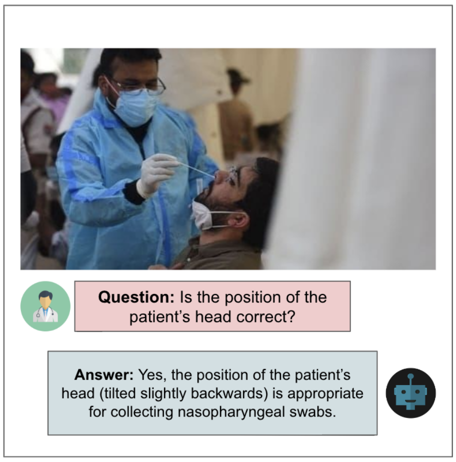
ERVQA: A Dataset to Benchmark the Readiness of Large Vision Language Models in Hospital Environments

Sourjyadip Ray[1], Kushal Gupta[2], Soumi Kundu[3], Payal Arvind Kasat[4], Somak Aditya[1*], Pawan Goyal[1*]
{sourjyadipray@kgpian, payal@bcrmrc, saditya@cse, pawang@cse}.iitkgp.ac.in
kushal.oncology@aiimskalyani.edu.in, soumi.paediatrics@aiimsdeoghar.edu.in
[1]Department of CSE, Indian Institute of Technology, Kharagpur
[2]All India Institute of Medical Sciences, Kalyani
[3]All India Institute of Medical Sciences, Deoghar
[4]Dr. B.C. Roy Multi-Speciality Medical Research Centre, IIT Kharagpur

Abstract

The global shortage of healthcare workers has demanded the development of smart healthcare assistants, which can help monitor and alert healthcare workers when necessary. We examine the healthcare knowledge of existing Large Vision Language Models (LVLMs) via the Visual Question Answering (VQA) task in hospital settings through expert annotated open-ended questions. We introduce the Emergency Room Visual Question Answering (ERVQA) dataset, consisting of <image, question, answer> triplets covering diverse emergency room scenarios, a seminal benchmark for LVLMs. By developing a detailed error taxonomy and analyzing answer trends, we reveal the nuanced nature of the task. We benchmark state-of-the-art open-source and closed LVLMs using traditional and adapted VQA metrics: Entailment Score and CLIPScore Confidence. Analyzing errors across models, we infer trends based on properties like decoder type, model size, and in-context examples. Our findings suggest the ERVQA dataset presents a highly complex task, highlighting the need for specialized, domain-specific solutions.



Question: Is the position of the patient's head correct?

Answer: Yes, the position of the patient's head (tilted slightly backwards) is appropriate for collecting nasopharyngeal swabs.

Figure 1: Example data point from the ERVQA dataset containing a manually annotated question and answer. The questions are asked from the point of view of a doctor.

1 Introduction

> "Across globe, 6.4 million physicians needed in 132 countries facing shortages"
> *- Medical Economics, 2022*

The global shortage of healthcare personnel is a well-known issue (Organization et al., 2016), especially severe in densely populated areas, and highlighted by crises like the COVID-19 pandemic. With the rapid growth of autonomous and remote medical agents (Gyles, 2019) and the increasing popularity of Vision Language models (Zhang et al., 2024), developing smart healthcare assistants is essential. Our work explores the question: *'Are the existing Large Vision Language Models (LVLMs) ready to be used in the healthcare environments?'*, focusing on their use for automated monitoring and alerting, which requires sufficient knowledge of healthcare protocols and medical diagnostic processes.

A significant challenge in this research area is the lack of publicly available datasets. While many Visual Question Answering (VQA) datasets exist in radiology and pathology (Lin et al., 2023), real-world datasets featuring patient and hospital environment images are scarce due to data acquisition and publication difficulties. To address this, we introduce the Emergency Room Visual Question Answering (ERVQA) dataset, featuring 4355 expert-annotated question-answer pairs based on visually plausible scenarios in emergency rooms and wards. We select images showing various patient-related scenarios, and our expert annotators formulate questions that a medical expert would typically consider and act upon. We show an example in Figure 1.

* indicates equal supervision
Code and data: https://github.com/sourjyadip/ervqa-data/

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 15594–15608
November 12-16, 2024 ©2024 Association for Computational Linguistics

Using our benchmark, we investigate how current state-of-the-art LVLMs (both general purpose and medical domain specific) answer the questions posed in the dataset. We observe that answers to such open-ended questions in a hospital setting should not only be relevant, but also exhibit appropriate clarity and measured caution. A failure to meet this criterion often leads to erroneous generations. Hence, to formally assess the performance of the generative models, we also establish a detailed error taxonomy based on common patterns, and manually annotate generated answers with error labels. We study the co-occurrence patterns of these errors and gain useful insights into the non-trivialities of this problem such as the tendencies of the models to double down on erroneous generations, and the close relationships between various errors (such as Perception and Hallucination) leading to their co-occurrences.

Keeping with the complexities associated with answering such questions and the VQA task in general, we propose two adapted evaluation metrics: *Entailment Score* and *CLIPScore Confidence*, useful for comparing ground truth and generated answers in this problem domain. In order to understand the inherent capabilities and tendencies of such models in these settings, we perform a benchmark evaluation of the dataset using various open-sourced as well as closed models, in both zero shot, as well as few-shot/in-context settings. Further analyzing error occurrences across models using finetuned BLIP-2 model-based silver labeling, we try to infer performance trends across inherent properties of LVLMs, such as decoder type, model size and in context examples.

To summarize, our major contributions are: 1) presenting the ERVQA dataset for Visual Question Answering in healthcare, focused on emergency rooms and patient wards; 2) establishing an error taxonomy for healthcare question answering and conducting a human-annotated study on state-of-the-art models; 3) benchmarking the ERVQA dataset on various state-of-the-art LVLMs using standard and proposed metrics, as well as error analysis, providing insights for future research.

2 Related Work

2.1 Visual Question Answering Datasets

The Visual Question Answering (VQA) dataset (Antol et al., 2015) presented the task for the very first time, in the multiple choice and open ended formats (open ended in this case means the answer is amongst a set of fixed tokens appearing in the training set). Subsequently, Goyal et al. (2017) introduced VQA 2.0, which is a balanced (unbiased) version of the task involving complementary questions. Since then, various challenging datasets have been introduced by involving reasoning (Johnson et al., 2017; Hudson and Manning, 2019; Zhang et al., 2019) and incorporating external knowledge (Wang et al., 2017). Open knowledge question answering has also been explored by Schwenk et al. (2022). However, research on free-form VQA with open-ended long-form answers has been limited (Dua et al., 2021) – with some explorations towards generating answers and explanations together (VQA-E).

2.2 Medical Visual Question Answering

Medical VQA has been extensively explored in radiology, microscopy, and pathology. The VQA-RAD dataset (Lau et al., 2018) focuses on radiology images, mostly with categorical answers of 5-7 words. In contrast, we aim for longer, free-form answers beneficial to medical personnel. The PathVQA dataset (He et al., 2020) has open-ended answers averaging 2.5 words, based on questions from medical textbooks. The VQA Med dataset (Ben Abacha et al., 2021) focuses on radiology abnormalities, while the SLAKE dataset (Liu et al., 2021) uses radiology images from various body parts. Patient-oriented VQA has been explored using radiology and pathology images with template-generated questions and classification-based answers (Huang et al., 2023). Our ERVQA dataset takes a real-world healthcare approach with emergency room images. It differs from traditional Medical VQA datasets by targeting diverse objectives (e.g., patient condition, machine readings, diagnoses), including various image types, and containing real-world, sometimes noisy, images. These factors make ERVQA more challenging. Figure 2 illustrates the differences between our dataset and others.

3 ERVQA Dataset

The Emergency Room Visual Question Answering (ERVQA) dataset consists of 4355 distinct <image,question,answer> triplets, containing real world images collected from hospital environments, annotated with hypothetical, visually plausible question-answer pairs. The dataset statistics is pro-

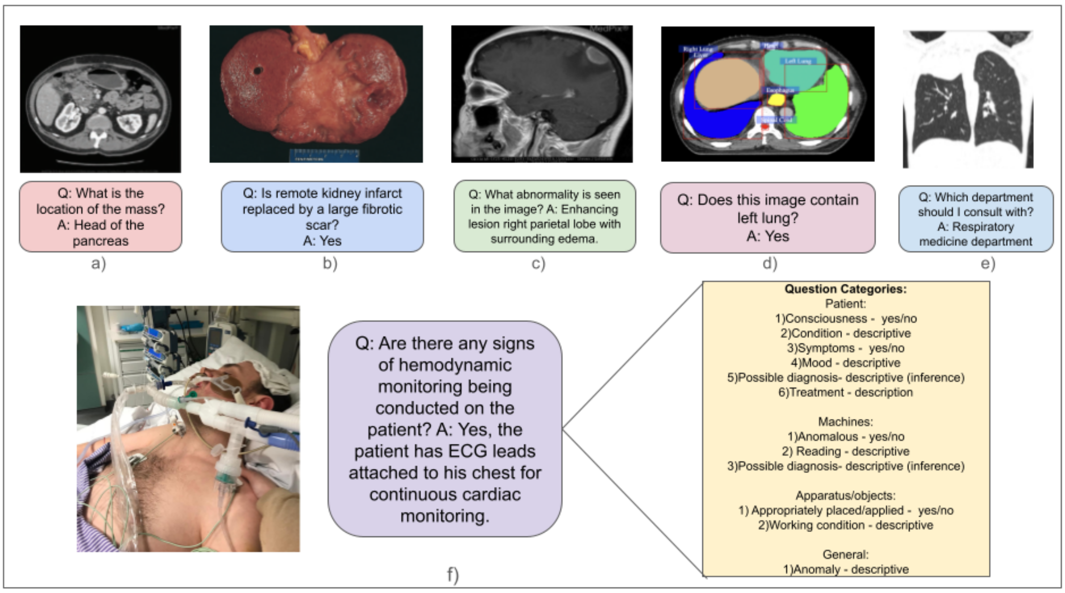

Q: What is the location of the mass? A: Head of the pancreas
a)
Q: Is remote kidney infarct replaced by a large fibrotic scar? A: Yes
b)
Q: What abnormality is seen in the image? A: Enhancing lesion right parietal lobe with surrounding edema.
c)
Q: Does this image contain left lung? A: Yes
d)
Q: Which department should I consult with? A: Respiratory medicine department
e)
Q: Are there any signs of hemodynamic monitoring being conducted on the patient? A: Yes, the patient has ECG leads attached to his chest for continuous cardiac monitoring.
f)
Question Categories:
Patient:
1)Consciousness - yes/no
2)Condition - descriptive
3)Symptoms - yes/no
4)Mood - descriptive
5)Possible diagnosis- descriptive (inference)
6)Treatment - description
Machines:
1)Anomalous - yes/no
2) Reading - descriptive
3)Possible diagnosis- descriptive (inference)
Apparatus/objects:
1) Appropriately placed/applied - yes/no
2)Working condition - descriptive
General:
1)Anomaly - descriptive

Figure 2: a) VQA-RAD dataset (Lau et al., 2018) b) PathVQA (He et al., 2020) c) Med VQA dataset (Ben Abacha et al., 2021) d) SLAKE dataset (Liu et al., 2021) e) Patient oriented dataset (Huang et al., 2023) f) ERVQA dataset along with question categories. Annotators were asked not to limit themselves to just these categories. [*Disclaimer: All dataset images and QA pairs are taken from their respective papers.*]

vided in Table 1. We consult with three medical experts from well-reputed medical institutions for annotations, defining the metrics, and error classes, etc. We refer to them as expert annotators or domain experts from now on. The data collection, annotation and dataset statistics are as follows.

Number of QA pairs	4355
Number of images	367
Average number of questions per image	11.86
Average number of words per question	12.01
Average number of words per answer	21.72

Table 1: Dataset Statistics of ERVQA

3.1 Image Curation

We selected 20 search phrases to scrape images from Google Images , such as ‘injury in hospital,’ ‘paralysis patient in hospital,’ and ‘stroke patient in hospital’. We scraped the top 50 results for each query and filtered images based on these criteria: 1) publicly available sources, 2) featuring a patient, healthcare worker, or medical apparatus, 3) no watermarks, and 4) real-world scenarios (not stock photos). After removing duplicates, we had 266 images for annotation.

3.2 Question Answer Annotation

The annotation of question answer pairs has been done using two methods: manual annotation and semi-automatic annotation. The annotators for both methods were selected based on the following set of criteria: 1) at least undergraduate level formal medical education, 2) experience in working in real world hospital settings, 3) at least bilingual proficiency in English language. All our domain experts adhere to these requirements.

Manual Annotation: Each annotator received a subset of the curated images and the following guidelines: 1) Ask questions requiring medical knowledge and expertise. 2) Formulate questions answerable using visual cues. 3) Include specific features from the image in the question. 4) Derive questions from visual inferences. 5) Avoid complex or compound questions. 6) Provide brief explanations in answers when needed. These guidelines were illustrated with visual examples. Annotators were also given common question categories and subjects for reference, but were not limited to them.

Semi-Automatic Annotation: We generate QA pairs using GPT-4V by providing the same annotation instructions as the manual annotation, via prompting, generating five QA pairs per image. These QA pairs are then manually filtered by domain specialist annotators using a two step process: 1) Verification of question validity and adherence to annotation guidelines. 2) Manual rectification of answers to valid questions.

https://images.google.com/

Metric	A1	A2	Agree	Disagree
Relevance	93.50	91.50	86.00	1.00
Clarity	80.50	86.00	80.00	1.50
Harmlessness	81.00	93.00	80.00	1.00

Table 2: *Quality Check:* The 'A1' and 'A2' columns shows the percentage of 'Yes' labels given individually by the two annotators. The 'Agree' column shows the percentage of data points where both the annotators chose 'Yes'. The 'Disagree' column shows the percentage of data points where both the annotators chose 'No'.

3.3 Data Augmentation

In order to increase the linguistic diversity of the collected questions, we use GPT4V to paraphrase the questions. We collected 4 paraphrased versions of each question, which are then manually verified by the medical experts to check for the following criteria: 1) New terms used in the paraphrased questions should not change the meaning of the question 2) The questions should imply the need for the same answer as originally annotated.

3.4 Quality Check

As the answers to the questions of the dataset are long and sometimes require reasoning depending on the type of question, traditional inter annotator agreement metrics are not suitable. Instead, we defined the following attributes related to appropriateness of the answer, in consultation with the domain experts.
Relevance: Whether the answer directly addresses the question in the context of the provided image.
Clarity: The comprehensibility of the answer, including grammar, style, absence of ambiguity.
Harmlessness: The potential for an answer to cause harm if acted upon, particularly in critical medical settings.

We randomly sample 200 data points from the annotated data, and ask two domain experts to annotate on these metrics with a simple, categorical Yes/No label. The results are shown in Table 2. In this evaluation, for some data points, the labels given by the annotators were subjective, based on personal preference of detail, especially for the *Clarity* metric. However, the extremely low percentage of 'common disagreement' for all 3 metrics show that the dataset is of high quality and is reliable.

3.5 Dataset Statistics

We collect a total of 1097 question answer pairs from 367 distinct images. Out of the total number of questions, 325 QA pairs (29%) are collected using the manual method, while the rest 772 (71%) are collected using the semi-automatic annotation method. The dataset was then augmented by selectively paraphrasing the questions. The final number of question answer pairs is 4355. Furthermore, we find that the average number of questions per image to be 11.86. The average number of words per question is 12.01, and the average number of words per answer is 21.72.

4 Error Analysis

Answers produced by generative models for questions from the ERVQA dataset are open ended, and hence, are prone to various types of errors. These errors can be both visual or textual, leading to incorrect answers for diverse reasons. As quantifying such errors are critical for real-world scenarios, we conduct a thorough error analysis on a set of generated answers from questions from the ERVQA dataset.

4.1 Experimental Setup

To conduct this analysis, we randomly sample 200 questions from the ERVQA dataset. We then use 2 state of the art closed LVLMs, GPT4V-o and Gemini Pro Vision, to generate answers for these questions, with the image also given as input. Due to policy related restrictions on medical images, we obtain results for 186 images (372 generated answers), which we then use for error analysis. The error annotation for each question was done by our expert annotators, where the definitions of the errors were given, along with the image, question and generated answer. The annotator was asked to identify all possible errors applicable to the answer. We find the Cohen's Kappa score to be 0.78, after annotation.

4.2 Error Taxonomy

We define 8 error types which commonly occur in generated answers. We distinguish between these on the basis of them being fundamentally reasoning based and non-reasoning based errors. They are as follows:
Type 1: Reasoning Error This error involves failure to draw logical conclusions from the image, inability to connect relevant visual cues to the question, or incorrect inference or assumption about the scene.
Non-Reasoning Errors:

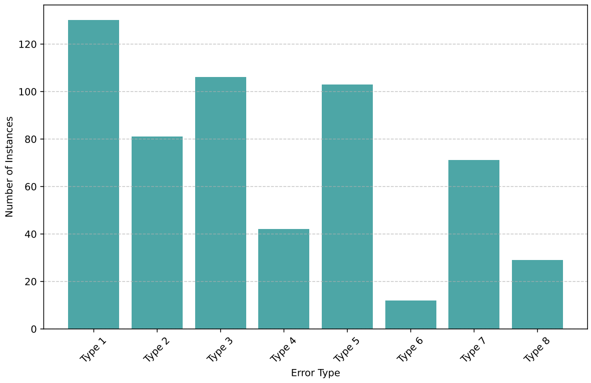


Figure 3: Distribution of error types amongst all 372 generated answers

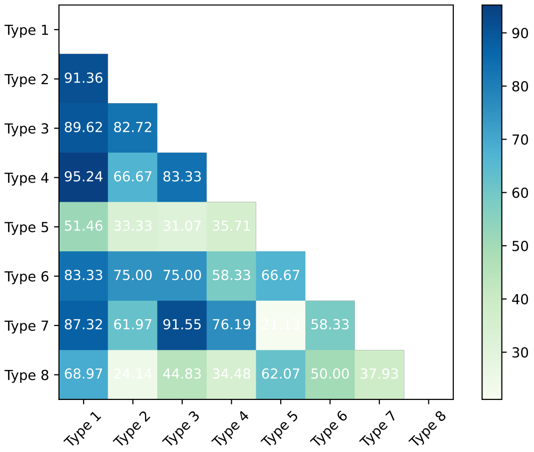


Figure 4: Error Co-occurrence Statistics

Type 2: Medical Factual Error This error involves using incorrect medical knowledge or terminology, misunderstanding medical procedures or techniques, misinterpreting medical equipment or devices, or incorrectly interpreting patient conditions or symptoms. It stems from deficiencies in the model's healthcare and medical knowledge base.

Type 3: Perception Error This error is caused by mis-identification of objects or elements in the image, failure to recognize or distinguish medical equipment or devices, or incorrect interpretation of visual cues (e.g., patient positioning, medical staff attire). These are attributed to the model's inability to interpret visual content.

Type 4: Coherence or Consistency Error This error involves contradictory statements within the same answer. inconsistencies between the generated answer and the image, or lack of coherence or logical flow in the answer.

Type 5: Specificity and Relevance Error This error is characterized by overly general or vague answers, answering irrelevant or unrelated aspects of the question, or failure to address the specific details or context of the question.

Type 6: Linguistic Error This error involves grammatical or syntactical mistakes, inappropriate or incorrect use of medical terminology, or unclear/confusing phrasing or sentence structure.

Type 7: Hallucination Error This error is caused by generating information or details not present in the image, introducing fabricated or imagined elements in the answer, or providing responses that are entirely unrelated to the image or question. This error is fundamentally linked to the model's generation process.

Type 8: Uncertainty and Confidence Error This error is characterized by failure to acknowledge the limitations or uncertainty of the generated answer, overconfidence or lack of appropriate uncertainty estimation, or inability to distinguish between confident and uncertain responses.

4.3 Error Occurrence and Co-occurrence

After collecting the error annotations for all 372 generated answers, from Figure 3 we find that *Type 1*, *Type 3* and *Type 5* occur the most, each occurring more than 100 times (> 26% of the generations!). *Type 1* errors occur due to the models making incorrect assumptions about factors like patient history, general hospital protocols, etc. *Type 3* errors mostly happen due to mis-identification of medical apparatuses, using incorrect naming of devices, and poor spatial understanding of the scene. *Type 5* errors occur usually due to the model not being able to understand the motive of the question, and resorting to generating unnecessarily long answers containing definitions and general protocols about the subject.

Co-occurrence Statistics: Figure 6 shows the error type co-occurrence statistics for the error annotations. We calculate occurrence between 2 errors in the following way: For an unordered pair of errors $\{e_i, e_j\}$ where $i \neq j$, we count the number of co-occurrences of e_i and e_j, divide it by $min\{|e_i|, |e_j|\}$, and multiply by 100 for scaling, as previously shown in Roy et al. (2024).

Analysis: From Figure 4, we see that Reasoning Error (*Type 1*) has a very high co-occurrence percentage with Medical Factual Errors (91.36%), Perception Errors (89.62%), Coherence/Consistency Errors (95.24%), Linguistic Errors (83.33%) and Hallucination Errors (87.32%). This shows that

GPT4Vo and Gemini Vision Pro will hallucinate, make linguistic and visual errors, and make unsure generations, and will try to ***double down*** on these errors, leading to reasoning errors in the generated answer. This tendency of overconfidence, has been previously shown in the case of LLMs (Xiong et al., 2023; Groot and Valdenegro-Toro, 2024). We also find from Figure 4, that Medical Factual errors (*Type 2*) are also closely related to both Perception Errors (82.72%) and Linguistic Errors (75%), thereby motivating the fact that both linguistic and visual improvements are crucial for medically factual generation. We also see that Perception Errors (*Type 3*) co-occur largely with Hallucination Errors (91.55%), due to the two being closely related.

5 Benchmark

In order to benchmark models against the ERVQA dataset, we choose to test on models which offer a wide range of variation in the vision encoder, LLM decoder, model size and finetuning strategies. We aim to explore the inherent capabilities of these models to provide error-free meaningful answers which are lexically and semantically close to the ground truth answers, void of the errors defined in Section 4, which is essential for their application in the healthcare domain.

5.1 Models

We experiment with both open sourced and closed models. Open source models include Llava 1.5 13B (Liu et al., 2024a) which consists of a CLIP ViT-L/14 encoder and Vicuna-7B decoder, Open-Flamingo 9B (Awadalla et al., 2023) consisting of CLIP ViT-L/14 encoder and Llama-7B decoder, MPLUG-Owl 7B (Ye et al., 2023) consisting of CLIP ViT-L/14 encoder and MPT-7B decoder, and Instruct-BLIP 7B model (Dai et al., 2024) containing a BLIP-2 based encoder and Vicuna-7B decoder. We experiment with Med-Flamingo 9B model (Moor et al., 2023) with Llama-7B decoder, finetuned on medical VQA datasets, to compare with its open domain counterpart Open-Flamingo. Although the medical data used is from the pathology/radiology domain (See Section 2), our interest lies in finding out whether it captures the nuances of answering medical questions. Finally, we also run experiments on closed models GPT4Vo by OpenAI (Achiam et al., 2023) and Gemini Pro Vision (Reid et al., 2024) by Google, to understand the capabilities of larger state of the art models trained on more training data.

5.2 Experimental Setup

Our goal is to understand the *inherent* capabilities of LVLMs with our challenging dataset. We run experiments in zero-shot and in-context few-shot (1-shot and 3-shot) settings to assess these models in low-resource scenarios. Llava-1.5 and InstructBLIP are tested only in zero-shot, and Open-Flamingo only in few-shot settings, due to limitations in their training strategies. The requirement for in-context examples for Open-Flamingo in 'zero-shot' settings is supported by (Awadalla et al., 2023). All the prompts, code and data have been made available for reproducibility (see Appendix).

5.3 Evaluation metrics

To evaluate generated answers against ground truth, we use traditional model-free metrics such as **BLEU-1** (Papineni et al., 2002), which compares unigrams, and **ROUGE-L** (Lin, 2004), which compares longest common subsequences. For semantic similarity, we use **SentenceBERT Similarity** (Reimers and Gurevych, 2019). However, these metrics do not fully capture two essential factors for VQA evaluation: a) whether the ground truth answer is entailed within the generated answer, and b) whether the generated answer aligns with the image. To address these issues for our dataset, we adapapt two model based evaluation metrics suitable for our problem domain:

Entailment Score: We use the *roberta-base-nli* model , trained on the SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2017) datasets. We define Entailment score as follows:

$$ES = p(entailment|ref, gen) \qquad (1)$$

We obtain ES by applying a softmax function on the entailment class logit obtained from the *roberta-base-nli* model, and we report the average across all the generations.

CLIPScore Confidence: CLIPScore has been originally introduced as a reference-free metric for evaluating the Image Captioning task (Hessel et al., 2021). We treat the reference and generated answers as if they were captions for the image, and calculate the CLIPScore using the rich cross-modal

`https://huggingface.co/cross-encoder/nli-roberta-base` image and text embeddings, as described in the paper. We. then calculate the reference-based CLIP-Score Confidence as follows:

$$CLIP-C = \frac{CLIP-S(img, gen)}{CLIP-S(img, ref) + CLIP-S(img, gen)} \tag{2}$$

This gives us an image-based comparison of the generated and reference answers.

5.4 Error Evaluation

Visual Question Answering using the ERVQA dataset is a complex task due to its sensitive applications, requiring the answers to be as error-averse as possible. Hence, evaluating models based on similarity metrics as well as occurrence of errors is of paramount importance. Since human annotations for all generations is expensive as well as cumbersome, we obtain silver-label error annotations using classification models for each type of error, and compare and analyse the performance based on percentage of error occurrences.

Error Classification Model: In order to build an error-wise dataset for building classification models, we first include all the manually annotated data obtained from Section 4 for each label. To balance the dataset, we introduce more erroneous examples by inducing errors to ground truth answers by prompting GPT4V with the error definition. Finally, for each class we obtain 200 positive and 200 negative samples.

We finetune a BLIP-2 model (Li et al., 2023) for error classification, where we pool the cross-modal QFormer outputs and attach a binary classification head to it. After training for 10 epochs with a learning rate of 1e-6, we save the model with the best performance against the validation set. We finally pass the model logits through a sigmoid layer to get the class predictions. We use a 5-fold cross validation setting where, across all classes, we report average macro F1 score on the best runs, to be 97.81.

6 Results and Analysis

6.1 Evaluation Metrics based Results

Table 3 presents the results across all models and settings based on the evaluation metrics described in Section §5.3. In the zero-shot setting, larger model sizes generally yield better results across all metrics. Llava-1.5 is the best-performing open-source model with the highest CLIP-C score. GPT4Vo provides the most semantically similar answers with high entailment scores. Among the 7B models, mPLUG-Owl performs best.

	BLEU	**ROUGE**	**SENT**	**ENT**	**CLIP-C**
Zero-shot					
mPLUG-Owl-7B	18.13	22.32	65.50	23.20	49.93
Instruct-BLIP-7B*	16.82	20.13	51.75	18.32	48.44
Med-Flamingo-9B	17.20	19.17	64.23	28.19	49.20
Llava1.5-13B*	20.11	22.78	66.00	19.03	**52.31**
GPT4Vo	16.63	20.72	69.34	38.98	51.02
Gemini Pro Vision	19.83	24.21	64.36	19.10	49.87
1-shot					
mPLUG-Owl-7B	18.74	23.01	69.30	24.47	50.98
Med-Flamingo-9B	20.76	25.62	61.12	10.49	48.95
Open-Flamingo-9B	18.01	**29.10**	64.34	11.18	48.98
GPT4Vo	22.25	26.00	76.79	**40.63**	51.31
Gemini Pro Vision	23.27	27.05	68.58	28.94	50.38
3-shot					
mPLUG-Owl-7B	20.03	26.67	64.83	20.04	50.11
Med-Flamingo-9B	20.21	25.34	61.23	8.58	49.56
Open-Flamingo-9B	17.54	28.70	63.91	9.07	49.01
GPT4Vo	23.66	27.10	**71.17**	38.22	51.03
Gemini Pro Vision	**24.02**	27.67	68.52	29.98	51.23

Table 3: Evaluation metrics based results for all models on zero-shot, 1-shot and 3-shot settings. **BLEU** denotes BLEU-1 score, **ROUGE** denotes ROUGE-L score, **SENT** denotes SentenceBERT Similarity, **ENT** denotes Entailment Score and **CLIP-C** denotes CLIP-Score Confidence. Bold figures indicate best results across all settings. * denotes models which have only been used for zero-shot evaluation.

In few-shot in-context settings, increasing the number of examples improves BLEU, ROUGE, and SENT scores, indicating better lexical pattern learning and word choice. However, open-source models show a decline in semantically relevant metrics like ENT and CLIP-C, indicating a tradeoff between lexical accuracy and semantic correctness. Closed models, on the other hand, improve in most metrics when provided with examples. Interestingly, there is no significant improvement from medical domain-specific fine-tuning, as seen with comparable results between Open-Flamingo and Med-Flamingo. This is likely due to Med-Flamingo being fine-tuned on Medical VQA datasets focused on pathology/radiology rather than the broader healthcare settings of ERVQA.

6.2 Error based Results

In general, the values of the embedding cosine-similarity based and lexical overlap based metrics reported for open-domain datasets, such as in Dua et al. (2021), are comparable with that of our experiments on ERVQA. However, in this section, we show that metric-based results do not reflect in the actual quality of the generated texts, with regards to errors. We obtain silver-labels for each error class using the classification model described in Section

https://huggingface.co/Salesforce/blip2-opt-2.7b

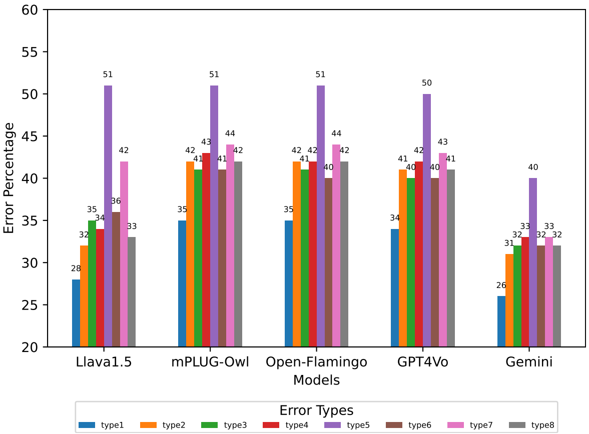


Figure 5: Model wise percentages of error occurrence for comparison on the basis of LLM Decoder type

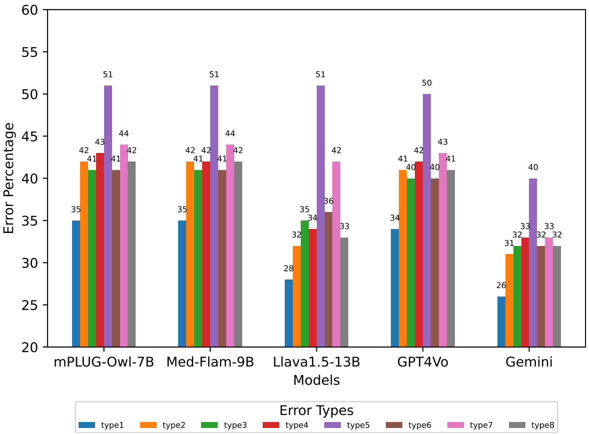


Figure 6: Model wise percentages of error occurrence for comparison on the basis of Model Size

§5.3 across all models and perform a comparison on the basis of percentage of error occurrence. We analyse our results based on variance in the following factors: a) LLM decoder type b) Model size and c) Number of In-context examples.

LLM decoder type based analysis: We consider 5 different best-performing LLM decoder based VLMs in the zero-shot setting for analysing the percentage of errors. Llava1.5 has a Vicuna-7B based decoder, mPLUG-Owl has a Llama-7B (Touvron et al., 2023) based decoder, Open-Flamingo has a MPT-7B (Team et al., 2023) based decoder, and the closed models (GPT4vo and Gemini Pro Vision) have undisclosed decoder models (Achiam et al., 2023) (Reid et al., 2024). According to Figure 5, among open source models, we notice lesser errors in the Llava1.5 model based on Vicuna. Overall, again Gemini gives the best results. Llama-7B, MPT-7B and GPT4Vo give similar results. An important point to note is that apart from Gemini, all other models give high percentage of Specificity/Relevance error (Type 5) (>50%), thus showing that these models have a tendency of generating long answers containing irrelevant details. In most other metrics, it is important to note that the Llava-1.5 model performs comparably to Gemini.

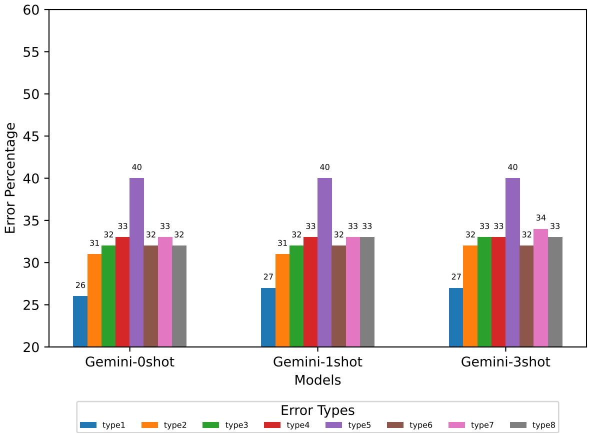


Figure 7: Model wise percentages of error occurrence for comparison on increasing in-context examples for Gemini Vision Pro

Model size based analysis: We conduct experiments on different models of various sizes in the zero-shot setting. Our models, in increasing order of size are as follows: mPLUG-Owl-7B < Med-Flamingo-9B < Llava1.5-13B< GPT4Vo Gemini Vision Pro. From Figure 6, although we see Llava-13B having lesser errors than smaller models, GPT4Vo gives more errors. Hence, error occurrence is invariant of model size.

Effect of In-context Few shot examples: From Table 2, we see that increasing the number of in-context examples leads to an increase in performance for the Gemini Vision Pro, except for the CLIP-C metric (there is an increase from zero-shot, but not uniform). Hence, we try to answer the question: *Does increase in performance in terms of evaluation metrics mean a decrease in percentage of errors?* We compare the error percentages across zero-shot, 1-shot and 3-shot settings for the Gemini Pro Vision model, and we find that the error occurrences are almost identical – meaning that the erroneous generations have little or no correlation to token-based and semantic metrics.

7 Conclusion and Future Work

Finally, we revisit our question: *"Are the existing Large Vision Language Models (LVLMs) ready to be used in healthcare environments?"* The ERVQA

https://lmsys.org/blog/2023-03-30-vicuna/

dataset is proposed as the first step towards addressing this. Our extensive error taxonomy and human evaluation reveal the complexity of generating accurate answers due to dense occurrence and co-occurrence patterns. Despite adapting semantically relevant VQA-specific evaluation metrics, improvement across these metrics doesn't necessarily reduce errors. The best-performing model, Gemini Vision Pro, still has an average error rate of 33% across all error classes, which is significant given the sensitive domain. Therefore, we conclude that **existing LVLMs are not yet ready for healthcare environments**. Our dataset is a valuable resource for evaluating and improving models and techniques in this field. We hope that the future work would focus on building more error-sensitive LVLMs, developing effective error mitigation strategies, and exploring video datasets for further advancement.

Limitations

Our work, although comprehensive and valuable, has the following limitations. First, the absence of existing high quality pre-traning data due to its highly confidential nature, has limited the knowledge base of existing LVLMs, we use for benchmarking our dataset. This affects the performance of models in consideration. Second, the doctors advising and helping with the annotations are from premium medical facilities from India. Their annotations are influenced by formal medical education and real-world experience in India. While this common background, along with strict guidelines ensure uniformity in their notions of relevance and accuracy of the answers, the experts were likely following medical protocols common in India. However, they might differ from doctors from other regions of the world in some cases, due to differences in medical standards and protocols, cultural and ethical considerations, resource availability and infrastructure, legal and regulatory frameworks and public health priorities particular in India (McPherson, 1989). A wider collaborative effort by medical experts from across the world, would significantly enhance this domain of study, both through diverse data and better modeling.

Ethics Statement

Since the annotations of the questions and answers, along with the error annotations, have been made by multiple domain experts, it is possible that the dataset reflects social, ethical and medical biases of the individuals.

All images scraped for the purposes of building this dataset are taken from Google Images search results, sourced from various news articles and similar publications. We do not claim any ownership of the images itself, but of the annotations done by the medical professionals. Image URLs are provided in the dataset files for downloading. To the best of the authors' knowledge, the images have been scraped from paywall-free articles. We plan to make the image links available on request after making the dataset publicly available, to ensure safe and responsible usage.

Furthermore, the questions and answers framed for this dataset are on the basis of plausible scenarios according to the image content. No real world medical records/data were used to frame the question answer pairs. Apart from the visual information in these images, we have not provided any resources for identification of the individuals such as patients/doctors/nurses and also institutions such as hospitals. We shall also provide an opt-out option for those who want certain data points censored/removed.

We use ChatGPT as a collaborative writing and coding assistant when required. However, all novelties including metric definitions, taxonomies, prompts, etc are completely made by the authors/collaborators.

Acknowledgements

The work was supported in part by a research grant from IIT KHARAGPUR AI4ICPS I HUB FOUNDATION. PI Pawan Goyal acknowledges the kind support by the grant given by I-Hub Foundation for Cobotics, IIT Delhi. PI Somak Aditya further acknowledges the support by SERB SRG/2022/000648.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A Hasan, and Henning Müller. 2021. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes*. 21-24 September 2021.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. The snli corpus.

Cheng-Han Chiang and Hung-yi Lee. 2023. A closer look into automatic evaluation using large language models. *arXiv preprint arXiv:2310.05657*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Radhika Dua, Sai Srinivas Kancheti, and Vineeth N Balasubramanian. 2021. Beyond vqa: Generating multi-word answers and rationales to visual questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1623–1632.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Tobias Groot and Matias Valdenegro-Toro. 2024. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. *arXiv preprint arXiv:2405.02917*.

Carlton Gyles. 2019. Robots in medicine. *The Canadian Veterinary Journal*, 60(8):819.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Jian Huang, Yihao Chen, Yong Li, Zhenguo Yang, Xuehao Gong, Fu Lee Wang, Xiaohong Xu, and Wenyin Liu. 2023. Medical knowledge-based network for patient-oriented visual question answering. *Information Processing & Management*, 60(2):103241.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *arXiv preprint arXiv:2305.13711*.

Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. 2023. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, page 102611.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Lei Liu, Xiaoyan Yang, Fangzhou Li, Chenfei Chi, Yue Shen, Shiwei Lyu Ming Zhang, Xiaowei Ma, Xiangguo Lyu, Liya Ma, Zhiqiang Zhang, et al. 2024b. Towards automatic evaluation for llms' clinical capabilities: Metric, data, and algorithm. *arXiv preprint arXiv:2403.16446*.

Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. 2024. Improving automatic vqa evaluation using large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4171–4179.

Klim McPherson. 1989. International differences in medical care practices. *Health Care Financing Review*, 1989(Suppl):9.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with dialogpt. *arXiv preprint arXiv:2006.12719*.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR.

World Health Organization et al. 2016. Global strategy on human resources for health: workforce 2030.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Soumyadeep Roy, Aparup Khatua, Fatemeh Ghoochani, Uwe Hadler, Wolfgang Nejdl, and Niloy Ganguly. 2024. Beyond accuracy: Investigating error types in gpt-4 responses to usmle questions. *arXiv preprint arXiv:2404.13307*.

Ananya B Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M Khapra. 2020. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer.

MosaicML NLP Team et al. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023. *URL www. mosaicml. com/blog/mpt-7b. Accessed*, pages 05–05.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. 2019. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5317–5327.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

A Appendix

A.1 Additional Related Work

There has been a rise in popularity of automatic evaluation of LLM generated text (Chiang and Lee, 2023). (Lin and Chen, 2023) discusses the use of prompt based evaluation of text using LLMs. (Liu et al., 2024b) discusses using automatic evaluation of clinical text. LLM based evaluation of VQA has also been explored in (Mañas et al., 2024). Other unsupervised methods like DEB (Sai et al., 2020) uses BERT as an evaluation model for adversarial text, while FED (Mehri and Eskenazi, 2020) uses an unsupervised method for measuring dialogue quality. We evaluate error occurrences using multi-modal representations from BLIP-2, making use of both text and visual modalities.

A.2 ERVQA Dataset Additional Details

During annotation with regards to the qualitative metrics: *Relevance*, *Clarity* and *Harmlessness*, the annotators sometimes disagree with each other. One such disagreement is shown in Figure 8.

The ERVQA dataset consists of a wide range of questions of varying types. A pie chart showing the various different types of questions are shown in Figure 9.

A.3 Examples of Different Errors

We demonstrate the different kinds of errors from the error taxonomy with an example image, question, erroneous answer and explanation. Please note that in some cases, while a single error is pointed out, there might be other errors within the answers, that have been avoided to prevent confusion in the reasoning process. This co-occurrence of errors makes the annotation as well the prediction task extremely hard and subjective. The errors are shown in Figures 10, 11, 12 and 13.

A.4 Training Details of Error Classifier Model

All the BLIP-2 based error classification models for the 8 types were trained on a single NVIDIA L40 GPU having 48GB GPU memory. We adopt a 5-fold cross validation strategy, and save the checkpoints with lowest Binary Cross Entropy Loss for the validation set. We fix the learning rate to 1e-6 for the Adam's optimizer after initial experimentation.

A.5 Reproducibility

For reproducibility of our zero-shot and few-shot in-context results, we provide the prompts used for generation. Please note we experiment with simpler prompts at first, and increase the complexity according to documentation and failure cases. Also, we make our datasets, error classifier and model code publicly available at `https://github.com/sourjyadip/ervqa-data/`.

A.5.1 GPT4Vo Prompts

Zero-shot: "Answer the following question for the given image: " + <question>
1-shot and 3-shot: We use the prompting strategy used in the OpenAI Forum.

A.5.2 Gemini Vision Pro Prompts

The images are provided to the model separately in batches.
Zero-shot: <question> + " Answer:" **1-shot:** "You are being provided with 1 example question and answer for the first image. Similarly, give an answer for the second question based on the second image. Q: "+ <question1> + "A: "+ <answer1> + " Q: "+ <question2> + "A: "
3-shot: "You are being provided with 3 example questions and answers for the first 3 images. Similarly, give an answer for the fourth question based on the fourth image. Q: "+ <question1> + "A: "+ <answer1> + " Q: "+ <question2> + "A: "+ <answer2> + " Q: "+ <question3> + "A: "+ <answer3> + " Q:" + <question> + "A: "

A.5.3 Llava Prompt

Zero-shot: "<image> \nUSER: " + <question> + "\nASSISTANT:"

A.5.4 Med-Flamingo Prompts

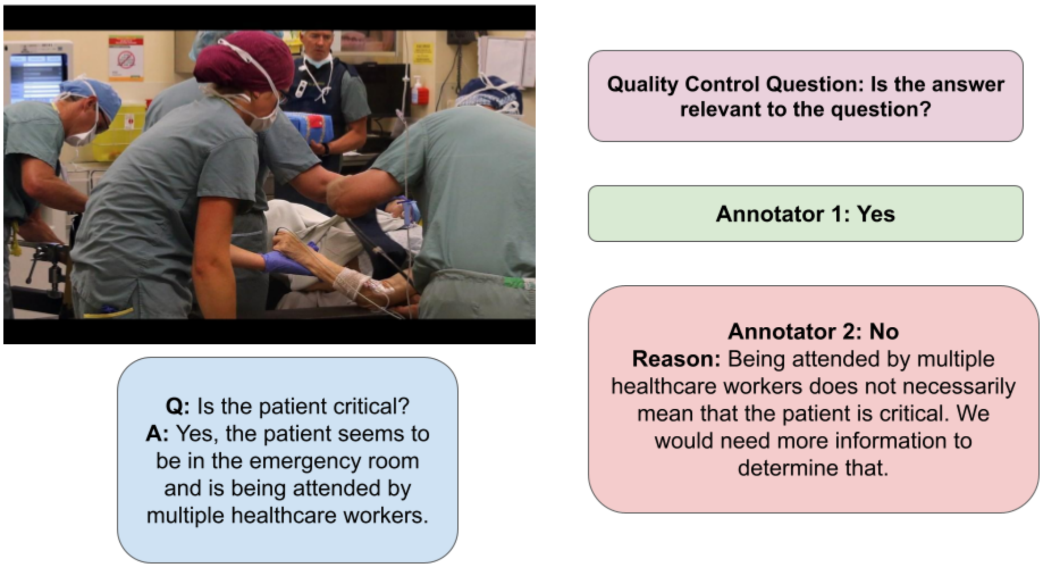
Zero-shot: "You are a helpful medical assistant. You are being provided with an image and a question about the image. Answer the question. <image>Question: " + <question> + "Answer: "
1-shot: "You are a helpful medical assistant. You are being provided with images, a question about the image and an answer. Follow the example and answer the second question. <image>Question: " + <question1> + "Answer: " + <answer1> + "<|endofchunk|>" + "<image>Question: " + <question> + " Answer:"
3-shot: "You are a helpful medical assistant. You are being provided with images, a question about the image and an answer. Follow the examples and answer the last question. <image>Question: " + <question1> + "Answer: " + <answer1> + "<|endofchunk|>" + "<image>Question: " + <question2> + "Answer: " + <answer2" + "<|endofchunk|>"+ "<image>Question: " + <question3> + "Answer: " + <answer3> + "<|endofchunk|>" + "<image>Question: " + <question> + " Answer:"

A.5.5 Open-Flamingo Prompts:

Zero-shot evaluation gives no result or generates some random tokens, which are incoherent. For **1-shot** and **3-shot**, we use the same prompts as Med-Flamingo.

A.5.6 mPLUG=Owl Prompts:

Zero-shot: "The following is a conversation between a curious human and AI assistant. The assistant gives helpful, detailed, and polite answers to the user's questions. Human: <image> Human: Is the facility adequate? AI: "
1-shot: 'The following is a conversation between a curious human and AI assistant. The assistant gives helpful, detailed, and polite answers to the users questions. Based on the question and answer given for the first image, give an answer for the



Q: Is the patient critical?
A: Yes, the patient seems to be in the emergency room and is being attended by multiple healthcare workers.

Quality Control Question: Is the answer relevant to the question?

Annotator 1: Yes

Annotator 2: No
Reason: Being attended by multiple healthcare workers does not necessarily mean that the patient is critical. We would need more information to determine that.

Figure 8: Disagreement on the basis of *Relevance*

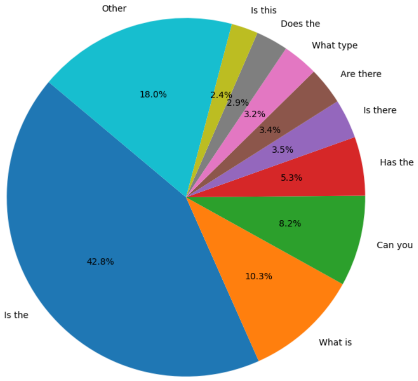


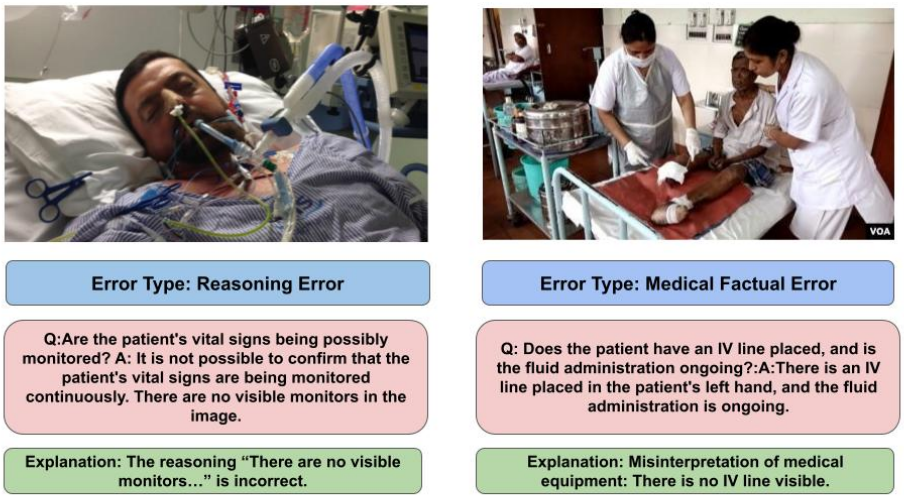
Figure 9: Distribution of the most common beginning bigrams for questions in the ERVQA dataset. The 'other' section is very diverse containing bigrams like *'How frequently', 'Please report, 'Which setup'*, etc

second question based on the second image Human: <image> Human:' + <question1> + ' AI: ' + <answer1> + 'Human: <image> Human:' + <question> + ' AI: '

3-shot: 'The following is a conversation between a curious human and AI assistant. The assistant gives helpful, detailed, and polite answers to the users questions. Based on the question and answer given for the first 3 images, give an answer for the fourth question based on the fourth image Human: <image> Human:' + <question1> + ' AI: ' + <answer1> + 'Human: <image> Human:' + <question2> + ' AI: ' + <answer2> + 'Human: <image> Human:' + <question3> + ' AI: ' + <answer3> + 'Human: <image> Human:' + <question> + ' AI: '

A.5.7 InstructBLIP Prompts

Zero-shot: <image> + <question>



Error Type: Reasoning Error

Q:Are the patient's vital signs being possibly monitored? A: It is not possible to confirm that the patient's vital signs are being monitored continuously. There are no visible monitors in the image.

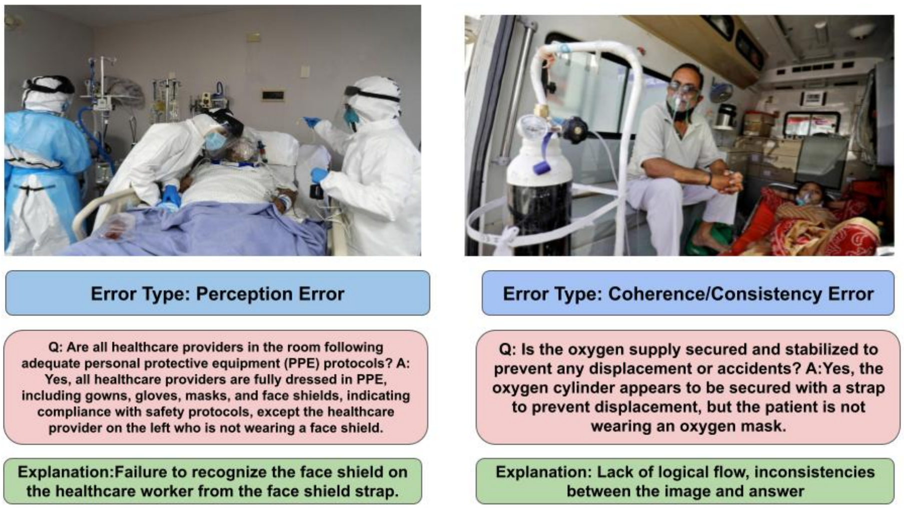
Explanation: The reasoning “There are no visible monitors…” is incorrect.

Error Type: Medical Factual Error

Q: Does the patient have an IV line placed, and is the fluid administration ongoing?:A:There is an IV line placed in the patient's left hand, and the fluid administration is ongoing.

Explanation: Misinterpretation of medical equipment: There is no IV line visible.

Figure 10



Error Type: Perception Error

Q: Are all healthcare providers in the room following adequate personal protective equipment (PPE) protocols? A: Yes, all healthcare providers are fully dressed in PPE, including gowns, gloves, masks, and face shields, indicating compliance with safety protocols, except the healthcare provider on the left who is not wearing a face shield.

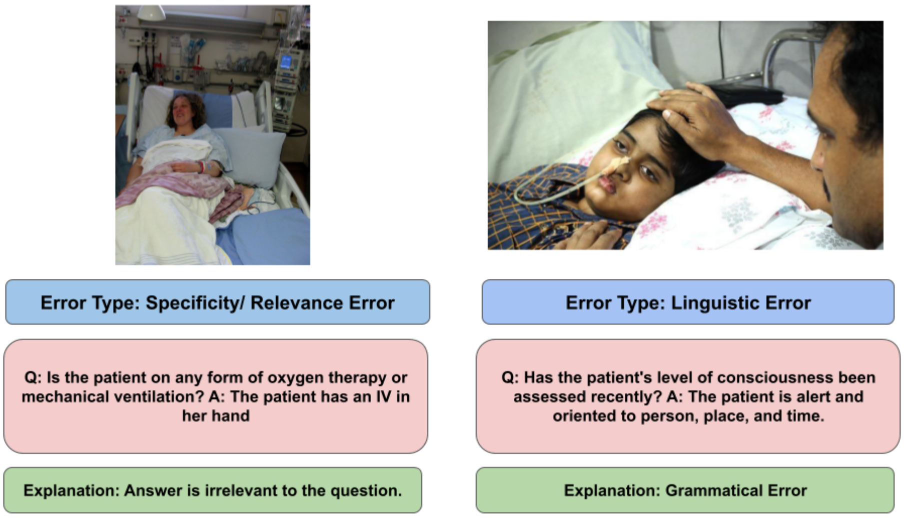
Explanation:Failure to recognize the face shield on the healthcare worker from the face shield strap.

Error Type: Coherence/Consistency Error

Q: Is the oxygen supply secured and stabilized to prevent any displacement or accidents? A:Yes, the oxygen cylinder appears to be secured with a strap to prevent displacement, but the patient is not wearing an oxygen mask.

Explanation: Lack of logical flow, inconsistencies between the image and answer

Figure 11



Error Type: Specificity/ Relevance Error

Q: Is the patient on any form of oxygen therapy or mechanical ventilation? A: The patient has an IV in her hand

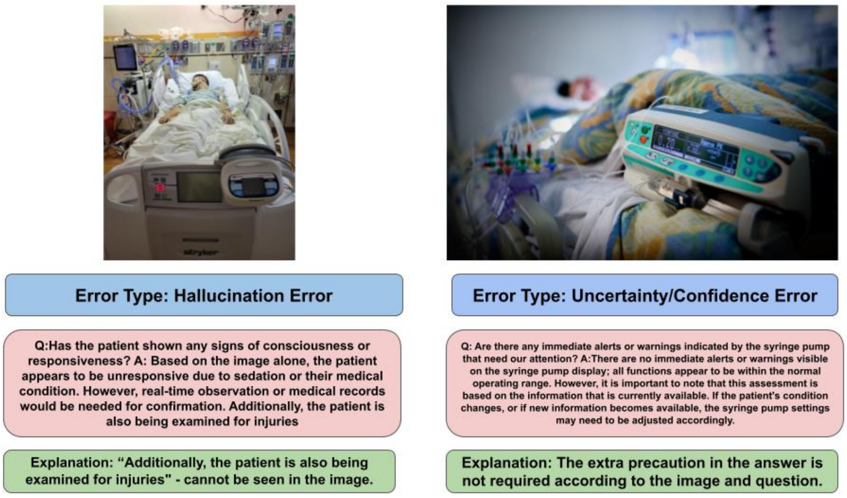
Explanation: Answer is irrelevant to the question.

Error Type: Linguistic Error

Q: Has the patient's level of consciousness been assessed recently? A: The patient is alert and oriented to person, place, and time.

Explanation: Grammatical Error

Figure 12



Error Type: Hallucination Error

Q:Has the patient shown any signs of consciousness or responsiveness? A: Based on the image alone, the patient appears to be unresponsive due to sedation or their medical condition. However, real-time observation or medical records would be needed for confirmation. Additionally, the patient is also being examined for injuries

Explanation: “Additionally, the patient is also being examined for injuries" - cannot be seen in the image.

Error Type: Uncertainty/Confidence Error

Q: Are there any immediate alerts or warnings indicated by the syringe pump that need our attention? A:There are no immediate alerts or warnings visible on the syringe pump display; all functions appear to be within the normal operating range. However, it is important to note that this assessment is based on the information that is currently available. If the patient's condition changes, or if new information becomes available, the syringe pump settings may need to be adjusted accordingly.

Explanation: The extra precaution in the answer is not required according to the image and question.

Figure 13