as the Turing Test. Six years later, John McCarthy coined the term and described it as "the science and engineering of making intelligent machines."

Early AI operated on simple conditional rules. Over time, technological advancements led to increasingly complex models capable of performing human-like functions (Kavlakoglu & Stryker, 2024). Various fields emerged to explore these advancements, with ML enabling computers to learn, evolving with the deeper complexity of DL, and recently leading into the generative and multimodal capabilities of FMs **(Figure 9)**.

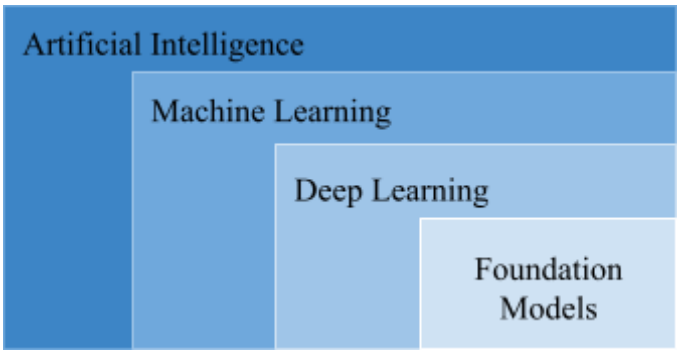


Figure 9. High-level diagram of AI subfields.

3.2. Machine Learning (ML)

ML is a subfield of AI that focuses on the development of algorithms capable of learning patterns and making predictions from data without explicit programming (Kavlakoglu & Stryker, 2024).

3.2.1. Classification

ML can be classified by learning strategy, model type, algorithm, or technique. The most common categorization distinguishes between supervised, unsupervised, and reinforcement learning, with some taxonomies further identifying semi-supervised learning (**Figure 10**).

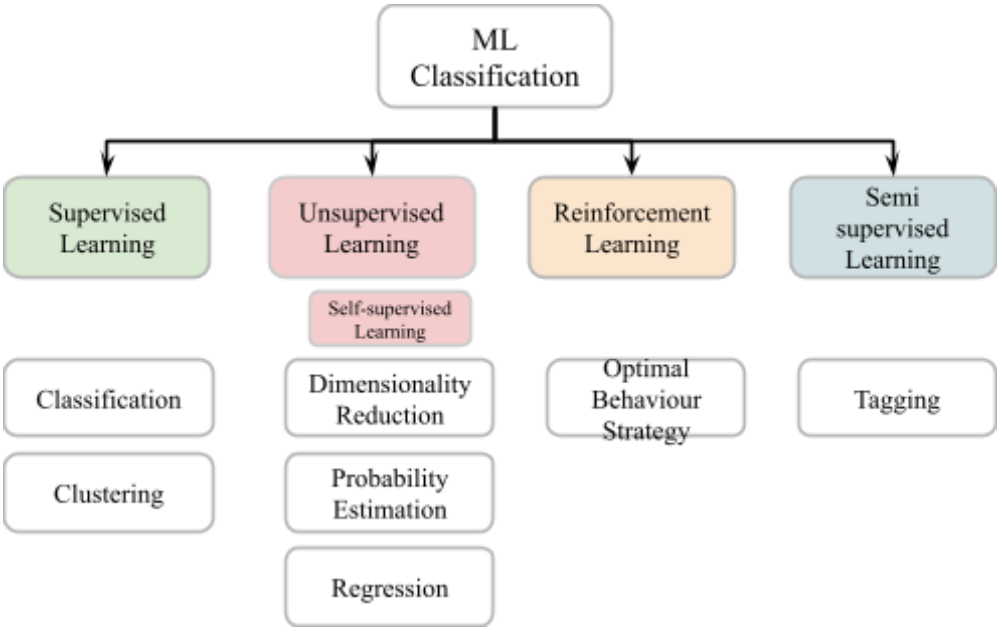


Figure 10. ML classification by learning technique and main applications. Inspired by Li, Lin,& Zeng, 2024.

Supervised Learning

In supervised learning, prediction models learn from labeled datasets. This method aims at learning the statistical mapping that best describes the relationship between inputs and outputs. Once trained, the model predicts outputs for new, unseen inputs by applying this learned mapping.

Unsupervised Learning

Unsupervised learning deals with unlabeled data, aiming to uncover inherent structures or distributions within the inputs. Without explicit output labels, the model identifies clusters, associations, or latent representations that reveal the data's underlying statistical laws (Li, Lin,& Zeng, 2024).

A subset of unsupervised learning, self-supervised learning leverages the unlabeled data itself to create its own output labels. It achieves this by defining pretext tasks where a portion of the data is used to predict another part, thereby training the model to learn useful representations without external labels. This approach allows the model to learn meaningful features from the data by solving these artificially constructed prediction problems (Bergmann, 2023).

Reinforcement Learning

Distinct from the previous, reinforcement learning frames learning as an agent's interaction with an environment. Actions that yield desirable outcomes receive rewards, whereas undesirable actions incur penalties. By trial and error, the agent learns a policy mapping states to actions that maximize cumulative reward over time.

Semi supervised Learning

Semi-supervised learning operates on a mixed dataset of a small labeled subset and a large unlabeled pool. By leveraging the abundant unlabeled data to inform or regularize the model, this paradigm seeks to achieve performance similar to fully supervised methods while significantly reducing labeling cost (Li, Lin,& Zeng, 2024)

3.2.2. Transfer Learning

Transfer learning is a branch of ML that studies applying knowledge gained from one task to a different but related one. Formally, the domain of a task consists of the data and the distribution that generates that data. There are at least two domains: a source domain from which knowledge is transferred, and a target domain where the learning is focused. The goal is to use the source domain data to learn a predictive function that minimizes prediction risk on the target domain (Wang & Chen, 2024). Effective transfer learning requires analyzing transferability, choosing the appropriate transfer technique, and selecting model parameters for best performance (**Figure 11**).

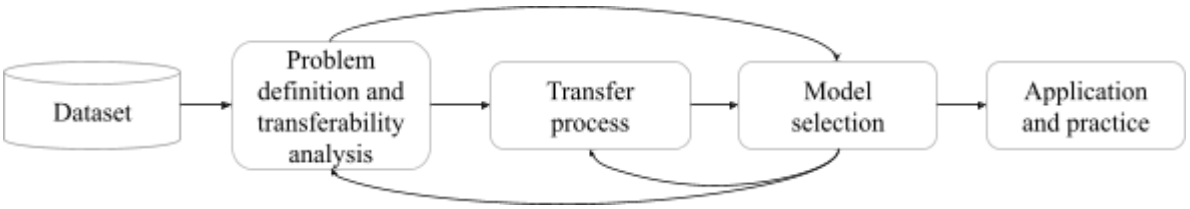


Figure 11. A complete transfer learning process (Wang & Chen, 2024).

3.2.3. ML techniques for breast cancer

Early breast cancer imaging analysis relied on expert designed image features paired with simple classifiers. Across all imaging modalities, support vector machines (SVM) and artificial neural networks (ANN) classifiers are some of the most established and applied. However, other classifiers including k-nearest neighbors (KNN), decision trees (DT), random forests (RF) and logistic regression (LR) have demonstrated comparable performance in certain studies (**Figure 12**). This highlights that the optimal choice of classifier is not

absolute, but instead must consider the properties of the data and the nature of the task. SVMs are emphasized in this review due to their proven effectiveness and adoption.

SVMs have proven highly effective for automated breast cancer diagnosis across diverse imaging modalities. In MG, SVMs have demonstrated high accuracy in tumor detection, density assessment, and mass classification. For example, Wajid and Hussain (2015) and Khalaf and Yassine (2015) applied SVMs to the MIAS, INBreast, and DDSM datasets, achieving up to 99% accuracy for abnormality assessment and 95.78% accuracy for cancer classification, respectively. In US, several studies have similarly shown reliable lesion detection and differentiation using SVMs: Prabusankarlal et al. (2015) reported 95.85% accuracy for breast mass detection and diagnosis, and Wu et al. (2015) extended these findings on a larger private cohort, achieving 96.67% accuracy in classification. Likewise, investigations on private MRI datasets report notably high diagnostic performance. Hassanien and Kim (2012) achieved 98% accuracy in distinguishing normal from abnormal tissue, while Soares et al. (2013) reported 94% accuracy for cancer detection. Finally, early SVM classifiers successfully differentiated cancer subtypes using HP imaging, as demonstrated by Brook et al. (2008). Taken together, these results underscore the robustness and versatility of SVM-based methods for accurate breast cancer diagnosis across multiple imaging platforms (Houssein, Emam, Ali, & Suganthan, 2021).

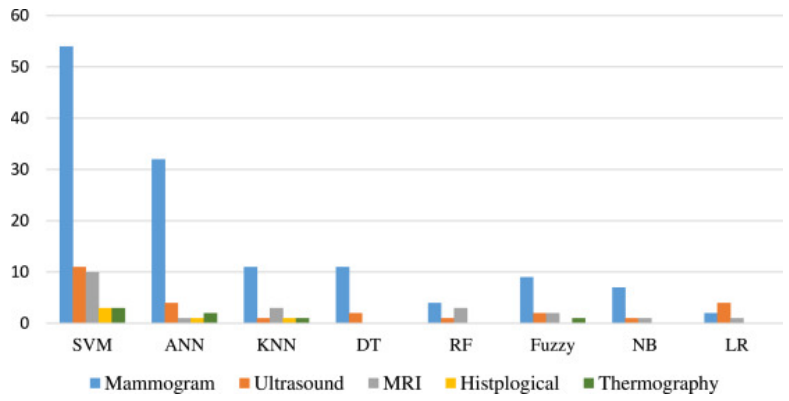


Figure 12. Number of papers using ML techniques per breast imaging modality between 2011 and 2020 based on the systematic review by Houssein, Emam, Ali, & Suganthan (2021).

3.3. Deep Learning (DL)

DL is a subset of ML that uses neural networks built with multiple layers to automatically learn from complex data. Its models consist of connected neurons across an input layer, many hidden layers and a final output layer (**Figure 13**). This arrangement allows the system to build hierarchical representations by extracting features directly from vast collections of unstructured or unlabeled data. Traditional ML depends on human selected traits while DL models learn those features on their own. It incorporates different learning strategies including semi supervised learning, self supervised learning, reinforcement learning and transfer learning (Kavlakoglu & Stryker, 2024).

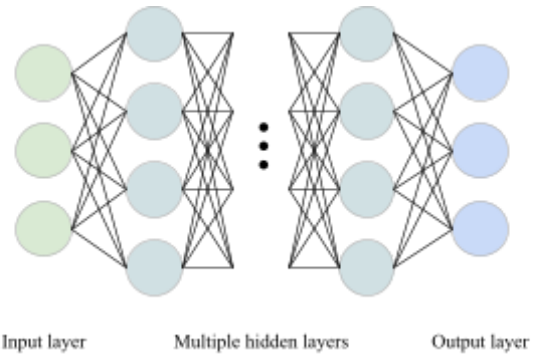


Figure 13. Typical architecture of DL neural networks.

3.3.1. Common Architectures

Different DL architectures are designed to capture underlying semantic information relevant to specific tasks. Understanding these architectures provides context for their specialized applications in medical imaging. Some of the basic DL algorithms include:

Convolutional Neural Networks (CNN)

CNNs are used for image processing and analysis. These are made up of convolutional layers, pooling layers, and fully connected layers that extract and process image features. Their efficiency in image classification and segmentation stems from parameter sharing and sparse connectivity.

Generative Adversarial Networks (GAN)

GANs are used for image generation tasks and are composed of two neural networks, a generator and a discriminator. The generator creates new data, while the discriminator evaluates whether the data is real or generated. Through this adversarial process, the generator improves its ability to produce realistic outputs.

Recurrent Neural Networks (RNN)

RNNs process sequential data by maintaining internal memory states that capture dependencies over time or space, making them effective for tasks such as 3D volumetric image analysis, natural language processing, or time series analysis.

Deep Reinforcement Learning (DRL)

DRLs combine DL with reinforcement learning to train agents that maximize rewards. These are specially useful for improving landmark detection and lesion segmentation tasks (Jiang et al., 2024).

3.3.2. DL techniques for breast cancer

DL has transformed breast cancer imaging by enabling the learning of rich feature representations directly from data. It has recently emerged as a key research focus, complementing traditional ML methods (**Figure 14**). Among the various DL architectures, CNNs have become the predominant approach for breast cancer imaging applications due to their significant ability to process and analyze medical images across different modalities. The following review focuses primarily on CNN based methods, as they are the most widely adopted DL technique in this domain.

CNNs have revolutionized automated breast cancer diagnosis by demonstrating exceptional performance across all major imaging modalities. In MG, CNNs have achieved notable accuracy in mass detection and classification. For instance, Chougrad et al. (2018) achieved 97.4% accuracy on the DDSM database, 95.5% accuracy on the INbreast database, and 96.60% accuracy on the BCDR database. Building on this, Al-antari and Kim (2020) proposed a system with DL classifiers to detect and classify lesions achieving F1 scores of 99.2% for DDSM and 98.02% for INbreast datasets.

Similarly, CNNs have demonstrated robust lesion detection and classification in US imaging. For instance, Han et al. (2017) trained a GoogleNet architecture on 7,408 ultrasound images, achieving 91.23% accuracy. Separately, Byra et al. (2019) utilized a transfer learning approach, reaching 88.7% accuracy on benign or malignant breast lesions within a 150 case test collection.

CNN approaches in MRI applications have shown promising results despite limited dataset sizes. A method developed by Feng et. al (2020) achieved an 85.0% accuracy on 100 MRI images for distinguishing benign from malignant lesions.

In HP imaging, CNNs have excelled in complex tissue analysis and cancer subtype classification. Yang et al. (2019) successfully classified breast tissue into four categories (normal, benign lesions, carcinoma in situ, and invasive carcinoma), achieving 91.75% accuracy. Similarly, Roy et al. (2019) developed patch-based CNN classifiers that achieved 90.0% accuracy for four class HP classification and 92.51% for binary classification tasks. (Houssein, Emam, Ali, & Suganthan, 2021)

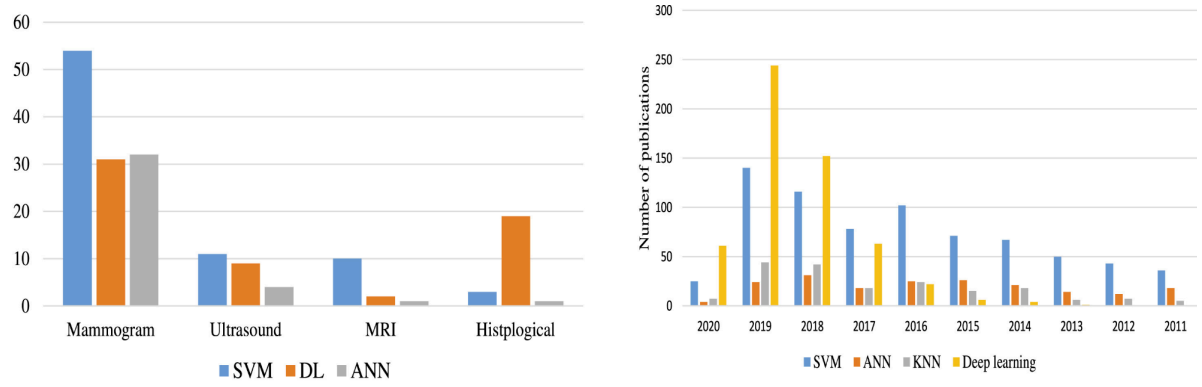


Figure 14. (left) Common ML and DL techniques for breast cancer imaging and (right) their annual publication trends for classification tasks based on the review by Houssein, Emam, Ali, & Suganthan (2021).

3.3.3. Limitations

Application of DL is challenging in clinical practice. It requires improving transparency, explainability, availability, accuracy, and performance. There are also ethical concerns, and regulatory requirements that need to be addressed.

Algorithmic challenges in DL include the lack of annotated data for training, integration into real world workflows, and transparency. Effective AI needs large annotated datasets that are labor intensive and costly to create. Integrating it to medical workflow requires extensive validation to ensure reproducibility and accuracy, and guidelines to be set for data and method normalization. Besides this, models often suffer from the “black box” problem, lacking necessary transparency for clinical adoption.

The development of effective models is also hindered by the interconnected challenges of limited and diverse data, ethical considerations, and medical data privacy and security. The integration of multimodal data is essential for advancing DL applications. Despite this, few DL models combine non imaging features with imaging data. Multicenter and high quality data is needed to maximize repeatability, and generalizability. However, the difference in acquiring and processing data results in significant heterogeneity. Sharing this data also raises privacy concerns, which call for collaborative and decentralized training methods (Jiang et al., 2024).

Chapter 4

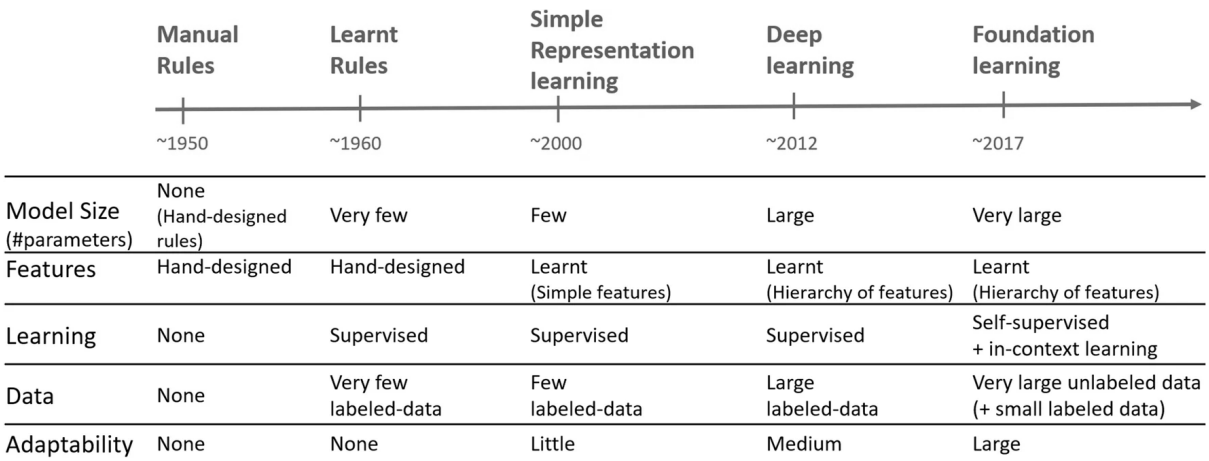
Foundation Models

4.1. Definition

FMs are first defined by Bommasani et al. (2021) as “any model that is trained on broad data that can be adapted to a wide range of downstream tasks.” Unlike DL models which often require large, task-specific, labeled datasets, FMs are pre-trained and available for fine tuning into various applications (Azad et. al 2023).

4.2. History

Early machine decision making began with expert systems: collections of hand coded rules that mapped inputs to outputs based on human expertise. Arthur Samuel (1959) ushered in the era of data driven learning by proposing that machines could learn without being explicitly programmed. Initially, such systems relied on expert defined features, but the advent of DL in the 2010s enabled hierarchical feature discovery directly from raw data, dramatically reducing the need for manual feature engineering (**Figure 15**).



	Manual Rules (~1950)	Learnt Rules (~1960)	Simple Representation learning (~2000)	Deep learning (~2012)	Foundation learning (~2017)
Model Size (#parameters)	None (Hand-designed rules)	Very few	Few	Large	Very large
Features	Hand-designed	Hand-designed	Learnt (Simple features)	Learnt (Hierarchy of features)	Learnt (Hierarchy of features)
Learning	None	Supervised	Supervised	Supervised	Self-supervised + in-context learning
Data	None	Very few labeled-data	Few labeled-data	Large labeled-data	Very large unlabeled data (+ small labeled data)
Adaptability	None	None	Little	Medium	Large

Figure 15. History of ML (Schneider, Meske, & Kuss., 2024).

DL’s modular architecture was flexible across different data types. Yet early neural networks were bottlenecked by the scarcity of labeled data given that supervised learning relied on this costly human annotation. Transfer learning partially alleviated this constraint by reusing representations from a “pre-trained” network on new tasks, fine-tuning only select layers. However, major progress arrived with self-supervised learning, which allowed models to learn from vast unlabeled corpora by solving surrogate tasks.

Self-supervised learning made scale the key factor driving progress. The introduction of the transformer architecture (Vaswani et al. 2017) facilitated the scaling of self-supervised training to massive datasets, leading to the emergence of FMs. These models demonstrate unprecedented expressiveness, multimodality, and the ability of solving new tasks without explicit retraining. FMs are typically refined via supervised fine-tuning, reinforcement learning from human feedback, or instruction tuning, enhancing alignment with user intent and ethical norms. Together, self-supervision, scale, and emergent in-context learning distinguish foundation models as the next stage in the evolution of ML (Schneider, Meske, & Kuss., 2024).

4.3. Characteristics

FMs have key features that differentiate them from traditional, specialized AI models. Foremost among these is *prompt sensitivity*, how interaction can influence the behaviour of AI systems. Prompting allows users to direct the system's output, ensuring it generates appropriate content. It is not limited to LLM[1], but also multimodal systems like image generation. Effective prompt design has shown to enhance performance by studies such as Kojima et al (2022). It can also help developers with fine tuning and defense against adversarial prompts.

Another key feature is the emergence of *in-context learning*, the ability to display new capabilities not explicitly programmed by developers. Training on large and diverse datasets enables these models to identify patterns and develop skills autonomously. As models continue to grow, the need for task specific tuning may diminish. However, there's an inherent uncertainty that raises concern about explainability, unpredictability, and the potential for unexpected behaviours.

Lastly, the importance of scale to model performance led to *homogenization*, the "consolidation of methodologies and models across AI applications and research communities (Schneider, Meske, & Kuss., 2024)." Factors such as the high cost of training and the monopoly of a few organizations on large proprietary datasets drive the centralization of a select set of models, which become the foundation for future AI systems. However, this raises concern on power centralization, dependencies, algorithmic monoculture, and the propagation of biases and undesirable behaviours across applications. It also impacts the roles of actors in developing AI, creating an ecosystem of FM providers, integrators, and end users, with implications for power dynamics and access to models (Schneider, Meske, & Kuss., 2024).

[1] Large Language Models (LLMs) are transformer-based FMs trained on large text corpora, enabling general-purpose language understanding and generation across diverse tasks.

4.4. Architecture

Most foundation models take advantage of transformer architectures. The transformer model is a DL architecture known for its ability to grasp complex relationships across different data modalities (**Figure 16**). This makes it a powerful tool for tasks that require an understanding of context and relationships between far-apart tokens.

1. **Input sequence:** the input for the model, a sequence of tokens.

2. **Embedding layer:** layer that converts each token in the input sequence into a fixed-size dense vector. Vectors capture the semantic meaning of tokens, and make them suitable for further processing.

3. **Positional encoding:** transformers do not have an inherent understanding of token order in the input sequence. This layer adds position information to the embedding vectors so the model can differentiate sequences with the same tokens in different order.

4. **Encoder blocks:** consists of a stack of identical blocks which process the entire sequence, each containing two main sub-layers to refine the representation of the input sequence:
    a. *Multi-head Attention*: a mechanism that allows the model simultaneous attention on different parts of the input sequence, capturing relationships between tokens regardless of their position to each other.
    b. *Feed-Forward Network*: a fully connected network to process the output of the attention mechanism, adding non-linearity and complexity to the model.

5. **Decoder blocks**: Similar to the encoder, with an additional attention layer focused on the encoder's output. It processes the target sequence during training and inference, refining its understanding of the sequence through this dual attention on input-output.
    a. *Masked Multi-Head Attention*: masks future tokens, preventing the decoder from attending them and ensuring predictions only rely on preceding tokens.
    b. *Encoder-Decoder Attention*: allows attention on relevant parts of the encoded input sequence.
    c. *Feed-Forward Network*: adds nonlinearity and complexity to the decoder's output.

6. **Linear layer**: maps the final decoder output to the target dimension.

7. **Softmax layer**: provides the probability distribution of the next token, derived from the linear layer's logits (unnormalized probabilities).

8. **Output sequence:** the stream of generated tokens (Singh, A., & Singh, K., 2025).

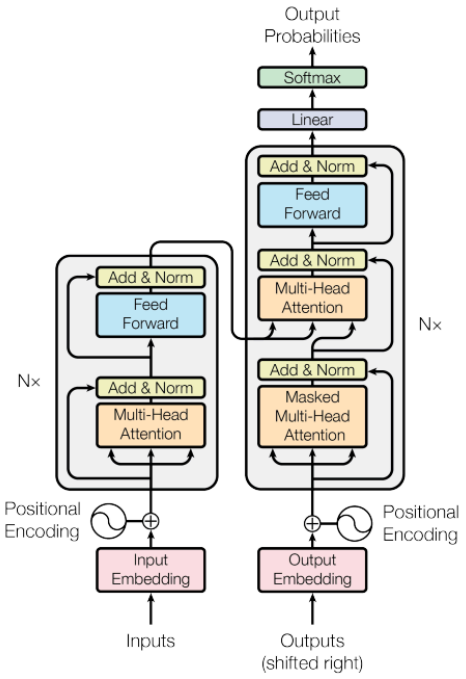


Figure 16. The Transformer - model architecture (Vaswani et al. 2017).

4.5. Training

4.5.1. Pre-training

In pre-training, a model typically leverages a Transformer architecture with self-attention layers to analyze a large corpus of unlabeled text. Using objectives such as masked language modeling, next sentence prediction, and causal language modeling, the model learns to capture the statistical patterns and linguistic structures inherent in natural language. This process cultivates a general language understanding and requires substantial computational resources over extended periods of time.

4.5.2. Fine-tuning

Following pre-training, fine-tuning adapts the pre-trained language model to specific downstream tasks by exploiting labeled, domain-specific datasets. Through back-propagation and gradient descent, the model's parameters are adjusted to minimize errors and align its output with task-specific objectives. This method leverages the model's established language understanding to adapt to the task using less data and reduced computational resources. However, thousands of labelled task-specific samples may still be required to achieve state-of-the-art performance.

4.5.3. Few-shot learning

Few-shot learning addresses scenarios where only a minimal number of labeled examples (typically one to five samples per class) are available. By employing meta-learning strategies, the model is first trained on a diverse set of classification tasks to develop transferable generalization capabilities. This meta-training enables rapid adaptation to new classes, mitigating overfitting in contexts where extensive labeled data is impractical.

4.5.4. Zero-shot learning

Zero shot learning enables the classification of text instances into categories that were not present during training. This method leverages semantic descriptions or class attributes, either by encoding both the text and the class information as embeddings or by using inference based approaches with large scale language models to assess compatibility through similarity metrics such as cosine similarity. In this way, the model is able to generalize to unseen classes by inferring relationships between textual representations and class semantics (Ferrari & Ginde, 2025).

4.6. Taxonomy

Azad et al (2024) proposed a methodical taxonomy to help researchers navigate the rapidly evolving field of foundation models in medical imaging. Their classification focuses on training strategies, but also factors in application areas, imaging modalities, specific organs of interest, and algorithms involved.

The taxonomy distinguishes between two model categories: Visually Prompted Models (VPM) and Textually Prompted Models (TPM) (**Figure 17**). VPMs are designed to handle visual inputs to guide their learning process. These models excel in tasks where visual prompts enhance image recognition and segmentation. On the other hand, TPMs leverage textual inputs to drive their learning and performance in visual recognition tasks. They combine textual and visual features through a fusion module to understand and process image-text pairs.

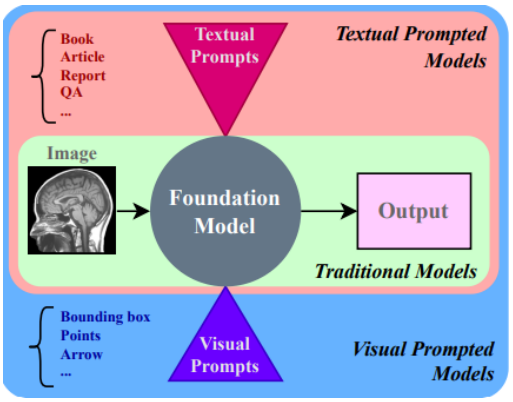


Figure 17. Textual and Visual Prompted model categories proposed by Azad et. al 2024.

Based on these categories, the taxonomy further classifies six distinct groups according to their objectives. TPMs encompass generative, conversational, contrastive, and hybrid forms, while VPMs are divided into adaptations and generalist forms (**Figure 18**).

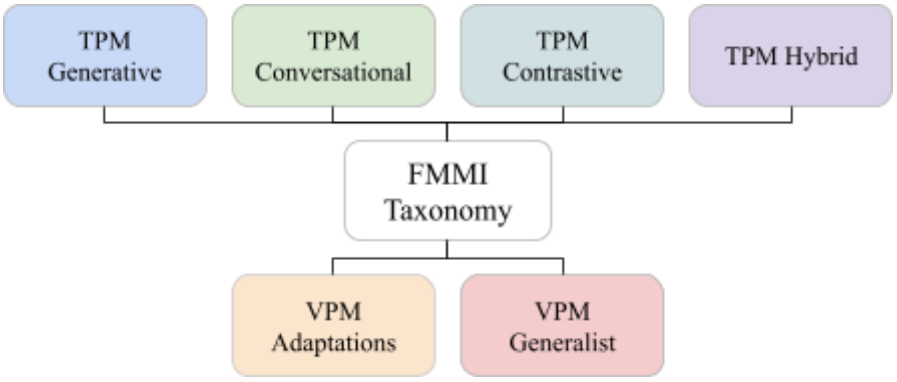


Figure 18. FM taxonomy in medical imaging. Inspired by Azad et al. (2024).

Models acquire an understanding of the relationship between vision and language through various pre-training objectives, which Azad et al. (2024) have broadly categorized as contrastive and generative. Contrastive objectives help models learn distinctive representations by pulling similar samples closer together in the feature space and pushing unrelated ones farther apart. Different loss functions are used to optimize different applications. For example, Image Contrastive Loss focuses on measuring and optimizing image similarity, while Image-Text Contrastive Loss aims to align image and text embeddings. On the other hand, generative objectives facilitate the learning of semantic features by training models to generate image or text data through various generation tasks. There also exist different loss functions to this objective. For instance, Masked Image Modelling enables the acquisition of cross-patch correlations by masking and reconstructing image patches, while Masked Language Modelling enhances language understanding by masking and predicting text tokens.

Contrastive TPMs excel at bridging the semantic gap between medical images and text by using contrastive learning. They are particularly useful in scenarios with limited labeled data, such as rare medical conditions or specialized imaging modalities. Generative TPMs focus on generating detailed responses and explanations for medical image-related queries. They aim to support clinical decisions by providing reasoning and interpretability. Hybrid TPMs combine generative and contrastive methodologies to integrate image-text tasks. They are adept at visual-questioning tasks and a valuable tool for quick diagnosis. Conversational TPMs enable interactive dialogues between professionals and AI systems. Experts can ask questions, seek explanations, and instruct on medical images.

Adaptation VPMs focus on extending medical imaging tasks. They are tailored for specific clinical applications and demonstrate robust generalization power. Generalist VPMs seek versatility by handling a wide spectrum of medical imaging tasks and data modalities. Their flexibility allows them to handle various tasks without the need for extensive retraining.

4.7. Downstream tasks and applications

Foundation models aim for broad capabilities that use combinations of different data types: text, image, video, and speech. For this reason, tasks and applications include and combine those of different and often overlapping areas of AI research. Text and image data are the most common and therefore NLP and Computer Vision have been the most explored. Tasks that require a single data type are defined as *unimodal*, while those integrating multiple data types are *multimodal*.

4.7.1. Natural Language Processing (NLP) tasks

NLP is the field of AI focused on enabling computers to understand, interpret, and generate human language. The following tasks focus on text data.

Reading comprehension

Evaluates a model's ability to read and comprehend a text passage to answer questions related to its content. Typically divided into four categories depending on the expected answer: cloze style (filling the blank), multiple choice, quoting a part of the text, and free-form answer. Achieving accuracy may require models to have certain world knowledge, process paraphrases, execute multi-sentence reasoning, and handle ambiguous or unanswerable queries.

Question answering

Assesses responses without context. In open cases, the model has access to a collection of knowledge without knowing where the answer appears. In closed cases, performance relies exclusively on knowledge acquired during the training phase.

Common Sense reasoning

Challenge models to apply real-world, common-sense reasoning rather than rely on memorized data. They cover scenarios like physical interactions and social situations. Tasks also cover problems requiring mathematical reasoning and natural-language inference, ensuring that models must understand and deduce rather than recall.

Natural Language generation

Aims to produce coherent, contextually appropriate text from structured or unstructured inputs. It includes tasks such as text summarization, code generation, machine translation, and writing tasks (Audiffren & Ostapuk, 2024).

4.7.2. Computer Vision tasks

Computer Vision is a field of AI that aims to enable computers to "see" and interpret the visual world, similar to how humans do. The following tasks focus on image data.

Image Classification

Assigns one or more semantic labels to an image. Models must cope with significant intra-class variability (pose, lightning, occlusion) and unknown “none of the above” cases, requiring large labeled datasets and architectures that generalize beyond training distributions.

Object Detection

Locates and classifies individual object instances within an image via bounding boxes. Variants include face detection and pedestrian detection. Detectors must handle objects at different scales, overlapped, occluded or in cluttered scenes.

4.7.3. Language and Vision tasks

This area of AI focuses on enabling machines to understand and interact with both images and text, integrating various AI components to handle them.

Visual Captioning

Generates a natural language description given an image. Requires the model to accurately detect the object, attributes, and their relations, then composing contextually appropriate sentences.

Text-to-image generation

Inverse of captioning, given a text prompt generates an image. This task requires the model to interpret the textual description and then synthesize a visually coherent image that matches the prompt.

Visual Questioning Answering (VQA) and Reasoning

Involve answering open or closed ended natural language questions about images. This requires a model to perform multimodal feature fusion, effectively combining information from both the image and the question. It often relies on attention mechanisms to focus on relevant parts of the image and question, and modular reasoning to break down complex queries into manageable steps (Szeliski, 2022).

4.8. Evaluation of FM

Given the wide range of downstream tasks FM can perform, it is necessary to use standardized evaluation metrics suited to each task to ensure accurate performance assessment. For tasks expecting answers that can be considered part of categories such as classification or VQA, binary and multiclass classification metrics effectively quantify the model's predictive accuracy. Tasks taking a free form generative approach, such as image captioning or text summarization, instead rely on alternative metrics like ROUGE, BLEU, METEOR, and CIDEr to evaluate the quality of generated text.

4.8.1. Binary classification

When evaluating a binary classifier, the predicted labels are compared against the actual true labels. This comparison yields four counts that make up the confusion matrix (**Table 1**): true positives (TP), where the model correctly predicted the positive class; true negatives (TN), where it correctly predicted the negative class; false positives (FP), instances incorrectly classified as positive; and false negatives (FN), instances incorrectly classified as negative.

Table 1. Confusion matrix.

	Predicted = Positive	Predicted = Negative
Actual = Positive	*True Positive* (TP)	*False Negative* (FN)
Actual = Negative	*False Positive* (FP)	*True Negative* (TN)

From these counts, several key performance metrics can be derived to evaluate different aspects of classifier performance. **Table 2** summarizes the most commonly used metrics, their formulas, and their insights.

Table 2. Summary of binary classification metrics. Inspired by (Ferrari & Ginde, 2025).

Metric	Formula	Key Insights
Precision	$P = \frac{TP}{TP + FP}$	Proportion of positive predictions that are correct, critical when false positives are costly as in medical diagnosis.
Recall	$R = \frac{TP}{TP + FN}$	Proportion of actual positives correctly identified, critical when false negatives are costly as in fraud detection.
F1-score	$F1 = \frac{2 \times P \times R}{P + R}$	Describes the balance between precision and recall, often used as summary measure for imbalanced datasets
Accuracy	$A = \frac{TP + TN}{TP + TN + FP + FN}$	Simple measure of overall correct classification, limited in scenarios with imbalanced class distributions

4.8.2. Multiclass classification

For multi-class classification problems involving K distinct classes ($k = 1, 2, 3, \ldots, K$), the metrics of Precision, Recall, and F1-score can be generalized through various averaging methods across these classes. Two common schemes are macro (unweighted) and micro (weighted by class size). Macro averaging provides equal weight to each class, making it valuable when all classes are considered equally important regardless of their frequency in the dataset. Conversely, micro averaging weights each prediction equally, giving more influence to classes with larger sample sizes and providing a measure more aligned with overall classification accuracy. Notably, micro precision, micro recall, and micro F1 are all measured using the same accuracy metric and thus have identical scores. **Table 3** summarizes the macro metrics, their formulas, and their insights.

Table 3. Summary of multiclass metrics. Inspired by (Ferrari & Ginde, 2025).

Metric	Formula	Key Insights
Macro Precision (mP)	$mP = \frac{\sum_{k=1}^{K} P_k}{K}$	Indicates the model's prediction correctness across all classes, penalizing poor performance on smaller classes.
Macro Recall (mR)	$mR = \frac{\sum_{k=1}^{K} R_k}{K}$	Shows if the model finds most true instances for all classes, crucial when missing a class is costly.
Macro F1 (mF1)	$mF1 = 2 \times \left(\frac{mP * mR}{mP^{-1} * mR^{-1}}\right)$	Emphasizes uniform performance across all classes

4.8.3. Text generation

Text generation tasks require different evaluation approaches than classifiers. These metrics compare generated text against one or more references, as described in Table 4.

Table 4. Summary of generation metrics. Inspired by (Ferrari & Ginde, 2025).

Metric	Key Insights
Recall-Oriented Understudy for Gisting Evaluation (ROUGE)	Evaluates summarization tasks by measuring the overlap between system generated and human text. Variants consider n-grams and longest common subsequences.
Bilingual Evaluation Understudy (BLEU)	Appropriate for evaluating machine translation. Calculates precision of n-grams in the generated text compared to the reference.
Metric for Evaluation of Translation with Explicit Ordering (METEOR)	Evaluates generated text by combining precision and recall and incorporating factors like synonym matches and alignment between the output and reference.
Consensus-Based Image Description Evaluation (CIDEr)	Evaluates the quality of captions of images. Considers both precision and recall, while also weighing saliency and rarity to capture relevant details.

4.9. Challenges

The large scale of FMs comes at a cost. These models face challenges similar to those previously discussed in DL. However, the pursuit of strong AI has not only intensified these existing issues but also introduced new ones. These challenges span various areas, including data diversity, design considerations, tuning complexities, theoretical understanding, environmental impact, and social implications.

Model performance can be enhanced with support for multimodal and multilingual data. Multimodal research has often focused on combinations of two modalities, such as text and image or text and audio. To develop effective multimodal FMs, it is crucial to create new datasets that integrate multiple modes. Similarly, multilingual models can benefit from tasks designed specifically for multilingual contexts. However, increasing the vocabulary size would require more parameters, creating an additional cost challenge.

The computational demands of FMs is a significant barrier to accessibility and innovation. As models continue to grow in size, additional research into model compression techniques is needed to reduce costs and facilitate wider participation in development. Another challenge is improving robustness in NLP to withstand adversarial inputs that manipulate output predictions. Unlike images, where transformations do not alter content meaning, even minor word substitutions can impact text semantics.

Achieving consistent model performance across both upstream and downstream tasks remains a fundamental challenge. Abnar et al. (2021) observed a nonlinear relationship between these tasks' performance, noting that increased training data and accuracy in pretraining does not necessarily imply improved downstream results. Additionally, the excess of self supervised tasks hinders establishing a clear relationship to downstream tasks. This creates ambiguity in how pretraining knowledge transfers to specific applications, making it difficult to determine which of them contribute meaningfully to downstream performance. As a result, models may learn representations that are overly broad or misaligned with the requirements of their target tasks.

A stronger theoretical understanding can better guide experimentation. There is currently a lack of profound theory to support tentative experiments. While some analyses attempt to understand phenomena like the collapse of pretraining and the generalization ability, a comprehensive theoretical foundation remains elusive. Moreover, semantic understanding poses a challenge, as it is unclear whether FMs genuinely grasp the meaning of language or simply rely on corpus learning. Although excelling in various datasets, they often struggle with stability and performance on domain-specific or smaller ones, failing to meet the purpose of human language use (Zhou, et al., 2024).

Lastly, ensuring responsible development and deployment of FMs requires addressing business, governance, ethical, and ecological challenges. As these models become integral parts of business processes, responsibilities and liabilities among stakeholders must be formalized. Organizations must also assess and mitigate their risks, adjusting AI management structures to consider issues such as privacy and copyright. Existing governance frameworks

can guide this adaptation. Regulatory bodies such as the EU AI Act (Regulation (EU) 2024/1689) are discussing the creation of ethical AI systems and promoting their responsible behaviour. Examining the broader economic and social implications, such as workforce dynamics and market competition will be important. In addition, the ecological footprint of training and deploying foundation models requires research into their environmental impact, including energy consumption, carbon emissions, and resource use. Solutions to improve sustainability may include optimizing model architecture, reducing redundancy in training data, researching model compression, and implementing structural changes like shared computing resources and federated learning networks (Schneider., Meske, & Kuss, 2024).

Chapter 5

Implementation

5.1. Model definition

As seen in the taxonomy, there exist several types of foundation models in medical imaging. The chosen model needed to address our initial hypothesis, enabling a comprehensive exploration into fine-tuning, performance, zero-shot learning, multimodality, and explainability. Potential limitations, such as GPU [2] availability, model size and licensing had to also be taken into consideration.

Generalist VPMs best explore the versatility of foundation models in integrating and utilizing diverse data types across various medical tasks. From these models, Zhang et al.'s (2024) BiomedGPT aligned the most with the purpose of the research. First, it is a fully transparent, open-source language-vision model licensed for academic research. The model checkpoints, datasets, and scripts used for preprocessing, training, fine tuning, and evaluation are accessible. Additionally, it achieved state-of-the-art results in 16 out of 25 experiments assessing its capabilities on both unimodal and multimodal tasks that included image classification, captioning, VQA, text summarization, and medical natural language inference. The performance was also evaluated by medical professionals.

Another key advantage of BiomedGPT is its lightweight architecture. It is available in three distinct sizes, referred to as BiomedGPT-S (tiny), BiomedGPT-M (medium), and BiomedGPT-B (base), along with instruction-tuned versions. This range of sizes allows for flexibility in environments with varying GPU accessibility and facilitates testing and tuning of performance effects. Overall, BiomedGPT is an accessible, well documented, and powerful foundation model useful for several research purposes.

The foundation for these capabilities is the model's multimodal transformer architecture (**Figure 19**), which is designed to handle 2D image and text data. A multimodal architecture differs from the traditional architecture in three key aspects. First, it is trained on datasets containing paired image-text examples to correlate visual features with descriptive text. The embedding space is therefore joint so both visual and textual data coexist and be processed. Lastly, it uses cross-attention layers to enable focusing on one modality while generating text (Singh, A., & Singh, K., 2025).

[2] A Graphics Processing Unit (GPU) is a specialized processor with a parallel architecture that accelerates the training of artificial intelligence models by efficiently performing the simultaneous matrix and vector operations essential for DL algorithms.

a

Model scale	#Parameters	Image projection		Representation size		Transformer block		
		Input size	Visual encoder	Hidden	Intermediate	Att. head	#Enc. layer	#Dec. layer
BiomedGPT-S	33 million	256 × 256	ResNet-50	256	1024	4	4	4
BiomedGPT-M	93 million	256 × 256	ResNet-101	512	2048	8	4	4
BiomedGPT-B	182 million	256 × 256	ResNet-101	768	3072	12	6	6

b

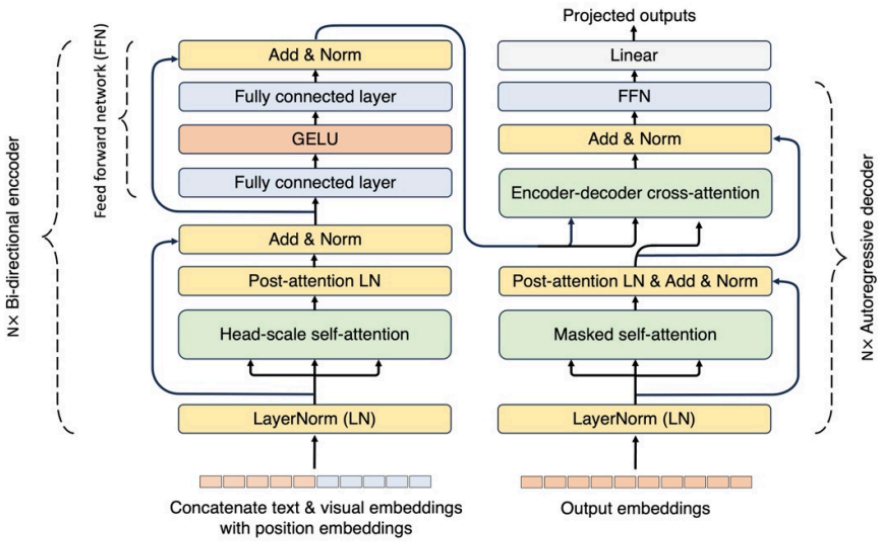


Figure 19. (a) BiomedGPT scale definition and (b) architecture (Zhang et al., 2024).

5.2. Model performance

BiomedGPT's performance is particularly notable when compared to other prominent large language models in the biomedical domain. It notably outperformed OpenAI's GPT-4 with vision (GPT-4V) in human evaluations specifically for radiology tasks. Furthermore, BiomedGPT surpassed Google's Med-PaLM M in both breast cancer diagnosis and medical VQA. This is a significant achievement, considering that Med-PaLM M is a much larger model, featuring 12 billion parameters.

BiomedGPT's performance across its various model sizes on diverse downstream tasks highlights its capabilities in different modalities. For binary image classification, BiomedGPT achieved accuracies of 97.0% and 89.7% on the SZ-CXR and MC-CXR datasets, surpassing the state of the art LightTBNet by 6.0% and 0.8%. In three class classification BiomedGPT-B surpassed the F1-Macro score of Med-PaLM M in both mass classification (scoring 57.2% vs 51.1%) and calcification classification (scoring 72.8% vs 67.9%).

On medical VQA tasks, BiomedGPT demonstrated solid accuracy for closed questions. It scored 88.0% on PathVQA and 81.3% on VQA-RAD, and set a new state-of-the-art score of 86.1% on SLAKE. However, the model is less effective with open-ended questions. Particularly, recording scores of 60.9% on PathVQA and 28.0% on VQA-RAD. Its performance on these was significantly lower, which can be attributed to its more limited model capacity and a lack of diverse conversational data during its training.

This can cause it to provide incomplete or syntactically awkward answers to more complex medical queries.

For medical image captioning, BiomedGPT achieved exceptional results on the PEIR GROSS dataset, surpassing the existing state-of-the-art with improvements of 8.1% in ROUGE-L, 0.5% in METEOR, and a significant 89.8 point increase in CIDEr scores. The model's performance on the IU X-RAY dataset revealed interesting trade-offs in its captioning strategy. While BiomedGPT achieved a leading CIDEr score of 40.1 (representing a 5.0-point improvement over the previous best model), it recorded lower METEOR and ROUGE-L scores of 12.9% and 28.5% respectively. This performance pattern reflects the model's training optimization, which prioritized capturing key visual elements over broader linguistic fluency. On the more challenging MIMIC-CXR dataset, BiomedGPT matched the leading model's METEOR score of 14.2%, though it fell short of the larger Med-PaLM M model in ROUGE-L (23.7% vs 26.2%) and CIDEr (14.7 vs 23.4) metrics.

Finally, in medical text summarization and NLP tasks, BiomedGPT demonstrated strong scaling properties and competitive performance despite its relatively compact size. For medical natural language inference using the MedNLI dataset, the model showed clear improvements with increased parameter count, achieving accuracies of 75.8%, 80.8%, and 83.8% across its three model variants. Notably, BiomedGPT-B achieved 83.8% accuracy while using only a quarter of the parameters of SciFive-Large (which achieved 86.6%), resulting in just a 2.8% performance gap despite the significant efficiency advantage. For text summarization tasks, BiomedGPT-B was evaluated across four benchmark datasets including MedQSum, HealthCareMagic, MIMIC-CXR, and MIMIC-III, demonstrating solid summarization capabilities across diverse medical text types from doctor-patient dialogues to radiology reports (Zhang et al., 2024).

5.3. Explored tasks

BiomedGPT addressed both unimodal applications, including classification, text summarization, and report generation, and multimodal tasks like VQA and captioning. BiomedGPT focused on the latter, and will therefore be the subject of these experiments. Specifically, VQA tasks were chosen due to their multimodal complexity, zero-shot capacities, and exploring the effect of instruction fine tuning on performance.

5.4. Dataset selection

For VQA tasks, BiomedGPT was fine-tuned with the PathVQA (He et al., 2020), and SLAKE (Liu et al., 2021) datasets. Additionally, zero-shot performance was evaluated using the VQA-RAD (Lau et al., 2018) dataset omitted from training. It is important to note that these datasets contain different open and closed questions for radiology images of various organs. Answers are generally short, and most questions are straightforward and not complex.

To evaluate zero-shot performance and fine-tuning across imaging modalities, we selected two datasets: BCDR-F01 (Moura et al., 2013) for mammography and a subset of

BreakHis for histopathology. To test the effect of instruction tuning, we will use a newly generated naive VQA dataset that mimics the instructions of VQA-RAD for BCDR-F0, and then compare it to a BreakHis VQA dataset provided by Hu et al. (2024) OmniMedVQA.

5.5. Evaluation metrics

BiomedGPT evaluates VQA tasks through weighted F1-score, accuracy and alignment accuracy. These same metrics are kept for consistency. In the case of VQA, accuracy measures the total cases where the answer is an exact match to the ground truth. Alignment accuracy measures if the answer is aligned to the question. For example, if a closed question is asked, the model answer is expected to align to yes/no answers.

5.6. Experiment hypotheses

Based on prior breast-imaging results with BiomedGPT and considering model architectures, sizes, fine-tuning approaches, and dataset complexity, the following hypotheses are proposed:

Model Size Effect

Increasing the model size will yield higher overall VQA accuracy on biomedical image-question pairs.

Zero-Shot Question Type Performance

In zero-shot settings, closed-ended questions will achieve higher accuracy than open-ended questions. Noticeably, performance will be worse on breast images compared to other regions.

Dataset Alignment and Instruction Tuning

Generating a VQA dataset closely matched in style and content to BiomedGPT's breast-imaging data, will improve both alignment and accuracy. In contrast, a more complex VQA dataset on a different modality such as BreakHis, will reduce alignment and accuracy but highlight the model's capacity and flexibility.

Fine-Tuning Trade-Offs

Fine-tuning non-instruction-tuned model variants on relevant VQA data will boost accuracy compared to their zero-shot baselines. However, without prior instruction tuning, answers are prone to be misaligned; achieving parity with instruction-tuned versions will require substantially more data or training epochs.

Chapter 6

Results and discussion

6.1. Development Environment

BiomedGPT underwent pre-training on a set of 10 NVIDIA A5000 GPUs, and the majority of its scripts employ a distributed launch configuration. Therefore, a GPU-enabled environment is required for optimal performance. However, given current GPU pricing, a physical setup may not be easily accessible. For this reason, a cloud environment was chosen, as it offers free (albeit limited) GPU access and the option to scale at a lower cost.

Different platforms were explored, but Google Colab was ultimately chosen. Its virtual machine offered consistent free tier GPU availability, integration with Drive, and 100 GB of storage, sufficient for most datasets. It also had limitations, most significantly a cumbersome Miniconda interaction and restricted terminal access.

The installation guide for BiomedGPT was adapted for this environment. The colab notebook *gcloud_conda_setup.ipynb* is provided as a guide to clone the repository, mount the Drive and install Miniconda to ensure its persistence. Miniconda was necessary to create an environment with the required Python and pip versions for the packages. The scripts could then be executed via the conda run command.

6.2. Data exploration

6.2.1. VQA-RAD

Before jumping into evaluating VQA zero-shot capacities, a data exploration is conducted on the dataset BiomedGPT evaluated it with. Lau et al. (2018) developed VQA-RAD, a dataset comprising 3,515 question-answer pairs linked to 315 radiology images of various organs. A random sample of the test set images is shown in **Figure 20**.

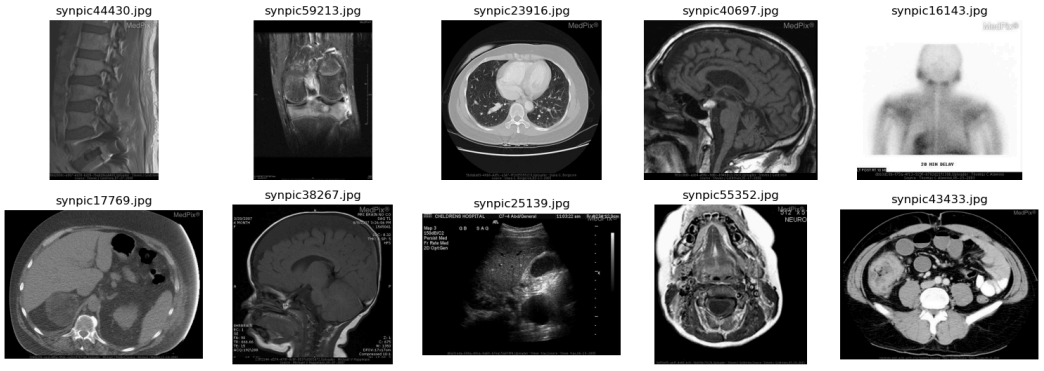


Figure 20. Random sample of the VQA-RAD test dataset.

VQA-RAD has two types of questions: closed (yes/no) and open questions. **Figure 21** shows the density distribution in the character length of questions and answers. Notice that most of the answers are short (below 40 characters), indicating a significant number of closed questions.

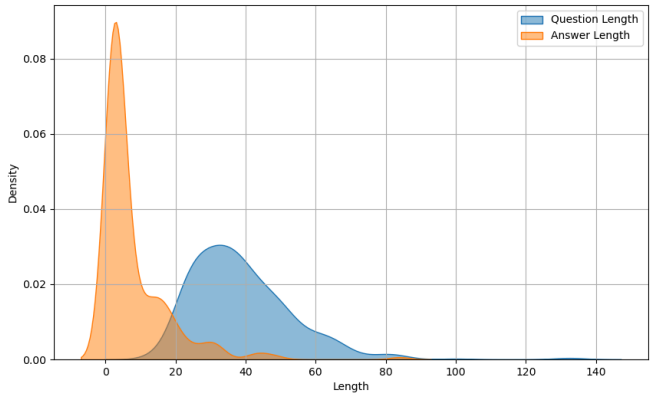


Figure 21. VQA-RAD test density of question and answer character length.

Considering that instruct tuned versions are used for evaluation, taking a look at the nature of the questions is important. **Figure 22** shows the distribution of the first and second words across the questions of the test set. Questions beginning with "*Is…*" are often related to closed, specific, and unambiguous answers. In contrast, questions starting with other words (e.g., what, are, can…) are typically correlated with open, more complex, and longer answers.

Figure 22. Distribution of the first and second words in the test questions.

6.2.2. BCDR-F01

Derived from Breast Cancer Digital Repository (BCDR), the BCDR-F01 dataset contains 200 biopsy-confirmed film MG lesions, equally split between 100 benign and 100 malignant cases (**Figure 23**). It comprises a total of 358 feature vectors, with 184 instances corresponding to the benign lesions and 174 to the malignant ones (Moura et al., 2013). For each lesion, the dataset provides a set of clinically relevant attributes, including six binary indicators of radiological findings (i.e., masses, microcalcifications, calcifications, axillary adenopathies, architectural distortions, and stroma), an ordinal measure of breast tissue density, and the patient's age at the time of examination. This rich yet concise set of features allows for targeted analysis of the relationship between observable characteristics and diagnostic outcomes.

This research could not identify a dedicated VQA dataset on the BCDR-F01 dataset. This presented an opportunity to showcase the zero-shot capabilities and flexibility of foundation models by generating a VQA using the dataset's clinically annotated features.

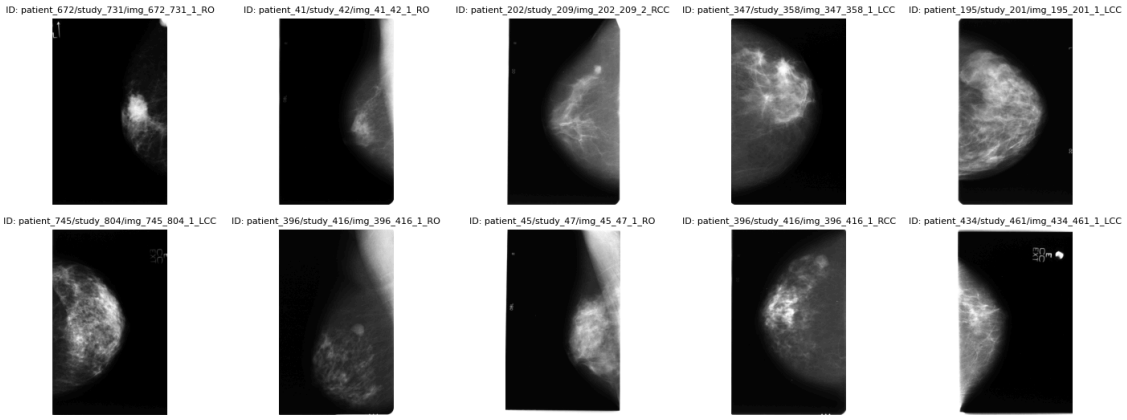


Figure 23. Random sample of the BCDR-F01 dataset.

6.2.3. BreakHis

The BreakHis dataset is a publicly accessible collection of breast cancer histopathological images introduced by Spanhol et al. (2016). It comprises a substantial number of microscopic images of breast tumor tissue, captured at various magnification factors (40x, 100x, 200x, and 400x) (**Figure 24**). Experiments used a subset including 241 benign and 443 malignant cases from the OmniMedVQA dataset by Hu et al (2024). Notice the significant difference in question structure and density compared to the one in VQA-RAD (**Figure 25**). All questions are open and consistent in format, falling into two categories: classification (malignant/benign) and modality (histopathology).

BiomedGPT was primarily pre-trained and fine-tuned using radiology imaging data. Considering the established success of histopathology in accurately classifying rare cancer subtypes, the aim of utilizing this dataset is to demonstrate the foundation model's inherent flexibility and its ability to generalize effectively in a zero-shot setting.

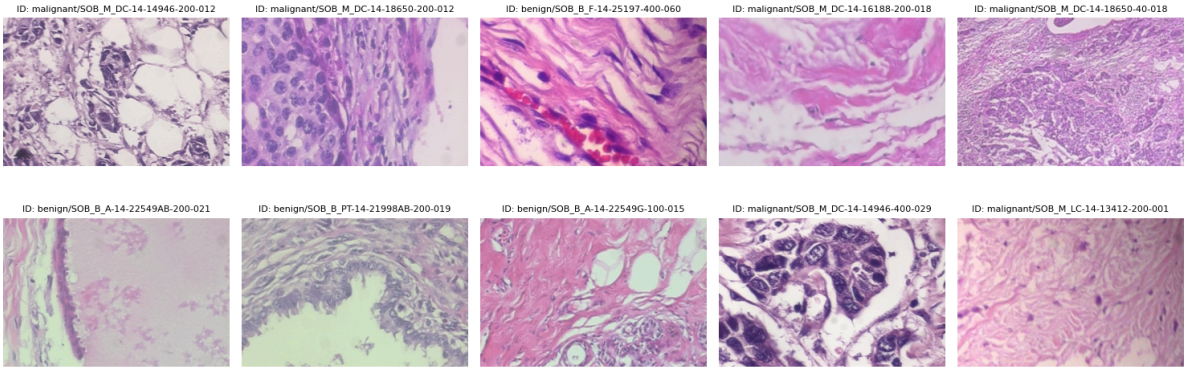


Figure 24. Random sample of the BreakHis subset dataset.

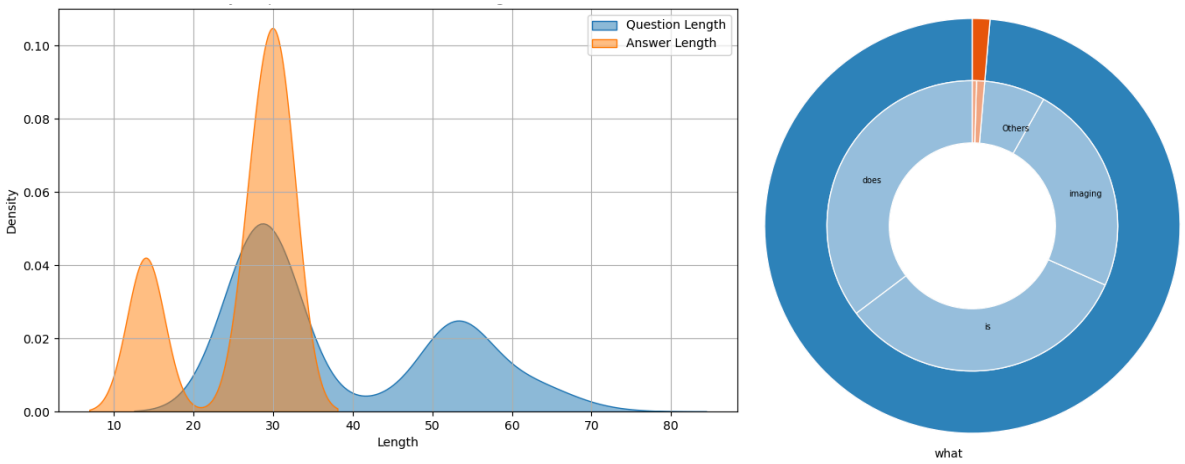


Figure 25. (left) OmniMedVQA BreakHis density of question and answer character length and (right) distribution of the first and second words in the test questions.

6.3. Data processing

Both the BCDR-F01 and OmniMedVQA BreakHis datasets are converted into BiomedGPT-compatible VQA formats by following a consistent, reproducible pipeline: each dataset is shuffled with a fixed random seed and stratified by answer labels to preserve the original distribution of question-answer pairs.

In the BCDR-F01 pipeline, closed questions use binary features (e.g., modality, lesion presence, organ, diagnosis) mapped directly to “yes” or “no,” while open questions draw on the corresponding class labels. All questions are generated from a predefined set of templates inspired on Path-VQA, VQA-RAD, and SLAKE. The resulting question-answer pairs are then split 70% train, 15% validation, and 15% test producing JSON files. These are then prepared into TSV and PKL files as BiomedGPT does for VQA-RAD to further fine tune.

Similarly, OmniMedVQA BreakHis entries, which already have paired questions and answers (e.g., malignancy status, subtype), are normalized and reformatted, then divided 70% train, 30% test.

6.4. Zero-shot inference

6.4.1. Method

BiomedGPT has the ability to answer biomedical questions in a free form manner at scale, without requiring retraining. This is a significant difference from earlier biomedical AI models such as BERT[3] or ViT[4] based models incapable of zero-shot prediction, or CLIP[5] based models that required a predefined answer (Zhang et al., 2024). As shown in **Figure 26**, BiomedGPT can generate answers by simply processing the input data.

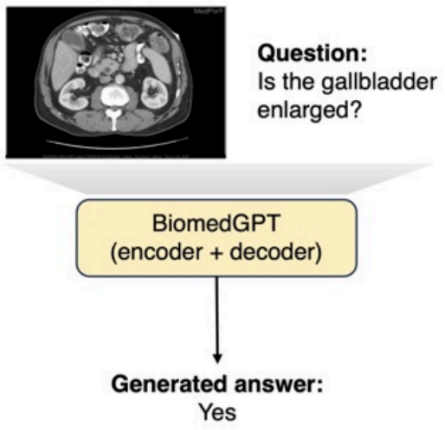


Figure 26. BiomedGPT-style zero-shot learning (Zhang et al., 2024).

6.4.2. Zero shot pipeline

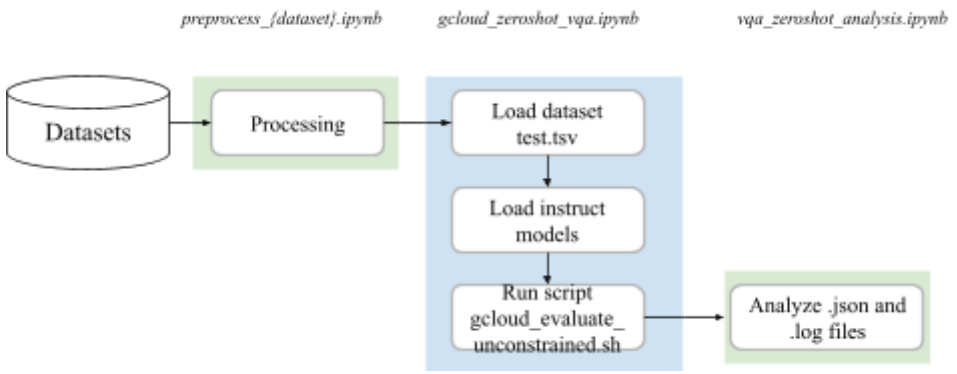


Figure 27. Zero-shot experiment pipeline across local (green) and cloud (blue) environments.

[3] Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained language model that learns deep bidirectional representations from unlabeled text using the Transformer architecture, enabling diverse NLP tasks (Devlin et al., 2019).

[4] Vision Transformer (ViT) applies the Transformer architecture, originally for NLP, directly to image classification by treating image patches as sequences (Dosovitskiy et al., 2021).

[5] Contrastive Language-Image Pre-training (CLIP) is an OpenAI model trained on a vast dataset of text-image pairs using contrastive learning, enabling cross-modal understanding and zero-shot capabilities (Radford et al., 2021).

As **Figure 27** illustrates, our zero-shot experiment pipeline involves both local and cloud environments. For each dataset, a specific local notebook, *preprocess_{dataset}.ipynb*, is used to transform the raw data into the structured training, validation, and test .tsv files, along with a .pkl file. The .tsv files contain the processed input data for BiomedGPT, while the accompanying .pkl file stores the answer mappings for model evaluation.

Zhang et al (2024) evaluate zero-shot inference using the test set and the script *evaluate_vqa_rad_unconstrained.sh*. For these experiments, we use *gcloud_evaluate_unconstrained.sh*, an adapted version designed for the cloud environment. This version changes the distributed launch so that it works with the GPU available from Google. It is also parameterized to facilitate testing different datasets.

The colab notebook *gcloud_zeroshot_vqa.ipynb* automatically downloads all three instruct models to run the script. Successful completion generates a log file with the details and score, and a csv file with the predictions made by the model. We analyze and plot the results locally using the notebook *vqa_zeroshot_analysis.ipynb*.

6.4.3. Results

VQA-RAD

The unique answer count and scores per model are shown in **Figure 28**. **Figure 29** presents a random image from the VQA-RAD test set, along with its question-answer pairs and the predictions from each model size. Details on the prediction distribution generated per model are shown in **Appendix 1**.

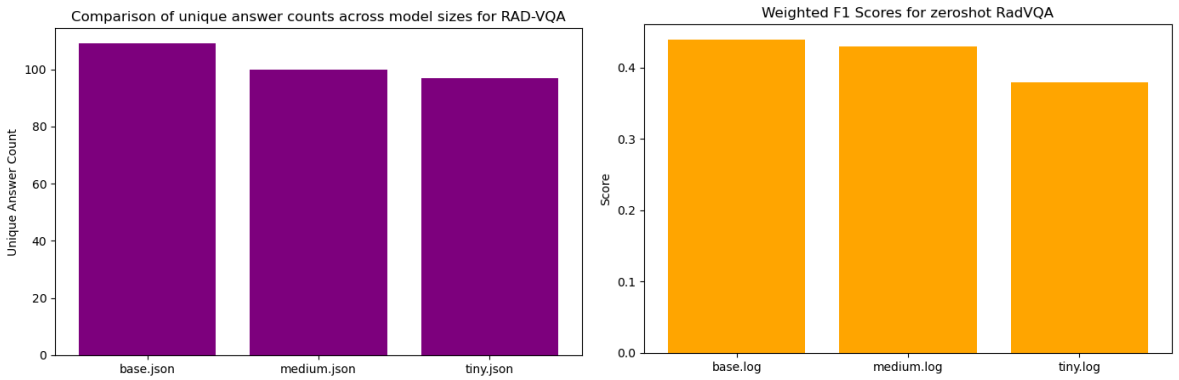


Figure 28. VQA-RAD unique answers (left) and weighted F1 scores (right) for instruct tuned models.

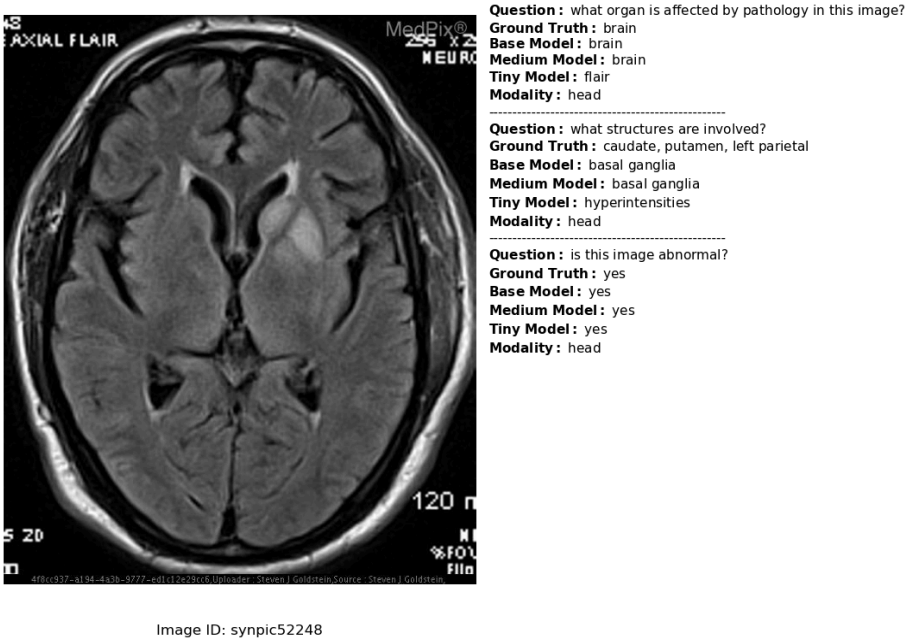
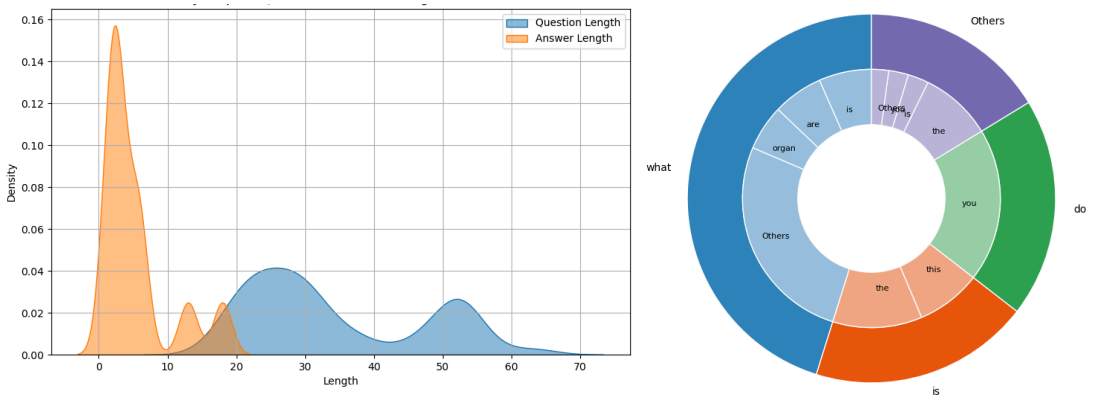


Figure 29. Comparison of BiomedGPT model inference with ground truth on a random VQA-RAD pair.

From these results one can observe that larger models show greater generalizability, evidenced by a higher count of unique answers. This is accompanied by an improvement in overall scoring. These results are in line with those reported by Zhang et al (2024). Do consider that parameters are kept unchanged from the repository, including the seed.

BCDR-F01

The generated dataset's distribution of open and closed questions, categorized by modality, lesion detection, organ, and diagnosis, closely mirrors that of VQA-RAD (**Figure 30**). Using the same zero-shot colab notebook, after the processed test.tsv file is uploaded, one can run the script parameterizing beam size[6] and specifying this new dataset.



[6] "beam size" refers to the number of top candidate sequences (answers) that a model keeps track of at each step of the decoding process. Exploring more potential answers can lead to higher quality results.

Figure 30. (left) BCDR-F01 test density of question and answer character length and (right) distribution of the First and Second words in the test questions.

The zero-shot experiment is repeated with the test set of this generated VQA dataset. The unique answer count and scores per model are shown in **Figure 31**. On the other hand, **Figure 32** presents a random image from the generated BCDR-F01 test set, along with its question-answer pairs and the predictions from each model size. Additionally, **Table 4** compares accuracy per question type and category. Details on the prediction distribution generated per model are shown in **Appendix 2**.

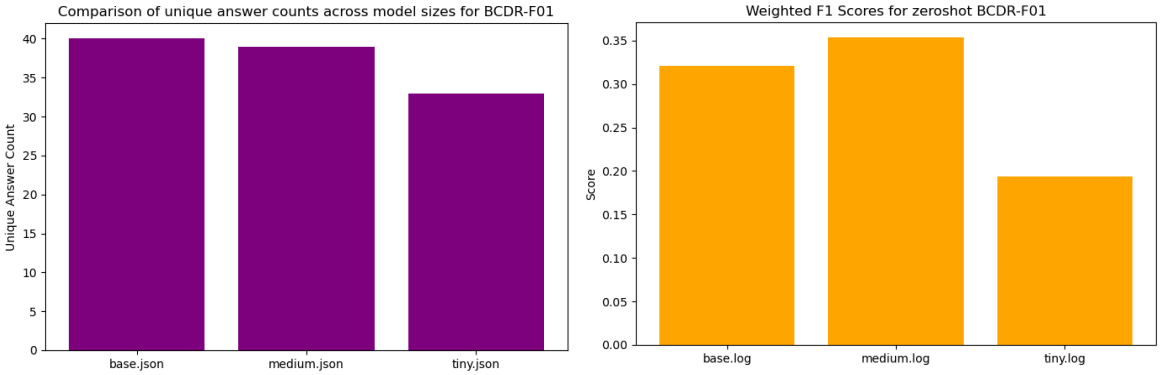


Figure 31. BCDR-F01 unique answers (left) and weighted F1 scores (right) for instruct tuned models.

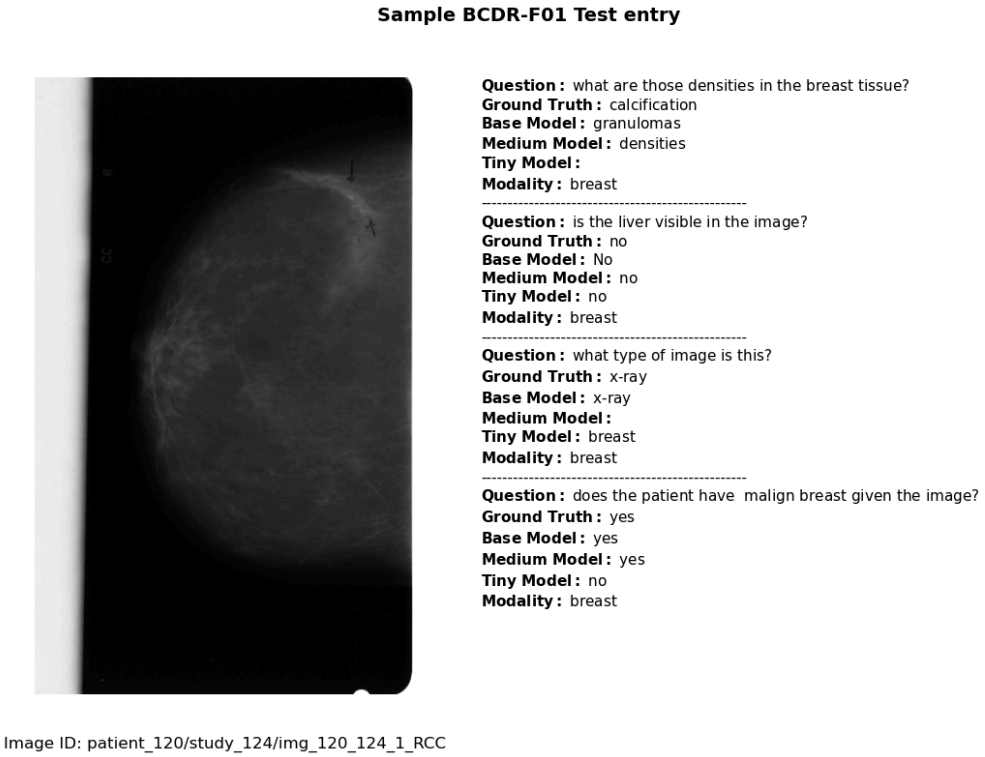


Figure 32. Comparison of model inference with ground truth on a random BCDR-F01 pair.

Table 4. Zero-shot accuracy for BCDR-F01 open and closed questions on a range of tasks.

	tiny_closed	tiny_open	medium_closed	medium_open	base_closed	base_open
calcification	0.30	0.00	0.92	0.00	0.87	0.07
classification	0.35	0.00	0.47	0.00	0.47	0.00
density	0.26	0.00	0.60	0.00	0.40	0.00
microcalcification	0.50	0.00	0.56	0.00	0.58	0.00
modality	0.23	0.06	0.87	0.00	0.47	0.59
organ	0.38	0.20	0.79	0.00	0.30	0.00

Once again we notice greater generalization on larger models. Interestingly, the medium model performed better in this seed. Both base and medium models had an impressive accuracy for closed questions, both noticeably higher than the tiny variant. There's a significant struggle for all model sizes to accurately answer open questions. Specially on classes that are rarely present in previous VQA datasets (density, microcalcification and 'breast' organ). However, the top answers on these questions reveal relevant instruction and alignment accuracy. As seen in **Appendix 3**, when asked to describe the density of the breast the medium model includes answers such as "high", "low", "relatively dense" or "relatively uniform". Similarly, base and medium variants answer with "carcinoma", "cancer", and "invasive ductal carcinoma" on classification.

BreakHis

Finally, we perform zero-shot on the OmniMedVQA BreakHis dataset. This test utilized a beam size of 10 to enable the evaluation of more comprehensive cancer subtyping answers. The unique answer count and scores per model are shown in **Figure 33**. In Figure **34,** a random image per ground truth is depicted, along with its question-answer pairs and the predictions from each model size.

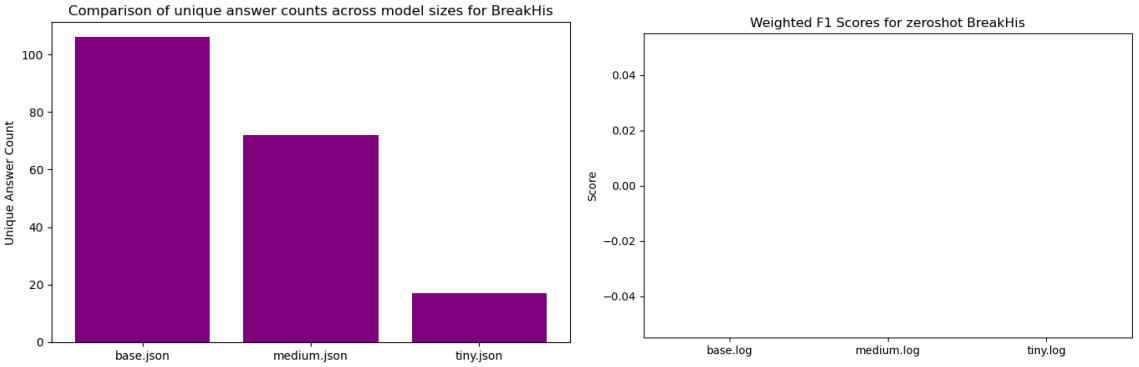


Figure 33. (Left) OmniMedVQA BreakHis unique answers and scores across instruct tuned models using BreakHis dataset. (Right) Note that on open-ended VQA, a larger beam size can decrease accuracy, as it produces more semantically varied answers that fail evaluation against ground truth.

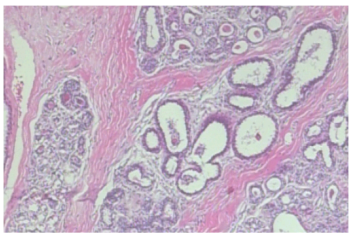


Image ID: benign/SOB_B_A-14-22549AB-40-012

Question : what imaging modality was used to capture this image?
Ground Truth : histopathology
Base Model : were obtained using light microscopy and
Medium Model : were obtained using 2x magnification.
Tiny Model :

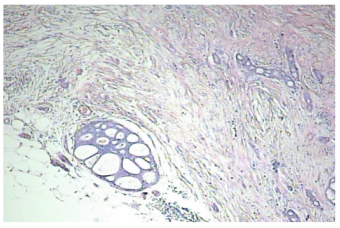


Image ID: benign/SOB_B_TA-14-21978AB-40-007

Question : what does this histological sample depict?
Ground Truth : benign breast histopathology
Base Model : of the ductal adenocarcinoma
Medium Model : showing a proliferation of lymphocytes and plasma cells
Tiny Model :

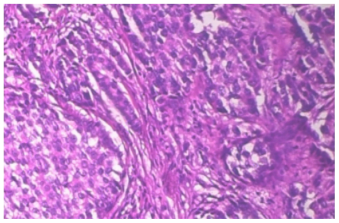


Image ID: malignant/SOB_M_DC-14-2773-100-026

Question : what does this image show?
Ground Truth : malignant breast histopathology
Base Model : of adenocarcinoma
Medium Model : of tumor cells
Tiny Model :

Figure 34. Comparison of model inference with the ground truths of OmniMedVQA BreakHis set. Note that no answers are reported by tiny model, possibly due to low answer confidence.

In this case, generalization across model sizes is more noticeable due to beam size. Given this parameter and considering all questions were open, the obtained 0 scores were expected. However, once again top answers reveal interesting instruction and alignment accuracy (**Appendix 4**). For modality, base and medium models recognize that images are obtained with a microscope, mention Haemotoxylin and Eosin, and attempt estimating the magnification. Similarly, these models mention “tissue”, “cells”, “ducts” and “breast” and attempt subtyping in some classification cases. While the answers include relevant breast cancer subtypes, they also contain unrelated subtypes from the lung, gastrointestinal, and brain domains, indicating the need for further refinement to improve prediction accuracy.

6.5. Fine-tuning

6.5.1. Method

BiomedGPT has an encoder-decoder cross-attention mechanism. Encoder-decoder LLMs are designed for sequence-to-sequence tasks. They translate sequences of textual input to sequences of textual output. In contrast to decoder-only LLMs, the input sequence does not represent a prompt but a genuine input that needs to be translated to a genuine output of unknown lengths. As shown in **Figure 35**, when working with pre-trained encoder-decoder LLMs, fine-tuning is usually done by fully fine-tuning the pre-trained LLM with a training dataset that contains input sequences and corresponding target sequences (Ferrari & Ginde, 2025).

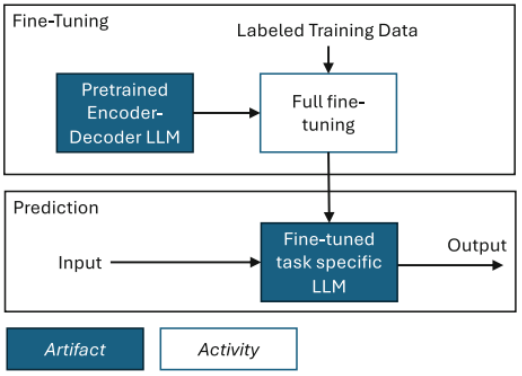


Figure 35. Schematic pipeline when using encoder-decoder LLMs (Ferrari & Ginde, 2025).

BiomedGPT adopts sequence to sequence learning for both pre-training and finetuning. Formally, given a sequence of tokens $x_{i,b}$ as input, where $i = 1,\ldots, I$ indexes the tokens in a data sample, and $b = 1,\ldots, B$ indexes a sample in the training batch. Let the model parametrized by θ autoregressive train by minimizing:

$$L_\theta(x_{1,1},\ldots, x_{i,b}) = -\sum_{b=1}^{B} log \prod_{i=1}^{I} p_\theta(x_{i,b}|x_{1,b},\ldots, x_{i-1,b}) = -\sum_{b=1}^{B}\prod_{i=1}^{I} logp_\theta(x_{i,b}|x_{<1,b}),$$

where x could refer to both linguistic and visual tokens in the context of BiomedGPT.

By minimizing the negative log-likelihood of each token conditioned on prior tokens, BiomedGPT learns to generate coherent answers from multimodal inputs. During fine-tuning on BCDR-F01, each example concatenates image embeddings with question tokens, and the model must sequentially predict answer tokens. Aligning the fine-tuning objective with pre-training ensures that learned biomedical language-vision correlations are effectively transferred to the VQA task, enabling accurate prediction of relevant responses.

6.5.2. *Fine tuning pipeline*

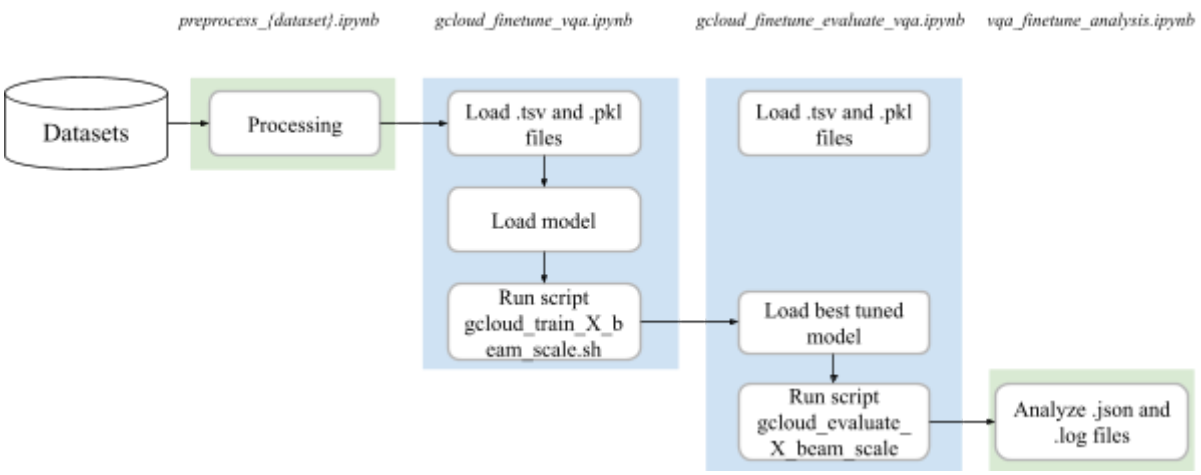


Figure 36. Fine tuning pipeline across local (green) and cloud (blue) environments.

Figure 36 shows our fine-tuning pipeline. Like zero-shot experiments, this requires processing the dataset and adapting the *train_vqa_rad_beam_scale.sh* script for the cloud environment. The colab notebook *gcloud_finetune_vqa.ipynb* is used to automate and facilitate loading the data and models to run it. An important difference is that training is performed on the base models rather than on the instruct variants used in the zero-shot experiments. Successful completion generates a file with the details and progression through epochs. It also saves 3 fine tuned models (last, best, and epoch).

For evaluation, the *evaluate_vqa_rad_beam_scale.sh* script is adapted into the cloud in *gcloud_evaluate_beam_scale.sh*. A colab notebook *gcloud_finetune_evaluate_vqa.ipynb* helps load the best tuned model checkpoint and test set to evaluate it with. A log and json file including the predictions are generated and analyzed similar to the zero-shot experiments using the local notebook *vqa_finetune_analysis.ipynb*.

6.5.3. *Results*

To study the effect of fine-tuning, we selected the medium and tiny non-instruct models as they performed the best and worst in the zero-shot experiments. It was also considered that fine-tuning takes much more time. For this reason, the max epochs were reduced from 100 to 15. This way, anyone can try these experiments on a free Colab session (4 hours max). Specifically, fine tuning the tiny variant took around 90 minutes, and the medium variant 110 minutes.

Figures 37 and **38** display the loss functions and scores per epoch for BiomedGPT tiny and BiomedGPT medium on the BCDR-F01 train and validation sets, respectively.

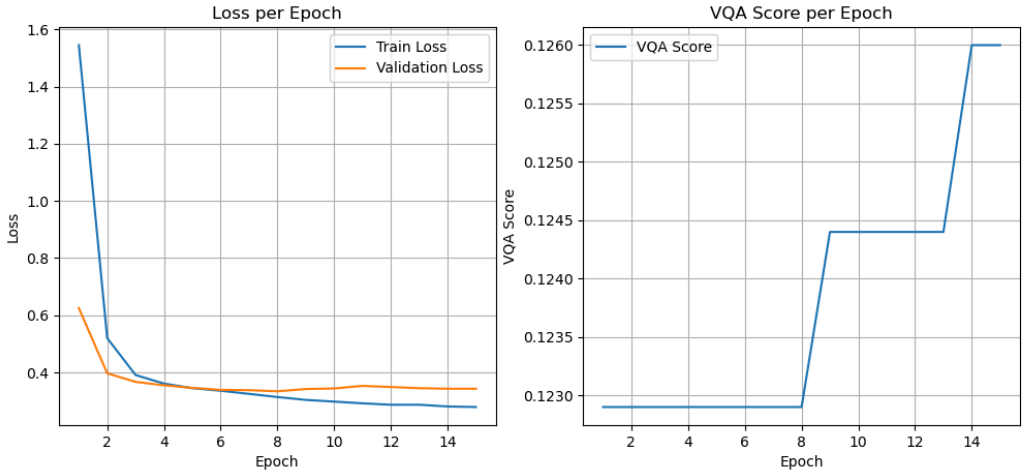


Figure 37. BiomedGPT tiny loss function and VQA score over 15 epochs on BCRD-F01

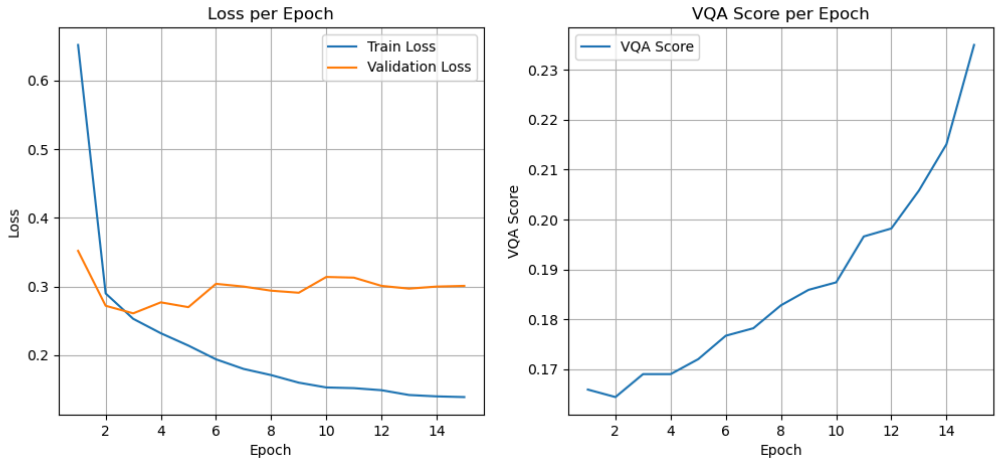


Figure 38. BiomedGPT medium loss functions and VQA score over 15 epochs on BCRD-F01

Comparing both models reveals important differences. Despite falling training loss in both models, the tiny version's VQA score remained stagnant compared to the medium version's progress. Both models also show a slight increase in the validation loss after the initial epochs. Although this signals overfitting, the performance on the VQA task continues improving. This suggests learning of relevant features, even if its general prediction accuracy on unseen data is not ideal.

After tuning, we evaluate both models as done in the zero-shot experiments. **Table 5** compares accuracy per category question type and category.

Table 5. Fine tune accuracy for BCDR-F01 open and closed questions on a range of tasks.

	tuned_tiny_closed	tuned_tiny_open	tuned_medium_closed	tuned_medium_open
calcification	0.00	0.00	0.14	0.38
classification	0.00	0.00	0.04	0.00
density	0.00	0.00	0.00	0.00
microcalcification	0.00	0.00	0.05	0.67
modality	0.00	0.81	0.34	0.00
organ	0.00	0.69	0.15	0.89

Compared to the zero-shot experiments, the fine-tuned models perform better on open-ended questions but underperform on closed ones. Additionally, the medium model demonstrates better generalization than the tiny variant. An analysis of the responses indicates instances of instruction misalignment, suggesting that instruction tuning plays a critical role in improving overall accuracy.

6.6. Overview

A high level overview of the experiments on BCDR-F01 are presented in **Table 6**.

Table 6. Summary of results with BCDR-F01 generated dataset

Dataset	Experiment	Model	Weighted F1 score	Accuracy		Instruction alignment accuracy
				Closed	Open	
BCDR-F01 generated VQA	zero shot	instruct BiomedGPT base	0.3210	0.3436	0.0431	Strong
		instruct BiomedGPT medium	.0.3533	0.7055	0.0000	Strong
		instruct BiomedGPT tiny	0.1935	0.5282	0.1108	Moderate
	fine tuning	tuned BiomedGPT medium (15 epoch)	0.2212	0.1166	0.3148	Weak
		tuned BiomedGPT tiny(15 epoch)	0.1244	0.0000	0.3000	None

Chapter 7

Conclusions and Future Work

7.1. Conclusions

By thoroughly researching FMs and experimenting on BiomedGPT, a prominent generalist biomedical FM, the primary objective of this research has been addressed. BiomedGPT has comparable performance on classification and lesion detection tasks to other FMs and ML/DL tasks. The experiments on a complex multimodal VQA task using both established and generated datasets show that BiomedGPT can effectively integrate visual and textual information to produce accurate, interpretable answers even under limited data conditions. Moreover, most of the secondary objectives were achieved with some limitations regarding availability, scope, and costs.

I. This research has successfully identified and implemented a generalist vision-language FM capable of processing both imaging and text data for unimodal and multimodal tasks.
II. To showcase the model's adaptability to imaging data, zero-shot and fine-tuning was performed on the BCDR-F01 and BreakHis datasets, which consist of mammograms and histopathology images, respectively.
III. Performance was assessed on VQA tasks that included questions on lesion, tumor, and cancer subtypes. Zero-shot performance showed promise, but fine tuning revealed instruction tuning is significant to reach these results.
IV. A concise and focused study was conducted to identify the key characteristics and performance differentiators of ML, DL, and FM in the context of breast imaging tasks. However, comparison to the explored task (i.e., VQA) was limited to BiomedGPT's recorded performance.
V. Potential for zero-shot learning on rare breast cancer subtypes was explored using the BreakHis dataset. However, the generation of answers including unrelated subtypes from other organs suggests that further refinement is necessary to improve specificity and clinical relevance.
VI. To analyze the interpretability and explainability of model outputs, BiomedGPT's responses to clinically grounded VQA prompts were examined, allowing assessment of how well the model aligns with diagnostic reasoning and whether its outputs can be meaningfully interpreted by clinicians.
VII. To investigate the potential for integrating multimodal data, we tackled the challenging task of VQA. Results show that comprehensive analysis is currently limited by the non complex nature of VQA datasets and reliance on instruction tuning.

Additionally, this research aims to advance the democratization of knowledge on FMs. My sincere gratitude to Zhang et al. (2024) for their fully transparent, detailed, and highly accessible model, which significantly aided my research, understanding, and experimentation. To broaden its impact, I have adapted their work to a free and accessible cloud platform to facilitate exploration to future researchers. The code organization of my work is shown in **Appendix 5**.

7.2. Limitations

A key limitation of this study is the nature of the datasets used for testing. Generated questions for BCDR-F01 are non complex nor diverse. Additionally, the use of older, non-digital mammography could impact the generalizability of the findings to modern imaging techniques. The subset of BreakHis from OmniMedVQA has similar issues on question quality and category diversity.

There are also practical limitations such as GPU access and costs. Results highlight the importance of model size, fine tuning strategies, and data volumes to reach state of the art performance. All of these incur a significant economic and time cost. While models with higher parameter size are available, open access is restricted to users with relevant academic credentials or behind a heavy paywall. Concerns on data governance are to be considered.

7.3. Future work

FMs can be easily adapted to several tasks and data modalities. In fact, BiomedGPT has already explored and proposed an architecture to handle 3D data. In this research only VQA was fully explored, but there are many other areas with relevant performance such as classification, text summarization, report generation, and captioning. Possible extensions of this work could focus on adapting and exploring these tasks on new (3D image) datasets, taking notice that an accurate evaluation may require medical validation. BiomedGPT was chosen due to its generative, generalist, and zero shot performance. However, as discussed in the taxonomy, there exist other FMs better suited for specific medical tasks. For instance, contrastive Generative TPMs can better approach rare disease diagnosis and zero-shot classification. Similarly, Hybrid TPMs combination of contrastive and generative pre training objectives adapt better to VQA and complex diagnosis tasks.

Bibliography

[1] Abnar S, Dehghani M, Neyshabur B, Sedghi H (2021) Exploring the limits of large scale pre-training. https://arxiv.org/abs/2110.02095

[2] Al-Antari, M. A., Han, S. M., & Kim, T. S. (2020). Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital X-ray mammograms. Computer methods and programs in biomedicine, 196, 105584.

[3] Arnold, M., Morgan, E., Rumgay, H., Mafra, A., Singh, D., Laversanne, M., Vignat, byra
J., Gralow, J. R., Cardoso, F., Siesling, S., & Soerjomataram, I. (2022). Current and future burden of breast cancer: Global statistics for 2020 and 2040. Breast (Edinburgh), 66, 15–23. https://doi.org/10.1016/j.breast.2022.08.010

[4] Audiffren, J. & Ostapuk, N. (2024) Tasks for LLMs and Their Evaluation. In *Large Language Models in Cybersecurity*. Springer. https://doi.org/10.1007/978-3-031-54827-7_6

[5] Azad, B., Azad, R., Eskandari, S., Bozorgpour, A., Kazerouni, A., Rekik, I., & Merhof, D. (2023). Foundational models in medical imaging: A comprehensive survey and future vision. arXiv. https://doi.org/10.48550/arXiv.2310.18689

[6] Bergmann, D. (2024). What is self-supervised learning?. IBM Think. https://www.ibm.com/think/topics/self-supervised-learning

[7] Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E et al (2021) On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258

[8] Brook, A., El-Yaniv, R., Isler, E., Kimmel, R., Meir, R., & Peleg, D. (2006). Breast cancer diagnosis from biopsy images using generic features and SVMs. IEEE Transactions on Information Technology in Biomedicine.

[9] Byra, M., Galperin, M., Ojeda‐Fournier, H., Olson, L., O'Boyle, M., Comstock, C., & Andre, M. (2019). Breast mass classification in sonography with transfer

learning using a deep convolutional neural network and color conversion. Medical physics, 46(2), 746-755.

[10] Chen, S., Cao, Z., Prettner, K., Kuhn, M., Yang, J., Jiao, L., Wang, Z., Li, W., Geldsetzer, P., Bärnighausen, T., Bloom, D. E., & Wang, C. (2023). Estimates and Projections of the Global Economic Cost of 29 Cancers in 204 Countries and Territories From 2020 to 2050. JAMA Oncology, 9(4), 465–472. https://doi.org/10.1001/jamaoncol.2022.7826

[11] Chougrad, H., Zouaki, H., & Alheyane, O. (2018). Deep convolutional neural networks for breast cancer screening. Computer methods and programs in biomedicine, 157, 19-30.

[12] Contra el Cáncer España. (2024). Breast Cancer (trans.). https://www.contraelcancer.es/es/todo-sobre-cancer/tipos-cancer/cancer-mama

[13] Cserni, G. (2020). Histological type and typing of breast carcinomas and the WHO classification changes over time. *Pathologica*, *112*(1), 25–41. https://doi.org/10.32074/1591-951X-1-20

[14] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186. https://aclanthology.org/N19-1423.pdf

[15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations. https://arxiv.org/abs/2010.11929

[16] European Parliament, & Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Official Journal of the European Union, L 1689. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

[17] Feng, H., Cao, J., Wang, H., Xie, Y., Yang, D., Feng, J., & Chen, B. (2020). A knowledge-driven feature learning and integration method for breast cancer diagnosis on multi-sequence MRI. Magnetic resonance imaging, 69, 40-48.

[18] Fernandes, R.C.M. (2022). Molecular Basis of Breast Cancer. In: Kim Hsieh, S.J., Morris, E.A. (eds) Modern Breast Cancer Imaging. Springer, Cham. https://doi-org.sire.ub.edu/10.1007/978-3-030-84546-9_1

[19] Ferrari, A., & Ginde, G. (2025). Handbook on Natural Language Processing for Requirements Engineering (1st ed. 2025.). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-73143-3

[20] Han, S., Kang, H. K., Jeong, J. Y., Park, M. H., Kim, W., Bang, W. C., & Seong, Y. K. (2017). A deep learning framework for supporting the classification of breast lesions in ultrasound images. Physics in Medicine & Biology, 62(19), 7714.

[21] Hassanien, A. E., & Kim, T. H. (2012). Breast cancer MRI diagnosis approach using support vector machine and pulse coupled neural networks. Journal of Applied Logic, 10(4), 277-284.

[22] He, X., Zhang, Y., Mou, L., Xing, E., & Xie, P. (2020). PathVQA: 30000+ Questions for Medical Visual Question Answering. arXiv. Retrieved from https://arxiv.org/abs/2003.10286v1

[23] Houssein, E. H., Emam, M. M., Ali, A. A., & Suganthan, P. N. (2021). Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review. Expert Systems with Applications, 167, 114161-. https://doi.org/10.1016/j.eswa.2020.114161

[24] Hu, Y., Li, T., Lu, Q., Shao, W., He, J., Qiao, Y., & Luo, P. (2024). OmniMedVQA: A New Large-Scale Comprehensive Evaluation Benchmark for Medical LVLM. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 22170–22183. https://doi.org/10.1109/CVPR52733.2024.02093

[25] Jiang, B., Bao, L., He, S., Chen, X., Jin, Z., & Ye, Y. (2024). Deep learning applications in breast cancer histopathological imaging: diagnosis, treatment, and prognosis. *Breast Cancer Research : BCR*, *26*(1), 137–17. https://doi.org/10.1186/s13058-024-01895-6

[26] Kaul, V., Enslin, S., & Gross, S. A. (2020). History of artificial intelligence in medicine. Gastrointestinal Endoscopy, 92(4), 807–812. https://doi.org/10.1016/j.gie.2020.06.040

[27] Karger, E.; Kureljusic, M. Artificial Intelligence for Cancer Detection—A Bibliometric Analysis and Avenues for Future Research. Curr. Oncol. 2023, 30, 1626–1647. https://doi.org/10.3390/curroncol30020125

[28] Kavlakoglu, E. & Stryker, C. (2024). What is AI?. IBM Think. https://www.ibm.com/think/topics/artificial-intelligence

[29] Khalaf, A. F., & Yassine, I. A. (2015, September). Novel features for microcalcification detection in digital mammogram images based on wavelet and statistical analysis. In 2015 IEEE international conference on image processing (ICIP) (pp. 1825-1829). IEEE.

[30] Lau, J. J., Gayen, S., Ben Abacha, A., & Demner-Fushman, D. (2018). A dataset of clinically generated visual questions and answers about radiology images. Scientific Data, 5(1), 180251–10. https://doi.org/10.1038/sdata.2018.251

[31] Li, H., Lin, L., & Zeng, H. (2024). *Machine Learning Methods* (1st ed. 2024.). Springer Nature Singapore. https://doi.org/10.1007/978-981-99-3917-6

[32] Liu, B., Zhan, L.-M., Xu, L., Ma, L., Yang, Y., & Wu, X.-M. (2021). Slake: A Semantically-Labeled Knowledge-Enhanced Dataset For Medical Visual Question Answering. *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 1650–1654. https://doi.org/10.1109/ISBI48211.2021.9434010

[33] Malik, S. S., Iqra, Akhtar, N., Fatima, I., Akram, Z., Masood, N., & Shakil Malik, S. (2020). Molecular Profiling of Breast Cancer in Clinical Trials: A Perspective. In *Essentials of Cancer Genomic, Computational Approaches and Precision Medicine* (pp. 313–332). Springer Singapore. https://doi.org/10.1007/978-981-15-1067-0_12

[34] Morris, E. A., & Kim Hsieh, S. J. (Eds.). (2022). Modern breast cancer imaging. Springer.

[35] Moura, D. C., López, M. A. G., Cunha, P., de Posada, N. G., Pollan, R. R., Ramos, I., Loureiro, J. P., Moreira, I. C., de Araújo, B. M. F., & Fernandes, T. C. (2013). Benchmarking datasets for breast cancer computer-aided diagnosis (CADx). In Lecture Notes in Computer Science (Vol. 8258, pp. 326-333).

Springer.

[36] Prabusankarlal, K. M., Thirumoorthy, P., & Manavalan, R. (2015). Assessment of combined textural and morphological features for diagnosis of breast masses in ultrasound. Human-centric Computing and Information Sciences, 5, 1-17.

[37] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. arXiv. https://arxiv.org/abs/2103.00020

[38] Roy, K., Banik, D., Bhattacharjee, D., & Nasipuri, M. (2019). Patch-based system for classification of breast histology images using deep learning. Computerized Medical Imaging and Graphics, 71, 90-103.

[39] Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. (1959). IBM Journal of Research and Development. https://doi.org/10.1147/rd.33.0210

[40] Schneider, J., Meske, C., & Kuss, P. (2024). Foundation Models: A New Paradigm for Artificial Intelligence. *Business & Information Systems Engineering*, *66*(2), 221–231. https://doi.org/10.1007/s12599-024-00851-0

[41] Shah, S. M., Khan, R. A., Arif, S., & Sajid, U. (2022). Artificial intelligence for breast cancer analysis: Trends & directions. *Computers in Biology and Medicine*,*142*,105221–105221. https://doi.org/10.1016/j.compbiomed.2022.105221

[42] Singh, A., & Singh, K. K. (2025). Multimodal Generative AI (1st ed. 2025.). Springer Nature Singapore. https://doi.org/10.1007/978-981-96-2355-6

[43] Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016). A Dataset for Breast Cancer Histopathological Image Classification. IEEE Transactions on Biomedical Engineering, 63(7), 1455–1462. https://doi.org/10.1109/TBME.2015.2496264

[44] Szeliski, R. (2022). Computer Vision : Algorithms and Applications (2nd ed. 2022.). Springer International Publishing. https://doi.org/10.1007/978-3-030-34372-9

[45] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., & Polosukhin, I. Attention Is All You Need. (n.d.). *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1706.03762

[46] Wajid, S. K., & Hussain, A. (2015). Local energy-based shape histogram feature extraction technique for breast cancer diagnosis. Expert Systems with Applications, 42(20), 6990-6999.

[47] Wang, J., & Chen, Y. (2023). Introduction to Transfer Learning : Algorithms and Practice (1st ed. 2023.). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-7584-4

[48] World Health Organization. (2025). *Cancer*. https://www.who.int/health-topics/cancer

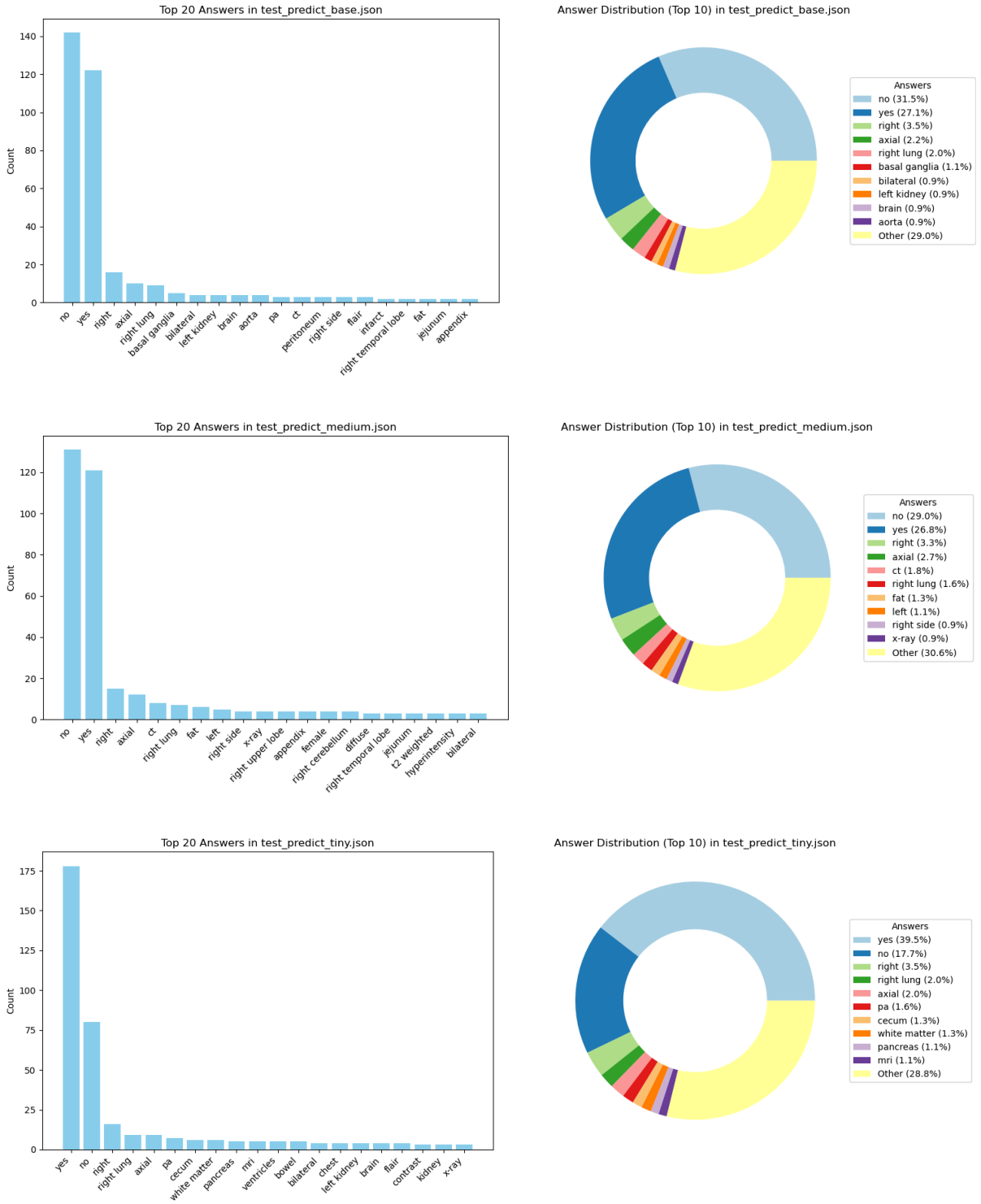
[49] Wu, W. J., Lin, S. W., & Moon, W. K. (2015). An artificial immune system-based support vector machine approach for classifying ultrasound breast tumor images. Journal of digital imaging, 28, 576-585.

[50] Yang, Z., Ran, L., Zhang, S., Xia, Y., & Zhang, Y. (2019). EMS-Net: Ensemble of multiscale convolutional neural networks for classification of breast cancer histology images. Neurocomputing, 366, 46-53.

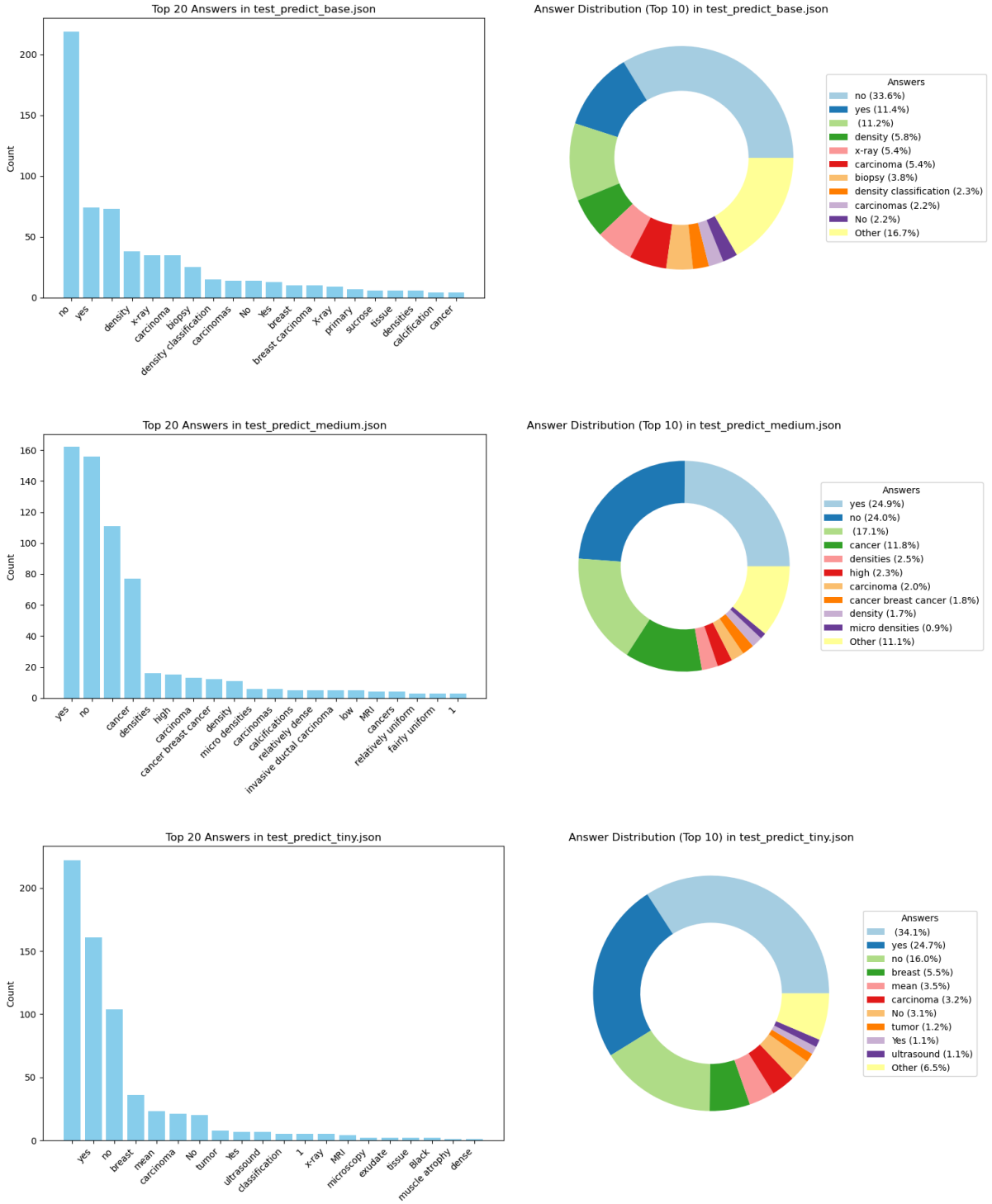
[51] Zhang, K., Zhou, R., Adhikarla, E., Yan, Z., Liu, Y., Yu, J., Liu, Z., Chen, X., Davison, B. D., Ren, H., Huang, J., Chen, C., Zhou, Y., Fu, S., Liu, W., Liu, T., Li, X., Chen, Y., He, L., … Sun, L. (2024). BiomedGPT: A Unified Biomedical Generative Pre-trained Transformer for Vision, Language, and Multimodal Tasks. Arxiv. https://arxiv.org/html/2305.17100v2

[52] Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., Peng, H., Li, J., Wu, J., Liu, Z., Xie, P., Xiong, C., Pei, J., Yu, P. S., & Sun, L. (2024). A comprehensive survey on pretrained foundation models: a history from BERT to ChatGPT. International Journal of Machine Learning and Cybernetics. https://doi.org/10.1007/s13042-024-02443-6

Appendix



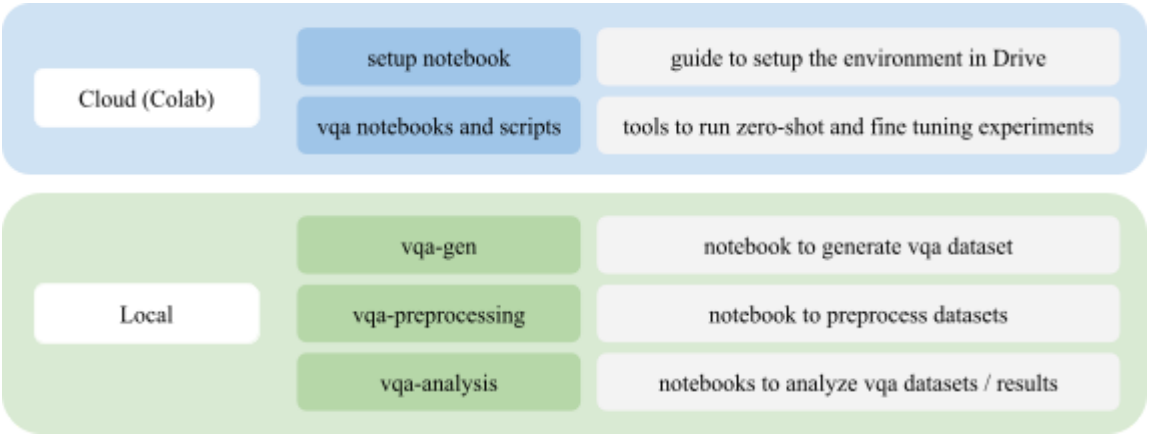
Appendix 1. Prediction distribution on zero shot VQA-RAD tests on instruct models. (left) Bar plot comparing the top 20 answers (right) Pie plot with the top 10 categories, ‘other’ being the grouping of the remaining



Appendix 2. Prediction distribution on zero shot BCDR-F01 tests on instruct models. (left) Bar plot comparing the top 20 answers (right) Pie plot with the top 10 categories, ‘other’ being the grouping of the remaining.

Appendix 3. Top 5 answers per category on open questions from the zero shot BCDR-F01 testset on instruct models.

Appendix 4. Prediction distribution on zero shot OmniMedVQA BreakHis tests on instruct models. (left) Bar plot comparing the top 20 answers (right) Pie plot with the top 10 categories, ‘other’ being the grouping of the remaining.



Appendix 5. Code structure for cloud and local environment use.