Consistency and Uncertainty: Identifying Unreliable Responses From Black-Box Vision-Language Models for Selective Visual Question Answering

Zaid Khan Yun Fu
Northeastern University

Abstract

The goal of selective prediction is to allow an a model to abstain when it may not be able to deliver a reliable prediction, which is important in safety-critical contexts. Existing approaches to selective prediction typically require access to the internals of a model, require retraining a model or study only unimodal models. However, the most powerful models (e.g. GPT-4) are typically only available as black boxes with inaccessible internals, are not retrainable by end-users, and are frequently used for multimodal tasks. We study the possibility of selective prediction for vision-language models in a realistic, black-box setting. We propose using the principle of neighborhood consistency to identify unreliable responses from a black-box vision-language model in question answering tasks. We hypothesize that given only a visual question and model response, the consistency of the model's responses over the neighborhood of a visual question will indicate reliability. It is impossible to directly sample neighbors in feature space in a black-box setting. Instead, we show that it is possible to use a smaller proxy model to approximately sample from the neighborhood. We find that neighborhood consistency can be used to identify model responses to visual questions that are likely unreliable, even in adversarial settings or settings that are out-of-distribution to the proxy model.

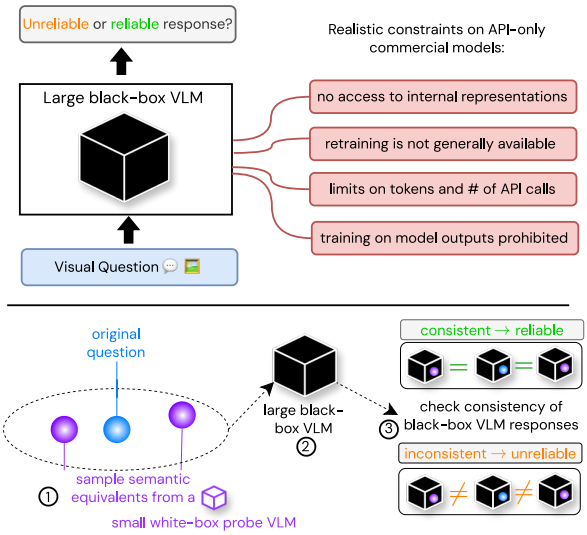


Figure 1. Identifying unreliable responses from an API-only black-box vision-language model (VLM) can be challenging because confidence scores are not always trustworthy, and more sophisticated methods for selective prediction require a level of access to the model that is unavailable. We explore the idea of model consistency to identify unreliable model responses in this realistic scenario: a reliable response is one that is consistent across questions that are semantically equivalent but different on the surface.

1. Introduction

Powerful commercial frontier models are sometimes only available as black boxes accessible through an API [2, 25]. When using these models in high-risk scenarios, it is preferable that the model defers to an expert or abstains from answering rather than deliver an incorrect answer [7]. Many approaches for selective prediction [7, 30] or improving the predictive uncertainty of a model exist, such as ensembling [14], gradient-guided sampling in feature space [11], retraining the model [27], or training a auxiliary module using model predictions [21]. Selective prediction has typically been studied in unimodal settings and/or for tasks with a closed-world assumption, such as image classification, and has only recently been studied for multimodal, open-ended tasks such as visual question answering [6, 29] (VQA).

In existing deployments, training data is private, model features and gradients are unavailable, retraining is not possible, the number of predictions may be limited by the API, training on model outputs is often prohibited, and queries are open-ended. In a black-box setting with realistic constraints, how do we identify unreliable predictions from a vision-language model?

An intuitive approach is to consider self-consistency: if a human subject is given two semantically equivalent questions, we expect the human subject's answers to the questions to be identical. A more formal notion of consistency is

that given a classifier $f(\cdot)$ and an point $\mathbf{x} \in \mathbb{R}^N$ in feature space, the classifier's predictions over an ϵ-neighborhood of $\mathbf{x}$ should be consistent with $f(\mathbf{x})$ for a small enough ϵ [11]. It is not straightforward to operationalize either of these notions. How can we scalably obtain "semantically equivalent" visual questions to an input visual question? Since we can't access the internal representations of a black-box model, how can we sample from the neighborhood of an input visual question?

First, we study selective prediction on VQA across in-distribution, out-of-distribution, and adversarial inputs using a large VLM. Next, we describe how rephrasings of a question can be viewed as samples from the ϵ-neighborhood of a visual question pair. We propose training a visual question generation model as a *probing model* to scalably and cheaply produce rephrasings of visual questions given answers and an image, allowing us to approximately sample from the neighborhood of a visual question pair. To quantify uncertainty in the answer to a visual question pair, we feed the rephrasings of the question to the black-box VLM, and count the number of rephrasings for which the answer of the VLM remains the same. Surprisingly, we show that consistency over model-generated "approximate rephrasings" is effective at identifying unreliable predictions of a black-box vision-language model, even when the rephrasings are not semantically equivalent and the probing model is an order of magnitude smaller than the black-box model.

Our approach is analogous to consistency over samples taken from the neighborhood of an input sample in feature space, but this method does not require access to the features of the vision-language model. Furthermore, it does not require a held-out validation set, access to the original training data, or retraining the vision-language model, making it appropriate for black-box uncertainty estimates of a vision-language model. We conduct a series of experiments testing the effectiveness of consistency over rephrasings for assessing predictive uncertainty using the task of selective visual question answering in a number of settings, including adversarial visual questions, distribution shift, and out of distribution detection.

Our contributions are:

- We study the problem of black-box selective prediction for a large vision-language model, using the setting of selective visual question answering.
- We show that on in-distribution data, a state-of-the-art large vision-language model is capable of identifying when it does not know the answer to a question, but this ability is severely degraded for out-of-distribution and adversarial visual questions.
- We propose identifying high-risk inputs for visual question answering based on consistency over samples in the neighborhood of a visual question.
- We show that consistency defines a different ordering than model confidence / uncertainty over instances in a dataset.
- We conduct a series of experiments validating the proposed method on in-distribution, out-of-distribution and adversarial visual questions, and show that our approach even works in the likely setting that the black box model being probed is substantially larger than the probing model.

We show that consistency over the rephrasings of a question is correlated with model accuracy on a question and can select slices of a test dataset on which a model can achieve lower risk, reject out of distribution samples, and works well to separate right from wrong answers, even on adversarial and out of distribution inputs. *Surprisingly, this technique works even though many rephrasings are not literally valid rephrasings of a question.* Our proposed method is a step towards reliable usage of vision-language models as an API. **Limitations:** Due to resource constraints, we study models that might now be considered relatively small ($\leq$ 13B parameters), and our VQG model is "small" ($<$ 700M).

2. Motivating Experiment

We empirically examine the predictive uncertainty of a large VLM through the lens of selective visual question answering. In contrast to the classical VQA setting where a model is forced to answer, a model is allowed to *abstain* from answering in selective VQA. For safety and reliability, it is important to examine both out-of-distribution and adversarial inputs, on which we expect that the VLM will have a high error rate if forced to answer every out-of-distribution or adversarial question posed to the model. However, because the VLM is allowed to abstain, in principle the model can achieve low risk (low error rate) on a *slice* of the dataset corresponding to questions that it knows the answer to. In a black-box setting, only the raw confidence scores for the answer candidates are likely to be available, so we use the confidence of the most likely answer as the uncertainty.

In Fig. 3, we plot the selective visual question answering performance of BLIP [16] finetuned on VQAv2 using the confidence scores on the validation sets of adversarial (AdVQA, Sheng et al. [24]), out-of-distribution (OKVQA, Marino et al. [20]) and in-distribution (VQAv2) datasets. For the in-distribution dataset (VQAv2), the model is quickly able to identify which questions it is likely to know the answer to, achieving nearly perfect accuracy by rejecting the most uncertain 40% of the dataset. However, for out-of-distribution and adversarial datasets, the model has a harder time – after rejecting 50% of the questions, the model still has an error rate of $\approx$ 40%. The reason for this is evident in Fig. 2, where we plot the distribution of confidence scores for incorrect and correct answers for OOD, in-distribution, and adversarial visual questions. For in-distribution visual questions, the confidence distribution is bimodal, and incorrect and correct answers are clearly separated by confidence. For OOD visual questions, many correctly answered ques-

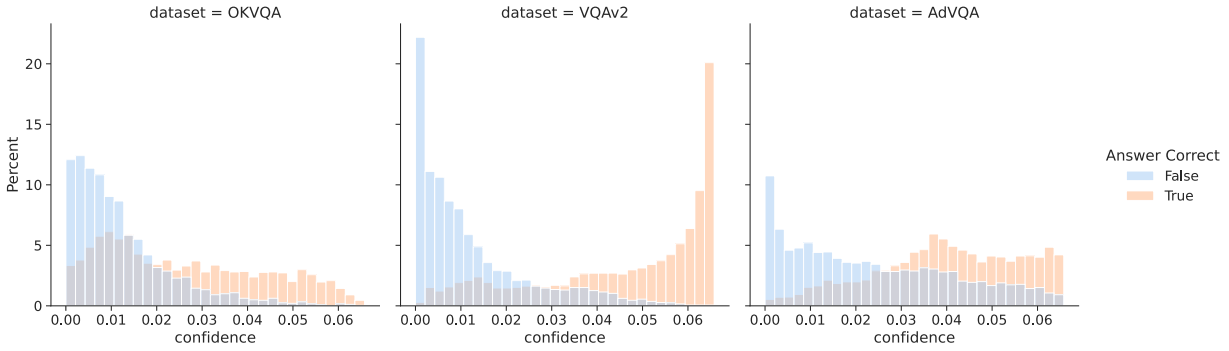


Figure 2. For out of distribution (OKVQA) and adversarial visual (AdVQA) questions, confidence scores alone do not work well to separate right from wrong answers — many correct answers are low confidence for OOD data, and many wrong answers are high confidence for adversarial data. **Note: Displayed confidence scores are raw. See Appendix for discussion on calibration.**

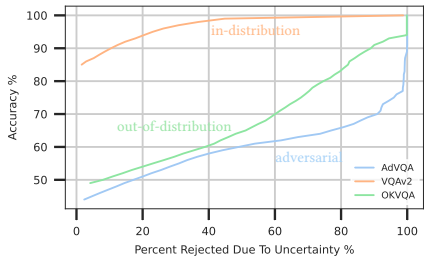


Figure 3. Selective VQA performance of a VLM (BLIP) on three datasets: adversarial (AdVQA), out-of-distribution (OKVQA), and in-distribution (VQAv2). On OOD and adversarial questions, the model has a harder time identifying which questions it should abstain from.

tions are low confidence and difficult to distinguish from incorrectly answered questions. A similar situation occurs for adversarial visual questions, in which many questions are incorrectly answered with high confidence.

Although the strategy of using model confidence *alone* to detect questions the model cannot answer is effective for in-distribution visual questions, this strategy fails on out-of-distribution and adversarial visual questions.

3. Method

3.1. Task Definition and Background

Given an image v and question q, the task of selective visual question answering is to decide whether a model $f_{VQA}(v, q)$ should predict an answer a, or abstain from making a prediction. A typical solution to this problem is to train a selection function $g(\cdot)$ that produces an abstention score $p_{\text{rej}} \in [0, 1]$. The simplest selection function would be to take the rejection probability $p_{\text{rej}} = 1 - p(a|q, v)$ where $p(a|q, v)$ is the model confidence that a is the answer, and then use a threshold τ so that the model abstains when $p_{\text{rej}} > \tau$ and predicts otherwise. A more complex approach taken by Whitehead et al. [29] is to train a parametric selection function $g(\mathbf{z}_v, \mathbf{z}_q; \theta)$ where $\mathbf{z}_v$ and $\mathbf{z}_q$ are the model's dense representations of the question and image respectively. The parameters θ are optimized on a held-out validation set, effectively training a classifier to predict when f_{VQA} will predict incorrectly on an input visual question v, q.

In the black box setting, access to the dense representations $\mathbf{z}_v, \mathbf{z}_q$ of the image v and question q is typically forbidden. Furthermore, even if access to the representation is allowed, a large number of evaluations of f_{VQA} would be needed to obtain the training data for the selection function. **Existing methods for selective prediction typically assume and evaluate a fixed set of classes, but for VQA, the label space can shift for each task (differing sets of acceptable answers for different types of questions) or be open-set.**

1. The approach should not require access to the black-box model's internal representations of v, q.
2. The approach should be model agnostic, as the architecture of the black-box model is unknown.
3. The approach should not require a large number of predictions from black-box model to train a selection function, because each usage of the black-box model incurs a financial cost, which can be substantial if large number of predictions are needed to train an auxiliary model.
4. Similarly, the approach should not require a held-out validation set for calibrating predictions, because this potentially requires a large number of evaluations of the black-box model.

3.2. Deep Structure and Surface Forms

Within the field of linguistics, a popular view first espoused by Chomsky [3] is that every natural language sentence has both a surface form and a deep structure. Multiple surface forms can be instances of the same deep structure. Simply put, multiple sentences that have different words arranged in different orders can mean the same thing. A rephrasing of a question corresponds to an alternate surface form, but the same deep structure. We thus expect that the answer to a rephrasing of a question should be the same as the original question. If the answer to a rephrasing is inconsistent with the answer to an original question, it indicates the model is sensitive to variations in the surface form of the original question. This indicates the model's understanding of the question is highly dependent on superficial characteristics, making it a good candidate for abstention — we hypothesize inconsistency on the rephrasings can be used to better quantify predictive uncertainty and reject questions a model has not understood.

3.3. Rephrasing Generation as Neighborhood Sampling

The idea behind many methods for representation learning is that a good representation should map multiple surface forms close together in feature space. For example, in contrastive learning, variations in surface form are generated by applying augmentations to an input, and the distance between multiple surface forms is minimized. In general, a characteristic of deep representation is that surface forms of an input should be mapped close together in feature space. Previous work, such as Attribution-Based Confidence [11] and Implicit Semantic Data Augmentation [28], exploit this by perturbing input samples in *feature space* to explore the neighborhood of an input. In a black-box setting, we don't have access to the features of the model, so there is no direct way to explore the neighborhood of an input in feature space. An alternate surface form of the input should be mapped close to the original input in feature space. Thus, a surface form variation of an input *should* be a neighbor of the input in feature space. *Generating* a surface form variation of a natural language sentence corresponds to *a rephrasing* of the natural language sentence. Since a rephrasing of a question is a surface form variation of a question, and surface form variations of an input should be mapped close to the original input in feature space, a rephrasing of a question is analogous to a sample from the neighborhood of a question. **We discuss this further in the appendix.**

3.4. Cyclic Generation of Rephrasings

A straightforward way to generate a rephrasing of a question is to invert the visual question answering problem, as is done in visual question generation. Let $p(V), p(Q), p(A)$ be the distribution of images, questions, and answers respectively. Visual question generation can be framed as approximating $p(Q|A,V)$, in contrast to visual question answering, which approximates $p(A|Q,V)$. We want to probe the predictive uncertainty of a black box visual question answering model $f_{BB}(\cdot)$ on an input visual question pair v, q where $v \sim p(V)$ is an image and $q \sim p(Q)$ is a question.. The VQA model f_{BB} approximates $p(A|Q,V)$. Let the answer a assigned the highest probability by the VQA model $f_{BB}(\cdot)$ be taken as the prospective answer. A VQG model $f_{VQG} \approx p(Q|A,V)$ can then be used to generate a rephrasing of an input question q. To see how, consider feeding the highest probability answer a from $f_{BB}(\cdot) \approx p(A|Q,V)$ into $f_{VQG}(\cdot) \approx p(Q|A,V)$ and then sampling a sentence $q' \sim f_{VQG} \approx p(Q|A,V)$ from the visual question generation model. In the case of an ideal $f_{VQG}(\cdot)$ and perfectly consistent $f_{BB}(\cdot)$, q' should be a generated question for which $p(a|q',v) \geq p(a_i|q',v) \forall a_i \in A$, with equality occurring in the case that $a_i = a$. So, q' is a question having the same answer as q, which is practically speaking, a rephrasing. We provide an algorithm listing in Algorithm 1.

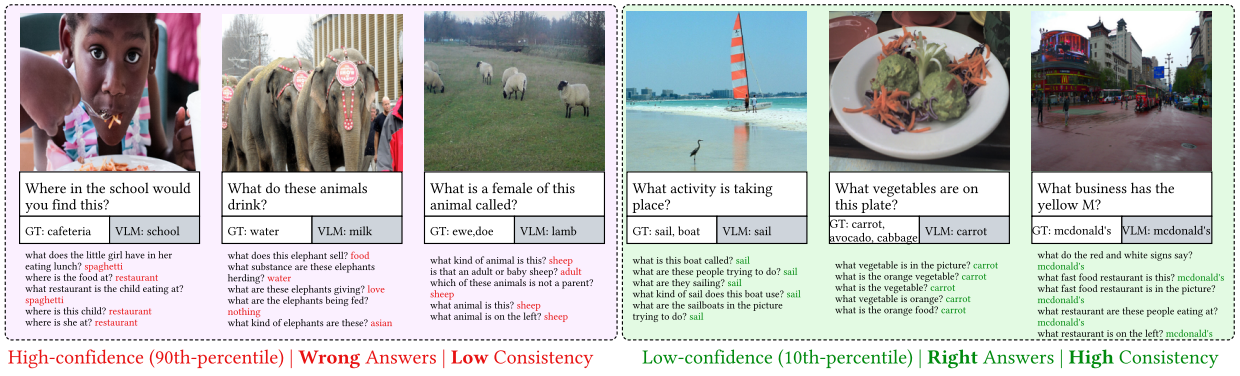
To summarize, we ask the black box model for an answer to a visual question, then give the predicted answer to a visual question generation model to produce a question q' conditioned on the image v and the answer a by the black box model, which corresponds to a question the VQG model thinks *should* lead to the predicted answer a. We assume the rephrasings generated by f_{VQG} are good enough, f_{BB} *should* be consistent on the rephrasings, and inconsistency indicates a problem with f_{BB}. In practice, each q' is not guaranteed to be a rephrasing (see Fig. 4) due to the probabilistic nature of the sampling process and because the VQG model is not perfect. The VQG model can be trained by following any procedure that results in a model approximating $p(a|q,v)$ that is an autoregressive model capable of text generation conditional on multimodal image-text input. The training procedure of the VQG model is an implementation detail we discuss in Sec. 3.5.

3.5. Implementation Details

We initialize the VQG model f_{VQG} from a BLIP checkpoint pretrained on 129m image-text pairs, and train it to maximize $p(a|q,v)$ using a standard language modeling loss. Specifically, we use

$$\mathcal{L}_{\text{VQG}} = -\sum_{n=1}^{N} \log P_\theta \left(y_n \mid y_{<n}, a, v\right) \tag{1}$$

where $y_1, Y_2, \ldots y_n$ are the tokens of a question q and a, v are the ground-truth answer and image, respectively, from a vqa triplet (v, q, a). We train for 10 epochs, using an AdamW [19] optimizer with a weight decay of 0.05 and decay the learning rate linearly to 0 from 2e-5. We use a batch size of 64 with an image size of 480×480, and train the model on



High-confidence (90th-percentile) | **Wrong** Answers | **Low** Consistency

Low-confidence (10th-percentile) | **Right** Answers | **High** Consistency

Figure 4. Examples showing the use of model-generated rephrasings to identify errors in model predictions with BLIP as the black box model f_{BB}. In the left panel, we show high-confidence answers that wrong, and identified by their low consistency across rephrasings. In the right panel, we show low-confidence answers that are actually correct, identified by their high-confidence across rephrasings.

the VQAv2 training set [9]. To sample questions from the VQG model, we use nucleus sampling [10] with a top-p of 0.9.

Algorithm 1: Probing Predictive Uncertainty of a Black-Box Vision-Language Model

Input: v, q, k
Data: f_{BB}, f_{VQG}
Result: $c \in \mathbb{Q}$: the consistency of f_{BB} over k rephrasings of v, q
$a_0 \longleftarrow f_{BB}(q, v)$;
$c \longleftarrow 0$;
for $i \leftarrow 0$ **to** k **do**
  $q' \longleftarrow nucleus\_sample(f_{VQG}(v, a_0))$;
  $a' \longleftarrow f_{BB}(q', v)$;
  if $a' = a_0$ **then**
    $c \longleftarrow c + 1$;
  end
end
return $c \div k$

4. Experiments

We conduct a series of experiments probing predictive uncertainty in a black-box visual question answering setting over two large vision-language models and three datasets. The primary task we use to probe predictive uncertainty is selective visual question answering, which we give a detailed description of in Sec. 3.5. **Futher qualitative examples and results can be found in the appendix.**

4.1. Experimental Setup

Black-box Models The experimental setup requires a black-box VQA model f_{BB} and a rephrasing generator f_{VQG}. We describe the training of the rephrasing generator f_{VQG} in Sec. 3.5. We choose ALBEF [15], BLIP [16], and BLIP-2[17] as our black-box models. ALBEF and BLIP have $\approx$ 200m parameters, while the version of BLIP-2 we use is based on the 11B parameter FLAN-T5 [4] model. ALBEF has been pretrained on 14m image-text pairs, while BLIP has been pretrained on over 100m image-text pairs, and BLIP-2 is aligned on 4M images. We use the official checkpoints provided by the authors, finetuned on Visual Genome [13] and VQAv2 [9] with 1.4m and 440k training triplets respectively.

Datasets We evaluate in three settings: in-distribution, out-of-distribution, and adversarial. For the in-distribution setting, we pairs from the VQAv2 validation set following the selection of [23]. For the out-of-distribution setting, we use OK-VQA [20], a dataset for question answering on natural images that requires outside knowledge. OK-VQA is an natural choice for a out-of-distribution selective prediction task, because many of the questions require external knowledge that a VLM may not have acquired, even through large scale pretraining. On such questions, a model that knows what it doesn't know should abstain due to lack of requisite knowledge. Finally, we consider adversarial visual questions in the AdVQA [24]. We use the official validation splits provided by the authors. The OK-VQA, AdVQA, and VQAv2 validation sets contain 5k, 10k, and 40k questions respectively.

4.2. Properties of Consistency

In this section we, analyze properties of the consistency. We are interested in:

1. Is increased consistency on rephrasings correlated with model accuracy on the original question?
2. What does the confidence distribution look like for different levels of consistency?

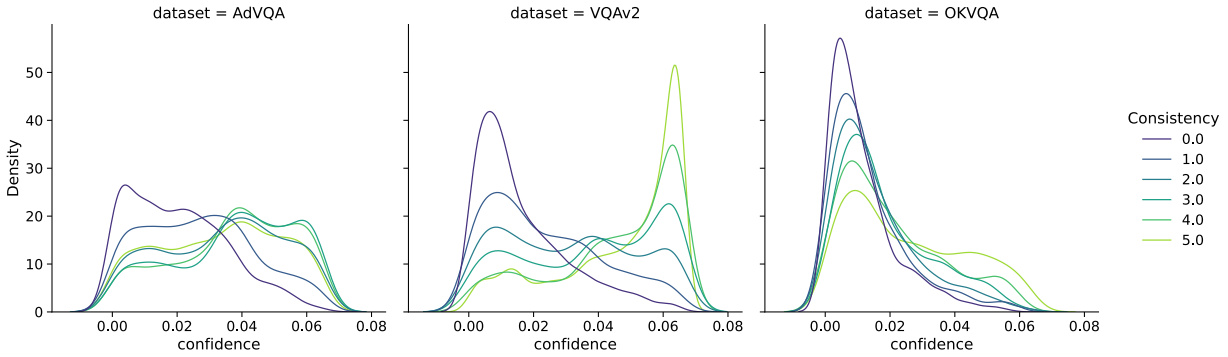


Figure 5. The distribution of confidence scores of f_{BB} at each level of consistency. While higher levels of consistency have a larger proportion of high confidence answers, they also retain a large number of low confidence answers, showing that consistency defines a different ordering over questions than confidence scores alone. BLIP is used as the black-box model f_{BB}.

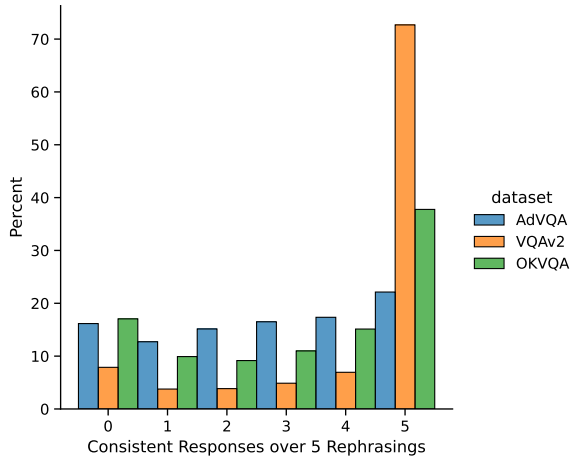


Figure 6. The percentage of each dataset at a given level of consistency. On a well-understood, in-distribution dataset (VQAv2), a large percentage of the questions are at a high consistency level.

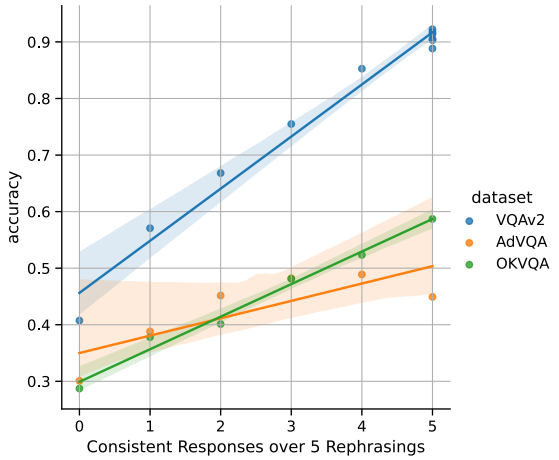


Figure 7. The accuracy of the answers of a VQA model (BLIP) plotted as a function of how consistent each answer was over up to 5 rephrasings of an original question.

3. What is the distribution of consistency across different datasets?

In Fig. 7 we plot the accuracy of the answers when f_{BB} is BLIP by how consistent each answer was over up to 5 rephrasings of an original question. We find that consistency over rephrasings is correlated with accuracy across all three datasets, through the correlation is weakest on adversarial data. Increased consistency on the rephrasings of a question implies lower risk on the original answer to the original question. Next, we examine how the distribution of model confidence varies across consistency levels in Fig. 5. Across all datasets, slices of a dataset at higher consistency levels also have a greater proportion of high-confidence answers, but *retain a substantial proportion of low confidence answers*. This clearly shows that consistency and confidence are not equivalent, and define different orderings on a set of questions and answers. Put another way, low confidence on a question does not preclude high consistency on a question, and similarly, high confidence on a question does not guarantee the model will be highly consistent on rephrasings of a question. Finally, plot the percentage of each dataset at a given level of consistency in Fig. 6. The in-distribution dataset, VQAv2, has the highest proportion of questions with 5 agreeing neighbors, with all other consistency levels making up the rest of the dataset. For the out-of-distribution dataset (OKVQA), a substantial proportion of questions ($\approx 40\%$) have five agreeing neighbors, with the rest of the dataset shared roughly equally between the other consistency levels. On the adver-

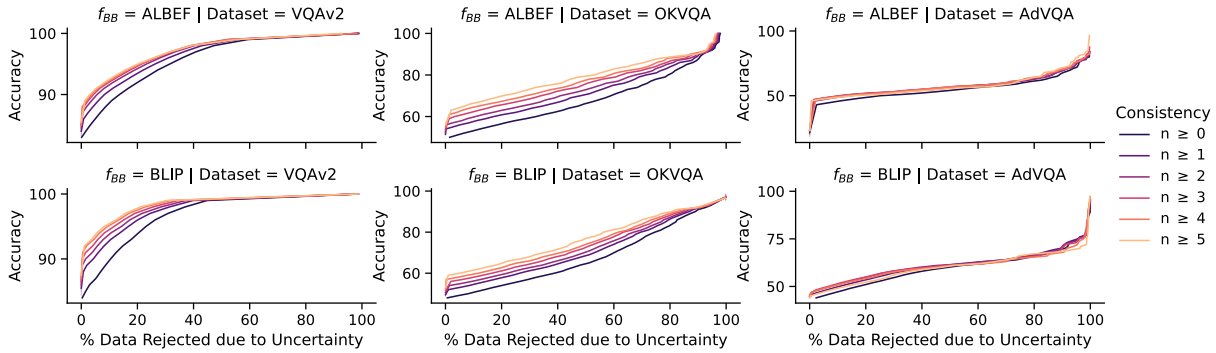


Figure 8. Risk-coverage curves at on slices of test datasets at different levels of consistency. A curve labeled $n \geq k$ shows the risk-coverage tradeoff for a slice of the target dataset where the answers of the model are consistent over at least k rephrasings of an original question. The $n \geq 0$ curve is the baseline. Higher consistency levels identify questions on which a model can achieve lower risk across all datasets.

f_{BB}	BLIP					ALBEF				
Risk Consistency	10.0	15.0	20.0	30.0	40.0	10.0	15.0	20.0	30.0	40.0
n ≥ 0	0.11	0.18	0.25	0.4	0.61	0.08	0.14	0.21	0.41	0.68
n ≥ 1	0.13	0.22	0.3	0.47	0.74	0.1	0.18	0.29	0.52	0.83
n ≥ 2	0.14	0.23	0.33	0.51	0.78	0.1	0.21	0.32	0.59	0.89
n ≥ 3	0.16	0.26	0.37	0.56	0.84	0.12	0.23	0.37	0.66	0.97
n ≥ 4	0.18	0.28	0.38	0.59	0.88	**0.13**	0.26	0.42	0.71	**1.0**
n ≥ 5	**0.19**	**0.31**	**0.44**	**0.65**	**0.95**	0.11	**0.33**	**0.47**	**0.8**	**1.0**

Table 1. OK-VQA coverage at a specified risk levels, stratified by consistency levels. $n \geq k$ indicates that the prediction of the model was consistent over at least k rephrasings of the question.

f_{BB}	BLIP					ALBEF				
Risk Consistency	20.0	30.0	40.0	50.0	56.0	20.0	30.0	40.0	50.0	60.0
n ≥ 0	0.01	0.09	0.51	0.83	0.98	0.0	0.07	0.24	0.75	1.0
n ≥ 1	0.01	**0.11**	0.58	0.9	**1.0**	0.01	0.09	0.29	0.86	1.0
n ≥ 2	0.01	0.1	**0.61**	**0.93**	**1.0**	0.01	0.09	**0.3**	**0.89**	1.0
n ≥ 3	0.01	0.1	0.58	**0.93**	**1.0**	0.02	0.11	**0.3**	**0.89**	1.0
n ≥ 4	0.01	0.08	0.55	0.92	**1.0**	0.02	0.11	**0.3**	0.87	1.0
n ≥ 5	0.01	0.04	0.53	0.87	**1.0**	**0.04**	**0.12**	0.27	0.84	1.0

Table 2. AdVQA coverage at a specified risk levels, stratified by consistency levels. $n \geq k$ indicates that the prediction of the model was consistent over at least k rephrasings of the question.

sarial dataset (AdVQA), the distribution is nearly flat, with equal slices of the dataset at each consistency level. One conclusion from this is that higher consistency is not necessarily rarer, and is highly dependent on how well a model understands the data distribution the question is drawn from.

4.3. Selective VQA with Neighborhood Consistency

We turn to the question of whether consistency over rephrasings is useful in the setting of selective visual question answering.

1. Can consistency select slices of a test dataset which a model understands well (achieves lower risk), or alternatively, identify questions the model doesn't understand, and should reject (high risk)?
2. How well does consistency over rephrasings work to identify low / high risk questions in out-of-distribution and adversarial settings?
3. What happens when the question generator is much smaller than the black-box model?

To analyze how useful consistency is for separating low-risk from high-risk inputs, we use the task of selective visual question answering. In Fig. 8 we plot risk-coverage curves for in-distribution, out-of-distribution, and adversarial visual questions. Each curve shows the risk-coverage tradeoff for questions at a level of consistency. For example, a curve labeled as $n \geq 3$ shows the risk-coverage tradeoff for questions on which 3 or more neighbors (rephrasings) were consistent with the original answer. Hence, the $n \geq 0$ curve is a baseline representing the risk-coverage curve for any question, regardless of consistency. If greater consistency over rephrasings is indicative over lower risk (and a higher probability the model knows the answer), we expect to see that the model should be able to achieve lower risk on slices of a dataset that the model is more consistent on. On in-distribution visual questions (VQAv2), the model achieves lower risk at equivalent coverage for slices of the dataset that have higher consistency levels. A similar situation holds for the out-of-distribution dataset, OKVQA, and the adversarial dataset AdVQA. In general, the model is able to achieve lower risk on slices of a dataset on which the consistency of the model over rephrasings is higher. In Tab. 2 and Tab. 1, we show risk-coverage information in tabular form for AdVQA and OK-VQA at specific risk levels. Finally, in Fig. 3, we show that our approach works even when there is a large size difference between the black-box model and the question generator.

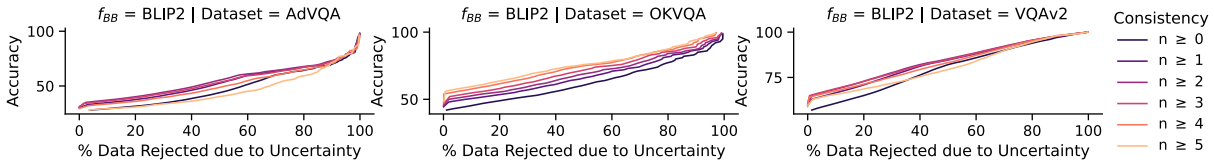


Figure 9. Risk-coverage curves when f_{VQG} (200m parameters) is substantially smaller than f_{BB} (11B). Even in this scenario, f_{VQG} can reliably identify high-risk questions based on consistency.

5. Related Work

5.1. Selective Prediction

Deep models with a reject option have been studied in the context of unimodal classification and regression [7, 8, 30] for some time, and more recently for the open-ended task of question answering [12]. Deep models with a reject option in the context of visual question answering were first explored by Whitehead et al. [29]. They take the approach of training a selection function using features from the model and a held-out validation set to make the decision of whether to predict or abstain. Dancette et al. [6] takes an alternate approach by training models on different dataset slices. The problem of eliciting truthful information from a language model [18] is closely related to selective prediction for VQA. In both settings, the model must avoid providing false information in response to a question.

5.2. Self-Consistency

Jha et al. [11] introduced the idea of using consistency over the predictions of a model to quantify the predictive uncertainty of the model. Their Attribution Based Confidence (ABC) metric is based on using guidance from feature attributions, specifically Integrated Gradients [26] to perturb samples in feature space, then using consistency over the perturbed samples to quantify predictive uncertainty. Shah et al. [23] show that VQA models are not robust to linguistic variations in a sentence by demonstrating inconsistency of the answers of multiple VQA models over human-generated rephrasings of a sentence. Similarly, Selvaraju et al. [22] show that the answers of VQA models to more complex reasoning questions are inconsistent with the answers to simpler perceptual questions whose answers should entail the answer to the reasoning question. We connect these ideas to hypothesize that inconsistency on linguistic variations of a visual question is indicative of more superifical understanding of the content of the question, and therefore a higher chance of being wrong when answering the question.

5.3. Robustness of VQA Models

VQA models have been shown to lack robustness, and severely prone to overfitting on dataset-specific correlations rather than learning to answer questions. The VQA-CP [1] task showed that VQA models often use linguistic priors to answer questions (e.g. the sky is usually blue), rather than looking at the image. Dancette et al. [5] showed that VQA models often use simple rules based on co-occurences of objects with noun phrases to answer questions. The existence of adversarial visual questions has also been demonstrated by [24], who used an iterative model-in-the-loop process to allow human annotators to attack state-of-the-art While VQA models are approaching human-level performance on the VQAv2 benchmark [9], their performance on more complex VQA tasks such as OK-VQA [20] lags far behind human performance.

6. Conclusion

The capital investment required to train large, powerful models on massive amounts of data means that there is a strong commercial incentive to keep the weights and features of a model private while making the model accessible through an API. Using these models in low-risk situations is not problematic, but using black-box models in situations where mistakes can have serious consequences is dangerous. At the same time, the power of these black-box models makes using them very appealing.

In this paper, we explore a way to judge the reliability of the answer of a black-box visual question answering model by assessing the consistency of the model's answer over rephrasings of the original question, which we generate dynamically using a VQG model. We show that this is analogous to the technique of consistency over neighborhood samples, which has been used in white-box settings for self-training as well as predictive uncertainty. We conduct experiments on in-distribution, out-of-distribution, and adversarial settings, and show that consistency over rephrasings is correlated with model accuracy, and predictions of a model that are highly consistent over rephrasings are more likely to be correct. Hence, consistency over rephrasings constitutes an effective first step for using a black-box visual question answering model reliably by identifying queries that a black-box model may not know the answer to.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4971–4980, 2017. 8

[2] Axel Brando, Damià Torres, Jose A. Rodríguez-Serrano, and Jordi Vitrià. Building uncertainty models on top of black-box predictive apis. *IEEE Access*, 8:121344–121356, 2020. 1

[3] N. Chomsky. *The Logical Structure of Linguistic Theory*. Springer, 1975. 4

[4] Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416, 2022. 5

[5] Corentin Dancette, Rémi Cadène, Damien Teney, and Matthieu Cord. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1554–1563, 2021. 8

[6] Corentin Dancette, Spencer Whitehead, Rishabh Maheshwary, Ramakrishna Vedantam, Stefan Scherer, Xinlei Chen, Matthieu Cord, and Marcus Rohrbach. Improving selective visual question answering by learning from your peers. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24049–24059, 2023. 1, 8

[7] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *NIPS*, 2017. 1, 8

[8] Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning*, 2019. 8

[9] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5, 8

[10] Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *ArXiv*, abs/1904.09751, 2019. 5

[11] Susmit Jha, Sunny Raj, Steven Fernandes, Sumit K Jha, Somesh Jha, Brian Jalaian, Gunjan Verma, and Ananthram Swami. Attribution-Based Confidence Metric For Deep Neural Networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 1, 2, 4, 8

[12] Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Annual Meeting of the Association for Computational Linguistics*, 2020. 8

[13] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2016. 5

[14] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NIPS*, 2016. 1

[15] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Neural Information Processing Systems*, 2021. 5

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022. 2, 5

[17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597, 2023. 5

[18] Stephanie C. Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Annual Meeting of the Association for Computational Linguistics*, 2021. 8

[19] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101, 2017. 4

[20] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3190–3199, 2019. 2, 5, 8

[21] Hussein Mozannar and David A. Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, 2020. 1

[22] Ramprasaath R. Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Tulio Ribeiro, Besmira Nushi, and Ece Kamar. Squinting at vqa models: Introspecting vqa models with sub-questions. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10000–10008, 2020. 8

[23] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *2019 Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 5, 8

[24] Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Alberto Lopez Magana, Wojciech Galuba, Devi Parikh, and Douwe Kiela. Human-adversarial visual question answering. 2021. 2, 5, 8

[25] Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. Black-box tuning for language-model-as-a-service. *ArXiv*, abs/2201.03514, 2022. 1

[26] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *ArXiv*, abs/1703.01365, 2017. 8

[27] Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. 2020. 1

[28] Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Cheng Wu, and Gao Huang. Implicit semantic data augmentation for

deep networks. In *Neural Information Processing Systems*, 2019. 4

[29] Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph E. Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision*, 2022. 1, 3, 8

[30] Liu Ziyin, Zhikang T. Wang, Paul Pu Liang, Ruslan Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. Deep gamblers: Learning to abstain with portfolio theory. *ArXiv*, abs/1907.00208, 2019. 1, 8