Know What You do Not Know: Verbalized Uncertainty Estimation Robustness on Corrupted Images in Vision-Language Models

Mirko Borszukovszki and **Ivo Pascal de Jong** and **Matias Valdenegro-Toro**
Department of Artificial Intelligence, Bernoulli Institute
University of Groningen, The Netherlands
Email: ivo.de.jong@rug.nl, m.a.valdenegro.toro@rug.nl

Abstract

To leverage the full potential of Large Language Models (LLMs) it is crucial to have some information on their answers' uncertainty. This means that the model has to be able to quantify how certain it is in the correctness of a given response. Bad uncertainty estimates can lead to overconfident wrong answers undermining trust in these models. Quite a lot of research has been done on language models that work with text inputs and provide text outputs. Still, since the visual capabilities have been added to these models recently, there has not been much progress on the uncertainty of Visual Language Models (VLMs). We tested three state-of-the-art VLMs on corrupted image data. We found that the severity of the corruption negatively impacted the models' ability to estimate their uncertainty and the models also showed overconfidence in most of the experiments.

(a) Noise severity 0; ✓ GPT-4V: "Tag on ear (95% confidence)"

(b) Noise severity 2; ✗ GPT-4V: "There is nothing unusual (95% confidence)"

Figure 1: Question: What is on the sheep? With small noise, GPT-4V is confidently incorrect.

1 Introduction

LLM-based AI assistants can help us with a wide variety of tasks. The responses generated by these models sound convincing and correct most of the time but it has been shown that they can confidently generate incorrect or even nonsensical answers. In the field of LLMs, this is known as hallucinations (Ji et al., 2023). Currently, the biggest problem with ChatGPT-like AI assistants is that they will generate real and hallucinated answers with the same degree of confidence, as seen in Figure 1. As there have already been examples of algorithmic biases with serious consequences in real-world applications of machine learning models (Angwin et al., 2016), with the rapid evolution of LLMs, it is likely that they will have increasingly more responsibilities in practical applications. There are multiple risks involved with deploying these models in high-stakes decisions in the real world (Weidinger et al., 2021; Echterhoff et al., 2024). We have to ensure that these models are well-calibrated, meaning that the model's confidence in a response accurately predicts the likelihood of the answer being correct.

Verbalized Uncertainty. LLMs generate text token by token, from a predefined vocabulary. At each step, the model generates a probability distribution over its vocabulary based on the input and the previously generated tokens and selects the next token from that probability distribution. In theory, the uncertainty of a given answer could be estimated by the combined probability of these tokens (Kuhn et al., 2023). Still, since these models are proprietary, we don't have access to these individual token probabilities so methods have been proposed to quantify the uncertainty of a response (Tian et al., 2023).

To estimate the model's uncertainty in a given answer, we could ask the model in our prompt to quantify it. This is known as verbalized uncertainty (Xiong et al., 2024). It has been shown that sometimes the models' verbalized confidence estimates are better calibrated than the conditional probabilities estimated via sampling (Tian et al., 2023).

Originally, these LLMs could only take in text input and produce text output. However, in the previous five years, multiple advancements were made to extend the capabilities of LLMs to the visual realm. These models can generate text answers from a prompt and an image, or even just

Original Image	Severity 1	Severity 2	Severity 3	Severity 4	Severity 5
Togetsukyo Bridge in Arashiyama.	Kyoto, Japan	Kyoto, Japan	Kyoto, Japan.	Lake Como in Italy.	Lake
Conf: 90%	**Conf:** 80%	**Conf:** 75%	**Conf:** 70%	**Conf:** 75%	**Conf:** 60%

Figure 2: Sample answer from Claude with Defocus Blur Corruption. **Question**: Where was this photo taken? **Correct Answer**: Japan, Kyoto, Arashiyama Area, the Bridge is named Togetsu-kyo Bridge (or Toei Bridge). It is clear how answers and confidence degrade with increasing corruption severity. Full answers in Table 5.

an image and are called Visual Language Models (VLMs). One of the first notable examples of these is ViLBERT (Lu et al., 2019) and two years later CLIP (Radford et al., 2021). For a more comprehensive overview of the evolution of VLMs, refer to Oza and Kambli (2024). Since then, some of the most widely used LLMs (ChatGPT, Gemini, Claude) have been upgraded with visual understanding. Since they were released in the last two years, there is still much to uncover in understanding their uncertainty.

Models and Corruptions. We tested three state-of-the-art VLMs on visual question-answering tasks where the images are corrupted with common corruptions taken from Michaelis et al. (2019). It is important to test if a model dealing with image data is robust to these corruptions, as they might not be present in the training set but are likely encountered in a practical application. A demonstration of these corruptions is shown in Figure 3.

There are five severity levels for each corruption each one adding more distortion to the image. This paper aims to answer the research question: *How does the severity level of the corruption impact the model's calibration, accuracy and confidence?* Ideally, as the corruptions become more and more severe and the model starts making mistakes, the confidence should go down along with the accuracy. However, there is evidence that LLMs exhibit overconfidence in their answers (Xiong et al., 2024; Groot and Valdenegro-Toro, 2024), suggesting that increasing severity will increase miscalibration in the models and that the decrease in accuracy will not be accompanied by lower confidence scores.

The three VLMs tested were: GPT-4 Vision (Achiam et al., 2023), Gemini Pro Vision (Team et al., 2023), and Claude 3 Opus (Anthropic, 2024). We tested all of them on the same image visual question answering tasks where the corruption levels progressively increased. The models were prompted to incorporate their level of uncertainty in their responses or express their answer as a 95% confidence interval. Figure 2 shows an example answer across corruption severities, with degrading answer quality and decreasing confidence, but still being overconfident.

2 Related Work

In this paper we estimate the model uncertainty by prompting. There is no consensus on the best method to elicit reliable confidence scores from LLMs. This is a problem as different methods yield different confidence scores so it is hard to compare the calibration of different models. Tian et al. (2023) examined various methods to extract confidence scores from the examined models and found that for models trained with Reinforcement Learning with Human Feedback (RLHF) (Ouyang et al., 2022), the verbalized confidence is better calibrated than other methods that for instance, estimate internal token probabilities by sampling. This finding makes verbalized uncertainty a viable option to estimate uncertainty in VLMs.

Even though estimating the model's confidence by prompting has some drawbacks, verbalized uncertainty is getting more attention and has also been examined by Xiong et al. (2024). Their work builds on Tian et al. (2023) as they investigate different prompting methods like chain-of-thought reasoning or top-k. Different prompting strategies yielded similar results: LLMs exhibit overconfidence and the majority of the models' confidence scores fall within the 80-100 range. This paper strengthens their findings and tests their "vanilla" prompting strategy on increasingly corrupted images.

Since uncertainty estimation is not often incorpo-

(a) Original image (b) JPEG compression (c) Defocus blur (d) Gaussian noise

Figure 3: Demonstration of the used corruptions on severity 5. **Question:** What kind of food is showcased in this photo? **Answer:** Japanese food. Also acceptable is that it is a food model, called Shokuhin Sampuru in Japanese.

(a) **Q:** What type of place is this? **A:** Savannah. (b) **Q:** What is in this photo? **A:** Tokyo Tower. (c) **Q:** How many birds are shown in this photo? **A:** 250-280.

Figure 4: Samples from the three tasks. (a) represents the "easy" task, (b) the "hard" task, (c) the "counting" task.

rated in computer vision applications (Valdenegro-Toro, 2021), there has not been much research published on the topic. The only paper that examined uncertainty estimation in VLMs is Groot and Valdenegro-Toro (2024). They also used verbalized confidence estimation on visual question-answering tasks and found that the models were poorly calibrated, showing severe overconfidence. We build on their research by introducing increasingly corrupted images in the dataset.

Most research (Ovadia et al., 2019; Hendrycks and Dietterich, 2019; Kadavath et al., 2022) has been focused on models applied in classification problems or when it comes to question answering, multiple choice or true/false questions. The main issue with this is that their methods for eliciting confidence scores are not applicable to state-of-the-art VLMs. While users would like to enjoy the benefits of well-calibrated models, they should not have to deal with the inner workings of the system and instead receive well-calibrated confidence scores in a verbalized form. We tested the models on more complex, open-ended questions which mimics the usage of these models in the real world. We combined the ideas from Hendrycks and Dietterich (2019) to test the models on increasingly corrupted images and Groot and Valdenegro-Toro (2024) to extend the research into VLMs where internal token probabilities are not available. With this paper, we aim to bridge the gap between uncertainty quantification on standard neural networks and VLMs. This is important due to the rapid advancement of VLMs, and the lack of research on their uncertainty calibration.

3 Methods

We tested the VLMs on three different datasets using three different corruptions. The specific details of the experiments, datasets, the used corruption techniques and the evaluation procedure are explained below.

3.1 Datasets and data

The three mentioned models were tested in three experiments:

1. ***Easy*** **visual question answering** evaluated on the popular visual question answering dataset (Antol et al., 2015; Goyal et al., 2017). From the testing part of this dataset, 36 randomly sampled images and the corresponding questions were selected. This dataset includes easier questions about images. Without any corruption added to the images, the models should be able to answer most of them.

2. ***Hard*** **visual question answering** evaluated on the Japanese Uncertain Scenes (JUS) dataset proposed by Groot and Valdenegro-Toro (2024). This dataset can be downloaded from a public GitHub repository[1]. This repository contains 29 "tricky" questions specifically designed to evaluate the model's ability to estimate their uncertainty.

3. The **Counting task** was also evaluated on the JUS dataset as it contains 13 challenging counting exercises. This is also not designed to evaluate the model's accuracy but rather to check its uncertainty estimates as most of them are nearly impossible to count precisely.

Figure 4 provides example images, questions and answers. For the selected images and the prompts taken from Groot and Valdenegro-Toro (2024), refer to Appendix B.

[1]https://github.com/ML-RUG/jus-dataset

3.2 Experiments

There were three types of corruption tested and five severity levels for each. Each model was tested on the original dataset and fifteen "corrupted" datasets for each task (3 corruptions, 5 severity levels). Since the models did not always adhere to the requested answer format and there could be multiple equally correct ways to answer an open question, all of the answers had to be manually checked which is the main reason for the low number of images in a particular dataset. Still, this project contains the results of more than 3700 answers across all models and corruptions, counterbalancing the low number of images in a single dataset.

Prompting the models with an image, the question plus the prompt from Appendix B to elicit verbalized confidence was automated using Python scripts and the APIs provided by OpenAI (GPT-4V), Google (Gemini Pro Vision) and Anthropic (Claude 3 Opus).

For each question, we recorded the confidence score from the model's answer. We also recorded if the answer was correct. Especially at higher severities, there were cases where the image was so distorted that the model refused to respond. For a well-calibrated model, this is a desired behaviour. Because of that, we cannot record that answer as incorrect, but we cannot mark it as correct either as the model did not answer the question. Since in the experiment, we need to measure the models' accuracy, we can only calculate it where each response is marked either correct/incorrect, so in these cases, the answer was not marked as either and no confidence score was recorded.

When the models provided an answer, it was always recorded and used for the analysis, and no data point had to be removed throughout the experiment. However, there were eight cases in the "easy" and "hard" visual question-answering experiments together where Gemini refused to respond due to the image being in conflict with its safety settings. Since there were no explicit images in any of the three datasets, this was most likely due to the model confusing a highly distorted image with explicit content. This confusion was only produced by Gemini and happened only with a small fraction of the tested images.

3.3 Image Corruptions

Michaelis et al. (2019) defines 15 types of corruption. They created multiple types of noise and blurring effects and other corruptions mimicking real-life distortions like fog, frost on the lens or snow. These were designed to benchmark neural networks' robustness to corrupted images.

From the 15 corruptions, we investigate three: Gaussian noise, defocus blur and JPEG compression. Different noise-based corruptions have very similar effects so we selected one of them. Gaussian noise or electronic noise is caused by high temperatures or poor lighting conditions (Boyat and Joshi). Since digital cameras are prone to this type of corruption, the robustness of VLMs against it needs to be tested. Blurring effects like zoom blur or motion blur were discarded as they may change the meaning of picture. Defocus blur does not introduce such ambiguities. Moreover, the most realistic corruption type was chosen from the 15 available ones: JPEG compression. We can safely assume that a VLM encounters images that are distorted due to the lossy nature of the JPEG compression algorithm as these types of digital images are very common.

For a demonstration of the different levels of the three tested corruptions, refer to Appendix C.

3.4 Evaluation Metrics

Apart from the accuracy and confidence scores, we measured the Expected Calibration Error (ECE) (Guo et al., 2017). The formula for calculating the ECE is:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|. \quad (1)$$

Where M is the number of bins, $|B_m|$ is the number of samples in the m-th bin, n is the total number of samples, $\text{acc}(B_m)$ is the accuracy of the m-th bin, and $\text{conf}(B_m)$ is the average confidence of the m-th bin. This takes the weighted average of the absolute difference between the accuracy and the average confidence of the bins.

This metric quantifies how much one can "trust" the model's confidence scores. The score can be in the range [0, 100] with the ideal ECE of a model being 0, which means that the confidence score accurately predicts the likelihood that the answer is correct.

4 Results

Here, we report the results of the three experiments. For visual question answering, we were mainly interested in how the ECE is affected by the increased

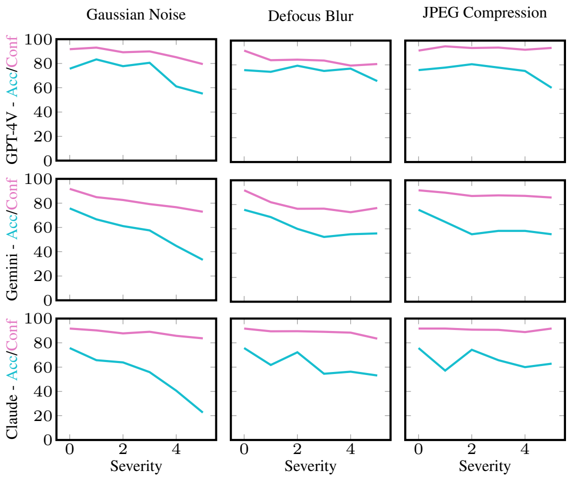


Figure 5: Accuracy and confidence plots for the three examined models and the three corruptions in the *easy* visual question answering experiment.

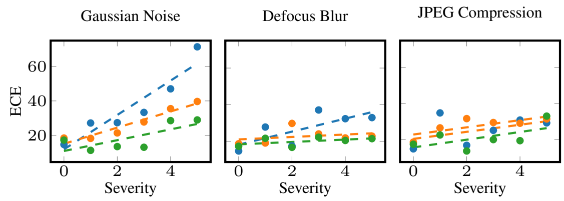


Figure 6: ECE vs severity level on different corruption types for Claude, Gemini, and GPT-4V in the *easy* visual question answering experiment.

	Claude	Gemini	GPT-4V
Gaussian Noise	**0.88**	**0.93**	0.53
Defocus Blur	0.54	0.11	0.28
JPEG Compression	0.21	0.58	0.36

Table 1: R^2 values of ECE vs. severity trends in Fig. 6.

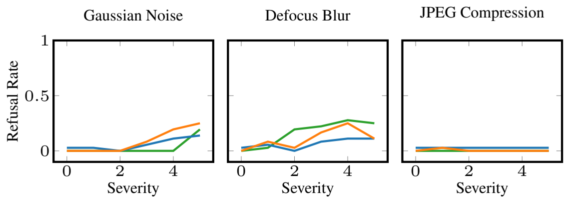


Figure 7: Refusal rates for Claude, Gemini, and GPT-V4 across different severity levels and corruptions in the *easy* visual question answering experiment.

corruption severity. Since the results seemed to increase linearly, we attempted to fit linear regression lines to the data points and calculated the coefficient of determination R^2 value to test the explanatory power of the linear models. We were also interested in whether there is any connection between a model's refusal rates and their performance. Especially at higher severity levels, refusing to answer can improve the model's performance. Refusing to answer is not as useful as a correct answer, but it is better than an incorrect guess. In the counting experiment, we only examined the change in accuracy as the corruption severity increased.

4.1 *Easy* Visual Question Answering

In this task, the models achieved fairly high accuracy scores on the dataset without any corruption. As the severity of the corruption increased, the models' accuracy started to degrade slightly, but the confidence remained fairly stable. The detailed results are illustrated in Figure 5.

We can see that for all models in all corruptions, the average confidence score was higher than the accuracy throughout all severity levels. This means that all models are overconfident. It can also be seen that the gap between the two lines widens as the severity increases. This is not apparent in all plots but is quite visible in the Gaussian noise column. To measure if this gap is actually increasing, we can calculate the ECE scores for each combination of model and corruption type in each severity level and see if it increases as we increase the severity. These ECE scores are visualized in Figure 6

The Expected Calibration Error increases as we increase the severity for all models and corruptions. However, this effect is very small for defocus blur and JPEG compression. The worst performance is shown by Claude 3 Opus with the Gaussian noise corruption. GPT-4V, on the other hand, outperforms the other two models in all three corruptions, achieving the lowest ECE scores across different severity levels.

Table 1 summarizes the R^2 values for the linear regression lines presented in Figure 6. The only high values are for Claude and Gemini for Gaussian noise corruption. From Figure 6 we can see that these are the two steepest lines in the plot, meaning that increasing severity had the most effect on the model's ECE in these two cases. The high R^2 values indicate that the increased severity explains a lot of variance in the ECE. The calibration errors are further investigated with calibration plots in Appendix A. These show that GPT-4V outperformed the other two models in all types of corruptions.

At higher severity levels, the models sometimes refused to answer and express their confidence score. The refusal rates are summarized in Figure 7.

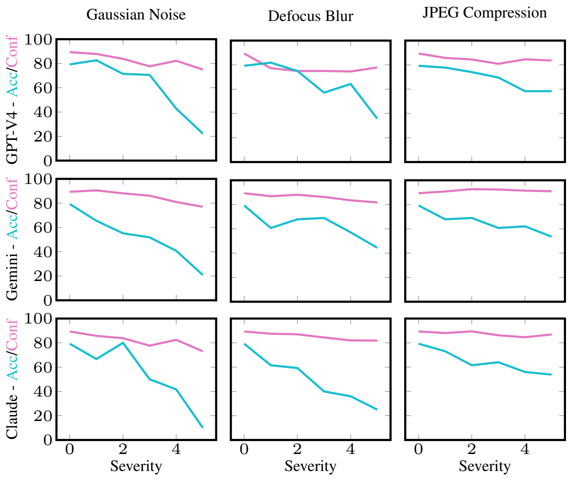


Figure 8: Accuracy and confidence plots for the three examined models and the three corruptions for the *hard* visual question answering experiment.

We can look at the relationship between this plot and Figures 6 and 12. All models show similar refusal rates for Gaussian noise but achieve different results in the ECE and the calibration plot. For defocus blur, GPT-4V's refusal rates are much higher than the other two models' and it outperforms them both in ECE scores and the calibration plot. The models showed very low refusal rates and their ECE performance is similar for JPEG compression but there are still small differences between the models in the calibration plot as GPT-4V still shows the best performance. Overall, we have some evidence suggesting that refusing to answer a question can prevent a model from making a wrong prediction, thus improving its accuracy and ECE score, but we see that this is not true for all corruption types.

4.2 *Hard* Visual Question Answering

To answer the research question, we need the models to make more and more mistakes as the severity increases to see if the confidence estimates also decrease. The models achieved fairly high accuracy scores for the easy dataset even at higher severity levels. The JUS dataset (Groot and Valdenegro-Toro, 2024) makes it possible to test the limits of these models by asking them nearly impossible questions. It should be emphasized that we are not interested in the actual accuracy of the models but rather in their calibration. Figure 8 summarizes the accuracy and confidence scores in this task.

Compared to the easy task in Figure 5, we see lower accuracy scores, but more importantly, it is more visible that the gap between accuracy and

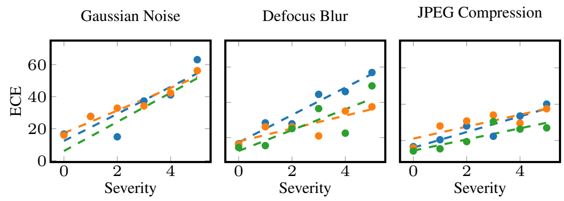


Figure 9: ECE scores for Claude, Gemini, and GPT-V4 for each corruption type for different severity levels in the *hard* visual question answering experiment.

	Claude	Gemini	GPT-V4
Gaussian Noise	0.77	0.94	0.87
Defocus Blur	0.95	0.70	0.68
JPEG Compression	0.84	0.71	0.73

Table 2: R^2 values of ECE vs. severity trends in Fig. 9.

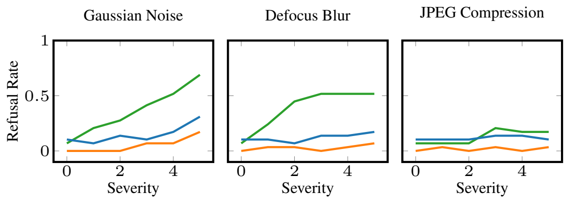


Figure 10: Refusal rates for Claude, Gemini, and GPT-V4 across different severity levels and corruptions in the *hard* visual question answering experiment.

confidence widens as we increase severity. This effect is apparent when we look at the ECE scores in Figure 9.

The most visible difference between Figures 6 and 9 is for JPEG compression and defocus blur corruptions. The models become more miscalibrated at higher severity levels. There is not as much difference between the models for Gaussian noise as in the easy dataset. We can also see that GPT-4V still achieves the lowest ECE scores, but the models show a more similar behaviour.

Table 2 shows the R^2 values for the linear regression lines in Figure 9. All of the values are around or above 0.7 indicating that the regression model explains the data well. This strengthens the visual intuition that we got from comparing Figures 5 and 8. We have stronger evidence that the models become more and more miscalibrated as we increase the severity of the corruption. The calibration plots shown in Appendix A show that GPT-4V again performed much better than the other two models.

As with the easy dataset, we can also examine the refusal rates in Figure 10. For Gemini and Claude, they are around the same as in the easier

	Claude	Gemini	GPT-4V
Gaussian Noise	0.13	0.61	0.22
Defocus Blur	0.35	0.47	0.13
JPEG Compression	0.31	0.02	0.31

Table 3: R^2 values of accuracy vs. severity trends in Fig. 11.

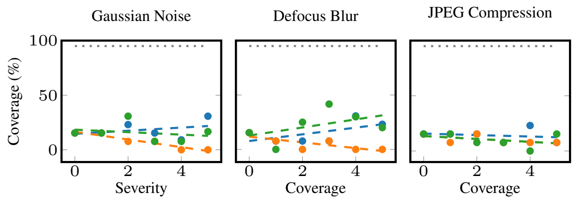


Figure 11: Coverage (confidence interval accuracy) scores for the counting experiment for Claude, Gemini, and GPT-V4. The dotted line at the top represents the 95% accuracy which would be expected for a perfectly calibrated model.

dataset but for GPT-4V, they are much higher. GPT-4V performed best on the hard dataset both in terms of ECE and the calibration plots, so we see that refusing to answer a question instead of making a wrong guess can improve a model's calibration.

4.3 Counting Task

The JUS dataset contains hard counting tasks that were evaluated using a different prompt described in Appendix B. In this task, the model was asked to output a 95% confidence interval. The answer was recorded as correct if that interval contained the actual prediction. There was one picture illustrated in Appendix C where there was no correct answer. It is impossible to count the bamboo trees without seeing their trunks as many of the visible branches could belong to the same bamboo tree. The results of this experiment are shown in Figure 11.

For a perfectly calibrated model, we would expect that a 95% confidence interval is correct 95% of the time. We can see that the models perform below 25% accuracy most of the time. There is not as much consistency in the linear regression lines as in the previous two tasks, most likely due to the models' poor performance on all severity levels and the high variance from the low number of test images.

We can look at Table 3 containing the R^2 values for the lines but we get much lower values than in the previous two tasks. The models are unable to answer the questions even on the original dataset, so increasing the severity of the corruption does not have an effect.

It is important to note that the models seldom refused to provide a response during this task. Out of the 208 times each model was queried (13 questions, 3 corruptions, 5 severity levels plus the original dataset), GPT-4V refused to answer 9 times, Gemini 2 times, and Claude 0 times. The low accuracies show that the models responded even when the 95% confidence interval was purely guessed.

Interestingly, the models often had an exact guess that was reasonably close to the right answer, but their confidence interval was so small that it almost never contained the true value. This shows the models' good visual capabilities even on hard images, but also signals their bad calibration as they were not capable of formulating an accurate 95% confidence interval based on a close estimate.

5 Discussion

Overall we found that increased corruption severity had a negative impact on the three examined models' accuracy and calibration. When the corruption level gradually became higher and the models' accuracy started to decrease, it was not accompanied by decreasing confidence scores. We also found that models that refuse to answer at a higher rate can achieve better accuracy and ECE scores. Our other main finding is that models are generally overconfident in their responses and output high confidence scores in most of their responses. This overconfidence was present in all three experiments but it was the most severe in the counting problems.

Interpretation of Results. Our findings about overconfident models and their high confidence scores are in line with Groot and Valdenegro-Toro (2024) and Xiong et al. (2024). Both of them found that the majority of the confidence scores of LLMs and VLMs fall within the [80, 100] range. While Xiong et al. (2024) looked at the performance of LLMs in different reasoning tasks (commonsense, arithmetic, symbolic), Groot and Valdenegro-Toro (2024) also examined the performance of VLMs in visual question answering tasks. Our results show that this characteristic of VLMs persists when they are tested on corrupted images. Higher corruption levels worsening calibration was also found by Hendrycks and Dietterich (2019) who tested different neural network architectures designed for image classification. We show that their findings can be extended to the realm of VLMs.

As mentioned before, one possible explanation for this overconfidence is the RLHF fine-tuning of these models. It rewards answers that sound more confident so the model learns to express its responses using confident language which influences the confidence scores of verbalized uncertainty.

Examining the the number of cases where the models refused to answer, we found that higher refusal rates can help the model's calibration as it is nearly impossible to give a correct answer to some highly distorted images. Especially with Gaussian noise, there were times when the models were fooled by the noise and output completely unrelated answers to the images. This happened less with JPEG compression as it was a less severe corruption than the other two. In case of defocus blur, the models were more likely to recognise the heavy blurring effect on the image and refuse to respond to the question. Images corrupted with Gaussian noise were less likely to be recognised as corrupted and in some cases they were even confused with pointillistic paintings which is a painting technique from the late 19th century using small colourful dots that form an image when viewed from a distance. For some examples of model responses, we refer to Appendix F.

Future Research. Apart from increasing the number of images in the dataset, there are other things that could be explored in the topic of uncertainty estimation in VLMs. Different prompting strategies, such as chain-of-thought reasoning or top-k explored by Xiong et al. (2024) could yield different results. These can be altered so the models are more restricted in their answers making automated data gathering easier. Apart from verbalized uncertainty, there exist other, sampling-based techniques for uncertainty estimation (Tian et al., 2023) that could be applied to VLMs.

The overconfidence of RLHF-based LLMs is present in multiple studies (Groot and Valdenegro-Toro, 2024; Xiong et al., 2024) but it would be interesting to explore if this overconfidence in VLMs could be treated with temperature scaling in the same way as in Kadavath et al. (2022). The APIs provided for the three investigated VLMs offer the ability to manipulate the model's temperature.

Michaelis et al. (2019) defines 15 corruption types, but we only tested three. Studying the effect of the others could reveal more differences between the models and their robustness to different corruptions.

6 Conclusions and Future Work

The key conclusions that we obtained from this work are the following:

- VLMs are overconfident. They often express their confidence in the range of [80, 100] even when this is not reflected in their accuracy.

- Increased corruption severity increases the ECE. When the models started making mistakes due to the increasingly corrupted images, their confidence did not decrease at the same pace which caused the ECE to go up. This is the main finding of the paper, as it answers our research question.

- There are differences in the calibration of state-of-the-art VLMs and the models are more robust to some corruptions than others. GPT-4V outperformed the other two models in the visual question-answering experiments, and JPEG compression was better handled by all of the models than Gaussian noise and defocus blur.

- Higher refusal rates can improve calibration. We see that when the model recognises that we are asking an impossible question and refuses to answer, it prevents itself from providing hallucinated answers and improves its calibration. GPT-4V also performed better in this regard than the other two models.

- VLMs were especially miscalibrated when they were asked to express their answer in a 95% confidence interval. Their accuracy in the counting experiment did not even come close to 95%, even when their initial guess for the exact number of objects was quite close to the answer.

From these results, we can see that there are many things that can be improved when it comes to the calibration of VLMs. In the current state of things, users are often presented with confident wrong answers which undermine the trust in these models. This paper contributes to the research in uncertainty estimation of VLMs and points out the shortcomings of these models with respect to their calibration. Better-calibrated models would be beneficial to millions of users as these models are already widely used by the general public.

7 Limitations

As the models output high confidence scores, the lower confidence bins were underrepresented in the calibration plots. This issue could be solved with more images in the datasets, but there were some limitations on the number of images that the models could be tested on. For the easy visual question answering experiment, there were tens of thousands more images available from the dataset by Antol et al. (2015) and Goyal et al. (2017). However, all of the images were used from the JUS dataset, which put a limit on the number of images in the hard VQA experiment and the counting experiment.

Since a correct answer to a question could be phrased in multiple ways, the answers had to be manually checked, which made the data-gathering process time-consuming. The used APIs also had a limit on the number of requests per minute, which prevented large-scale testing.

One way to automate the check for the correctness of the answers could be to use an LLM to check the semantic equivalence of the correct reference answer and the response provided by the model. This was not a suitable approach in this paper, as these methods still have limited correlation (Spearman rank correlation $\rho < 0.7$) with manual assessment (Mañas et al., 2024). Another way could be to use better prompts that restrict the model to one or two-word answers that are easier to check automatically but that would put a limit on the complexity of the tested questions.

Lastly, we found that some models are more likely to refuse to answer for images with severe corruptions. These samples can therefore not be included in the answer, which can decrease error rate and decrease the recorded miscalibration. This may be considered acceptable, as not giving an answer may be preferable over a random guess, but it is not as good as giving a correct answer. There are methods that attempt to minimize the refusal rate and get more correct answers (Srinivasan et al., 2024), but those were not considered in this study.

8 Ethical Considerations

In this paper we show that VLMs verbalized uncertainty is prone to severe degradation under input image corruption, which adds to other common problems with VLMs like incorrect predictions and hallucinations, this raises ethical concerns on their use, as these models are effectively not able to identify when they do not know or cannot answer a prompt.

Users of these models are recommended to always double check with a human any kind of output that is given by VLM, as they are not trustworthy, and when used for critical applications, humans can be hurt.

Our work shows that GPT4V is able to detect Gaussian noise corruptions, which indicates that its developers might have included this in its training set, but not other kinds of corruptions, showcasing the limitations of VLMs as only capabilities considered during training are available during inference/deployment. More research is needed to further detect other kinds of input image corruptions.

References

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

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. *And it's biased against blacks. ProPublica*, 23:77–91.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

AK Boyat and BK Joshi. A review paper: Noise models in digital image processing. arxiv 2015. *arXiv preprint arXiv:1505.03489*.

Jessica Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in high-stakes decision-making with llms. *arXiv preprint arXiv:2403.00811*.

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

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

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

Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. 2019. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32.

Jay Oza and Gitesh Kambli. 2024. Pixels to phrases: Evolution of vision language models. *Authorea Preprints*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Raghavi Chandu. 2024. Selective" selective prediction": Reducing unnecessary abstention in vision-language reasoning. *arXiv preprint arXiv:2402.15610*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Matias Valdenegro-Toro. 2021. I find your lack of uncertainty in computer vision disturbing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1263–1272.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

A Calibration Plots

Since one metric like the ECE can hide the nuances in the model's behaviour, we can make a calibration plot for each corruption. We calculate the model's average confidence in a confidence bin and plot its accuracy along the y-axis.

Figure 12 illustrates the calibration of the three models in each corruption type for the "easy" visual question answering. Since the calibration plots for one specific severity level and corruption type are too noisy due to the low number of data points and the uneven distribution of the confidence scores, we plotted the calibration using all severity levels of a given corruption. Even with this adjustment, most of the bins contained one or two data points so we decided to use four equal bins covering the [0, 100] interval.

GPT-4V outperformed the other two models in all types of corruption. It is especially close to the dashed line indicating perfect calibration, in the defocus blur corruption. In the confidence bins where there were enough data points, indicated by the small error bars, the models show overconfidence as those points lie below the dashed line.

There are points below 50% confidence where the error bar is very large or zero. These points occur because the models tend to output high confidence scores so lower bins have few or no data points. If there are only one or two correct answers in a confidence bin and no other data points, then the accuracy for that bin will be 100% and the standard error will be undefined or zero since it is calculated as the standard deviation divided by the square root of the number of samples.

Figure 13 illustrates the calibration for the "hard" visual question answering. There we can again see that GPT-4V performed much better than the other two models. It should be noted that the problem of confidence bins with low or zero number of data points is still present, so the same bin size had to be used.

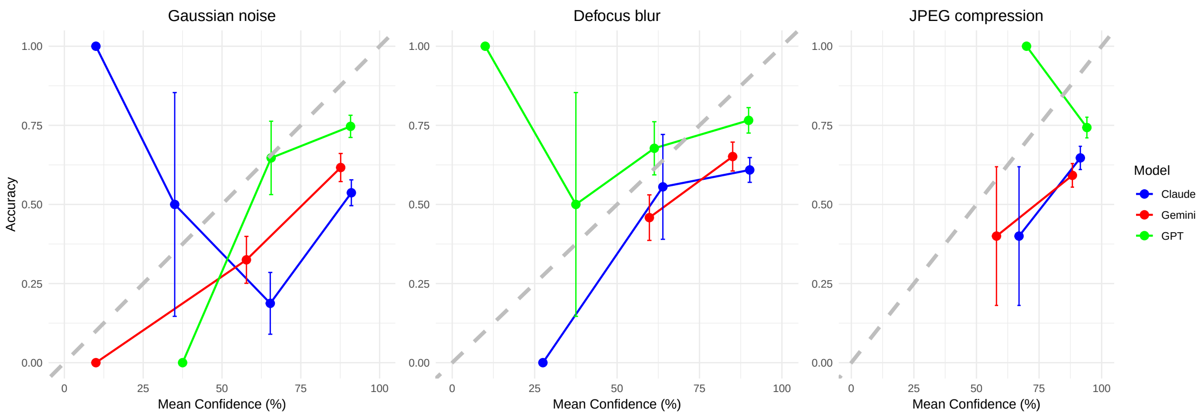


Figure 12: Calibration plots for the three examined corruption types and three models in the easy visual question answering experiment. The error bars represent the standard error and the dashed line indicates perfect calibration.

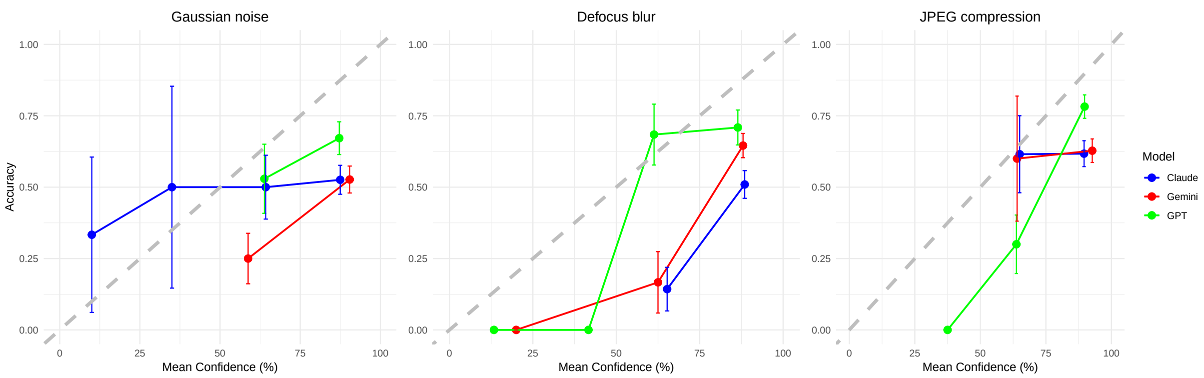


Figure 13: Calibration plots for the three examined corruption types and three models in the hard visual question answering experiment with the error bars calculated using the standard error.

B Image data for the easy VQA and model questions

To reduce the size of the Easy VQA dataset to be able to manually assess the responses we randomly selected a subset of the images and their questions. The selected images can be found in Table 4.

To elicit verbalized confidence we expanded the questions used. We appended either:

'*Moreover, please rate your confidence in your answer between 0 and 100%. The answer should be in the format: "Answer (confidence%)".*' or

'*Provide your actual prediction. Moreover, please express your estimate as a 95% confidence interval. This means you should provide a range within which you are 95% confident the true value lies. Format your answer as: "[Lower Bound, Upper Bound]", where the lower bound is the start of the range and the upper bound is the end of the range. Ensure that this interval reflects a 95% confidence level based on your estimation.*', depending on whether the question was a counting task.

C Demonstration of corruptions and an impossible question

Figure 14 illustrates the effect of the used corruptions at different severity levels. In Figure 15, we give an example of a question that even a human should not be able to answer. As mentioned in subsection 4.3, we expect the model not to answer. It should be noted that some of the images are so distorted at higher severity levels that they also become impossible to answer.

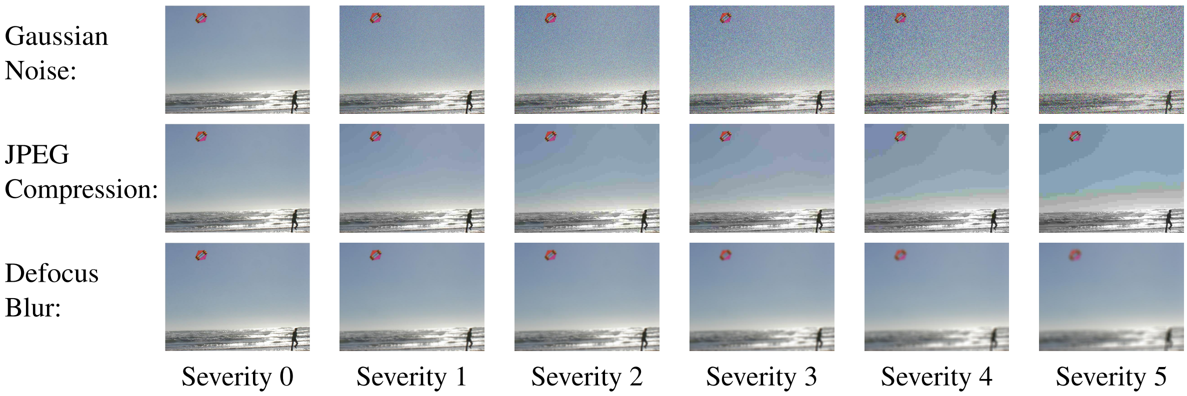


Figure 14: Demonstration of different severity levels for the three tested corruptions

Figure 15: Impossible question: How many bamboo trees are shown in this photo?

Table 4: Randomly selected images and their corresponding questions for "easy" VQA.

Image ID	Question
COCO_test2015_000000341181	Is the kitchen well lit?
COCO_test2015_000000244073	What color is the plane?
COCO_test2015_000000415036	Is there a plug near the bed?
COCO_test2015_000000551714	Is the person wearing gloves?
COCO_test2015_000000084296	How many chairs are around the table?
COCO_test2015_000000512556	What is the bench made of?
COCO_test2015_000000358972	What type of boat is that?
COCO_test2015_000000154340	Are they preparing food?
COCO_test2015_000000473114	Is this computer equipment?
COCO_test2015_000000444844	Is this apartment completely empty?
COCO_test2015_000000459379	What is the boy doing?
COCO_test2015_000000003004	What is the animal doing?
COCO_test2015_000000515370	What shape is the kite?
COCO_test2015_000000066725	What appliance is pictured?
COCO_test2015_000000121284	What time is it?
COCO_test2015_000000420197	What is beside the dog?
COCO_test2015_000000471029	Is the man right-handed?
COCO_test2015_000000112870	What is on the shelf above the toilet?
COCO_test2015_000000343994	What is in the baby's mouth?
COCO_test2015_000000351008	Is the horse running down the street?
COCO_test2015_000000334624	Is the man listening to something on his smartphone?
COCO_test2015_000000177197	What color is the batter wearing?
COCO_test2015_000000407045	Is the bench brown?
COCO_test2015_000000110643	How many birds?
COCO_test2015_000000517475	Are they using foil paper?
COCO_test2015_000000268054	Is the oven on?
COCO_test2015_000000068573	Is this indoors?
COCO_test2015_000000262294	What type of sweater is the man wearing?
COCO_test2015_000000206488	Is there money on the table?
COCO_test2015_000000066282	What type of place is this?
COCO_test2015_000000166735	Are there people in the boat?
COCO_test2015_000000068702	Are there picture frames in this picture?
COCO_test2015_000000475609	Has the ball been thrown?
COCO_test2015_000000434294	Is the boy wearing shoes?
COCO_test2015_000000373079	Is the person wearing a wedding band?
COCO_test2015_000000022109	What is on the sheep?

D Frequency of correct and incorrect answers in each confidence bin

In Figures 16 and 17 'gn' stands for Gaussian noise, 'db' stands for defocus blur and 'jc' stands for JPEG compression. There are two main things that we can observe from the two figures which influence the calibration plots. The proportion of correct and incorrect answers does not reflect the confidence bin and most of the responses fall into higher confidence bins.

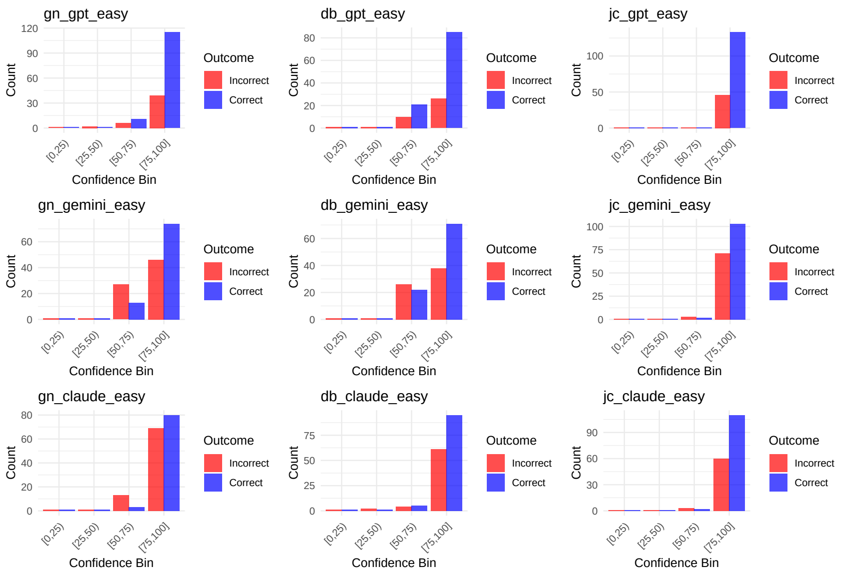


Figure 16: Histograms of confidence scores in the easy visual question answering experiment with the red and blue bars indicating the proportion of correct and incorrect responses

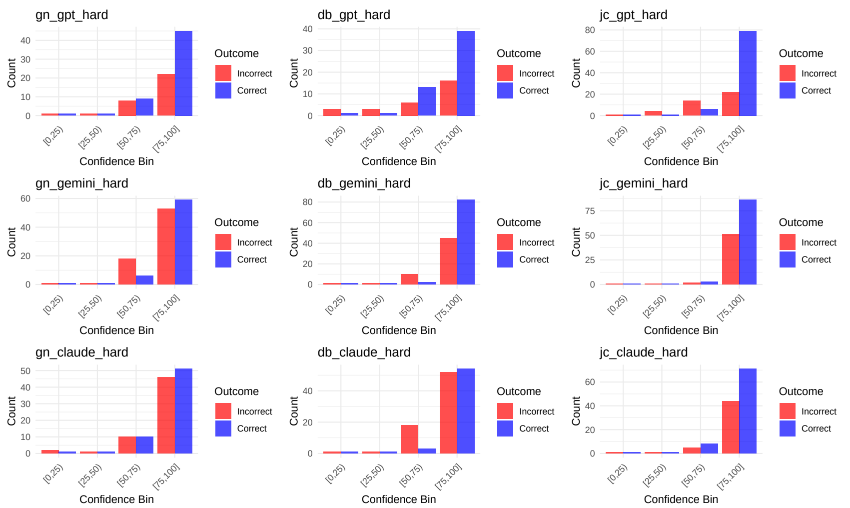


Figure 17: Histograms of confidence scores in the hard visual question answering experiment with the red and blue bars indicating the proportion of correct and incorrect responses

E Frequency of all answers in each confidence bin

In the 6 Figures below, we see the frequency of confidence scores for each model in the two VQA tasks. While models give more low-confidence answers in the hard VQA task, the overwhelming majority of answers fall into the higher confidence bins.

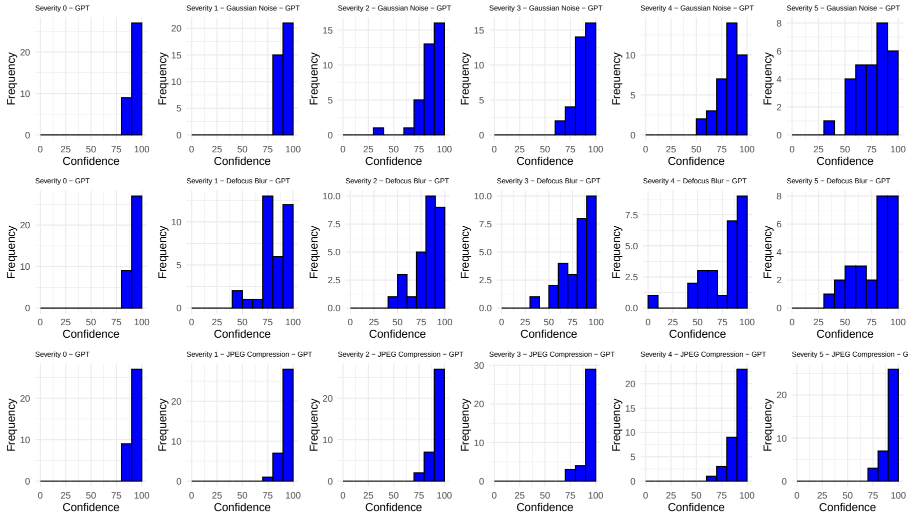


Figure 18: Confidence histograms GPT-4V easy VQA

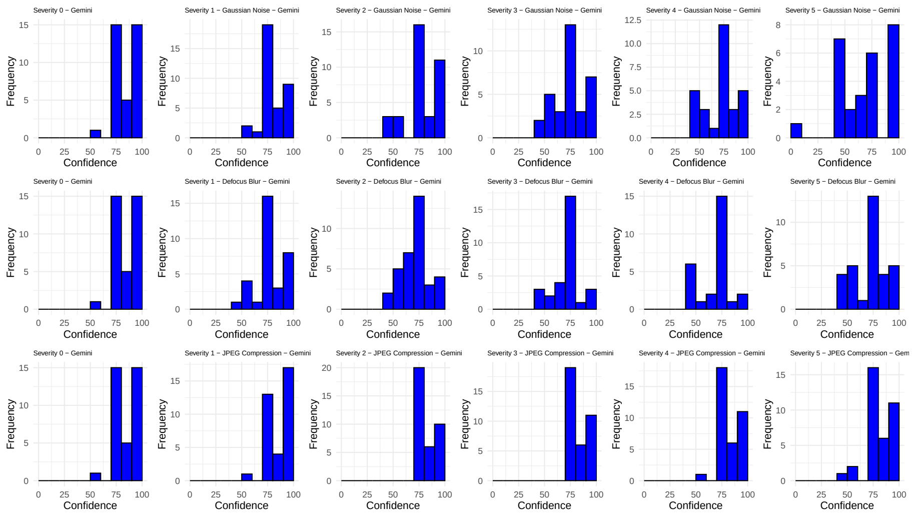


Figure 19: Confidence histograms Gemini easy VQA

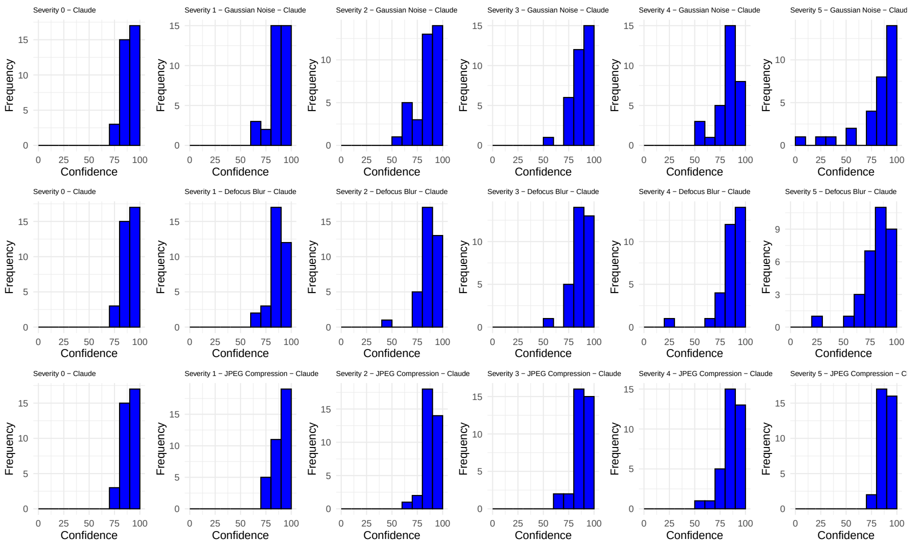


Figure 20: Confidence histograms Claude easy VQA

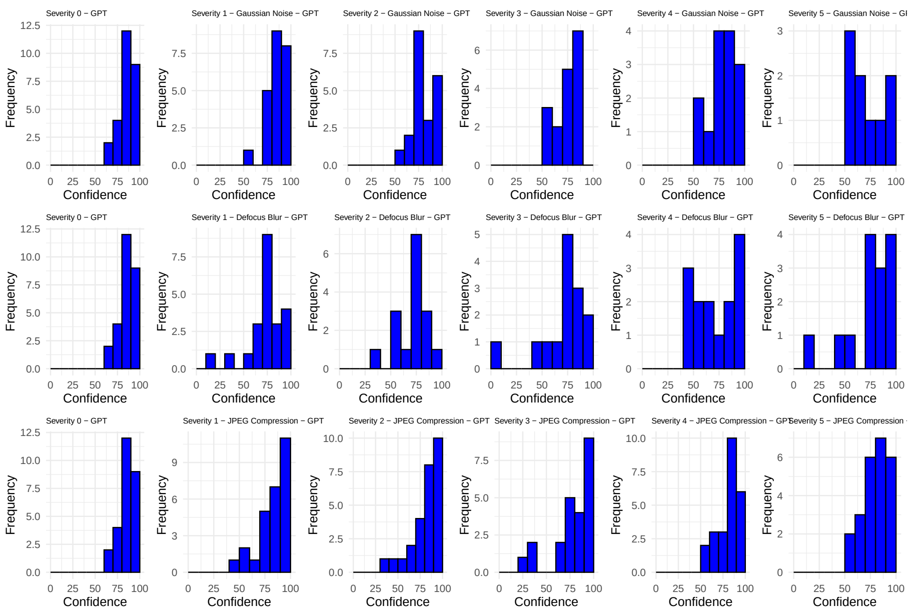


Figure 21: Confidence histograms GPT-4V hard VQA

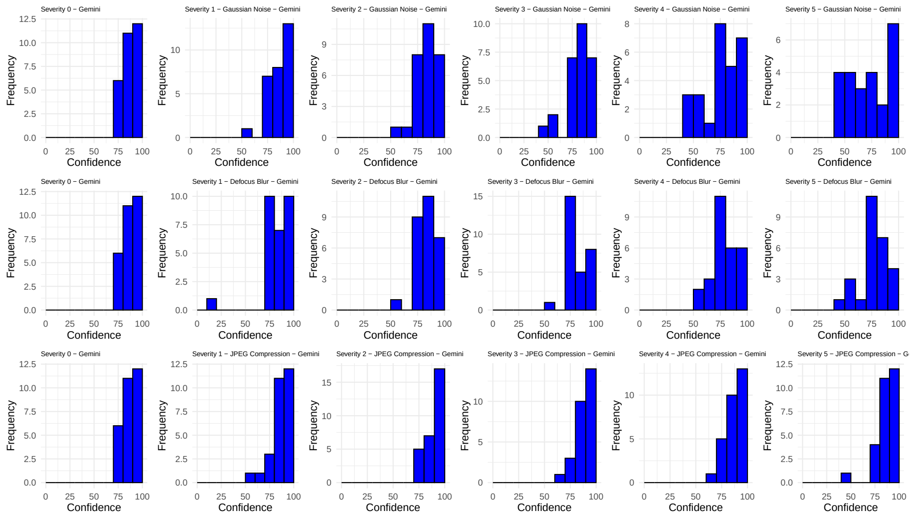


Figure 22: Confidence histograms Gemini hard VQA

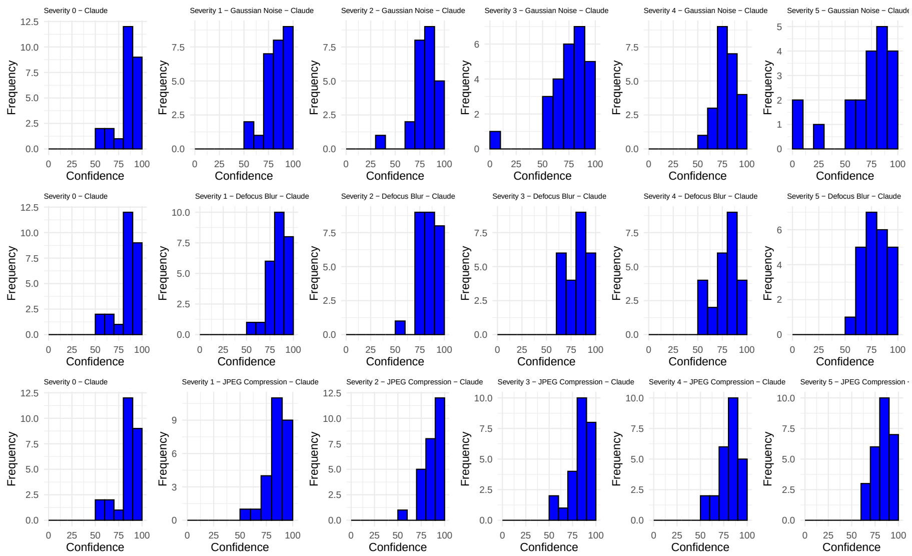


Figure 23: Confidence histograms Claude hard VQA

F Responses under noise

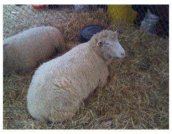
In Table 5 we show how Claude even identifies the exact bridge in Kyoto but as the severity of the blur increases, it gives incorrect or more and more general answers instead of refusing to answer.

Table 5: Claude answers with defocus blur
Question: Where was this photo taken?
Correct answer: Japan, Kyoto, Arashiyama Area, the Bridge is named Togetsu-kyo Bridge (or Toei Bridge)

Severity 0	Severity 1	Severity 2	Severity 3	Severity 4	Severity 5
					
Based on the wooden bridge structure spanning the water and the mountains in the background, this photo appears to have been taken in Kyoto, Japan at the famous Togetsukyo Bridge in Arashiyama (confidence 90%).	Kyoto, Japan (80%)	Kyoto, Japan (75%)	Based on the mountainous landscape and the traditional wooden houses along the riverbank, this photo appears to have been taken in Kyoto, Japan (70% confidence).	Based on the mountainous landscape and the body of water, which appears to be a lake, this photo was likely taken at Lake Como in Italy (75% confidence).	Lake (60%)

In Table 6 we show responses from GPT-4V. It first answers the question correctly and then as the severity increases, gets confused with the added noise and in the end it refuses to answer.

Table 6: GPT-4V answers with Gaussian noise. **Question**: What is on the sheep? **Correct answer**: An eartag

Severity 0	Severity 1	Severity 2	Severity 3	Severity 4	Severity 5
					
Tag on ear (95%)	There is an ear tag on the sheep (confidence 100%).	There is nothing unusual on the sheep; it appears to be just wool-covered, typical for a sheep. Answer (confidence 95%).	There is nothing unusual on the sheep; it looks like a typical sheep with its wool coat. Answer (100%).	The sheep is wearing a blue blanket or cover (90% confidence).	It is difficult to accurately discern specifics about the sheep due to the distorted and noisy nature of the image. The image quality suggests it may have been altered or degraded, mostly showing pixelated and multicolored noise patterns rather than clear objects or textures. Thus, I cannot definitively make out detailed attributes such as what might be on the sheep. (Confidence: 30%)