# [TINY] VISION LANGUAGE MODELS CAN IMPLICITLY QUANTIFY ALEATORIC UNCERTAINTY.

**Xi Wang**
Johns Hopkins University
xwang457@cs.jhu.edu

**Eric Nalisnick**
Johns Hopkins University
nalisnick@jhu.edu

## ABSTRACT

Recent advances in vision language models (VLMs), such as GPT-4o, have revolutionized visual reasoning by enabling zero-shot task completion through natural language instructions. In this paper, we study VLMs' ability to detect input ambiguities, i.e., aleatoric uncertainty. Our key finding is that VLMs can effectively identify ambiguous inputs by simply including an instruction to output "Unknown" when uncertain. Through experiments on corrupted ImageNet and "OOD" detection tasks, we demonstrate that VLMs successfully reject uncertain inputs while maintaining high accuracy on confident predictions. This capability for implicitly quantifying uncertainty emerges without additional training or in-context learning, distinguishing VLMs from traditional vision models that often produce overconfident predictions on ambiguous inputs.

## 1 INTRODUCTION

The ability to quantify input uncertainty is crucial for the reliability of machine learning systems. An ideal system should be capable of signaling uncertainty in two key scenarios (Hüllermeier & Waegeman, 2021): a) when the inherent ambiguity of the input makes it impossible to provide a meaningful answer (aleatoric uncertainty), and b) when the input exceeds the model's capabilities (epistemic uncertainty).

Traditional vision models often struggle with uncertainty quantification (Nixon et al., 2019; Guo et al., 2017; Minderer et al., 2021; Gal & Ghahramani, 2016; Ovadia et al., 2019), largely due to their training regime: These models are typically small-scale and trained on specific, highly curated datasets for single tasks. Consequently, when encountering test samples with features absent from their training data, they tend to produce incorrect predictions with high confidence, potentially compromising downstream decision-making processes.

Vision language models (Liu et al., 2023; Gao et al., 2023; Liu et al., 2024, VLMs), represent a paradigm shift in visual reasoning. Unlike their traditional counterparts, VLMs undergo self-supervised pre-training on multi-billion-sample datasets. This extensive pre-training enables them to perform diverse tasks in a zero-shot manner, requiring only an image and a natural language task description at inference time. However, despite their increasing adoption in real-world applications, their capability to model uncertainty remains poorly understood.

As such, in this paper, we study whether the latest vision language models (VLMs), e.g. GPT-4o, can effectively capture *aleatoric uncertainty*. Our key finding is that both commercial and open-source VLMs can identify aleatoric uncertainty by simply including *one additional line of prompt* that allows the model to respond with `Unknown` when facing ambiguous inputs, which confirms that VLMs are capable of quantifying aleatoric uncertainty *implicitly* using natural language (in contrast to explicit quantification through outputting a numerical value). This ability emerges without requiring any additional fine-tuning process or few-shot examples. This differs VLMs from classic vision models, which require alternative training routines or objectives to avoid producing overconfident but incorrect predictions when faced with uncertain inputs.

## 2 EVALUATION TASK AND METHODS

**Evaluation task** To evaluate the ability of VLMs to model aleatoric uncertainty, we selected a subset of the validation set of ImageNet (Deng et al., 2009) that overlaps with CIFAR-

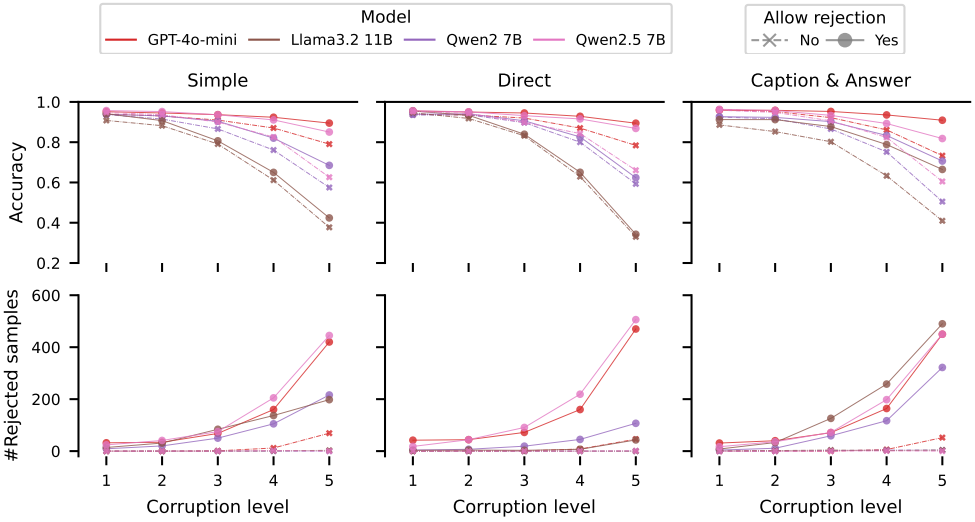


Figure 1: **VLMs can detect noisy inputs and reject to answer them.** Evaluated on the same 1,000 samples subset of Gaussian noise-corrupted ImageNet under various corruption intensities (x-axis) and instructional prompt (columns), allowing rejection option allows models to maintain high accuracy across increasing corruption levels (top row, solid v.s. dashed lines). The improvement stems from selectively rejecting more ambiguous samples as noise increases (bottom row), demonstrating VLMs' ability to recognize uncertainty. Notably, Llama3.2 and Qwen2 show the strongest improvement with Caption & Answer prompting but minimal gains with Direct prompting (second column, brown and purple lines), suggesting these models require *explicit* textual descriptions to assess uncertainty.

10 (Krizhevsky et al., 2009) categories and prompt the model to classify the image into one of the {`airplane`,...,`truck`} (i.e. CIFAR-10 classes). Then, instead of directly feeding the VLMs with standard clean images, we tweak the inputs with the following two types of ambiguous inputs

- **Gaussian Noise-Corrupted ImageNet** We evaluate VLMs using ImageNet-C's (Hendrycks & Dietterich, 2019) Gaussian noise corruption at various intensity levels. When images become visually indistinguishable due to high noise levels, we expect models to signal uncertainty rather than attempt classification.
- **"Out-of-Distribution" ("OOD") categories** We test the models using images from categories outside the CIFAR-10 classification set. In these cases, where none of the available classification options are applicable, we expect models to decline to provide an answer rather than force-fitting an inappropriate category.

Note that these two tasks are traditionally used to evaluate the out-of-distribution (OOD) robustness of models trained on clean classification datasets. Specifically, they assess OOD generalization (Liu et al., 2021) and detection (Hendrycks et al., 2018), which relate to epistemic uncertainty. However, we argue for a different interpretation in the context of VLMs. Given that VLMs are exposed to a wide range of noisy images and diverse concepts during their pre-training phase, these inputs are unlikely to be truly out-of-distribution for them. Therefore, we consider these tasks as evaluations of aleatoric uncertainty rather than epistemic uncertainty.

**Uncertainty quantification through rejection option** Given an input, classic vision models provide a numerical predictive distribution vector over each candidate option, and one can utilize certain statistics from the vector, such as maximum value, entropy, or variance, to estimate the amount of uncertainty, however, this is not the case for VLMs, which typically provide an answer in the format of free-form language. As such, to allow the VLM to express its uncertainty, we ask the VLMs to output `Unknown` when it is unsure about the input, i.e. either because the input is too noisy to read or when it is unclear whether the input belongs to the provided CIFAR-10 categories. Notice that we choose not to prompt the VLM to output a numerical confidence score for measuring uncertainty, in that it has been shown to be challenging for LLM to output a well-calibrated numerical score (Xiong et al., 2023). To be more specific, to allow the VLM to say `Unknown`, we include the following prompts in the inputs for the two tasks considered:

| Model | Simple | | Direct | | Caption & Answer | |
|---|---|---|---|---|---|---|
| | **Precision** ↑ | **Recall** ↑ | **Precision** ↑ | **Recall** ↑ | **Precision** ↑ | **Recall** ↑ |
| GPT-4o-mini | 0.964 | 0.974 | 0.974 | 0.975 | 0.988 | 0.974 |
| Llama 3.2 | 0.991 | 0.718 | 0.994 | 0.692 | 0.994 | 0.897 |
| Qwen 2 | 0.964 | 0.757 | 0.994 | 0.730 | 0.992 | 0.917 |
| Qwen 2.5 | 0.977 | 0.965 | 0.991 | 0.968 | 0.992 | 0.972 |

Table 1: **VLMs can detect "OOD" inputs and reject to answer them.** When classifying inputs into CIFAR-10 categories, but presented with images outside the 10 classes ("OOD" inputs), VLMs successfully reject them, shown by the near-perfect recall rate while maintaining high precision, indicating that the models are not over-refusing. GPT-4o-mini and Qwen 2.5 maintain strong recall rates regardless of prompting method, while Llama 3.2 and Qwen 2 show significantly improved recall only with Caption & Answer prompting. This pattern mirrors the findings in Fig. 1, suggesting these models require explicit textual descriptions to effectively assess visual uncertainty.

```
...if you find an image very ambiguous and cannot confidently classify it
, return "unknown" as the label...
...if the image does not clearly belong to any of the 10 classes provided
, classify it as "unknown"...
```

## 3 Experiments

**Choice of models and decoding methods** Four VLMs are considered for our experiments Llama 3.2 (Dubey et al., 2024, 11B), Qwen 2 (Yang et al., 2024, 7B), Qwen 2.5 (Team, 2024, 7B), and GPT-4o-mini (Achiam et al., 2023). For all models, we included only one image in a query at a time and we used greedy decoding for generating answers.

**Instructional prompt** In addition to instructional prompt that allows the model to say `Unknown`, we additionally include prompts that tell the model to generate the answer in a certain way. To be more specific, we considered three regimes,:

- **Simple** Simple and standard way: we prompt the model to provide step-by-step reasoning (Wei et al., 2022) and then provide a classification answer.
- **Direct** We prompt the model to *only output the classification answer*, which prohibits the model from generating explicit intermediate verbal reasoning steps including image caption.
- **Caption & answer** We explicitly prompt the model to always first caption the image, then answer the question using both the caption and the image.

**Evaluation metrics** For ImageNet-C classification, we checked the improvement of accuracy when the rejection option is enabled, ideally, when the rejection option is enabled, we expect samples kept unrejected to have high accuracy, implying that the model only attempts to classify images it is certain about. For detecting "OOD" categories, we used standard binary classification metrics, where we consider the true "OOD" samples out of CIFAR-10 categories as positive samples and samples inside as negative samples. Then we compute precision, which measures the portion of true "OOD" samples in all reported "OOD" samples, and **recall**, which measures the portion of "OOD" samples correctly detected by the model.

**Results** On ImageNet-C (Fig. 1), models show high accuracy across corruption levels by effectively rejecting ambiguous inputs. Similarly, for "OOD" detection (Table. 1), VLMs achieve near-perfect precision and recall, successfully identifying unclassifiable inputs without over-rejecting valid ones. Notably, Llama3.2 and Qwen2 show a distinct pattern across both tasks: they perform poorly with Direct prompting but improve significantly when first generating image captions (Caption & Answer), suggesting these models require an explicit textual intermediate step to assess visual uncertainty.

## 4 Future work

In the current experiments, we only considered performing uncertainty quantification implicitly by prompting the VLMs to output `Unknown`. Future work could look into whether one can adopt LLM uncertainty quantification techniques that provide a *numerical score*, such as `[IDK]` token (Cohen et al., 2024) or semantic entropy (Kuhn et al., 2023; Kossen et al., 2024), on VLMs.

## REFERENCES

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

Roi Cohen, Konstantin Dobler, Eden Biran, and Gerard de Melo. I don't know: Explicit modeling of uncertainty with an [idk] token. *arXiv preprint arXiv:2412.06676*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

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

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

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927*, 2024.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024.

Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34:15682–15694, 2021.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR Workshops*, 2019.

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

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

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

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.