# Data Exploration of OmniMed-VQA

## Introduction

`OmniMedVQA` is a large-scale multimodal medical dataset with 88,996 QA items spanning 42 datasets [1]. It combines over 82,000 unique medical images with visual question answering (VQA) items covering five primary question types: Disease Diagnosis, Anatomy Identification, Modality Recognition, Other Biological Attributes, and Lesion Grading. This exploration summarizes dataset statistics, visual examples, modalities, and key challenges for modeling.

## Data Cleaning and Schema Consistency

While inspecting the JSON files, we found that `Chest CT Scan.json` contains a single entry using the key `modality` instead of `modality_type`.

We automatically correct this entry so that `modality` is renamed to `modality_type` for consistency.

## Dataset Statistics

### Number of Samples

- QA Items: 88,996
- Unique images: 82,059
- Datasets represented: 42

### Question Type Distribution

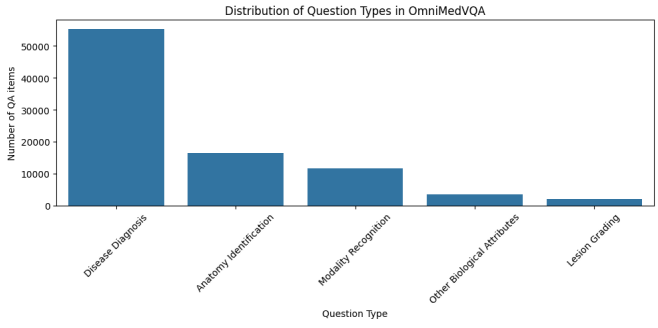



Figure 1: Question Type Distribution

The dataset is heavily dominated by Disease Diagnosis (55,387 items), followed by Anatomy Identification (16,448) and Modality Recognition (11,565).

Less common types such as Other Biological Attributes (3,498) and Lesion Grading (2,098) may require special attention during modeling to avoid underfitting.

### Ground Truth Answer Distribution

Some question types are heavily skewed toward a few frequent answers:

- Disease Diagnosis: `No` and `No, It's normal.` account for ~7,400 QA items.
- Modality Recognition: `MR` and `CT` dominate.

Some question types also contain answers that appear very rarely (sometimes only once). For example:

- Modality Recognition: "Histopathology." appears 8 times.
- Disease Diagnosis: "Fundus neoplasm" appears once.

This sparsity could make learning on rare classes challenging.

Some semantically identical answers differ in punctuation, capitalization, or minor wording, e.g.:

- `x_ray.` vs `X-ray`
- `Dermoscopic imaging` vs `Dermoscopy` vs `Dermoscopy.`
- `Fundus photography` vs `fundus photography.` vs `fundus photography`

Preprocessing steps such as lowercasing, stripping punctuation, and mapping variants to canonical forms may be beneficial. Despite the long-tail and answer variability, all major modalities and question types are represented, which is promising for building a generalizable multimodal model.

### Dataset-Level Distribution

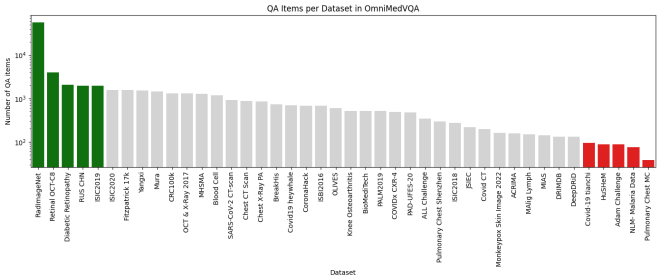



Figure 2: Number of QA Items per Dataset

While RadImageNet alone contributes 56,697 QA items (>60% of the total), several datasets at the bottom (e.g., Pulmonary Chest MC with 38 items) are very small. This imbalance in datasets isn't necessarily an issue as long as all modalities are adequately represented.

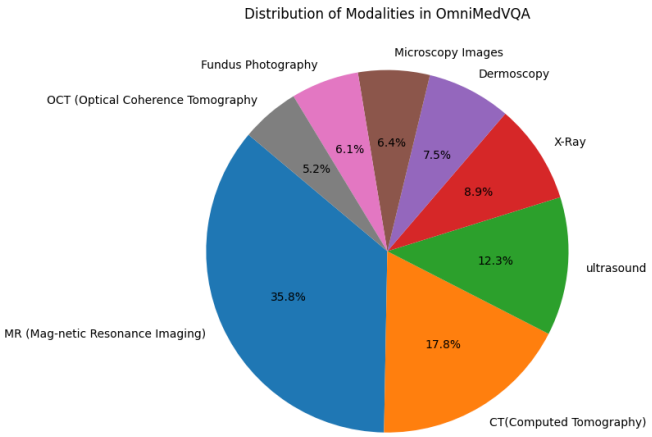



Figure 3: Distribution of Modalities

### Modalities

The OmniMedVQA dataset includes 8 distinct modalities. While MR dominates with ~35.8% of QA items, followed by CT (~17.8%) and Ultrasound (~12.3%), the less frequent modalities such as OCT (5.2%), Fundus Photography (6.1%), and Microscopy Images (7.5%) still have a substantial number of QA items (4,646–5,680), which should be sufficient for model training.

Although there is a skew toward MR and CT, all clinically relevant modalities are represented, reducing the risk that models will completely ignore underrepresented modalities. However, care may still be needed to ensure that rare modalities are weighted during training or evaluation.

## Visual Question Answering Examples

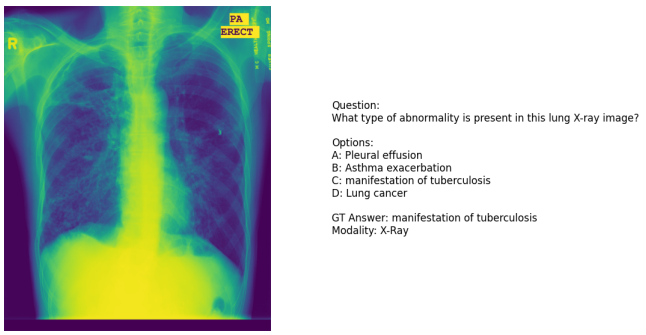



Figure 4: Modailty question from pulmonary chest dataset

Within each dataset, there is a diversity in the types of questions being asked for a given image. The first image presents a modality question, which asks about the imaging technique used (such as X-ray, CT, or MR), while the second image is a disease diagnosis question. All VQA items include multiple options relevant to the question posed, the ground truth, and the modality for the image.

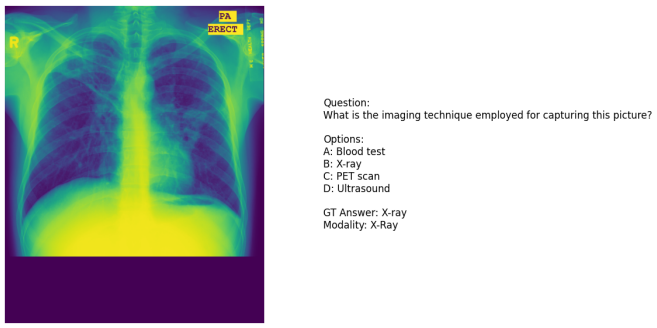



Figure 5: Disease question from pulmonary chest dataset

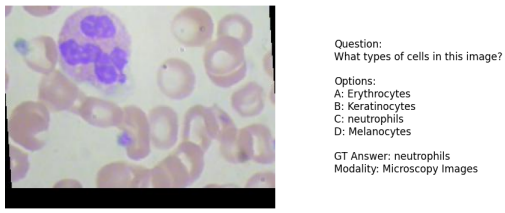
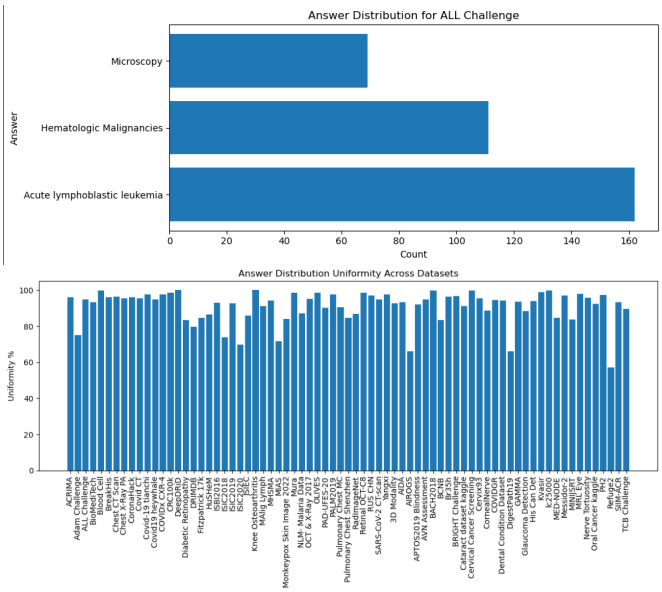


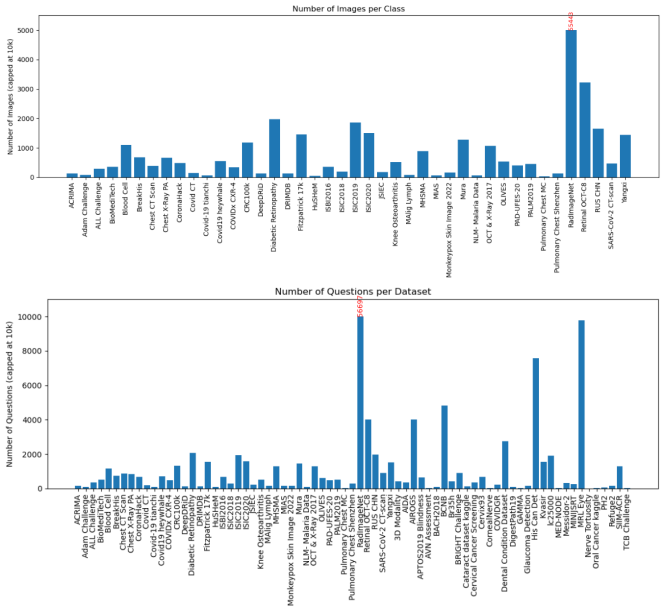

Figure 6: Anatomy question from blood cell datset

Different modalities for imaging are used across all datasets but the same problem type can appear across different datasets. Questions are framed similarly across datasets creating consistency in the OmniMedVQA dataset. We also see that the quality of imaging can vary between datasets. Some of the images are high quality images using technology such as X-ray or a CT scan. But some of the images are camera pictures of human subjects with some background of the environment. In this example it is a picture taken of the output from a microscope. From the examples seen the text length of questions are short and direct with answer options being just a few words.

## Challenges




Many of the questions show highly skewed answer distributions, with one option dominating while others are rarely chosen. This strong answer bias makes the data harder to interpret and can lead models to overfit to the majority choice rather than learning meaningful patterns.




Even in the questions and images we can see large numbers of class imbalance. For example, RadImageNet accounts for 55k questions and images with the next closest one being 10k and 3k per plot respectively. Such imbalance makes it harder for models to learn meaningful patterns from underrepresented classes.

## References


[1] Y. Hu *et al.*, "OmniMedVQA: A new large-scale comprehensive evaluation benchmark for medical LVLM," *arXiv preprint arXiv:2402.09181*, 2024.