The Role of Model Confidence on Bias Effects in Measured Uncertainties

Anonymous ACL submission

Abstract

With the growing adoption of Large Language Models (LLMs) for open-ended tasks, accurately measuring epistemic uncertainty, a key indicator of a model's lack of knowledge or confidence, has become crucial for ensuring reliable outcomes. However, quantifying epistemic uncertainty in such tasks remains challenging due to the presence of aleatoric uncertainty, which arises from inherent randomness among multiple valid answers. Building on previous work showing that LLMs are more likely to copy information from input when model confidence is low, we empirically analyze how text-based and image-based biases in input affect the behavior of GPT-4o and Qwen2-VL across varying confidence levels in Visual Question Answering (VQA) tasks. Our findings reveal that all considered biases induce greater changes in measured uncertainties, when model confidence after bias mitigation is lower. Moreover, lower model confidence leads to greater underestimation of epistemic uncertainty (i.e. overconfidence) due to the presence of bias, whereas it has no significant effect on the direction and smaller effect on the magnitude of aleatoric uncertainty changes. Based on these observations, we hypothesize that biases degrade the ranking performance of measured uncertainty, motivating our exploration of bias mitigation as a potential uncertainty quantification approach. This approach improves uncertainty quantification in the presence of aleatoric uncertainty with GPT-4o.

1 Introduction

Robust quantification of Large Language Models' (LLMs) confidence in their answers is vital for trust and safety in critical applications. Overestimating confidence can lead to erroneous decisions, while underestimating it prevent effective utilization of their capabilities (Hendrycks et al., 2021).

Much of the existing literature leverages uncertainty to estimate a model's confidence in its answers (Guo et al., 2017). General uncertainty quantification typically accounts for both epistemic and aleatoric uncertainties. However, only epistemic uncertainty indicates a model's confidence, as it reflects the limitations of the underlying knowledge. In contrast, aleatoric uncertainty stems from the irreducible randomness of the true answer distribution. An example of epistemic uncertainty and aleatoric uncertainty is provided in Figure 1.

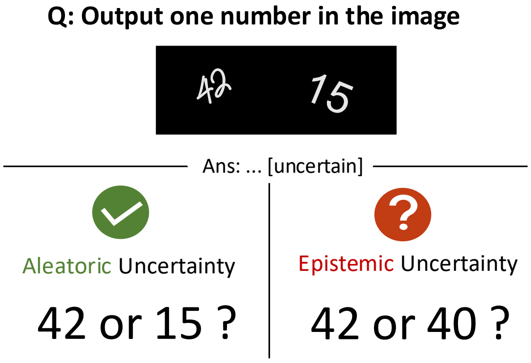


Figure 1: An example of aleatoric uncertainty and epistemic uncertainty. Both 42 and 15 are valid answers, while 40 reflects the model's lack of knowledge about the '42' in the image.

Earlier research predominantly tackles the aleatoric uncertainty from different phrasings of the same semantic meaning, often by semantic equivalence calculations (Kuhn et al., 2023; Farquhar et al., 2024; Lin et al., 2023). Recent work has shifted focus towards more general scenarios (Ahdritz et al., 2024; Yadkori et al., 2024), where multiple distinct semantic meanings are valid (Jiang et al., 2022; Jia et al., 2024; Barandas et al., 2024). These two studies concurrently find that models are more likely to copy information from prompts under high epistemic uncertainty, which indicates low model confidence, compared to high aleatoric uncertainty. As repeating provided information reflects confirmation bias (Nickerson, 1998; Shi et al., 2024), we hypothesize that models tend to rely more on bias rather than semantic meaning when their confidence without bias is lower. Thus, biases (Wang et al., 2023; Liu et al., 2024; Gavrikov et al., 2024; Ye et al., 2024) have a greater impact in lower-confidence scenarios.

To test our hypothesis, we construct visual-language datasets in which questions admit multiple semantically valid answers from VL_Checklist (Zhao et al., 2022) and CREPE (Ma et al., 2023). These datasets contain images with multiple correct and multiple incorrect descriptions, allowing us to analyze how the impact of text-based and image-based biases on epistemic and aleatoric uncertainty changes with different confidence levels.

Our findings show that for GPT-4o (Hurst et al., 2024) and Qwen2-VL (Wang et al., 2024), lower model confidence after bias mitigation correlates with stronger bias effects, estimated by the absolute change in measured uncertainties with and without bias. However, its influence on aleatoric uncertainty is weaker compared to epistemic uncertainty. Since greater overconfidence in lower-confidence instances, a phenomenon observed in human behavior (Sulistyawati et al., 2011), can undermine the ranking performance of the measured entropy, we further explore the impact of these directional changes. We find that epistemic uncertainty tends to be significantly more underestimated (overconfidence) with the bias when the model is actually less confident without the bias. However, model confidence has no significant impact on the directional changes in aleatoric entropy caused by biases.

Therefore, we explore bias mitigation as a potential uncertainty quantification method, which does not require access to the internal model state or additional data training. Our results show that removing text-based biases boosts the Area Under the Receiver Operating Characteristic (AUROC) by approximately 7% with GPT-4o.

2 Related Work

Uncertainty Quantification with a Single Valid Answer. Traditional machine learning models treat uncertainty as a measure of confidence when a single valid answer exists for each question (Hendrycks and Gimpel, 2016; Lakshminarayanan et al., 2017). This approach applies to tasks focused on predicting a single output (Nguyen and O'Connor, 2015; Guo et al., 2017; Wang et al., 2022). For instance, in single-choice classification problems like MMLU (Hendrycks et al., 2020), studies (Rae et al., 2021; Kadavath et al., 2022) show that LLMs are generally well-calibrated. However, as models grow and train on more diverse data, they are increasingly used for open-ended questions with multiple valid answers.

Reinforcement Learning with Human Feedback (RLHF) has complicated uncertainty estimation (Ouyang et al., 2022). Studies (Xiong et al., 2023; Zhou et al., 2024) show that RLHF-trained LLMs often overestimate their certainty, raising concerns about the reliability of self-reported uncertainty. This overconfidence highlights the potential misalignment between self-perception and actual knowledge. Moreover, Huang et al. (2023a) and Feng et al. (2024) found that self-reflection alone is insufficient for accurately assessing uncertainty.

Jiang et al. (2023) found that rephrasing and reordering prompts improve uncertainty quantification in single-answer settings. While their approach overlaps with ours in textual perturbation, we extend the analysis to commonly studied biases, including image-based biases, and examine multi-answer scenarios where epistemic and aleatoric uncertainty coexist. More importantly, we examine how biases affect these uncertainties across confidence levels, offering insights for improving uncertainty quantification.

Uncertainty Quantification with a Single Semantic Valid Answer. Prior work on uncertainty estimation in LLMs with aleatoric uncertainty mainly addresses variability in natural language generation, where multiple semantically equivalent outputs are valid. Common benchmarks like CoQA (Reddy et al., 2019), TriviaQA (Joshi et al., 2017), and AmbigQA (Min et al., 2020) mostly assume a single valid reference answer, with few allowing multiple.

Various techniques have been proposed to quantify uncertainty under such settings, including training auxiliary classifiers (Kamath et al., 2020; Cobbe et al., 2021) and leveraging internal model states (Ren et al., 2022; Burns et al., 2022; Lin et al., 2023).

Semantic equivalence has proven to be effective for uncertainty estimation in tasks with a single semantically valid answer. Kuhn et al. (2023) and Farquhar et al. (2024) used it to reduce aleatoric uncertainty from phrasing variability of the same semantic meaning. Research by Huang et al. (2023b) observed that sample-based methods outperform single-inference approaches for measuring epistemic uncertainty.

Building on these findings, we shift focus from phrasing variations to the complexity of multiple semantically valid answers, aiming to understand uncertainty estimation when validity spans diverse interpretations.

Uncertainty Quantification with Multiple Semantic Valid Answers. Uncertainty estimation becomes more complex with multiple semantically valid answers. Ahdritz et al. (2024) tackled this by assuming larger models capture aleatoric uncertainty, while a smaller model head is trained to predict it, requiring access to internal states. They also found that LLMs are more likely to copy information from the inputs when epistemically uncertain compared to aleatorically uncertain. Yadkori et al. (2024) built on similar findings by using mutual information to estimate epistemic uncertainty, measuring answer distribution dependency on provided hints. Their iterative prompting method requires second-order information computation.

This growing body of research underscores the need for more robust methods to distinguish epistemic uncertainty from aleatoric uncertainty in tasks with multiple semantically valid answers. We extend this by analyzing how bias effects on measured uncertainty vary with model confidence to improve uncertainty estimation.

In addition, unlike Yadkori et al. (2024), who used preselected multi-label queries with high entropy (> 0.7) constructed from WordNet (Fellbaum, 1998), where LLMs achieve near-perfect performance, we use unfiltered datasets to better reflect real-world challenges.

3 Bias Effects on Measured Uncertainties

Ahdritz et al. (2024) and Yadkori et al. (2024) both found that LLMs are more likely to copy input information when experiencing high epistemic uncertainty but not high aleatoric uncertainty, resembling confirmation bias. Inspired by these findings, we analyze how biases impact measured epistemic and aleatoric uncertainty across different confidence levels to provide further insights into distinguishing these two types of uncertainty.

Prompt Template
You are given an image and a set of descriptions. Your task is to evaluate each description and determine whether it is true based on the image.
Below are the descriptions:
{Label_0}: {Option_0}
{Label_1}: {Option_1}
{Label_2}: {Option_2}
{Label_3}: {Option_3}
Provide one index of the descriptions that are true, regardless of the number of descriptions that you believe are true. Return your response as a single index without any additional explanations or text. Here is an example format for your response:
0
Use the provided format and structure for your response.

Table 1: The Vanilla Prompt used to obtain greedy outputs from Large Language Models for evaluating their correctness. An example is provided in Appendix A.1.

3.1 Epistemic and Aleatoric Uncertainty

Epistemic uncertainty arises from uncertainty in distinguishing correct from incorrect predictions, reflecting the model's lack of knowledge or confidence. On the other hand, aleatoric uncertainty stems from uncertainty among multiple valid answers and represents irreducible randomness. We leverage ground-truth information to quantify a model's confidence after bias mitigation and evaluate the bias effects on measured epistemic and aleatoric uncertainty.

We define epistemic entropy as the entropy calculated from the probability of making a correct prediction (the sum of the probabilities of all valid answers) and the probabilities of each incorrect prediction, where i represents individual predicted outcomes:

$$P(\text{correct}) = \sum_{i \in \text{correct}} P(i) \tag{1}$$

$$\begin{aligned} \text{Epistemic Entropy} = &- P(\text{correct}) \log P(\text{correct}) \quad (2) \\ &- \sum_{i \notin \text{correct}} P(i) \log P(i) \quad (3) \end{aligned}$$

Aleatoric entropy is defined as the entropy computed from the probability distribution between the correct answers:

$$\text{Aleatoric Entropy} = - \sum_{i \in \text{correct}} \frac{P(i)}{P(\text{correct})} \log \frac{P(i)}{P(\text{correct})} \tag{4}$$

Consequently, the total measured entropy, based on all four options, can be decomposed into epistemic and aleatoric entropy as follows:

$$\text{Entropy} = \text{Epistemic Entropy} + P(\text{correct}) \times \text{Aleatoric Entropy} \tag{5}$$

We use epistemic entropy as the estimated epistemic uncertainty, and aleatoric entropy as the estimated aleatoric uncertainty in this paper.

Building on the proven effectiveness of semantic equivalence in addressing phrasing variability using LLM-based Natural Language Inference (Farquhar et al., 2024), this paper focuses on the challenge of multiple valid answers with distinct semantic meanings. We adopt a multiple-choice format where two semantically distinct options can be correct, providing a conceptual framework for analysis without first resolving semantic equivalence. For the generalization of uncertainty quantification from classification tasks to open-ended generation tasks, refer to Appendix B of Jiang et al. (2023).

With many top-performing models being closed-source, understanding their behavior without internal model states is crucial. We examine how input biases affect measured epistemic and aleatoric uncertainty, focusing on observable behaviors. Our analysis applies to both open- and closed-source models, offering insights into bias effects on measured uncertainties.

3.2 Biases

We consider three text-based biases and three image-based biases for our analysis. The text-based biases include:

Phrasing Bias. LLMs often rely on spurious linguistic correlations, making predictions without fully understanding context (Wang et al., 2021; Si et al., 2023). Since linguistic cues vary by input, we mitigate phrasing bias by rephrasing prompts while preserving semantic meaning to average out probability shifts caused by bias.

Positional Bias. LLMs are known to exhibit sensitivity to the positional arrangement of input options (Wang et al., 2023; Liu et al., 2024). We shuffle the positions of the options to neutralize the probability shift from positional bias across prompts.

Label Bias. While label bias falls under linguistic features like phrasing bias, shuffling assigned labels offers a more targeted intervention than general paraphrasing. Liu et al. (2024) highlighted its significant impact in GPT-3.5 and GPT-4.

While image-based biases are often reduced during the training stage through image perturbations (Shorten and Khoshgoftaar, 2019), we remain interested in exploring whether insights from text-based biases can also be applied to image-based biases. The three image-based biases we consider are:

Shape Bias. The shape bias of vision models has been discussed in several studies (He et al., 2023; Gavrikov et al., 2024), where models rely on shape cues to generate their outputs.

Orientation Bias. The orientation of images can influence the predictions of vision models, a phenomenon known as orientation bias (Henderson and Serences, 2021; Ye et al., 2024).

Low-level Feature Bias. Injecting noise into images mitigates biases by reducing reliance on low-level features, such as texture, background artifacts, lighting, and contrast (Shorten and Khoshgoftaar, 2019).

More details of prompts perturbation to mitigate biases are provided in Appendix A.1. We assess bias impact by measuring the absolute change in epistemic and aleatoric entropy between predictions from a biased prompt and averaged distributions from shuffled-bias prompts. We quantify bias-induced overconfidence by subtracting entropy from a biased prompt from that of the averaged distribution across perturbed prompts. We use linear regression analysis to examine the correlation, with confidence levels after bias mitigation as the independent variable and two types of bias effects as the dependent variables.

3.3 Uncertainty Quantification

We extend our experiments to explore bias mitigation as a potential uncertainty quantification method, estimating uncertainty without ground truth by reducing input biases, as illustrated in Figure 2.

Unlike the mutual information approach proposed by Yadkori et al. (2024), which adds information to induce confirmation bias, our method operates in a smaller search space by directly targeting biases in default inputs, eliminating the need for broader searches. While our methods have some overlap with the methods proposed by Jiang et al. (2023), our work focuses on a distinct setting where aleatoric uncertainty is present. Besides, we aim to emphasize the impact of broader biases, including image-based biases, on measured entropy.

4 Experiments

Dataset. We use the VL_checklist and CREPE datasets, which contain numerous images with human-verified positive and negative descriptions. In contrast, some datasets (Thrush et al., 2022; Tong et al., 2024) contain image descriptions but lack multiple correct and incorrect ones per image, while others (Ray et al., 2023; Liu et al., 2023) include only a limited number. For our analysis, we randomly select two correct and two incorrect descriptions and present them in a random order to ensure unbiased LLM evaluation.

These datasets evaluate more advanced model capabilities, compositional reasoning (Hua et al.,

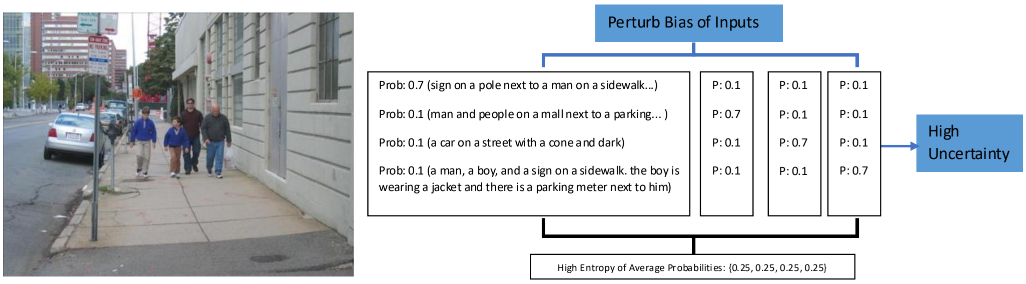

Perturb Bias of Inputs
Prob: 0.7 (sign on a pole next to a man on a sidewalk...)
Prob: 0.1 (man and people on a mall next to a parking...)
Prob: 0.1 (a car on a street with a cone and dark)
Prob: 0.1 (a man, a boy, and a sign on a sidewalk. the boy is wearing a jacket and there is a parking meter next to him)
P: 0.1
P: 0.7
P: 0.1
P: 0.1
P: 0.1
P: 0.1
P: 0.7
P: 0.1
P: 0.1
P: 0.1
P: 0.1
P: 0.7
High Uncertainty
High Entropy of Average Probabilities: {0.25, 0.25, 0.25, 0.25}

Figure 2: Perturb prompts to shuffle bias factors to achieve estimates of uncertainty without bias.

2024), compared to early multi-label datasets such as WordNet, where current LLMs achieve near-perfect performance. To balance data coverage and budget, we create 1,000 questions from 1,000 images per dataset.

Evaluation Metrics. We use linear regression coefficients and p-values to analyze bias impact on measured epistemic and aleatoric entropy across bias-mitigated confidence levels. The coefficients indicate how increases in model confidence influence bias-induced changes in measured uncertainties. A positive coefficient suggests greater bias effects with higher confidence, while a negative coefficient implies that higher confidence reduces bias impact. P-values assess statistical significance, with low values (typically <0.05) indicating a meaningful effect and high values suggesting the effect may be due to chance.

We adopt AUROC for uncertainty quantification, following prior studies (Band et al., 2022; Kuhn et al., 2023; Lin et al., 2023; Farquhar et al., 2024). As demonstrated by McDermott et al. (2024), AUROC is robust to class imbalance and effectively captures the ranking performance of uncertainty estimations.

Models. Given the popularity and diverse applications of the GPT series, we select the latest stable version of GPT-4o (‘gpt-4o-2024-11-20’) available at the time of our experiments. Additionally, we extend our empirical analysis to the open-source LLM Qwen2-VL (‘Qwen2-VL-72B-Instruct-GPTQ-Int4’).

Experimental Settings. Farquhar et al. (2024) found that sampling settings, like temperature and top-P, minimally affect sampling-based uncertainty quantification. Based on this, we fix generation parameters (temperature = 0.9, top-P = 1) to ensure consistency and avoid unnecessary tuning. We run ten shuffled prompts, aligning with the sample sizes used in previous sampling-based methods (Huang et al., 2023b; Kuhn et al., 2023; Farquhar et al., 2024) and the per-iteration sample count in iterative-based methods (Yadkori et al., 2024).

With OpenAI’s closed-source LLMs now providing top-20 token probabilities, we use this to compute prediction probabilities across four options, rather than approximating via sampling (Farquhar et al., 2024). We approximate unbiased model confidence by summing correct options from averaging probabilities across 10 shuffled prompts, reducing bias reliance. We also extend our experiments by approximating model confidence with epistemic entropy from the average probabilities, presented in Appendix A.3.

Following Kuhn et al. (2023) and Farquhar et al. (2024), we approximate greedy decoding by setting a very low temperature (1e-15) using a single-run output as the model’s ‘best generation’ for correctness labels in uncertainty quantification. Despite potential output variations in closed-source LLMs even at zero temperature, this approach ensures consistency and aligns with established research. Table 1 shows the prompt used to obtain the model’s answer for correctness evaluation.

5 Results and Analysis

5.1 Relationship Between Model Confidence and Bias Effects

We use unbiased model confidence, the sum of the unbiased probabilities of correct options, as the independent variable and analyze its impact on changes in measured epistemic and aleatoric entropy with and without bias. A greater change indicates a stronger bias impact on the model’s behavior. Our results, derived from two models and two datasets as shown in Table 2, reveal consistent patterns across all biases:

Lower model confidence correlates with greater bias influence. When the model exhibits lower

Dataset	Bias	Metrics	GPT-4o			Qwen2-VL		
			Epistemic	Aleatoric	Ratio Epi./Ale.	Epistemic	Aleatoric	Ratio Epi./Ale.
VL_Checklist	Phrasing	Coefficients	- 0.2300	- 0.0579	3.97	- 0.0332	- 0.0123	2.70
		P-value	(< 0.001)	(0.006)		(< 0.001)	(0.079)	
	Positional	Coefficients	- 0.6098	- 0.0629	9.69	- 0.1571	- 0.0844	1.86
		P-value	(< 0.001)	(0.111)		(< 0.001)	(< 0.001)	
	Label	Coefficients	- 0.3572	- 0.0911	3.92	0.0602	0.0757	0.80
		P-value	(< 0.001)	(0.005)		(0.001)	(< 0.001)	
	Shape	Coefficients	- 0.1679	- 0.0707	2.37	- 0.0664	- 0.0081	8.20
		P-value	(< 0.001)	(< 0.001)		(< 0.001)	(0.042)	
	Orientation	Coefficients	- 0.1746	- 0.0671	2.60	- 0.1073	- 0.0230	4.67
		P-value	(< 0.001)	(< 0.001)		(< 0.001)	(0.157)	
	Low-level Feature	Coefficients	- 0.1466	- 0.0457	3.21	- 0.0493	- 0.0214	2.30
		P-value	(< 0.001)	(0.004)		(< 0.001)	(0.026)	
CREPE	Phrasing	Coefficients	- 0.1149	- 0.0481	2.39	- 0.0025	- 0.0011	2.27
		P-value	(< 0.001)	(< 0.001)		(0.521)	(0.804)	
	Positional	Coefficients	- 0.2914	- 0.1162	2.51	0.0192	0.0525	0.37
		P-value	(< 0.001)	(< 0.001)		(0.307)	(0.002)	
	Label	Coefficients	- 0.1663	- 0.1147	1.45	0.0638	0.0407	1.57
		P-value	(< 0.001)	(< 0.001)		(< 0.001)	(0.001)	
	Shape	Coefficients	- 0.0952	- 0.0215	4.43	- 0.0196	- 0.0188	1.04
		P-value	(< 0.001)	(0.042)		(0.013)	(0.018)	
	Orientation	Coefficients	- 0.0797	- 0.0347	2.30	- 0.0320	- 0.0106	3.02
		P-value	(< 0.001)	(0.006)		(0.004)	(0.324)	
	Low-level Feature	Coefficients	- 0.0919	- 0.0336	2.74	- 0.0202	- 0.0044	4.59
		P-value	(< 0.001)	(0.002)		(0.002)	(0.466)	

Table 2: Both GPT-4o and Qwen2-VL exhibit greater sensitivity to bias at lower confidence levels, as reflected in absolute changes in both epistemic and aleatoric entropy with and without biases. This is supported by the consistent negative coefficients. Additionally, the impact of bias on epistemic entropy is more strongly correlated with the model confidence compared to aleatoric entropy, as indicated by coefficient Ratio Epi./Ale. greater than one and the relatively lower statistical significance of p-values for aleatoric entropy.

confidence without bias, its outputs tend to be more susceptible to bias, as evidenced by consistently negative coefficients for GPT-4o and only three exceptions among 12 points in Qwen-2.

The impact of bias on measured epistemic entropy is more strongly correlated with the model confidence than measured aleatoric entropy. This is evidenced by the consistently higher coefficients for epistemic entropy compared to aleatoric entropy, as indicated by Ratio Epi./Ale. greater than one for GPT-4o with only two exceptions for Qwen2-VL. In some cases, the impact of bias on aleatoric uncertainty appears unrelated or nearly unrelated to the unbiased model confidence, as indicated by large p-values for aleatoric entropy, which shows no statistical significance.

Similar results are obtained using debiased epistemic entropy as the approximated confidence, as shown in Appendix A.3.

5.2 Relationship Between Model Confidence and Bias-Induced Overconfidence

While lower model confidence leads to greater bias-induced changes, the direction of this change is crucial. If bias increases under-confidence (raising measured entropy), instances with already low confidence without the bias will exhibit even lower confidence with the bias, whereas higher-confidence instances will remain relatively higher with a less pronounced reduction. Consequently, the ranking of measured entropy remains unaffected by the presence of the bias. However, greater over-confidence in lower-confidence instances disrupts the ranking of measured entropy.

Therefore, we further use model confidence to analyze its impact on the entropy reduction, subtracting measured entropy derived from a single biased prompt from that calculated from bias-shuffled prompts. Our results, derived from two models and two datasets as shown in Table 3, reveal consistent patterns across all biases:

Lower model confidence is associated with greater underestimation of epistemic entropy in the presence of bias. When model confidence after bias mitigation is lower, measured epistemic entropy decreases more with bias than without it. This is evidenced by the consistently negative coefficients associated with the measured epistemic entropy reduction caused by bias, with the majority of p-values indicating statistical significance.

Model confidence has no significant effect on the direction of aleatoric entropy changes caused by bias. This is supported by the inconsistent coefficient directions associated with the measured

Dataset	Bias	Metrics	GPT-4o			Qwen2-VL		
			Epistemic	Aleatoric	Ratio Epi./Ale.	Epistemic	Aleatoric	Ratio Epi./Ale.
VL_Checklist	Phrasing	Coefficients	- 0.1651	0.0157	10.52	- 0.0158	- 0.0198	0.80
		P-value	(< 0.001)	(0.547)		(0.042)	(0.032)	
	Positional	Coefficients	- 0.7585	- 0.0499	15.2	- 0.1827	- 0.0722	2.53
		P-value	(< 0.001)	(0.285)		(< 0.001)	(0.021)	
	Label	Coefficients	- 0.3811	- 0.0898	4.24	-0.0338	-0.0233	1.45
		P-value	(< 0.001)	(0.030)		(0.280)	(0.355)	
	Shape	Coefficients	- 0.1542	- 0.0344	4.48	- 0.0620	- 0.0013	47.69
		P-value	(< 0.001)	(0.156)		(< 0.001)	(0.938)	
	Orientation	Coefficients	- 0.1441	- 0.0181	7.96	- 0.1309	- 0.0235	5.57
		P-value	(< 0.001)	(0.433)		(< 0.001)	(0.264)	
	Low-level Feature	Coefficients	- 0.1188	- 0.0121	9.82	- 0.0257	- 0.0011	23.36
		P-value	(< 0.001)	(0.525)		(< 0.009)	(0.921)	
CREPE	Phrasing	Coefficients	- 0.1019	0.0184	5.54	- 0.0242	0.0097	2.49
		P-value	(< 0.001)	(0.268)		(< 0.001)	(0.109)	
	Positional	Coefficients	- 0.3929	- 0.0772	5.09	- 0.0951	0.0392	2.43
		P-value	(< 0.001)	(0.024)		(< 0.001)	(0.083)	
	Label	Coefficients	- 0.2641	- 0.1082	2.44	- 0.0152	0.0184	0.83
		P-value	(< 0.001)	(< 0.001)		(0.430)	(0.261)	
	Shape	Coefficients	- 0.0580	0.0068	8.52	- 0.0147	- 0.0082	1.79
		P-value	(< 0.001)	(0.605)		(0.163)	(0.421)	
	Orientation	Coefficients	- 0.0586	- 0.0206	2.84	- 0.0776	- 0.0095	8.17
		P-value	(< 0.001)	(0.169)		(< 0.001)	(0.495)	
	Low-level Feature	Coefficients	- 0.0741	- 0.0181	4.09	- 0.0152	- 0.0079	1.92
		P-value	(< 0.001)	(0.172)		(0.062)	(0.285)	

Table 3: Both GPT-4o and Qwen2-VL exhibit greater overconfidence in measured epistemic entropy due to bias when their confidence is lower, demonstrated by the negative coefficients and statistically significant p-values. In contrast, model confidence has no significant effect on the direction of aleatoric entropy changes caused by bias, supported by the inconsistent coefficient directions and statistically insignificant p-values.

aleatoric entropy reduction due to the presence of bias, with the majority of p-values showing no statistical significance.

Similar results are obtained using debiased epistemic entropy as the approximated confidence, as shown in Appendix A.3.

5.3 Uncertainty Quantification Through Bias Mitigation

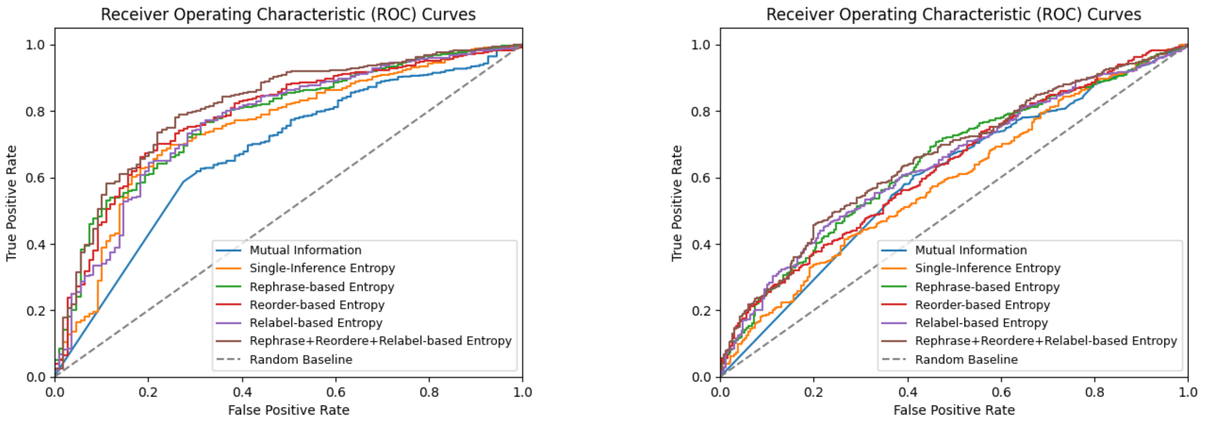
Section 5.2 shows that biases systematically distort the ranking of epistemic entropy and, consequently, measured entropy (Equation 5), while their impact on aleatoric entropy is much less and non-directional. Therefore, we hypothesize that when model confidence (self-perception) ranking aligns with the ranking of probability of being correct (true knowledge) (Farquhar et al., 2024) and bias is considerable, mitigating bias can improve uncertainty quantification, making it more robust to the presence of aleatoric uncertainty. Since GPT-4o shows higher coefficients than Qwen2-VL in both Table 2 and 3, indicating greater bias-induced noise, we focus our experiments on GPT-4o.

Baselines. We focus on **Entropy** for uncertainty quantification, as it has been proven to be a strong baseline in recent studies on methods applicable to closed-source LLMs (Kuhn et al., 2023; Farquhar et al., 2024; Yadkori et al., 2024). We also include other commonly used baselines, namely the **probability of the prediction (Prob)** and the **number of answers (#Answers)**. Additionally, we incorporate the recently proposed **Mutual Information** approach by Yadkori et al. (2024), which leverages prompt perturbation to quantify confirmation bias.

Since closed-source models may yield varying token probabilities for the same prompt under identical decoding, we introduce an additional baseline: averaging probabilities over multiple samples of the same prompt. This **Repetitive-based** approach allows us to investigate whether improved performance arises from better probability estimation due to increased sampling.

Analysis. As shown in Table 4, we observe that repetitive-based samplings have minimal improvement over single-inference estimations.

Mitigating biases consistently improves performance across all baselines considered. While no single bias mitigation method demonstrates a clear advantage over the others, summing the entropy obtained from each bias removal leads to further performance gains. This finding aligns with the results reported by Jiang et al. (2023), despite being applied in a distinct setting where aleatoric uncertainty is present. The similar accuracies achieved by perturbed prompts are presented in Appendix A.2, indicating that the observed improvement is



(a) ROC curves on VL_Checklist.

(b) ROC curves on CREPE.

Figure 3: Comparison of ROC curves for the text-based bias mitigation methods and baselines on two datasets using GPT-4o. The high prevalence of identical Mutual Information estimates makes it less suitable when a high abstention rate is required. The bias mitigation approach maintains robustness across different thresholds.

Methods	**# Inference**	**VL_Checklist**	**CREPE**
Mutual Information	20	0.6782	0.5973
Repetitive-based #Answers	10	0.6763	0.5821
Rephrased-based #Answers (proposed)	10	0.7328	0.6106
Single-inference Prob	1	0.7349	0.5801
Repetitive-based Prob	10	0.7233	0.6017
Rephrase-based Prob (proposed)	10	0.7762	0.6513
Single-inference Entropy	1	0.7492	0.5870
Repetitive-based Entropy	10	0.7412	0.6084
Rephrase-based Entropy (proposed)	10	0.7779	0.6442
Reorder-based Entropy (proposed)	10	0.7844	0.6299
Relabel-based Entropy (proposed)	10	0.7665	0.6406
Rephrase+Reorder+Relabel-based Entropy (proposed)	10*3	**0.8123**	**0.6588**
Resize-based Entropy (proposed)	10	0.7605	0.6219
Rotate-based Entropy (proposed)	10	0.7565	0.6204
Noise-based Entropy (proposed)	10	0.7535	0.6252
Resize+Rotate+Noise-based Entropy (proposed)	10*3	*0.7699*	*0.6287*

Table 4: This table presents the AUROC scores for epistemic uncertainty quantification using different methods with GPT-4o. While the Repetitive-based method shows minimal improvement, bias mitigation methods based on any single bias consistently enhance performance on both datasets. Furthermore, combining methods based on different biases further improve performance over individual methods.

not attributable to prompt quality.

Among the bias mitigation strategies, combining three text-based methods yields the greatest performance improvement, increasing AUROC by 6.39% on VL_Checklist and 7.18% on CREPE. In contrast, combining three image-based methods results in a relatively modest improvement, with 2.07% on VL_Checklist and 4.17% on CREPE. This smaller gain likely results from training image augmentation already mitigating image-based biases, reducing the need for additional bias correction at inference. Combining image- and text-based bias mitigation yields no further gains, suggesting text-based corrections capture most biases affecting uncertainty estimation. These findings highlight that bias removal is not only crucial for ensuring fairness in predictions but also essential for accurately quantifying uncertainty when bias is significant.

The low performance of the Mutual Information method can be attributed to the concentration of its values as shown in Figure 3, a challenge shared by the # Answers method. Specifically, the large number of identical Mutual Information values limits its ability to differentiate within a significant subset, particularly among instances with low uncertainty estimates, as reflected in the low AUROC score. This limitation reduces its suitability for high-stakes applications that potentially demand a high rate of abstention. In contrast, the bias mitigation approach based on text maintains robustness across different thresholds.

6 Conclusion

While entropy decomposes into epistemic and aleatoric components, our findings show that lower model confidence amplifies bias effects on measured entropy. Notably, bias influence on epistemic entropy is more sensitive to model confidence than on aleatoric entropy.

Furthermore, while model confidence has no significant effect on the direction of aleatoric entropy changes caused by bias, we observe that lower model confidence leads to a greater underestimation of epistemic entropy in the presence of bias.

We improve uncertainty quantification by removing three text-based biases and three image-based biases in AUROC with GPT-4o, though image-based bias removal has a smaller effect, likely due to existing image perturbation during training.

Limitations

Reliance on Token Probabilities. While OpenAI provides token probabilities for its closed-source models, other LLMs impose stricter limitations. Some return only the predicted token's probability without alternatives, while others, like Gemini, limit usage to one query per day. These constraints hinder the entropy-based uncertainty quantification method we use, which may require more samples to approximate the token probabilities.

Increase in Inference Cost. While bias mitigation enhances the robustness of uncertainty quantification, it comes at the expense of the increased number of inferences. Shuffling prompts to account for each individual bias requires multiple model queries, increasing costs compared to single-inference methods.

References

Gustaf Ahdritz, Tian Qin, Nikhil Vyas, Boaz Barak, and Benjamin L Edelman. 2024. Distinguishing the knowable from the unknowable with language models. *arXiv preprint arXiv:2402.03563*.

Neil Band, Tim GJ Rudner, Qixuan Feng, Angelos Filos, Zachary Nado, Michael W Dusenberry, Ghassen Jerfel, Dustin Tran, and Yarin Gal. 2022. Benchmarking bayesian deep learning on diabetic retinopathy detection tasks. *arXiv preprint arXiv:2211.12717*.

Marília Barandas, Lorenzo Famiglini, Andrea Campagner, Duarte Folgado, Raquel Simão, Federico Cabitza, and Hugo Gamboa. 2024. Evaluation of uncertainty quantification methods in multi-label classification: A case study with automatic diagnosis of electrocardiogram. *Information Fusion*, 101:101978.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Christiane Fellbaum. 1998. Wordnet: An electronic lexical database. *MIT Press google schola*, 2:678–686.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*.

Paul Gavrikov, Jovita Lukasik, Steffen Jung, Robert Geirhos, Bianca Lamm, Muhammad Jehanzeb Mirza, Margret Keuper, and Janis Keuper. 2024. Are vision language models texture or shape biased and can we steer them? *arXiv preprint arXiv:2403.09193*.

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.

Xilin He, Qinliang Lin, Cheng Luo, Weicheng Xie, Siyang Song, Feng Liu, and Linlin Shen. 2023. Shift from texture-bias to shape-bias: Edge deformation-based augmentation for robust object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1526–1535.

Margaret Henderson and John T Serences. 2021. Biased orientation representations can be explained by experience with nonuniform training set statistics. *Journal of Vision*, 21(8):10–10.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2021. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*.

Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Hang Hua, Yunlong Tang, Ziyun Zeng, Liangliang Cao, Zhengyuan Yang, Hangfeng He, Chenliang Xu, and Jiebo Luo. 2024. Mmcomposition: Revisiting the compositionality of pre-trained vision-language models. *arXiv preprint arXiv:2410.09733*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023a. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2023b. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Zixia Jia, Junpeng Li, Shichuan Zhang, Anji Liu, and Zilong Zheng. 2024. Combining supervised learning and reinforcement learning for multi-label classification tasks with partial labels. *arXiv preprint arXiv:2406.16293*.

Jyun-Yu Jiang, Wei-Cheng Chang, Jiong Zhong, Cho-Jui Hsieh, and Hsiang-Fu Yu. 2022. Uncertainty in extreme multi-label classification. *arXiv preprint arXiv:2210.10160*.

Mingjian Jiang, Yangjun Ruan, Sicong Huang, Saifei Liao, Silviu Pitis, Roger Baker Grosse, and Jimmy Ba. 2023. Calibrating language models via augmented prompt ensembles.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. *arXiv preprint arXiv:2006.09462*.

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

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

Xinyi Liu, Pinxin Liu, and Hangfeng He. 2024. An empirical analysis on large language models in debate evaluation. *arXiv preprint arXiv:2406.00050*.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. 2023. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921.

Matthew McDermott, Lasse Hyldig Hansen, Haoran Zhang, Giovanni Angelotti, and Jack Gallifant. 2024. A closer look at auroc and auprc under class imbalance. *arXiv preprint arXiv:2401.06091*.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. *arXiv preprint arXiv:2004.10645*.

Khanh Nguyen and Brendan O'Connor. 2015. Posterior calibration and exploratory analysis for natural language processing models. *arXiv preprint arXiv:1508.05154*.

Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, and 1 others. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan A Plummer, Ranjay Krishna, and Kate Saenko. 2023. Cola: How to adapt vision-language models to compose objects localized with attributes. *arXiv preprint arXiv:2305.03689*, 2.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. 2022. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations*.

Li Shi, Houjiang Liu, Yian Wong, Utkarsh Mujumdar, Dan Zhang, Jacek Gwizdka, and Matthew Lease. 2024. Argumentative experience: Reducing confirmation bias on controversial issues through llm-generated multi-persona debates. *arXiv preprint arXiv:2412.04629*.

Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48.

Chenglei Si, Dan Friedman, Nitish Joshi, Shi Feng, Danqi Chen, and He He. 2023. What spurious features can pretrained language models combat?

Ketut Sulistyawati, Christopher D Wickens, and Yoon Ping Chui. 2011. Prediction in situation awareness: Confidence bias and underlying cognitive abilities. *The International Journal of Aviation Psychology*, 21(2):153–174.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. 2021. Identifying and mitigating spurious correlations for improving robustness in nlp models. *arXiv preprint arXiv:2110.07736*.

Yuxia Wang, Daniel Beck, Timothy Baldwin, and Karin Verspoor. 2022. Uncertainty estimation and reduction of pre-trained models for text regression. *Transactions of the Association for Computational Linguistics*, 10:680–696.

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

Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. 2024. To believe or not to believe your llm. *arXiv preprint arXiv:2406.02543*.

Wenqian Ye, Guangtao Zheng, Yunsheng Ma, Xu Cao, Bolin Lai, James M Rehg, and Aidong Zhang. 2024. Mm-spubench: Towards better understanding of spurious biases in multimodal llms. *arXiv preprint arXiv:2406.17126*.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. 2022. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*.

Kaitlyn Zhou, Jena D Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. *arXiv preprint arXiv:2401.06730*.

A Appendix

A.1 Details of Prompt Design

Table 5 gives an example of vanilla prompt we used in our experiments.

Prompt Example
You are given an image and a set of descriptions. Your task is to evaluate each description and determine whether it is true based on the image.
Below are the descriptions:
0: person sitting in a boat with a paddle in the water. there is another paddle and boat in the water. the boat has writing on the side of it.
1: person wearing shirt and captain on boat in water
2: a boat with a paddle and captain on it, in dioxide
3: captain of ground with yacht in water
Provide one index of the descriptions that are true, regardless of the number of descriptions that you believe are true. Return your response as a single index without any additional explanations or text.
Here is an example format for your response:
0
Use the provided format and structure for your response.

Table 5: The Vanilla Prompt example used to obtain greedy outputs.

Phrasing Bias. We utilize GPT-4o to help paraphrase our default prompt shown in Table 1 while keeping the options unchanged. Table 10 lists all the rephrased prompts used in our experiments to perturb bias related to phrasing.

Positional Bias. To perturb positional bias, we shuffle the assignments of option_0, option_2, option_3, and option_4 in the prompt template shown in Table 1, while keeping the four labels in their natural order: 0, 1, 2, 3.

Label Bias. To perturb label bias, we maintain the original positions of the options but shuffle the labels assigned to Label_0, Label_1, Label_2, and Label_3, such as 2, 0, 3, 1.

Shape Bias. We resize images across different inputs by varying the length-to-width ratio from 0.5 to 1.5, intentionally distorting the shapes of objects in the images.

Orientation Bias. We rotate images across different inputs by varying the rotated degrees from -45° to 45°. The rotation angles are kept relatively small to preserve the overall spatial relationships within the images.

Low-level Feature Bias. We add random Gaussian noise with mean=0 and std=25 to the images across different inputs to disrupt local features while preserving their overall semantic meaning.

A.2 Accuracy Comparison Between Default Prompt and Single Perturbed Prompt

Model	Dataset	Bias	Accuracy (%)
GPT-4o	VL_Checklist	Default	89.1
		Phrasing	86.5
		Positional	85.8
		Label	83.6
		Shape	87.5
		Orientation	86.5
		Low-level Feature	86.7
	CREPE	Default	73.3
		Phrasing	73.7
		Positional	71.7
		Label	70.7
		Shape	73.1
		Orientation	72.9
		Low-level Feature	72.8
Qwen2-VL	VL_Checklist	Default	92.1
		Phrasing	82.1
		Positional	82.8
		Label	77.9
		Shape	82.2
		Orientation	81.4
		Low-level Feature	81.5
	CREPE	Default	78.7
		Phrasing	78.5
		Positional	78.7
		Label	77.9
		Shape	76.7
		Orientation	75.6
		Low-level Feature	74.9

Table 6: This table presents the accuracy achieved by the default prompt and the average accuracy achieved by each perturbed prompt with regard to each bias.

Table 6 presents the accuracy comparison between the default prompt with greedy generation and each single bias-perturbed prompt used in our sampling method. The ranking of prompt performance does not correlate with their effectiveness in uncertainty quantification, indicating that the improvements in uncertainty quantification cannot be attributed to prompt quality.

A.3 More Empirical Results

Dataset	GPT-4o	Qwen2-VL
VL_Checklist	1.01	1.06
CREPE	1.27	1.22

Table 7: This table presents the ratio of Epistemic entropy to Aleatoric entropy across both datasets and models using the default prompt. Ratios closer to one indicate that aleatoric entropy is comparable in magnitude to epistemic entropy.

Table 7 shows that the magnitude of aleatoric entropy is comparable that of epistemic entropy.

We further validate our empirical findings by using the epistemic entropy after bias reduction, calculated from the average probabilities of ten shuffled prompts, as an approximation of the underlying model confidence. The results remain consistent with those obtained when approximating model confidence using the sum of the probabilities of correct options from the average probabilities.

More specifically, the effects of bias, measured by changes in measured uncertainties, are more pronounced when model confidence is lower; in other words, when debiased epistemic entropy is higher. This is evidenced by consistently positive and statistically significant coefficients for changes in measured epistemic uncertainty due to biases in GPT-4o. Qwen2-VL follows the same pattern, with two exceptions: Label bias in VL_Checklist and Positional bias in CREPE. For aleatoric uncertainty, GPT-4o also shows predominantly positive coefficients, whereas Qwen2-VL exhibits inconsistent coefficient directions with much smaller values, as indicated by Epi./Ale. ratios greater than one-except for the same two exceptions. These results are detailed in Table 8.

Lower model confidence is more strongly associated with greater underestimation of measured epistemic uncertainty, whereas it has no significant effect on the direction of changes in measured aleatoric uncertainty. This is supported by the consistently positive and largely significant coefficients for the decrease in measured epistemic uncertainty, while the coefficients for the decrease in measured aleatoric uncertainty are predominantly insignificant except the same two Qwen2-VL cases.

Dataset	Bias	Metrics	GPT-4o			Qwen2-VL		
			Epistemic	Aleatoric	Ratio Epi./Ale.	Epistemic	Aleatoric	Ratio Epi./Ale.
VL_Checklist	Phrasing	Coefficients	0.2622	0.0739	3.55	0.0347	- 0.0056	6.20
		P-value	(< 0.001)	(< 0.001)		(<0.001)	(0.234)	
	Positional	Coefficients	0.4719	0.0379	12.45	0.1326	- 0.0654	2.03
		P-value	(< 0.001)	(0.087)		(< 0.001)	(< 0.001)	
	Label	Coefficients	0.2999	0.0575	5.22	-0.0255	-0.0828	0.31
		P-value	(< 0.001)	(0.004)		(0.064)	(< 0.001)	
	Shape	Coefficients	0.2023	0.0822	2.46	0.0644	0.0144	4.47
		P-value	(< 0.001)	(< 0.001)		(< 0.001)	(0.132)	
	Orientation	Coefficients	0.2126	0.0876	2.43	0.0916	0.0316	2.90
		P-value	(< 0.001)	(< 0.001)		(< 0.001)	(0.005)	
	Low-level Feature	Coefficients	0.1851	0.0536	3.45	0.0476	0.0205	2.32
		P-value	(< 0.001)	(< 0.001)		(< 0.001)	(0.003)	
CREPE	Phrasing	Coefficients	0.1825	0.0558	3.27	0.0067	- 0.0020	3.30
		P-value	(< 0.001)	(< 0.001)		(0.043)	(0.614)	
	Positional	Coefficients	0.3344	0.0476	7.03	0.0139	-0.0508	0.27
		P-value	(< 0.001)	(0.024)		(0.356)	(< 0.001)	
	Label	Coefficients	0.2129	0.0721	2.95	- 0.0744	- 0.0676	1.10
		P-value	(< 0.001)	(< 0.001)		(< 0.001)	(< 0.001)	
	Shape	Coefficients	0.1694	0.0423	4.00	0.0173	- 0.0029	5.97
		P-value	(< 0.001)	(< 0.001)		(0.011)	(0.675)	
	Orientation	Coefficients	0.1723	0.0689	2.50	0.0227	- 0.0084	2.70
		P-value	(< 0.001)	(< 0.001)		(0.020)	(0.364)	
	Low-level Feature	Coefficients	0.1565	0.0517	3.03	0.0184	0.0064	2.88
		P-value	(< 0.001)	(< 0.001)		(0.001)	(0.227)	

Table 8: Both GPT-4o and Qwen2-VL exhibit greater changes in measured entropy due to bias when the true confidence, approximated by epistemic entropy derived from the average probabilities of shuffled prompts, is lower. This is indicated by the consistently positive coefficients for epistemic entropy, and much lower coefficients for aleatoric entropy as ratios greater than one. Additionally, the impact of bias on epistemic entropy is more strongly correlated with the model confidence than aleatoric entropy.

Dataset	Bias	Metrics	GPT-4o			Qwen2-VL		
			Epistemic	Aleatoric	Ratio Epi./Ale.	Epistemic	Aleatoric	Ratio Epi./Ale.
VL_Checklist	Phrasing	Coefficients	0.1537	0.0187	8.22	0.0230	- 0.0071	3.24
		P-value	(< 0.001)	(0.374)		(< 0.001)	(0.250)	
	Positional	Coefficients	0.4874	0.0330	14.8	0.1311	- 0.0449	2.92
		P-value	(< 0.001)	(0.229)		(< 0.001)	(0.024)	
	Label	Coefficients	0.2942	0.0486	6.05	0.0267	-0.0070	3.81
		P-value	(< 0.001)	(0.059)		(0.211)	(0.685)	
	Shape	Coefficients	0.1277	0.0438	2.92	0.0387	- 0.0033	47.69
		P-value	(< 0.001)	(0.022)		(< 0.001)	(0.779)	
	Orientation	Coefficients	0.1590	0.0289	5.50	0.0883	0.0108	8.18
		P-value	(< 0.001)	(0.117)		(< 0.001)	(0.457)	
	Low-level Feature	Coefficients	0.1219	0.0192	6.35	0.0272	- 0.0080	3.4
		P-value	(< 0.001)	(0.236)		(< 0.009)	(0.333)	
CREPE	Phrasing	Coefficients	0.1577	- 0.008	197.13	0.0116	0.0070	1.66
		P-value	(< 0.001)	(0.961)		(0.023)	(0.183)	
	Positional	Coefficients	0.4043	0.0327	12.36	0.0975	- 0.0433	2.25
		P-value	(< 0.001)	(0.230)		(< 0.001)	(0.017)	
	Label	Coefficients	0.2890	0.0863	3.35	0.0171	- 0.0419	0.41
		P-value	(< 0.001)	(< 0.001)		(0.505)	(0.005)	
	Shape	Coefficients	0.1425	0.0108	13.19	0.0282	- 0.0060	4.70
		P-value	(< 0.001)	(0.436)		(0.002)	(0.496)	
	Orientation	Coefficients	0.1478	0.0579	2.55	0.0738	- 0.0022	33.55
		P-value	(< 0.001)	(< 0.001)		(< 0.001)	(0.857)	
	Low-level Feature	Coefficients	0.1299	- 0.0083	15.65	0.0186	0.0033	5.64
		P-value	(< 0.001)	(0.561)		(0.010)	(0.613)	

Table 9: Both GPT-4o and Qwen2-VL exhibit greater overconfidence in measured epistemic entropy due to bias when their confidence is lower, supported by positive coefficients and statistically significant p-values. In contrast, model confidence has no significant effect on the direction of aleatoric entropy changes caused by bias, as the directions of coefficients are inconsistent and p-values are not statistically significant.

Prompt Template 1
You are given an image and a set of descriptions. Your task is to evaluate each description and determine whether it is true based on the image.
Below are the descriptions:
<Options >
Provide one index of the descriptions that are true, regardless of the number of descriptions that you believe are true. Return your response as a single index without any additional explanations or text. Here is an example format for your response:
0
Use the provided format and structure for your response.

Prompt Template 2
You are presented with an image and a list of descriptions. Your task is to assess each description and judge if it is true based on the image.
The descriptions are listed below:
<Options >
Indicate one index of the descriptions that are true, regardless of how many you think are correct. Return your response as a single index without any additional explanations or text. Here is an example format for your response:
0
Use the provided format and structure for your response.

Prompt Template 3
You have an image and several descriptions. Your task is to evaluate each description and determine its validity based on the image.
Below are the descriptions:
<Options >
List one index of the descriptions that are true, even if multiple descriptions seem accurate. Return your response as a single index without any additional explanations or text. Here is an example format for your response:
0
Use the provided format and structure for your response.

Prompt Template 4
Given an image and a set of descriptions, your task is to evaluate each description and determine if it is true based on the image.
Here are the descriptions:
<Options >
Provide one index of the descriptions that are true, even if multiple descriptions are accurate. Respond with a single index without any additional explanations or text. Here is an example format for your response:
0
Use the provided format and structure for your response.

Prompt Template 5
You have an image and a series of descriptions. Your task is to evaluate each description to determine its truthfulness based on the image.
Below are the descriptions:
<Options >
Indicate one index of the true descriptions, even if there are multiple true descriptions. Return your response as a single index without any additional explanations or text. Here is an example format for your response:
0
Use the provided format and structure for your response.

Prompt Template 6
Given an image and several descriptions, your task is to evaluate each description and determine whether it is true based on the image.
Here are the descriptions:
<Options >
Provide one index of the true descriptions, even if multiple descriptions are valid. Return your response as a single index without any additional explanations or text. Here is an example of how your response should look:
0
Use the provided format and structure for your response.

Prompt Template 7
You are provided with an image and a series of descriptions. Evaluate each description to determine if it is true based on the image.
Below are the descriptions:
<Options >
Provide one index of the descriptions that are true, even if there are multiple descriptions that seem valid. Return your response as a single index without any additional explanations or text. Here is an example format for your response:
0
Use the provided format and structure for your response.

Prompt Template 8
Your task is to evaluate an image and a set of descriptions to determine if each description is true based on the image.
Here are the descriptions:
<Options >
Provide an index of the true description(s), even if multiple descriptions seem correct. Return your response as a single index without any additional explanations or text. Here is an example format for your response:
0
Use the provided format and structure for your response.

Prompt Template 9
You have been given an image and a list of descriptions. Your task is to evaluate each description and determine if it is true based on the image.
The descriptions are as follows:
<Options >
Provide one index of the descriptions that are true, even if you think more than one description is correct. Return your response as a single index without any additional explanations or text. Here is an example format for your response:
0
Use the provided format and structure for your response.

Prompt Template 10
You've been presented with an image alongside a series of descriptions. Your objective is to assess each description to determine its accuracy based on the image.
The descriptions are listed below:
<Options >
You need to identify one description that is true, regardless of how many you think are correct. Please format your response as a single index without any additional explanations or text. Here is an example of how your response should look:
0
Ensure you adhere to this format and structure in your response..

Table 10: The ten prompts used to average the output distribution of Large Language Models in order to reduce phrasing bias through paraphrasing.