Selective Prediction For Open-Ended Question Answering in Black-Box Vision-Language Models

Zaid Khan **Yun Fu**
Northeastern University
Boston, USA
{khan.za,yun.fu}@northeastern.edu

Abstract

When mistakes have serious consequences, reliable use of a model requires understanding when the predictions of the model are trustworthy. One approach is selective prediction, in which a model is allowed to abstain if it is uncertain. Existing methods for selective prediction require access to model internals, retraining, or large number of model evaluations, and cannot be used for black box models available only through an API. This is a barrier to the use of powerful commercial foundation models in risk-sensitive applications. Furthermore, existing work has largely focused on unimodal foundation models. We propose a method to improve selective prediction in a black box vision-language model by measuring consistency over the neighbors of a visual question. Although direct sampling of the neighborhood is not possible, we propose using a probing model as a proxy. We describe experiments testing the proposed method on in-distribution, out-of-distribution and adversarial questions. We find that the consistency of a vision-language model across rephrasings of a visual question can be used to identify and reject high-risk visual questions, even in out-of-distribution and adversarial settings, constituting a step towards safe use of black-box vision-language models.

1 Introduction

Foundation models are sometimes only available as black boxes accessible through an API [1, 2] for commercial reasons, risk of misuse, or privacy considerations. A black box model is difficult to use safely for high-risk scenarios in which it is preferable that a model defers to an expert or abstains from answering rather than deliver an incorrect answer [3]. Many approaches for selective prediction [3, 4] or improving the predictive uncertainty of a model exist, such as ensembling [5], gradient-guided sampling in feature space [6], retraining the model [7], or training a auxiliary module using model predictions [8]. 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 [9] (VQA). Despite the progress in selective prediction, *current methods are not appropriate for models available only in a black-box setting, such as models-as-a-service*, where access to the internal representations is not available, retraining is infeasible, and each evaluation is expensive.

Black-box predictive uncertainty has been studied previously, but existing methods require a large number of evaluations to build an auxiliary model [2, 8], which can be prohibitively expensive when each evaluation has a non-neglible financial cost, or are designed for tasks with a closed-world assumption [10] with a small label space. Furthermore, while predictive uncertainty for *unimodal* large language models has been the subject of significant study [11–13], the predictive uncertainty of vision-language models (VLMs) has been studied only by Whitehead et al. [9], but their evaluation focuses on a white-box setting and smaller VLMs without web-scale pretraining. Black-box tuning of large models for

R0-FoMo: Workshop on Robustness of Few-shot and Zero-shot Learning in Foundation Models at NeurIPS 2023.

increased performance [1] is possible, but little is known about improving or understanding predictive uncertainty for large black-box models. In this paper, we consider selective prediction for large, black-box VLMs, which implies training data is private, model features and gradients are unavailable, and ensembling / retraining are not possible, all of which are typical features of models-as-a-service. We hypothesize that we can apply the principle of consistency over neighborhood samplings [6] used in white-box settings for *black box uncertainty estimates for visual question answering*, by using question generation to approximate sampling from the neighborhood of an input question without access to the features. We describe how rephrasings of a question can be viewed as samples from the neighborhood of a visual question pair. We propose using a visual question generation model as a *probing model* to produce rephrasings of questions given an initial answer from the black-box VLM, 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. This 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.

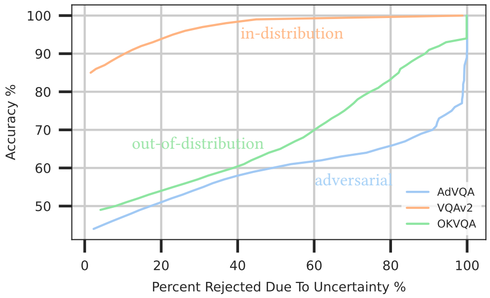


Figure 1: Selective VQA performance of a VLM (BLIP) on three datasets: adversarial (AdVQA), out-of-distribution (OKVQA), and in-distribution (VQAv2). On OOD and adversarial questions, the model has a harder time identifying which questions it should abstain from.

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 propose identifying high-risk inputs for visual question answering based on consistency over samples in the neighborhood of a visual question.
- 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 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.

2 Method

2.1 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. [9] 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.** Other constraints are:

1. No access to the black-box model's internal representations of v, q.
2. Model agnostic, as the architecture of the black-box model is unknown.

2.2 Deep Structure and Surface Forms

Within linguistics, a popular view espoused by Chomsky [14] is that every natural language sentence has a surface form and a deep structure. Multiple surface forms can be instances of the same deep structure: 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. Thus, 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, 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.

2.3 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 [6] and Implicit Semantic Data Augmentation [15], 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.**

2.4 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

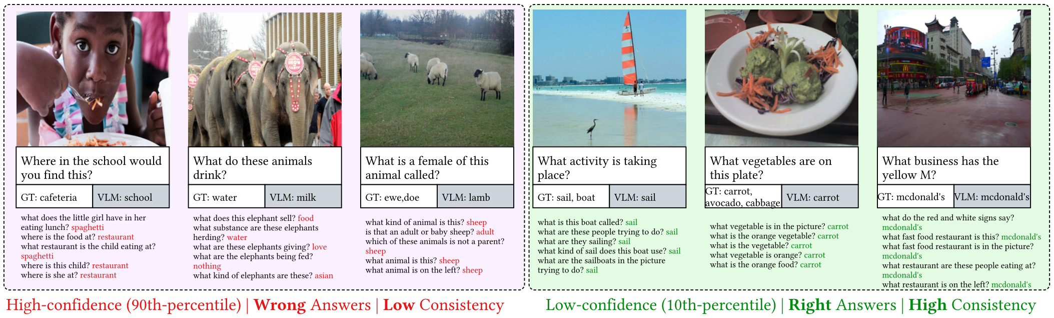

Where in the school would you find this?
GT: cafeteria
VLM: school
what does the little girl have in her eating lunch? spaghetti
where is the food at? restaurant
what restaurant is the child eating at? spaghetti
where is this child? restaurant
where is she at? restaurant
What do these animals drink?
GT: water
VLM: milk
what does this elephant sell? food
what substance are these elephants herding? water
what are these elephants giving? love
what are the elephants being fed? nothing
what kind of elephants are these? asian
What is a female of this animal called?
GT: ewe,doe
VLM: lamb
what kind of animal is this? sheep
is that an adult or baby sheep? adult
which of these animals is not a parent? sheep
what animal is this? sheep
what animal is on the left? sheep
What activity is taking place?
GT: sail, boat
VLM: sail
what is this boat called? sail
what are these people trying to do? sail
what are they sailing? sail
what kind of sail does this boat use? sail
what are the sailboats in the picture trying to do? sail
What vegetables are on this plate?
GT: carrot, avocado, cabbage
VLM: carrot
what vegetable is in the picture? carrot
what is the orange vegetable? carrot
what is the vegetable? carrot
what vegetable is orange? carrot
what is the orange food? carrot
What business has the yellow M?
GT: mcdonald's
VLM: mcdonald's
what do the red and white signs say? mcdonald's
what fast food restaurant is this? mcdonald's
what fast food restaurant is in the picture? mcdonald's
what restaurant are these people eating at? mcdonald's
what restaurant is on the left? mcdonald's
High-confidence (90th-percentile) | Wrong Answers | Low Consistency
Low-confidence (10th-percentile) | Right Answers | High Consistency

Figure 2: Examples showing the use of 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 are 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.

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.

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. 2) 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. 2.5.

2.5 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 [16] 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 the VQAv2 training set [17]. To sample questions from the VQG model, we use nucleus sampling [18] with a top-p of 0.9.

3 Experiments

The primary task we use to probe predictive uncertainty is selective visual question answering, which we give a detailed description of in Sec. 2.5. **Futher qualitative examples and results can be found in the appendix.**

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. 2.5. We choose ALBEF [19], BLIP [20], and BLIP-2[21] as our black-box models. ALBEF and BLIP have ≈ 200m parameters, while the version of BLIP-2 we use is based on the 11B parameter FLAN-T5 [22] 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

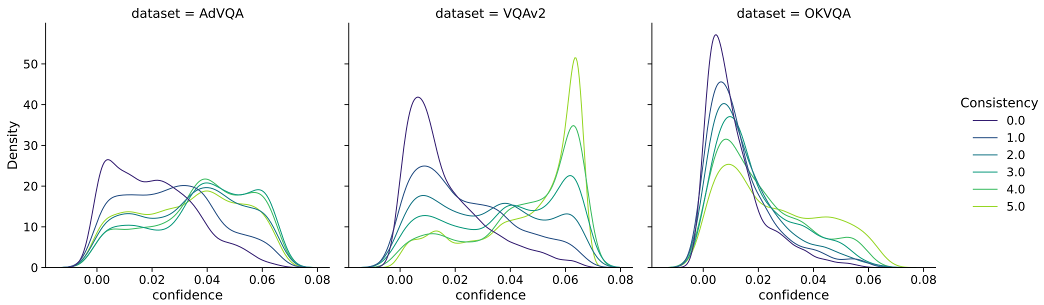


Figure 3: The distribution of confidence scores of f_{BB} at each level of consistency. While higher levels of consistency have a larger proportion of high confidence answers, they also retain a large number of low confidence answers, showing that consistency defines a different ordering over questions than confidence scores alone. BLIP is used as the black-box model f_{BB}.

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**

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 1: OKVQA (left, OOD) AdVQA (right, adversarial) coverage at a specified risk levels, stratified by consistency levels. $n \geq k$ means that the prediction of the model was consistent over at least k rephrasings of the question. The **bolded** numbers indicate which consistency level maximizes coverage at a specified risk level. In all cases, choosing consistency levels strictly greater than 0 is the optimal strategy to maximize coverage @ risk.

aligned on 4M images. We use the official checkpoints provided by the authors, finetuned on Visual Genome [23] and VQAv2 [17] 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 [24]. For the out-of-distribution setting, we use OK-VQA [25], 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 AdVQA [26]. 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.

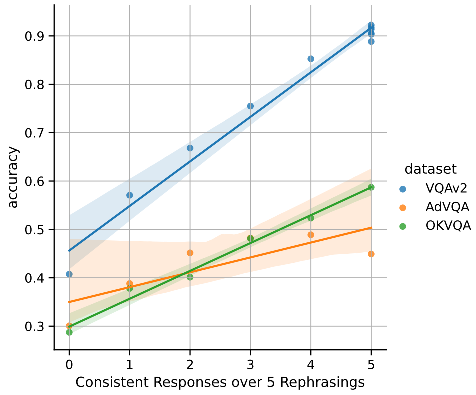


Figure 4: The accuracy of the answers of a VQA model (BLIP) plotted as a function of how consistent each answer was over up to 5 rephrasings of an original question. Consistency is correlated with accuracy.

3.1 Selective VQA with Neighborhood Consistency

In Fig. 4 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. Next, we examine how the distribution of model confidence varies across consistency levels in Fig. 3. 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.

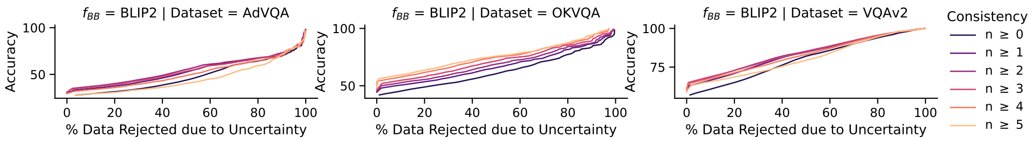


Figure 5: Risk-coverage curves when f_{VQG} (200m parameters) is substantially smaller than f_{BB} (11B). Even in this scenario, f_{VQG} can reliably identify high-risk questions based on consistency. A curve labeled $n \geq k$ shows the risk-coverage tradeoff for a slice of the target dataset where 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.

We turn to the question of whether consistency over rephrasings is useful in the setting of selective visual question answering. 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. 5 we plot risk-coverage curves for 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, even when there is large size difference between the black-box model and the question generator.

4 Related Work

Selective Prediction Deep models with a reject option have been studied in the context of unimodal classification and regression [3, 4, 27] for some time, and more recently for the open-ended task of question answering [13]. Deep models with a reject option in the context of visual question answering were first explored by Whitehead et al. [9]. They take the approach of training a selection function using featueres from the model and a held-out validation set to make the decision of whether to predict or abstain. The problem of eliciting truthful information from a language model [28] is closely related to selective prediction for VQA. In both settings, the model must avoid providing false information in response to a question.

Self-Consistency Jha et al. [6] 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 [29] to perturb samples in feature space, then using consistency over the perturbed samples to quantify predictive uncertainty. Shah et al. [24] 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. [30] 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 Conclusion

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.

References

[1] 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.

[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. doi: 10.1109/ACCESS.2020.3006711.

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

[4] 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.

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

[6] 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*, volume 32. Curran Associates, Inc., 2019.

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

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

[9] 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.

[10] Jakob Smedegaard Andersen, Tom Schöner, and Walid Maalej. Word-level uncertainty estimation for black-box text classifiers using RNNs. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5541–5546, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.484. URL `https://aclanthology.org/2020.coling-main.484`.

[11] Zhengbao Jiang, J. Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2020.

[12] Saurav Kadavath, Tom Conerly, Amanda Askell, T. J. Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zachary Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, John Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom B. Brown, Jack Clark, Nicholas Joseph, Benjamin Mann, Sam McCandlish, Christopher Olah, and Jared Kaplan. Language models (mostly) know what they know. *ArXiv*, abs/2207.05221, 2022.

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

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

[15] 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.

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

[17] 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.

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

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

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

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

[28] 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.

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

[30] 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.

[31] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

[32] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Neural Information Processing Systems*, 2019.

[33] Hong Liu, Jianmin Wang, and Mingsheng Long. Cycle self-training for domain adaptation. In *Neural Information Processing Systems*, 2021.

[34] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 625–632, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. doi: 10.1145/1102351.1102430. URL `https://doi.org/10.1145/1102351.1102430`.

[35] Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. *ArXiv*, abs/1904.01685, 2019.

f_{BB}	BLIP										ALBEF									
Risk	0.0	5.0	10.0	15.0	20.0	25.0	30.0	35.0	40.0	45.0	0.0	5.0	10.0	15.0	20.0	25.0	30.0	35.0	40.0	45.0
Consistency																				
n ≥ 0	0.0	0.0	0.11	0.18	0.25	0.32	0.4	0.49	0.61	0.77	0.02	0.03	0.08	0.14	0.21	0.3	0.41	0.53	0.68	0.85
n ≥ 1	0.0	0.0	0.13	0.22	0.3	0.38	0.47	0.59	0.74	0.89	0.02	0.04	0.1	0.18	0.29	0.4	0.52	0.66	0.83	0.97
n ≥ 2	0.0	0.0	0.14	0.23	0.33	0.42	0.51	0.63	0.78	0.94	0.03	0.04	0.1	0.21	0.32	0.45	0.59	0.73	0.89	**1.0**
n ≥ 3	0.0	0.0	0.16	0.26	0.37	0.45	0.56	0.68	0.84	**1.0**	0.03	0.05	0.12	0.23	0.37	0.51	0.66	0.83	0.97	**1.0**
n ≥ 4	0.0	0.0	0.18	0.28	0.38	0.48	0.59	0.74	0.88	**1.0**	**0.04**	**0.06**	**0.13**	0.26	0.42	0.55	0.71	0.88	**1.0**	**1.0**
n ≥ 5	0.0	0.0	**0.19**	**0.31**	**0.44**	**0.54**	**0.65**	**0.8**	**0.95**	**1.0**	**0.04**	**0.06**	0.11	**0.33**	**0.47**	**0.63**	**0.8**	**0.93**	**1.0**	**1.0**

Table 2: More granular risk-coverage data for OK-VQA.

f_{BB}	BLIP									ALBEF								
Risk	20.0	25.0	30.0	35.0	40.0	45.0	50.0	55.0	56.0	20.0	25.0	30.0	35.0	40.0	45.0	50.0	55.0	60.0
Consistency																		
n ≥ 0	0.01	**0.04**	0.09	0.23	0.51	0.69	0.83	0.95	0.98	0.0	0.04	0.07	0.12	0.24	0.46	0.75	0.92	1.0
n ≥ 1	0.01	**0.04**	**0.11**	**0.27**	0.58	0.76	0.9	**1.0**	**1.0**	0.01	0.05	0.09	0.15	0.29	0.55	0.86	**1.0**	1.0
n ≥ 2	0.01	**0.04**	0.1	0.25	**0.61**	0.79	**0.93**	**1.0**	**1.0**	0.01	0.05	0.09	0.15	**0.3**	0.59	**0.89**	**1.0**	1.0
n ≥ 3	0.01	**0.04**	0.1	0.25	0.58	**0.8**	**0.93**	**1.0**	**1.0**	0.02	0.06	0.11	0.17	**0.3**	**0.6**	**0.89**	**1.0**	1.0
n ≥ 4	0.01	0.02	0.08	0.24	0.55	0.77	0.92	**1.0**	**1.0**	0.02	0.06	0.11	0.16	**0.3**	**0.6**	0.87	**1.0**	1.0
n ≥ 5	0.01	0.01	0.04	**0.27**	0.53	0.72	0.87	**1.0**	**1.0**	**0.04**	**0.07**	**0.12**	**0.18**	0.27	0.53	0.84	**1.0**	1.0

Table 3: More granular risk-coverage data for AdVQA.

A Detailed Risk-Coverage Data

In Tabs. 2 to 5, we show more granular risk-coverage curves across all three evaluated datasets and both black-box models.

B Inference Details

For both BLIP and ALBEF, we follow the original inference procedures. Both models have an encoder-decoder architecture and VQA is treated as a text-to-text task. We use the rank-classification approach [31] to allow the autoregressive decoder of the VLM to predict an answer for a visual question. Concretely, let $\mathcal{A} = \{a_1, a_2, a_3, \dots a_k\}$ be a list of length k for a dataset consisting of the most frequent ground-truth answers. These answer lists are standardized and distributed by the authors of the datasets themselves. We use the standard answer lists for each dataset. Next, let v, q be a visual question pair and let f_{BB} be a VQA model. Recall that f_{BB} is a language model defining a distribution $p(a|q, v)$, and is thus able to assign a score to each $a_i \in \mathcal{A}$. We take the highest probability a_k

$$\max_{a_k \in \mathcal{A}} f_{BB}(v, q, a_k) = \max_{a_k \in \mathcal{A}} p(a_k|v, q) \tag{2}$$

as the predicted answer for a question. This is effectively asking the model to rank each of the possible answer candidates, turning the open-ended VQA task into a very large multiple choice problem. Note that the highest probability $a_k \in \mathcal{A}$ is *not* necessarily the answer that would be produced by $f_{BB} \sim p(a|v, q)$ in an unconstrained setting such as stochastic decoding. However, for consistency with previous work, we use the rank classification approach.

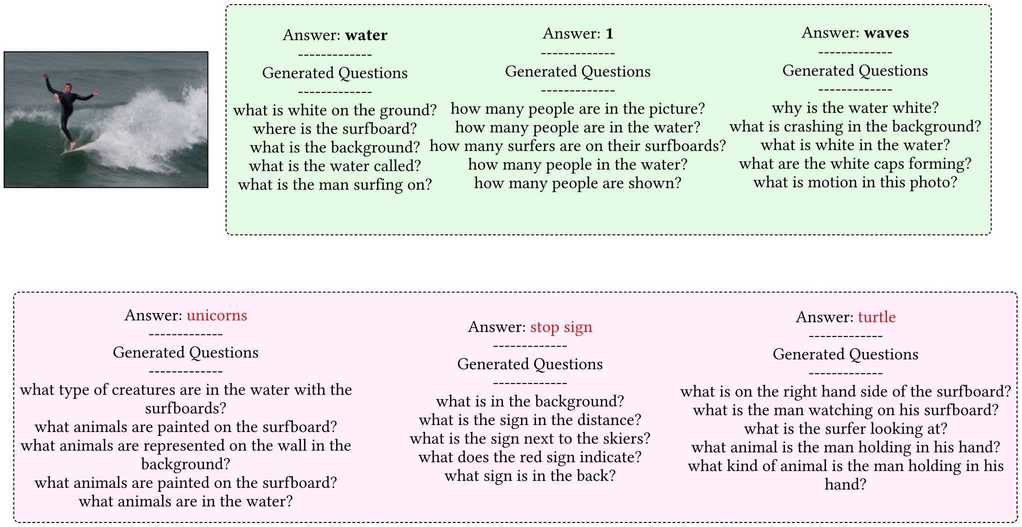
Visual question answering is thus treated differently when using large autoregressive vision-language models compared to non-autoregressive odels. In traditional approaches, VQA is treated as a classification task, and a standard approach used in older, non-autoregressive vision-language models such as ViLBERT [32] is to train a MLP with a cross-entropy loss with each of the possible answers as a class.

f_{BB}	BLIP														
risk	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	11.0	12.0	13.0	14.0
Consistency															
n ≥ 0	0.01	0.55	0.63	0.69	0.74	0.77	0.8	0.82	0.85	0.88	0.9	0.91	0.93	0.95	0.97
n ≥ 1	0.01	0.6	0.69	0.76	0.8	0.83	0.86	0.9	0.92	0.94	0.96	0.98	0.99	**1.0**	**1.0**
n ≥ 2	0.01	0.63	0.72	0.78	0.83	0.86	0.89	0.92	0.94	0.96	0.98	**1.0**	**1.0**	**1.0**	**1.0**
n ≥ 3	0.01	0.66	0.75	0.81	0.85	0.88	0.92	0.94	0.96	0.98	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
n ≥ 4	0.01	0.68	0.77	0.83	0.87	0.91	0.93	**0.96**	0.98	0.99	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
n ≥ 5	0.01	**0.7**	**0.79**	**0.84**	**0.88**	**0.92**	**0.94**	**0.96**	**0.99**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**

Table 4: Granular risk-coverage data for VQAv2 with BLIP as f_{BB}.

f_{BB}	ALBEF														
risk	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	11.0	12.0	13.0	14.0
Consistency															
n ≥ 0	0.01	0.55	0.63	0.69	0.74	0.77	0.8	0.82	0.85	0.88	0.9	0.91	0.93	0.95	0.97
n ≥ 1	0.01	0.6	0.69	0.76	0.8	0.83	0.86	0.9	0.92	0.94	0.96	0.98	0.99	**1.0**	**1.0**
n ≥ 2	0.01	0.63	0.72	0.78	0.83	0.86	0.89	0.92	0.94	0.96	0.98	**1.0**	**1.0**	**1.0**	**1.0**
n ≥ 3	0.01	0.66	0.75	0.81	0.85	0.88	0.92	0.94	0.96	0.98	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
n ≥ 4	0.01	0.68	0.77	0.83	0.87	0.91	0.93	**0.96**	0.98	0.99	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
n ≥ 5	0.01	**0.7**	**0.79**	**0.84**	**0.88**	**0.92**	**0.94**	**0.96**	**0.99**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**

Table 5: Granular risk-coverage data for VQAv2 with ALBEF as f_{BB}.



Answer: **water**
Generated Questions
what is white on the ground?
where is the surfboard?
what is the background?
what is the water called?
what is the man surfing on?

Answer: **1**
Generated Questions
how many people are in the picture?
how many people are in the water?
how many surfers are on their surfboards?
how many people in the water?
how many people are shown?

Answer: **waves**
Generated Questions
why is the water white?
what is crashing in the background?
what is white in the water?
what are the white caps forming?
what is motion in this photo?

Answer: unicorns
Generated Questions
what type of creatures are in the water with the surfboards?
what animals are painted on the surfboard?
what animals are represented on the wall in the background?
what animals are painted on the surfboard?
what animals are in the water?

Answer: stop sign
Generated Questions
what is in the background?
what is the sign in the distance?
what is the sign next to the skiers?
what does the red sign indicate?
what sign is in the back?

Answer: turtle
Generated Questions
what is on the right hand side of the surfboard?
what is the man watching on his surfboard?
what is the surfer looking at?
what animal is the man holding in his hand?
what kind of animal is the man holding in his hand?

Figure 6: The rephrasing generator f_{VQG} can hallucinate questions that imagine not present in the context of the image.

C Hallucinations

We describe a peculiar mode of the rephrasing generator f_{VQG} in this section. When an answer is out-of-context for a given image, the rephrasing generator f_{VQG} will generate questions premised on the out-of-context answer. For example, in Fig. 6, we show that if an out-of-context answer such as "unicorn" for the surfing image in Fig. 6 is provided to f_{VQG} for cycle-consistent rephrasing generation, f_{VQG} will generate questions such as "what animals are in the water", assuming that there are unicorns in the water, though this is implausible. A more correct question would have been something such as "what animals are not present?" A likely reason f_{VQG} cannot handle these cases well is because f_{VQG} is trained on a VQA dataset to approximate $p(q|v, a)$, and traditional VQA datasets have very few counterfactual questions such as these.

This is not specific to the f_{VQG} used in our framework, and should apply to any question generator trained in this manner. It does reveal that even large VLMs pretrained on a massive amount of image-text pairs have a superficial understanding of counterfactuals, and possibly other properties of language.

D Are the rephrasings really rephrasings?

As visible in Fig. 2, some of the rephrasings are not literally rephrasings of the original question. It may be more correct to call the rephrasings pseudo-rephrasings, in the same way that generated labels are referred to as pseudolabels in the semi-supervised learning literature [33]. However, the pseudo-rephrasings seem to be *good enough* that inconsistency over the pseudo-rephrasings indicates potentially unreliable predictions from f_{BB}.

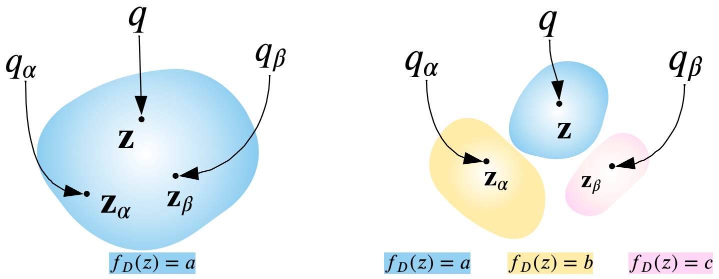


Figure 7: See Appendix D for an explanation of the figure.

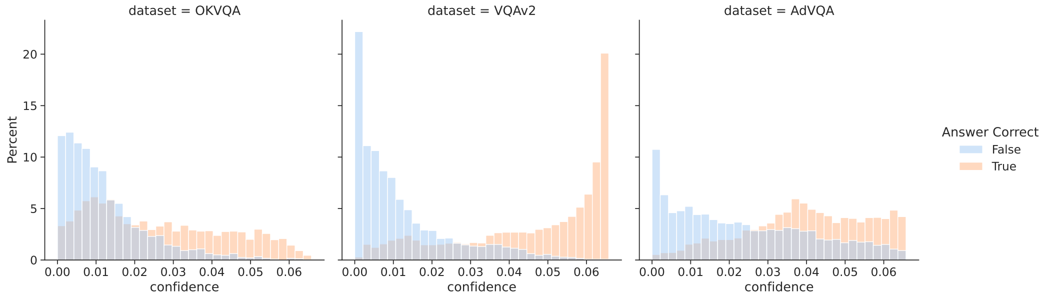


Figure 8: For out of distribution (OKVQA) and adversarial visual (AdVQA) questions, confidence scores alone do not work well to separate right from wrong answers — many correct answers are low confidence for OOD data, and many wrong answers are high confidence for adversarial data.

Why does this work? Decompose f_{BB} as $f_{BB} = f_D(f_E(v, q))$, where $f_E(v, z) = \mathbf{z}$ is the encoder that maps a visual question pair v, q to a dense representation $\mathbf{z}$, and $f_D(\mathbf{z}) = a$ is the decoder that maps the dense representation $\mathbf{z}$ to an answer. For two rephrasings q_α, q_β of a question q, the model will be consistent over the rephrasings if all the rephrasings are embedded onto a subset of the embedding space that f_D assigns the same answer a. This is the situation we depict on the left side of Fig. 7.

On the other hand, if q_α and q_β are embedded into parts of the embedding space that f_D assigns them different answers, the answers will not be consistent (right side of Fig. 7). Thus, whether a q_α, q_β are linguistically valid rephrasings does not matter so much as if q_α, q_β *should* technically have the same answer as the original question q. Of course, it is true that the answer to a linguistically valid rephrasing should be the same as the same as the answer to the question being rephrased. However, for any question, there are many other questions that have the same answer but are *not* rephrasings of the original question.

E Calibration

The confidence scores in Figs. 3 and 8 are the raw scores from the logits of the VQA model, in this case BLIP. Recall that the models under consideration are autoregressive models that approximate a probability distribution $p(a|v, q)$, where a can take on an infinite number of values — the model must be able to assign a score to any natural language sentence. The raw distribution of confidence scores is clearly truncated in the sense that all scores appear to lie in the interval $[0, 0.07]$. We apply temperature scaling [34] to assess how well the confidence scores are calibrated. In temperature scaling, the logits of a model are multiplied by a parameter τ. This is rank-preserving, and yields confidence scores that are more directly interpretable. In our case, we can use it to rescale the model logits into the interval $[0, 1]$ and analyze the *Adaptive Calibration Error* [35] of the model's predictions. We grid search the τ that minimizes the Adaptive ECE directly on the model predictions, and show the results in Tabs. 6 to 8. The Adaptive Calibration Error is lowest on the in-distribution dataset, highest on the adversarial dataset, and second highest on the out-of-distribution dataset. Notably, the model

	Raw Confidence	Accuracy	Scaled Confidence	Error
percentile				
0	0.020	0.477	0.390	0.087
10	0.022	0.507	0.430	0.077
20	0.024	0.540	0.473	0.067
30	0.026	0.573	0.522	0.051
40	0.029	0.604	0.577	0.026
50	0.032	0.647	0.643	0.004
60	0.036	0.699	0.723	0.024
70	0.041	0.766	0.819	0.053
80	0.047	0.831	0.934	0.104
90	0.054	0.909	1.000	0.091

Table 6: Calibration of BLIP on OK-VQA. For scaling, a temperature of 19.9 is used.

	Raw Confidence	Accuracy	Scaled Confidence	Error
percentile				
0	0.042	0.837	0.841	0.004
10	0.047	0.898	0.926	0.028
20	0.051	0.938	1.000	0.062
30	0.055	0.968	1.000	0.032
40	0.058	0.984	1.000	0.016
50	0.060	0.994	1.000	0.006
60	0.062	0.998	1.000	0.002
70	0.064	0.999	1.000	0.001
80	0.065	1.000	1.000	0.000
90	0.065	0.999	1.000	0.001

Table 7: Calibration of BLIP on VQAv2. For scaling, a temperature of 19.3 is used.

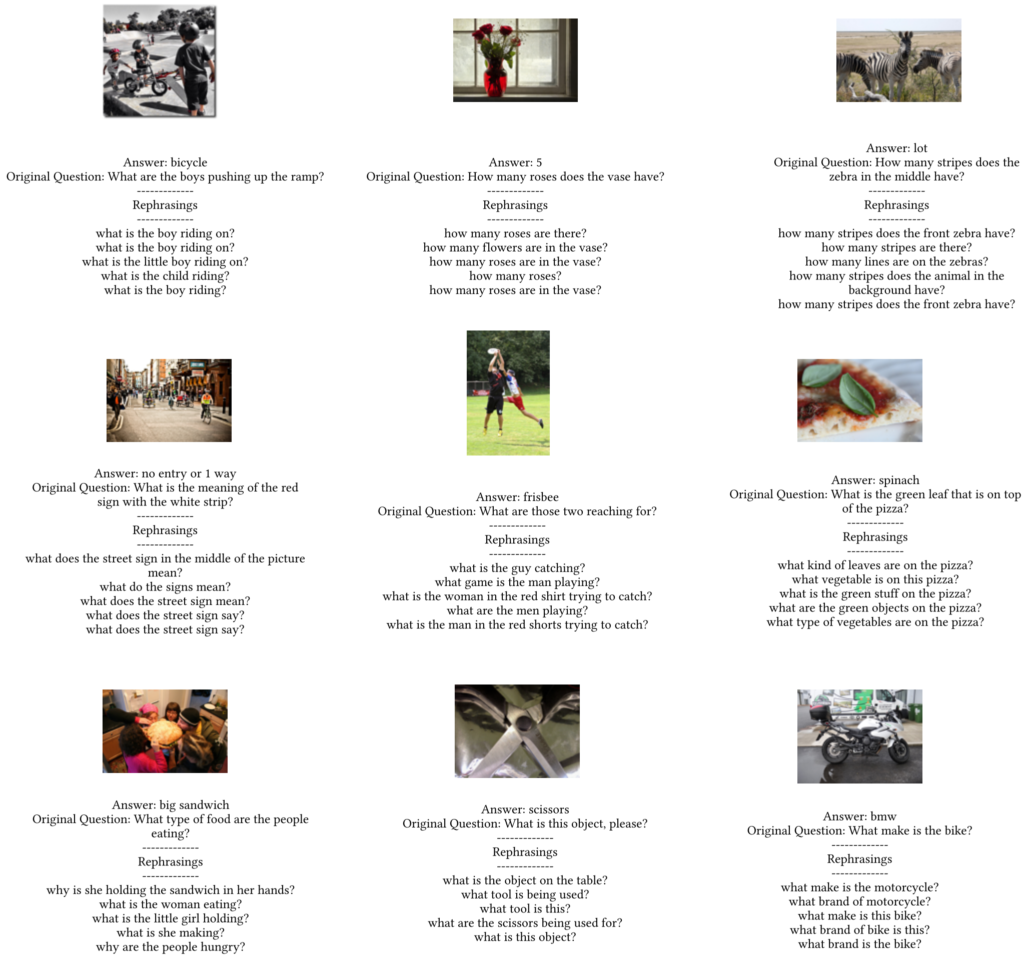
is systematically overconfident on adversarial samples, but not on out-of-distribution samples. This suggests that calibration is not the *only* problem in selective prediction.

F More Rephrasings Examples

We show more examples of generated rephrasings by Fig. 9.

	Raw Confidence	Accuracy	Scaled Confidence	Error
percentile				
0	0.032	0.430	0.637	0.206
10	0.035	0.472	0.703	0.231
20	0.039	0.510	0.769	0.259
30	0.042	0.547	0.834	0.287
40	0.045	0.580	0.897	0.317
50	0.048	0.601	0.956	0.355
60	0.051	0.618	1.000	0.382
70	0.055	0.636	1.000	0.364
80	0.058	0.655	1.000	0.345
90	0.062	0.693	1.000	0.307

Table 8: Calibration of BLIP on AdVQA. For scaling, a temperature of 12.5 is used.



Answer: bicycle
Original Question: What are the boys pushing up the ramp?

Rephrasings
----
what is the boy riding on?
what is the boy riding on?
what is the little boy riding on?
what is the child riding?
what is the boy riding?

Answer: 5
Original Question: How many roses does the vase have?

Rephrasings
----
how many roses are there?
how many flowers are in the vase?
how many roses are in the vase?
how many roses?
how many roses are in the vase?

Answer: lot
Original Question: How many stripes does the zebra in the middle have?

Rephrasings
----
how many stripes does the front zebra have?
how many stripes are there?
how many lines are on the zebras?
how many stripes does the animal in the background have?
how many stripes does the front zebra have?

Answer: no entry or 1 way
Original Question: What is the meaning of the red sign with the white strip?

Rephrasings
----
what does the street sign in the middle of the picture mean?
what do the signs mean?
what does the street sign mean?
what does the street sign say?
what does the street sign say?

Answer: frisbee
Original Question: What are those two reaching for?

Rephrasings
----
what is the guy catching?
what game is the man playing?
what is the woman in the red shirt trying to catch?
what are the men playing?
what is the man in the red shorts trying to catch?

Answer: spinach
Original Question: What is the green leaf that is on top of the pizza?

Rephrasings
----
what kind of leaves are on the pizza?
what vegetable is on this pizza?
what is the green stuff on the pizza?
what are the green objects on the pizza?
what type of vegetables are on the pizza?

Answer: big sandwich
Original Question: What type of food are the people eating?

Rephrasings
----
why is she holding the sandwich in her hands?
what is the woman eating?
what is the little girl holding?
what is she making?
why are the people hungry?

Answer: scissors
Original Question: What is this object, please?

Rephrasings
----
what is the object on the table?
what tool is being used?
what tool is this?
what are the scissors being used for?
what is this object?

Answer: bmw
Original Question: What make is the bike?

Rephrasings
----
what make is the motorcycle?
what brand of motorcycle?
what make is this bike?
what brand of bike is this?
what brand is the bike?

Figure 9: More examples of generated rephrasings.