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

Supplementary Material

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 1. 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 2. More granular risk-coverage data for AdVQA.

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**

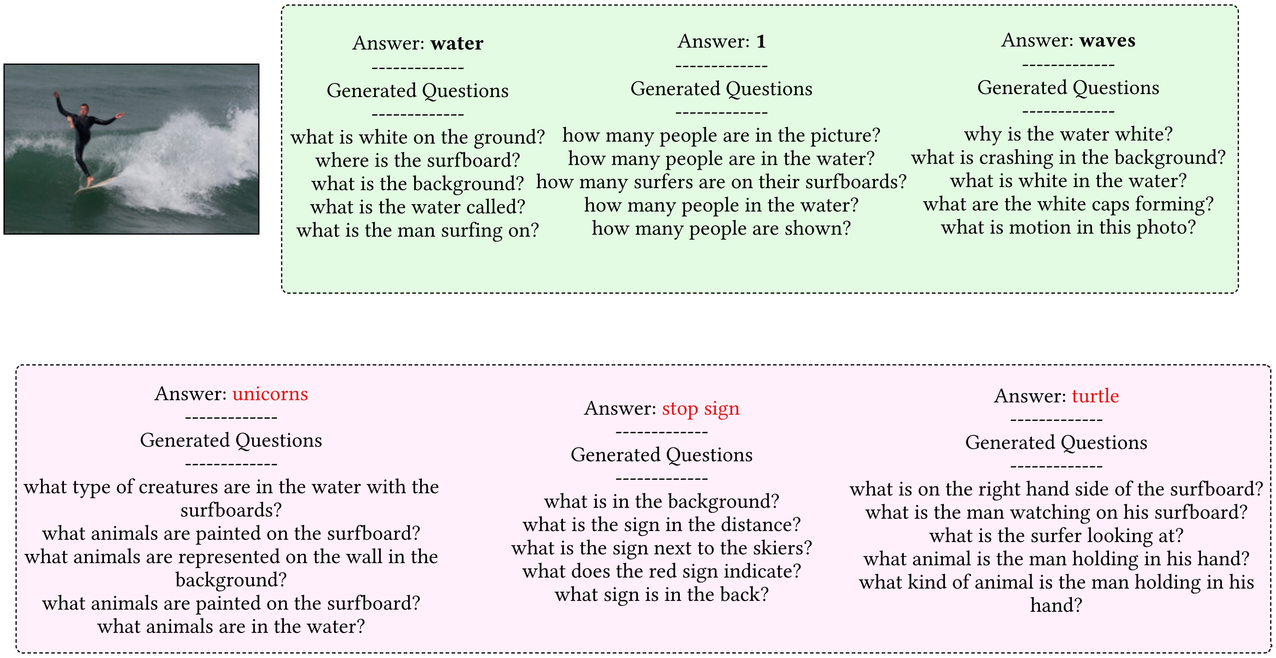
Table 3. Granular risk-coverage data for VQAv2 with BLIP as f_{BB}.

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 4. Granular risk-coverage data for VQAv2 with ALBEF as f_{BB}.

A. Detailed Risk-Coverage Data

In Tabs. 1 to 4, we show more granular risk-coverage curves across all three evaluated datasets and both black-box models.



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

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 [1] to allow the autoregressive decoder of the VLM to predict an answer for a visual question. Concretely, let $\mathcal{A} = \{a_1, a_2, a_3, \ldots a_k\}$ 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

$$\max_{a_k \in \mathcal{A}} f_{BB}(v, q, a_k) = \max_{a_k \in \mathcal{A}} p(a_k|v, q) \tag{1}$$

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 [3] is to train a MLP with a cross-entropy loss with each of the possible answers as a class.

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. 1, we show that if an out-of-context answer such as "unicorn" for the surfing image in Fig. 1 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

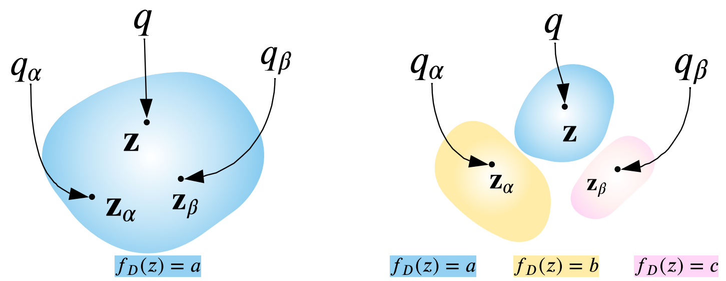


Figure 2. See Suppl. D for an explanation of the figure.

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 **??**, 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 [2]. However, the pseudo-rephrasings seem to be *good enough* that inconsistency over the pseudo-rephrasings indicates potentially unreliable predictions from f_{BB}.

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. 2.

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. 2). 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 **????** 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 [4] 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* [5] 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. 5 to 7. 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 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. 3.

	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 5. 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

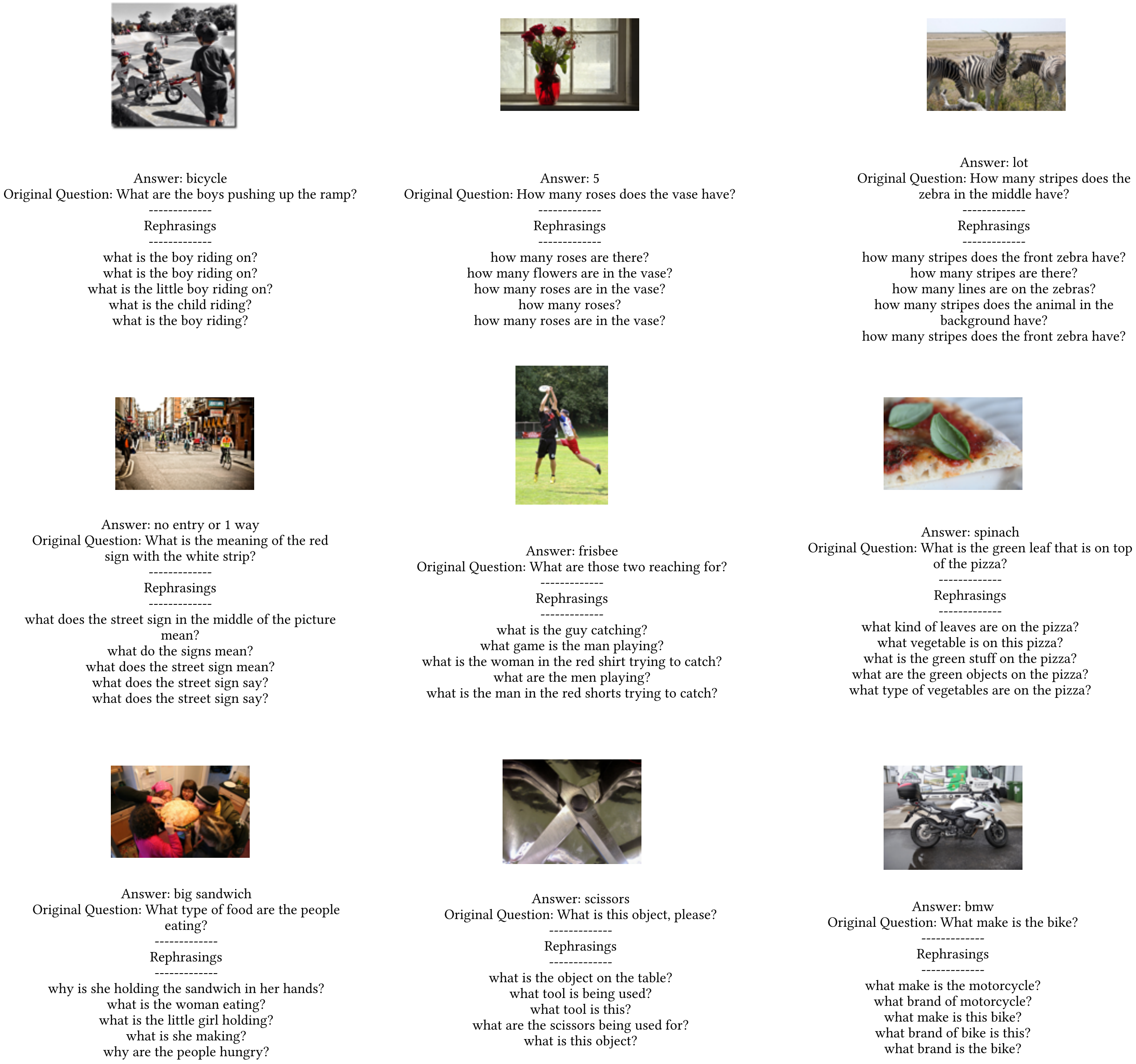
Table 6. Calibration of BLIP on VQAv2. For scaling, a temperature of 19.3 is used.

	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 7. Calibration of BLIP on AdVQA. For scaling, a temperature of 12.5 is used.

References

[1] 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



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 3. More examples of generated rephrasings.

learners. *ArXiv*, abs/2005.14165, 2020. 3

[2] Hong Liu, Jianmin Wang, and Mingsheng Long. Cycle self-training for domain adaptation. In *Neural Information Processing Systems*, 2021. 4

[3] 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. 3

[4] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, page 625–632, New York, NY, USA, 2005. Association for Computing Machinery. 4

[5] Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. *ArXiv*, abs/1904.01685, 2019. 4