The Role of Model Confidence on Bias Effects in Measured Uncertainties for Vision-Language Models

Xinyi Liu[1] Weiguang Wang[2] Hangfeng He[1]
[1]University of Rochester [2]Johns Hopkins University
xinyi.liu1@simon.rochester.edu, www@jhu.edu
hangfeng.he@rochester.edu

Abstract

With the growing adoption of Large Language Models (LLMs) for open-ended tasks, accurately assessing epistemic uncertainty, which reflects a model's lack of knowledge, has become crucial to ensuring reliable outcomes. However, quantifying epistemic uncertainty in such tasks is challenging due to the presence of aleatoric uncertainty, which arises from multiple valid answers. While bias can introduce noise into epistemic uncertainty estimation, it may also reduce noise from aleatoric uncertainty. To investigate this trade-off, we conduct experiments on Visual Question Answering (VQA) tasks and find that mitigating prompt-introduced bias improves uncertainty quantification in GPT-4o. Building on prior work showing that LLMs tend to copy input information when model confidence is low, we further analyze how these prompt biases affect measured epistemic and aleatoric uncertainty across varying bias-free confidence levels with GPT-4o and Qwen2-VL. We find that all considered biases have greater effects in both uncertainties when bias-free model confidence is lower. Moreover, lower bias-free model confidence is associated with greater bias-induced underestimation of epistemic uncertainty, resulting in overconfident estimates, whereas it has no significant effect on the direction of bias effect in aleatoric uncertainty estimation. These distinct effects deepen our understanding of bias mitigation for uncertainty quantification and potentially inform the development of more advanced techniques. [1]

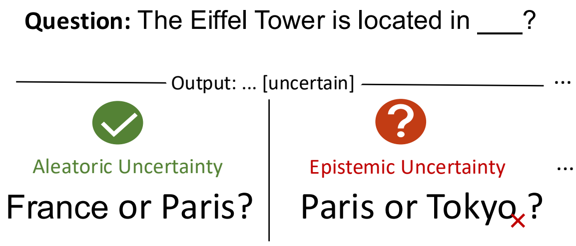


Figure 1: Uncertainty between valid answers (e.g., France and Paris) reflects aleatoric uncertainty, while uncertainty between Paris and Tokyo reflects epistemic uncertainty due to the model's lack of knowledge.

1 Introduction

Robust quantification of Large Language Models' (LLMs) confidence in their answers is vital for trust and safety in critical applications (Hendrycks et al., 2021; Rudner and Toner, 2024). Without effective confidence ranking, accurate predictions may be overlooked, while inaccurate predictions may be prioritized and lead to harmful outcomes (Geifman and El-Yaniv, 2017).

Much of the existing literature leverages uncertainty to approximate a model's confidence in its answers (Guo et al., 2017; Malinin and Gales, 2020). Model uncertainty can stem from aleatoric uncertainty, epistemic uncertainty, or both. Importantly, only epistemic uncertainty is indicative of the model's confidence, since it captures the limitations of the model's knowledge. In contrast, aleatoric uncertainty stems from the irreducible randomness of the true answer distribution and persists even if the model has perfect knowledge. As such, the true goal of "uncertainty quantification" is to quantify the epistemic uncertainty. When two predictions exhibit similar total uncertainty, the one driven by aleatoric uncertainty indicates a more knowledgeable and confident model than one dominated by epistemic uncertainty. Figure 1 illustrates this distinction through an example where the model is uncertain for different underlying reasons.

Traditional uncertainty quantification methods, however, typically estimate total uncertainty. This is because they often operate under the single-answer assumption, where aleatoric uncertainty is absent. Yet in real-world scenarios with multiple valid answers, distinguishing between the two becomes crucial.

[1]https://github.com/XinyiLiu0227/Uncertainty_Quantification_Bias

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 20297–20313
November 4-9, 2025 ©2025 Association for Computational Linguistics

In settings where each question has only one valid answer and uncertainty is thus purely epistemic, it may be intuitive that the presence of bias (Ye et al., 2024a; Yang et al., 2024; Seo et al., 2022), namely spurious features that models rely on without understanding the true semantic meanings, can lead to inaccurate uncertainty estimation based on biased generation probabilities. Therefore, mitigating bias can improve the effectiveness of uncertainty quantification based on generation probabilities (Jiang et al., 2023). However, the potential presence of aleatoric uncertainty introduces additional complexity. Bias may also reduce aleatoric uncertainty by concentrating probability mass on a single or smaller subset of valid answers. In such cases, bias may reduce the noise introduced by aleatoric uncertainty, potentially facilitating a clearer estimation of epistemic uncertainty.

We investigate whether mitigating prompt-introduced biases can enhance uncertainty quantification with the presence of aleatoric uncertainty, using GPT-4o, one of the most advanced multimodal LLMs. These biases arise from arbitrary and unavoidable choices in spurious features that do not alter the underlying semantics when using a single prompt, such as phrasing, answer position, verbalizer assignment, and image shape (Wang et al., 2023; Liu et al., 2024; Gavrikov et al., 2024; Ye et al., 2024b). Our results show that bias mitigation consistently enhances uncertainty quantification with the presence of aleatoric uncertainty, without requiring access to the internal model state. Specifically, removing text-based biases boosts AUROC (Hanley and McNeil, 1983; McDermott et al., 2024) by approximately 7%. Motivated by this, we further examine how bias affects epistemic and aleatoric uncertainty separately.

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 (Ahdritz et al., 2024; Yadkori et al., 2024) has shifted focus towards more general scenarios, where multiple distinct semantic meanings are valid (Jiang et al., 2022; Jia et al., 2024; Barandas et al., 2024). These two studies find that models are more likely to copy information from prompts under high epistemic uncertainty than under high aleatoric uncertainty, which may be interpreted as a form of confirmation bias (Nickerson, 1998; Shi et al., 2024). Therefore, we hypothesize that the impact of the

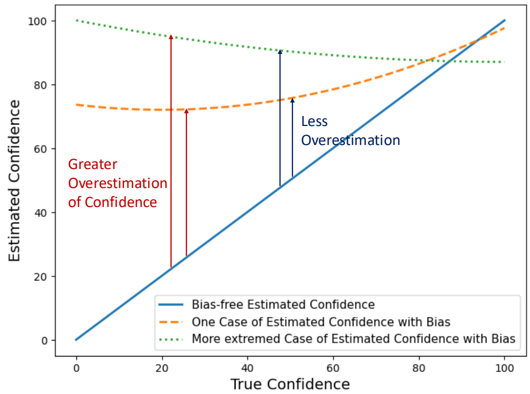


Figure 2: Systematically greater overestimation of confidence in lower-confidence instances can flatten the estimated confidence curve, undermining ranking robustness. Sometimes it even reverses the correct order.

prompt-introduced biases examined in our earlier experiments on epistemic uncertainty amplifies with lower bias-free model confidence, whereas its impact on aleatoric uncertainty remains relatively insensitive to confidence levels.

Most multi-label Natural Language Processing datasets were introduced early and are now well studied, allowing LLMs to achieve near-perfect performance with minimal uncertainty (Yadkori et al., 2024). We therefore construct visual-language datasets where LLM performance is not yet saturated, enabling analysis of both text-based and image-based prompt biases.

For both closed-source model GPT-4o (Hurst et al., 2024) and open-source model Qwen2-VL (Wang et al., 2024), our findings show that lower bias-free model confidence correlates with stronger bias effects, estimated by the absolute change in both epistemic and aleatoric uncertainty measured with and without bias. However, this correlation is notably weaker for aleatoric uncertainty than for epistemic uncertainty.

As illustrated in Figure 2, assigning more inflated confidence scores to lower-confidence instances undermines the robustness of ranking by measured confidence. Notably, this pattern has been observed in human behavior (Sulistyawati et al., 2011). In extreme cases, such distortions may even reverse the correct ranking: when the bias-free confidence in A exceeds that in B, the biased estimates incorrectly favor B over A. To better understand this directional distortion in confidence estimation, we examine how bias interacts with different sources of uncertainty. We find that bias-induced underestimation of epistemic uncertainty is greater when the model is genuinely less confident, resulting in overconfident estimates. In contrast, the bias-induced directional shift in aleatoric uncertainty estimation shows no significant association with confidence.

The distinct effects of bias on epistemic and aleatoric uncertainty deepen our understanding of bias mitigation in uncertainty quantification. This understanding may also guide the development of more advanced methods.

2 Related Work

Uncertainty Quantification with a Single Valid Answer. Traditional machine learning models treat total uncertainty as a measure of confidence when each question has a single valid answer (Hendrycks and Gimpel, 2016; Lakshminarayanan et al., 2017; Guo et al., 2017; Wang et al., 2022). 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.

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 confidence, raising concerns about the reliability of self-reported uncertainty. 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 partially overlaps with ours in textual perturbation, we extend the analysis to multi-answer scenarios that involve aleatoric uncertainty and additional prompt-introduced biases, including image-based biases. Crucially, we further examine how these biases affect the two uncertainties differently across varying confidence levels, offering a deeper understanding of the bias mitigation method.

Uncertainty Quantification with a Single Semantic Valid Answer. Prior work on LLM uncertainty with aleatoric components mainly focuses on variability in generating semantically equivalent outputs, using benchmarks such as CoQA (Reddy et al., 2019), TriviaQA (Joshi et al., 2017), and AmbigQA (Min et al., 2020).

Proposed techniques include 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), requiring additional training or model access. Semantic equivalence has proven to be effective in reducing aleatoric uncertainty from phrasing variability without access to internal model states (Kuhn et al., 2023; Farquhar et al., 2024). Research by Huang et al. (2023b) observed that sample-based methods outperform single-inference approaches.

Building on these findings, we shift focus from phrasing variation to the challenge of multiple semantically valid answers, aiming to capture the distinct characteristics of epistemic and aleatoric uncertainty.

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. They also observed that LLMs are more likely to copy input information 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 through iterative prompting.

This growing body of work underscores the need to distinguish epistemic from aleatoric uncertainty with multiple semantically valid answers. We extend this by analyzing how biases introduced by relying on a single prompt affect these two measured uncertainties across different model confidences. In addition, Yadkori et al. (2024) preselected multi-label queries with high entropy (> 0.7) from the WordNet dataset (Fellbaum, 1998), where LLMs achieve near-perfect performance. This approach results in instances with high total uncertainty but correct outputs, which may not reflect real-world data distributions. We use unfiltered datasets to better capture practical challenges.

3 The Role of Bias in Uncertainty Quantification

While bias might add noise to epistemic uncertainty estimation, it also may reduce the noise introduced by aleatoric uncertainty. We evaluate this trade-off using GPT-4o, one of the most advanced multimodal LLMs, to assess whether mitigating the prompt-introduced biases improves uncertainty quantification under aleatoric uncertainty.

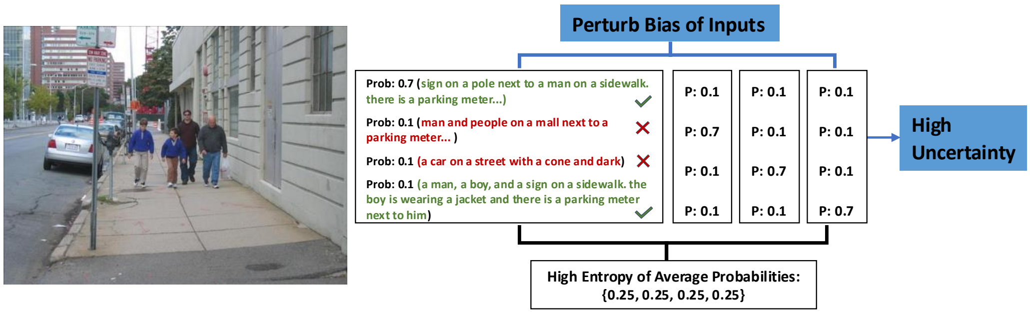


Figure 3: Perturb prompts to shuffle bias factors to estimate bias-free uncertainty.

Ahdritz et al. (2024) and Yadkori et al. (2024) both found that LLMs are more likely to copy input information under high epistemic uncertainty but not high aleatoric uncertainty. Inspired by these findings, we further analyze how these prompt-introduced biases impact each type of uncertainty estimation separately, aiming to provide deeper insight.

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. In contrast, aleatoric uncertainty stems from uncertainty among multiple valid answers and exists even with perfect world knowledge.

Building on the proven effectiveness of semantic equivalence in addressing phrasing variability, particularly the use of LLM-based Natural Language Inference (Farquhar et al., 2024), we focus on the challenge of multiple valid answers with distinct meanings. To address this, we adopt a multiple-choice format with two semantically distinct correct options and two incorrect ones. This design provides sufficient data for analysis while offering a clear conceptual framework, without introducing additional variance from semantic-equivalence resolution. For generalizing uncertainty quantification from classification to open-ended generation, please refer to Appendix B of Jiang et al. (2023).

In uncertainty quantification (see Section 3.3), ground-truth information is unavailable. However, for analyzing bias impact, we use ground-truth labels to quantify epistemic and aleatoric uncertainty separately. We estimate epistemic and aleatoric uncertainty using epistemic entropy and aleatoric entropy, respectively. We define epistemic entropy as the entropy over the probability of a correct prediction (i.e., the summed probabilities of all valid answers) and the individual probabilities of each incorrect prediction. Let i denote a potential output, and "correct" the set of valid answers:

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

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

Aleatoric entropy is defined as the entropy over the normalized distribution of correct answers:

$$\text{Aleatoric Entropy} = - \sum_{i \in \text{correct}} \frac{P(i)}{P(\text{correct})} \log \frac{P(i)}{P(\text{correct})} \tag{3}$$

Consequently, the total entropy over the full output distribution, which is commonly used to estimate model uncertainty, can be decomposed into epistemic and aleatoric entropy as follows. A detailed proof is provided in Appendix A.1.

$$\text{Entropy} = \text{Epistemic Entropy} + P(\text{correct}) \times \text{Aleatoric Entropy} \tag{4}$$

3.2 Prompt-Introduced Biases

We consider three text-based biases and three image-based biases. 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). 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 positions 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.

Although image-based biases are often reduced through image perturbations during training

(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., 2024b).

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

More details of prompts perturbation strategies to mitigate biases are provided in Appendix A.2.

3.3 Uncertainty Quantification in the Presence of Aleatoric Uncertainty

We explore bias mitigation for uncertainty quantification, aiming to estimate a model's confidence in its outputs without ground truth access by reducing prompt-introduced biases, as depicted in Figure 3.

Unlike the mutual information approach proposed by the recent work (Yadkori et al., 2024), which injects hints into prompts to measure copying behavior, our method operates in a smaller search space by directly targeting biases in default prompts, avoiding broader searches. Specifically, we address both text- and image-based biases unavoidably introduced by a single prompt, as identified in prior work (Wang et al., 2023; Liu et al., 2024; Gavrikov et al., 2024; Ye et al., 2024b).

3.4 Bias Effects on Measured Uncertainties

As many top-performing models are closed-source, understanding their behavior as observable without internal states is crucial. We examine how prompt-introduced biases affect measured epistemic and aleatoric uncertainty, offering insights that can be leveraged for both open- and closed-source models.

To assess the impact of bias, we compare entropy values from single prompts to those averaged over multiple bias-shuffled prompts (see Figure 3). Specifically, we measure: (1) bias effect as the *absolute change* in epistemic and aleatoric entropy, and (2) bias-induced underestimation as the *decrease* in entropy from the averaged distribution to the single prompt. While the averaged distribution across bias-shuffled prompts may not be entirely bias-free, it is relatively bias-reduced reference (Wang et al., 2023; Liu et al., 2024) and we refer to it as "bias-free" for convenience.

We perform two separate linear regressions to examine the relationship between bias-free confidence levels (independent variable) and each of the two bias effect measures (dependent variable).

4 Experiments

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.2.

Dataset. We use the VL_checklist (Zhao et al., 2022) and CREPE datasets (Ma et al., 2023), 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. 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., 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 adopt the AUROC metric for uncertainty quantification, following prior studies (Band et al., 2022; Kuhn et al., 2023; Lin et al., 2023; Farquhar et al., 2024). AUROC measures how well confidence scores rank instances with correct versus incorrect predictions and is robust to class imbalance (McDermott et al., 2024).

For further analysis, we use linear regression coefficients and p-values to examine how bias-free model confidence influences bias-induced changes in measured epistemic and aleatoric uncertainty. Regression coefficients indicate the direction and magnitude of this relationship: a positive coefficient suggests greater bias effects at higher confidence levels, 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 rather than one due to chance.

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

Due to budget constraints, we evaluate only two randomly selected biases for Claude-3.7-Sonnet ('claude-3-5-sonnet-20241022') on CREPE, with the results presented in Appendix A.5.

Experimental Settings. With OpenAI's closed-source LLMs now providing top-20 token probabilities, we compute prediction probabilities across all options directly for GPT-4o. When token probabilities are not exposed, a common strategy is to approximate them through sampling, which has been shown effective (Farquhar et al., 2024). Following this approach, we approximate the token probabilities for Claude-3.7-Sonnet using 10 sampled prompts.

Bias-free model confidence is estimated by summing the probabilities of correct options averaged across bias-shuffled prompts. We further extend our experiments by approximating model inconfidence using bias-free epistemic entropy (higher entropy indicates lower confidence), with the results presented in Appendix A.5.

Following Kuhn et al. (2023) and Farquhar et al. (2024), we approximate greedy decoding by using a single output generated at a very low temperature (1e-15) as the model's 'best generation' for assigning correctness labeling, using the prompt shown in Table 1. While closed-source LLMs may still exhibit variation at zero temperature, this approach remains consistent with established research.

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) for sampling from bias-shuffled prompts to ensure consistency and avoid unnecessary tuning. We run ten shuffled prompts for each type of bias, 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).

5 Results and Analysis

5.1 Uncertainty Quantification Through Bias Mitigation

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

When model confidence (self-perception) aligns with its true knowledge (Kadavath et al., 2022; Farquhar et al., 2024), it serves as a good estimate of the probability of correctness. As shown in Equation (4), the model's total uncertainty incorporates both epistemic uncertainty that indicates model confidence, and aleatoric uncertainty which does not. We use GPT-4o to evaluate the trade-off that bias mitigation introduces between these two types of uncertainty for uncertainty quantification.

Baselines. We focus on **Entropy** as our main baseline, given its strong performance in recent studies targeting closed-source LLMs (Kuhn et al., 2023; Farquhar et al., 2024; Yadkori et al., 2024). We also include two commonly used baselines: the **Prob** (probability of the prediction) and the **#Answers** (number of answers), as well as the recently proposed **Mutual Information** approach (Yadkori et al., 2024), which adopts iterative prompting to estimate confidence based on the model's tendency to copy provided hints.

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	***	**		***	ns	
	Positional	Coefficients	- 0.6098	- 0.0629	**9.69**	- 0.1571	- 0.0844	**1.86**
		P-value	***	ns		***	***	
	Label	Coefficients	- 0.3572	- 0.0911	**3.92**	0.0602	0.0757	0.80
		P-value	***	**		***	***	
	Shape	Coefficients	- 0.1679	- 0.0707	**2.37**	- 0.0664	- 0.0081	**8.20**
		P-value	***	***		***	*	
	Orientation	Coefficients	- 0.1746	- 0.0671	**2.60**	- 0.1073	- 0.0230	**4.67**
		P-value	***	***		***	ns	
	Low-level Feature	Coefficients	- 0.1466	- 0.0457	**3.21**	- 0.0493	- 0.0214	**2.30**
		P-value	***	**		***	*	
CREPE	Phrasing	Coefficients	- 0.1149	- 0.0481	**2.39**	- 0.0025	- 0.0011	**2.27**
		P-value	***	***		ns	ns	
	Positional	Coefficients	- 0.2914	- 0.1162	**2.51**	0.0192	0.0525	0.37
		P-value	***	***		ns	**	
	Label	Coefficients	- 0.1663	- 0.1147	**1.45**	0.0638	0.0407	**1.57**
		P-value	***	***		***	***	
	Shape	Coefficients	- 0.0952	- 0.0215	**4.43**	- 0.0196	- 0.0188	**1.04**
		P-value	***	*		*	*	
	Orientation	Coefficients	- 0.0797	- 0.0347	**2.30**	- 0.0320	- 0.0106	**3.02**
		P-value	***	**		**	ns	
	Low-level Feature	Coefficients	- 0.0919	- 0.0336	**2.74**	- 0.0202	- 0.0044	**4.59**
		P-value	***	**		**	ns	

Table 3: Both GPT-4o and Qwen2-VL exhibit greater bias impact at lower confidence levels, as reflected in absolute changes in both epistemic and aleatoric entropy with and without bias. This is supported by the consistent negative coefficients. Moreover, the bias impact on epistemic uncertainty correlates more strongly with confidence than on aleatoric uncertainty, as indicated by coefficient Ratio Epi./Ale.> 1 (**bolded**) and the relatively lower statistical significance of p-values for aleatoric entropy. (***p $\leq$ 0.001, **p $\leq$ 0.01, *p $\leq$ 0.05, ns=not significant p > 0.05)

To address potential variation in token probabilities under identical decoding in closed-source models, we also introduce a **Repetitive-based** baseline that averages probabilities over multiple runs of the same prompt. This allows us to examine whether performance gains stem from better probability estimation simply through repeated sampling.

Analysis. As shown in Table 2, we observe that simple Repetitive-based samplings have minimal improvement over single-inference estimations.

Bias mitigation consistently improves performance across all baselines. While no single bias mitigation method clearly outperforms the others, summing the entropy obtained from each bias removal leads to further performance gains. Similar accuracies across the ten bias-shuffled prompts shown in Appendix A.3 suggest that the improvement is not due to prompt quality differences.

Among 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 comparison, combining three image-based methods yields more modest improvement (2.07% and 4.17%, respectively), likely because image perturbation during training has already mitigated much of the image-based bias. 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 important for fairness but also critical for quantifying (epistemic) 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 4 in Appendix, a limitation shared by the #Answers baseline. Specifically, the prevalence of identical Mutual Information values, especially in low-uncertainty instances, limits its discriminative power and results in a low AUROC score. This makes it less suitable for high-stakes applications that demand a high abstention rate. In contrast, the text-based bias mitigation approaches remain robust across different thresholds.

5.2 Relationship Between Model Confidence and Bias Impact

We compute bias-free model confidence using the sum of the bias-free probabilities of correct options, which serves as the independent variable. We then examine its relationship to absolute changes in measured epistemic and aleatoric entropy, comparing outputs with and without bias. Larger change indicates stronger bias impact. Results from two models and two datasets, as shown in Table 3, reveal consistent patterns across all biases:

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

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	***	ns		*	*	
	Positional	Coefficients	- 0.7585	- 0.0499	**15.2**	- 0.1827	- 0.0722	**2.53**
		P-value	***	ns		***	*	
	Label	Coefficients	- 0.3811	- 0.0898	**4.24**	-0.0338	-0.0233	**1.45**
		P-value	***	*		ns	ns	
	Shape	Coefficients	- 0.1542	- 0.0344	**4.48**	- 0.0620	- 0.0013	**47.69**
		P-value	***	ns		***	ns	
	Orientation	Coefficients	- 0.1441	- 0.0181	**7.96**	- 0.1309	- 0.0235	**5.57**
		P-value	***	ns		***	ns	
	Low-level Feature	Coefficients	- 0.1188	- 0.0121	**9.82**	- 0.0257	- 0.0011	**23.36**
		P-value	***	ns		***	ns	
CREPE	Phrasing	Coefficients	- 0.1019	0.0184	**5.54**	- 0.0242	0.0097	**2.49**
		P-value	***	ns		***	ns	
	Positional	Coefficients	- 0.3929	- 0.0772	**5.09**	- 0.0951	0.0392	**2.43**
		P-value	***	*		***	ns	
	Label	Coefficients	- 0.2641	- 0.1082	**2.44**	- 0.0152	0.0184	0.83
		P-value	***	***		ns	ns	
	Shape	Coefficients	- 0.0580	0.0068	**8.52**	- 0.0147	- 0.0082	**1.79**
		P-value	***	ns		ns	ns	
	Orientation	Coefficients	- 0.0586	- 0.0206	**2.84**	- 0.0776	- 0.0095	**8.17**
		P-value	***	ns		***	ns	
	Low-level Feature	Coefficients	- 0.0741	- 0.0181	**4.09**	- 0.0152	- 0.0079	**1.92**
		P-value	***	ns		ns	ns	

Table 4: Both GPT-4o and Qwen2-VL exhibit greater bias-induced underestimation of epistemic uncertainty estimation 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 bias effect in aleatoric entropy, supported by mostly insignificant p-values and mixed coefficient signs. The coefficient ratio Epi./Ale. > 1 is **bolded**. (***$p \leq 0.001$, **$p \leq 0.01$, *$p \leq 0.05$, ns=not significant $p > 0.05$)

free confidence, its outputs tend to be more sensitive to bias, as evidenced by consistently negative coefficients for GPT-4o with only three exceptions in Qwen-2.

Bias impact on epistemic uncertainty estimates is more strongly correlated with model confidence than on aleatoric uncertainty estimates. This is evidenced by 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 bias impact on aleatoric uncertainty shows no significant correlation with bias-free model confidence, as indicated by large p-values ($p > 0.05$).

Similar results are obtained using bias-free epistemic entropy as the approximated model inconfidence, as shown in Appendix A.5. The extended experiments on Claude-3.7-Sonnet with two randomly selected biases also align with these findings and are presented in the same appendix.

5.3 Relationship Between Model Confidence and Bias-Induced Overconfidence

While lower model confidence leads to greater bias-induced changes, the direction of change is crucial. Greater under-confidence (i.e. overestimation of uncertainty) in lower bias-free confidence instances improves the robustness of estimated confidence ranking under estimation noise by amplifying the contrast between instances with low and high bias-free confidence. However, greater over-confidence in lower-confidence instances hurts the ranking performance of estimated confidence (see Figure 2).

Therefore, we further examine how model confidence relates to bias impact on entropy reduction, subtracting measured entropy from a single prompt from that of bias-shuffled prompts. Results from two models and two datasets, as shown in Table 4, reveal consistent patterns across all biases:

Lower model confidence is associated with greater bias-induced underestimation of epistemic entropy (i.e., overconfidence). When bias-free model confidence is lower, bias causes a larger reduction in epistemic entropy. This is evidenced by consistently negative coefficients for epistemic entropy reduction, with the majority of p-values indicating statistical significance.

Model confidence has no significant effect on the direction of bias-induced aleatoric entropy changes. This is supported by the predominance of non-significant p-values and inconsistent coefficient signs for aleatoric entropy reduction.

Using bias-free epistemic entropy to approximate model inconfidence yields similar results, as shown in Appendix A.5. The extended experiments on Claude-3.7-Sonnet with two randomly selected biases also align with these findings and are presented in the same appendix.

6 Conclusion

Removing three text-based biases and three image-based biases improves uncertainty quantification in the presence of aleatoric uncertainty, as measured by AUROC on GPT-4o. However, the improvement from image-based bias removal is smaller, likely due to existing image perturbation during training.

While uncertainty decomposes into epistemic and aleatoric components, our findings show that lower model confidence amplifies bias effects on measured uncertainties, with a greater amplification observed on epistemic than on aleatoric uncertainty. Moreover, while model confidence does not significantly affect the direction of bias-induced changes in measured aleatoric uncertainty, lower model confidence is associated with greater bias-induced underestimation of epistemic uncertainty (i.e. overconfidence).

Future work may leverage the distinct bias effects on these two types of uncertainty across varying confidence levels to develop more advanced techniques for disentangling them. In addition, our analysis could be extended to pure-text datasets once more challenging or less exposed multi-label benchmarks in LLMs pretraining become available.

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*.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30.

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.

James A Hanley and Barbara J McNeil. 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3):839–843.

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.

Andrey Malinin and Mark Gales. 2020. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*.

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*.

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*.

Tim GJ Rudner and Helen Toner. 2024. Key concepts in ai safety: Reliable uncertainty quantification in machine learning. *CSET Issue Briefs*.

Seonguk Seo, Joon-Young Lee, and Bohyung Han. 2022. Information-theoretic bias reduction via causal view of spurious correlation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2180–2188.

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*.

Yu Yang, Eric Gan, Gintare Karolina Dziugaite, and Baharan Mirzasoleiman. 2024. Identifying spurious biases early in training through the lens of simplicity bias. In *International conference on artificial intelligence and statistics*, pages 2953–2961. PMLR.

Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. 2024a. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*.

Wenqian Ye, Guangtao Zheng, Yunsheng Ma, Xu Cao, Bolin Lai, James M Rehg, and Aidong Zhang. 2024b. 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 Mathematical Proof of Equation (4)

The entropy over the full distribution is

$$\text{Entropy} = -\sum_{i} P(i) \log P(i) \tag{5}$$

$$= -\sum_{i \in \text{correct}} P(i) \log P(i) - \sum_{i \notin \text{correct}} P(i) \log P(i) \tag{6}$$

The aleatoric entropy is defined as the entropy over the conditional distribution among correct options:

$$\text{Aleatoric Entropy} = -\sum_{i \in \text{correct}} \frac{P(i)}{P(\text{correct})} \log \left(\frac{P(i)}{P(\text{correct})} \right), \tag{7}$$

where $P(\text{correct}) = \sum_{i \in \text{correct}} P(i)$.

Multiply both sides by $P(\text{correct})$ to get:

$$P(\text{correct}) \cdot \text{Aleatoric Entropy}$$

$$= -\sum_{i \in \text{correct}} P(i) \log \left(\frac{P(i)}{P(\text{correct})} \right) \tag{8}$$

$$= -\sum_{i \in \text{correct}} P(i) \log P(i) + \sum_{i \in \text{correct}} P(i) \log P(\text{correct}) \tag{9}$$

$$= -\sum_{i \in \text{correct}} P(i) \log P(i) + P(\text{correct}) \log P(\text{correct}) \tag{10}$$

Substitute this back into the total entropy:

$$\text{Entropy}$$

$$= -\sum_{i \in \text{correct}} P(i) \log P(i) - \sum_{i \notin \text{correct}} P(i) \log P(i) \tag{11}$$

$$= \left[-\sum_{i \in \text{correct}} P(i) \log P(i) + P(\text{correct}) \log P(\text{correct}) \right] \tag{12}$$

$$- P(\text{correct}) \log P(\text{correct}) - \sum_{i \notin \text{correct}} P(i) \log P(i) \tag{13}$$

$$= P(\text{correct}) \cdot \text{Aleatoric Entropy} + \tag{14}$$

$$\underbrace{\left[-P(\text{correct}) \log P(\text{correct}) - \sum_{i \notin \text{correct}} P(i) \log P(i) \right]}_{\text{Epistemic Entropy}} \tag{15}$$

$$= P(\text{correct}) \cdot \text{Aleatoric Entropy} + \text{Epistemic Entropy} \tag{16}$$

A.2 Details of Prompt Design

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

Phrasing Bias. We utilize GPT-4o to help paraphrase our default prompt shown in Table 1 while keeping the options unchanged. Table 11 lists all the rephrased prompts used in our experiments to perturb bias related to phrasing.

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.

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.3 Accuracy Comparison Between Default Prompt and Single Perturbed Prompt

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.

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.

A.4 Details of Uncertainty Quantification Performance

Figure 4 shows the ROC curves for text-based bias mitigation and baselines, providing more details of their performance across different threshold regions.

A.5 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.

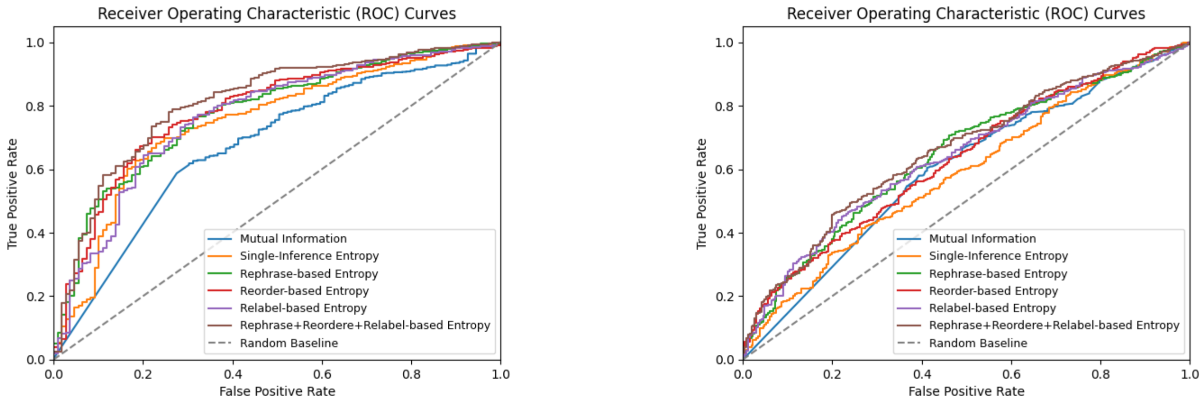
Relative Magnitudes of Epistemic and Aleatoric Entropy. Table 7 shows that the magnitude of aleatoric entropy is comparable to that of epistemic entropy.

Bias-free Epistemic Entropy as Model Confidence 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 bias-induced changes in measured epistemic uncertainty on GPT-4o. Qwen2-VL follows the same pattern, with exceptions for Label bias. 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 with the same two exceptions, and nonsignificant p-values. These results are detailed in Table 9.

Lower model confidence is more strongly associated with greater underestimation of measured epistemic uncertainty, whereas it has no significant effect on the direction of bias-induced 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 for the same two Qwen2-VL cases.

Extension to Claude-3.7-Sonnet. Due to budget constraints, we randomly selected two biases for extended experiments on Claude-3.7-Sonnet with CREPE. Table 8 presents the results, which align with the findings observed on other models. More specifically, stronger bias effects on epistemic uncertainty estimation are associated with lower bias-free model confidence, as shown by negative and statistically significant coefficients. In contrast, the relationship for aleatoric uncertainty is weaker, as indicated by non-significant p-values and coefficients with mixed signs.



(a) ROC curves on VL_Checklist.

(b) ROC curves on CREPE.

Figure 4: 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.

Bias	Metrics	Absolute Change			Underestimation		
		Epistemic	Aleatoric	Ratio Epi./Ale.	Epistemic	Aleatoric	Ratio Epi./Ale.
Phrasing	Coefficients	-0.1579	0.0111	**14.23**	-0.1238	- -0.0184	**6.73**
	P-value	***	ns		***	ns	
Label	Coefficients	-0.5420	0.2012	**2.69**	-0.5681	0.2094	**2.71**
	P-value	***	**		***	**	

Table 8: This table reports the impact of phrasing and label bias on CREPE with Claude-3.7-Sonnet. Stronger bias effects on epistemic uncertainty estimation correspond to lower bias-free model confidence, as indicated by the negative coefficients with significant p-values. For aleatoric uncertainty, this relationship is weaker, reflected in the non-significant p-values and mixed coefficient signs.

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	***	***		***	ns	
	Positional	Coefficients	0.4719	0.0379	**12.45**	0.1326	- 0.0654	**2.03**
		P-value	***	ns		***	***	
	Label	Coefficients	0.2999	0.0575	**5.22**	-0.0255	-0.0828	0.31
		P-value	***	**		ns	***	
	Shape	Coefficients	0.2023	0.0822	**2.46**	0.0644	0.0144	**4.47**
		P-value	***	***		***	ns	
	Orientation	Coefficients	0.2126	0.0876	**2.43**	0.0916	0.0316	**2.90**
		P-value	***	***		***	**	
	Low-level Feature	Coefficients	0.1851	0.0536	**3.45**	0.0476	0.0205	**2.32**
		P-value	***	***		***	**	
CREPE	Phrasing	Coefficients	0.1825	0.0558	**3.27**	0.0067	- 0.0020	**3.30**
		P-value	***	***		*	ns	
	Positional	Coefficients	0.3344	0.0476	**7.03**	0.0139	-0.0508	0.27
		P-value	***	*		ns	***	
	Label	Coefficients	0.2129	0.0721	**2.95**	- 0.0744	- 0.0676	**1.10**
		P-value	***	***		***	***	
	Shape	Coefficients	0.1694	0.0423	**4.00**	0.0173	- 0.0029	**5.97**
		P-value	***	***		*	ns	
	Orientation	Coefficients	0.1723	0.0689	**2.50**	0.0227	- 0.0084	**2.70**
		P-value	***	***		*	ns	
	Low-level Feature	Coefficients	0.1565	0.0517	**3.03**	0.0184	0.0064	**2.88**
		P-value	***	***		***	ns	

Table 9: Both GPT-4o and Qwen2-VL exhibit greater bias impact at lower confidence levels, as reflected in absolute changes in both epistemic and aleatoric entropy with and without bias. This is supported by the consistent positive coefficients. Moreover, the bias impact on epistemic uncertainty correlates more strongly with confidence than on aleatoric uncertainty, as indicated by coefficient Ratio Epi./Ale.> 1 (**bolded**) and the relatively lower statistical significance of p-values for aleatoric entropy. (***p $\leq$ 0.001, **p $\leq$ 0.01, *p $\leq$ 0.05, ns=not significant p > 0.05)

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	***	ns		***	ns	
	Positional	Coefficients	0.4874	0.0330	**14.8**	0.1311	- 0.0449	**2.92**
		P-value	***	ns		***	*	
	Label	Coefficients	0.2942	0.0486	**6.05**	0.0267	-0.0070	**3.81**
		P-value	***	ns		ns	ns	
	Shape	Coefficients	0.1277	0.0438	**2.92**	0.0387	- 0.0033	**47.69**
		P-value	***	*		***	ns	
	Orientation	Coefficients	0.1590	0.0289	**5.50**	0.0883	0.0108	**8.18**
		P-value	***	ns		***	ns	
	Low-level Feature	Coefficients	0.1219	0.0192	**6.35**	0.0272	- 0.0080	**3.4**
		P-value	***	ns		***	ns	
CREPE	Phrasing	Coefficients	0.1577	- 0.008	**197.13**	0.0116	0.0070	**1.66**
		P-value	***	ns		*	ns	
	Positional	Coefficients	0.4043	0.0327	**12.36**	0.0975	- 0.0433	**2.25**
		P-value	***	ns		***	*	
	Label	Coefficients	0.2890	0.0863	**3.35**	0.0171	- 0.0419	0.41
		P-value	***	***		ns	**	
	Shape	Coefficients	0.1425	0.0108	**13.19**	0.0282	- 0.0060	**4.70**
		P-value	***	ns		**	ns	
	Orientation	Coefficients	0.1478	0.0579	**2.55**	0.0738	- 0.0022	**33.55**
		P-value	***	***		***	ns	
	Low-level Feature	Coefficients	0.1299	- 0.0083	**15.65**	0.0186	0.0033	**5.64**
		P-value	***	ns		**	ns	

Table 10: 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. The coefficient ratio Epi./Ale. > 1 is **bolded**. (***$p \leq 0.001$, **$p \leq 0.01$, *$p \leq 0.05$, ns=not significant $p > 0.05$)

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