Auditing the Performance and Calibration of Multi-Modal Large Language Models

Brendan Kennedy[1] **Lauren Phillips**[1] **Sai Munikoti**[1]

Sameera Horawalavithana[1] **Ian Stewart**[1] **Karl Pazdernik**[1,2]

[1] Pacific Northwest National Laboratory

[2] Department of Statistics, North Carolina State University

Abstract

The impressive accuracy scores of Multi-modal Large Language Models (MLLMs) on visual multiple-choice question answering (MCQA) tasks only begins to measure their readiness for sensitive domains such as medicine, scientific research, and multi-modal analytics. Here, we conduct an analysis using uncertainty quantification (UQ) methods for text generation to probe the robustness and calibration underlying the strong performance of MLLMs on image QA benchmarks. Among several findings, we show that model calibration shifts drastically when comparing UQ metrics in the classification versus the open-ended, generative setting increasingly employed by MLLMs. Based on our analysis, we suggest that the path to robust, deployable MLLMs requires not only achieving high accuracy on benchmarks, but also improving performance and calibration on challenging, open-ended tasks across the multi-modal spectrum.

1 Introduction

Multi-modal Large Language Models (MLLMs) offer the potential for breakthrough progress in information technology (Tian et al., 2024), science (Horawalavithana et al., 2024; Birhane et al., 2023), education (Bewersdorff et al., 2024; Kasneci et al., 2023), and medicine (Panagoulias et al., 2024; Huang et al., 2023). However, achieving this promise doesn't just require high accuracy on benchmarks, but also requires increased deep understanding of model failures and limitations. One area of research that could help to solve this need is uncertainty quantification (UQ; Xia et al., 2025). UQ methods provide measures of model uncertainty which enables model developers and end users to more accurately predict when a model should or should not be trusted, and can even be used to predict when a model is hallucinating (Farquhar et al., 2024).

UQ is a serious challenge for both LLMs and MLLMs, as such models' large size and complex, generative outputs prevent the employment of many traditional UQ strategies that are designed for smaller, non-generative paradigms, such as conformal prediction sets (Vovk et al., 2005; Shafer and Vovk, 2008), neural network dropout methods (Gal and Ghahramani, 2016), and Bayesian Neural Networks (Neal, 1992). To address this challenge, recent work has been proposed to quantify uncertainty for text generation through post hoc, sampling-based strategies, characterizing a model's response to an input according to the dispersion observed in its outputs over multiple sampled or perturbed generations (e.g., Kuhn et al., 2023). However, such methods have yet to be applied to multi-modal models, the performance of which is currently assessed solely through multiple-choice question answering benchmarks.

In this work, we conduct a comparative analysis of uncertainty in closed (i.e., multiple choice) versus open-ended question-answering. We investigate whether our estimates of model calibration — the extent to which UQ measurements align with model accuracy or performance in a generative task — change when shifting from UQ in the classification setting to UQ in the generative setting. Prior work performing UQ on MLLMs are restricted to the classification setting (i.e., MCQA

Workshop "Towards Trustworthy Predictions: Theory and Applications of Calibration for Modern AI" at AISTATS 2026, Tangier, Morocco.

tasks), enabling the use of traditional UQ methods such as conformal modeling (Kostumov et al., 2024; Ren et al., 2023). However, measuring a model's uncertainty, and thereby its calibration, in the classification setting ignores the additional uncertainty that comes with generating text, a task with complexity orders of magnitude higher. Contrasting with prior work, here we apply semantic entropy (Kuhn et al., 2023) to MLLMs for the first time, allowing a direct comparison between UQ in the classification setting and in the generative setting. We observe that not only are these UQ metrics poorly correlated, but models become less calibrated with respect to generative UQ.

2 Related Work

2.1 Uncertainty and LLMs

The majority of methods proposed for UQ in text generation, as well as emerging work in confidence calibration of LLMs, function by sampling multiple times from an LLM and characterizing the variance observed in the set of generations. For example, Gao et al. (2024) introduced a perturbation-based approach that applied semantics-preserving alterations to prompts while measuring the consistency of outputs. Similarly, Kuhn et al. (2023) proposed semantic entropy, which performed clustering over the output generations using a pre-trained natural language inference model, which predicts whether two texts are semantically entailed, contradictory, or neutral. Entropy is calculated over the aggregate token probabilities within each cluster. Aichberger et al. (2024) built on the idea of semantic entropy by developing a sampling approach that produces maximally (semantically) diverse samples.

A related paradigm that also samples multiple times from a model is conformal language modeling. Quach et al. (2024) proposed an algorithm to generate a conformal set of responses from a model with probabilistic guarantees on coverage of these generations. Ulmer et al. (2024) built on this work further with "non-exchangeable" conformal sets, using a nearest neighbor technique to ensure statistical guarantees even when i.i.d. assumptions are violated. In a related space, Gui et al. (2024) proposed conformal alignment, an adaptation of the conformal framework to alignment with reference data.

2.2 UQ Evaluations for Multi-modal Large Language Models

Recent work has begun to address the fact that UQ has yet to receive much attention in the multi-modal space, but has largely focused on classification tasks such as MCQA. Chandu et al. (2024) introduced a new benchmark for uncertainty of LVLMs that systematically creates unanswerable or ambiguous questions. Their work showed that LVLMs are generally not able to recognize when they are unable to answer a question. In a related vein, Kostumov et al. (2024) considered UQ in MCQA using conformal predictions, again finding a generally poor level of calibration in LVLMs. Similarly, Groot and Valdenegro-Toro (2024) took the idea of "verbalizing uncertainty" — that is, using prompting to ask the model to voice it's certainty in its prediction — and applied it to several textual and visual classification tasks, again finding poor calibration and model overconfidence.

3 Methods

3.1 Uncertainty Quantification Methods

3.1.1 Classification uncertainty

Classification-based UQ methods rely on access to class probabilities. While this is straightforward for classification models, generative models such as LLMs are increasingly applied to classification tasks. In prior works, the answer is extracted from the generation, for example by searching for the text "The answer is B." (e.g., Liu et al., 2024) or by assuming the first generated token is the answer (e.g., Kostumov et al., 2024). Thus, the process of extracting class probabilities for a model's uncertainty estimate is not straight-forward. We measure uncertainty in the classification setting by first extracting and normalizing conditional token probabilities from the language model. To extract classification probabilities, we extract the probabilities of the "answer token" in the MCQA setting. This differs from previous strategies in MLLM UQ (Kostumov et al., 2024), which applied the standard approach of extracting answer token probabilities at the *first* position while generating. We use regular expressions to identify the span of tokens containing the answer (e.g., "The answer is A", "(B)", or "C"), extract the model's logit output at the position of the output token, and normalize the raw logits from each answer token in order to obtain class probabilities.

3.1.2 Generative uncertainty

In our analyses, we apply the semantic entropy (SE) method from Kuhn et al. (2023) to quantify uncertainty of MLLMs on text generation tasks. At a high level, SE samples generations from a model multiple times, clusters the generated texts, and computes the entropy over members of clusters. Intuitively, a model that generates responses that can be binned into a single, homogeneous cluster can be seen as having high

certainty — i.e., it generates the same, or similar, responses — while a model that generates responses that are assigned randomly to multiple clusters has lower certainty. An advantage of SE is that Farquhar et al. (2024) extended the approach to be model-agnostic, allowing it to be applied to a variety of text generation models, including proprietary models.

The original formulation of SE estimates the likelihood that a sequence generated by an LLM belong to a given class (semantic cluster). Lower values of SE (with the smallest possible value of 0) indicate that a high amount of probability mass is assigned to a small set of clusters, while higher values (with the maximum being $log(k)$, for k generated samples) indicate that probability is distributed evenly over clusters of similar size. In Farquhar et al. (2024), for the set of clusters C and a generated sequence $\mathbf{x}$, SE is estimated as:

$$SE(x) \approx -\sum_{i=1}^{|C|} P(C_i|\mathbf{x}) \log P(C_i|\mathbf{x}) \tag{1}$$

In the original work, clustering for SE is done using a pre-trained Natural Language Inference (NLI) model. The NLI task involves taking two sentences and predicting whether or not the first sentence entails, or implies the correctness of, the second sentence, if they are contradictory, or if there is no relation (neutral). To judge that two model responses t_1 and t_2 are in the same cluster, the output of the NLI model must show that t_1 entails t_2 and vice versa.

We adopt a simplification of SE, which we refer to as "Cluster Entropy" (CE), which extends SE to black-box models where probabilities are not accessible. Farquhar et al. (2024) originally introduced this modification as "discrete" SE, disregarding the text probabilities altogether, approximating $P(C_i|\mathbf{x})$ to be the proportion of all the sampled answers which belong to cluster C_i. They report almost identical performance between CE and SE, and we also find this to be the case. To keep our evaluation consistent between open-source and proprietary models, we utilize CE for all of our experiments.

4 Results

Here, we directly contrast model accuracy and uncertainty calibration in MCQA versus open-ended question answering. To make this comparison, we take the visual MCQA dataset ScienceQA (Saikh et al., 2022) and create an open-ended version, $\text{ScienceQA}_{\text{Open}}$ by removing the list of options from all prompts. In doing so, we are able to directly compare the differences in performance and uncertainty calibration between different task formulations of the same data domain.

4.1 Implementation Details

We experiment with the 7-billion parameter LLaVA 1.5 vision language model and the ScienceQA (Lu et al., 2022) dataset. ScienceQA is a common dataset for multi-modal evaluation because it tests a wide breadth of domain knowledge as well as requires reasoning to be performed over both text and image data. We adapt each question in ScienceQA to produce a parallel open ended question (excluding certain question formats that cannot be answered in the open-ended setting; see Appendix for details).

We compute uncertainty separately for classification and generation in both ScienceQA tasks. For classification, we use the normalized token probability described in Section 3.1.1 as the confidence metric in the predicted class (i.e., the class whose letter had the highest assigned probability). For generation, we compute cluster entropy (CE) according to Section 3.1, and perform nucleus sampling 10 times from the model with temperature set to 1.0, following Farquhar et al. (2024). To form text clusters, we experimented with multiple strategies including the original semantic entailment approach, but present results using a textual similarity score, Rouge-L (see Appendix for details).

We followed Kuhn et al. (2023) in using Rouge-L to grade generation accuracy, with a Rouge-L score of greater than 0.6 being considered a positive generation. Accuracy in the classification case was computed between the correct option and the most likely token among the answer tokens (i.e., the set of characters 'A,' 'B,' 'C,' ...).

To compare the two metrics' calibration, we first calculate the CE for the model's sampled outputs or the token probability for the classification score. We create the ROC by varying the threshold at which we predict the model to be accurate and plotting how this affects the true and false positive rates. Higher AUROC scores indicate that the metric is predictive of model accuracy, with an AUROC of 0.5 indicating that the predictor has no value as a predictor of the target class.

4.2 Findings

Below, we report on the performance and calibration of LLaVA on ScienceQA across classification and generative formulations. In Table 1, we report accuracy on ScienceQA in the classification (ACC_{class}) and generation ($\text{ACC}_{generation}$) setting, separated by the number of options in the original question. Throughout our analysis, one of the most striking observations has

Table 1: Accuracy metrics for assessing LLaVA on ScienceQA. 90% bootstrapped confidence intervals are shown in brackets.

# Options	ACC_{class}	$\text{ACC}_{generation}$
2 (n =651)	0.69 [0.66, 0.72]	0.26 [0.23, 0.29]
3 (n =422)	0.61 [0.57, 0.65]	0.35 [0.31, 0.39]
4 (n =748)	0.63 [0.60, 0.60]	0.53 [0.50, 0.56]
5 (n =44)	0.07 [0.02, 0.14]	0.02 [0, 0.07]
n =1865	0.63 [0.62, 0.66]	0.38 [0.36, 0.40]

Table 2: Calibration Metrics for LLaVA on ScienceQA, showing misalignment between direct classification UQ via probability and generation-based UQ (via CE).

# Options	ECE	AUROC_{prob}	AUROC_{CE}
2 ($n = 651$)	0.16	0.75	0.46
3 ($n = 422$)	0.18	0.69	0.53
4 ($n = 748$)	0.18	0.78	0.74
5 ($n = 44$)	0.45	0.50	0.27
$n = 1865$	0.18	0.76	0.61

to do with the discrepancy in performance and model calibration based on the number of options in a question. This finding is in line with previous work showing fragility of multiple-choice LLMs due to arbitrary ordering of answers (Zheng et al., 2023; Pezeshkpour and Hruschka, 2024). Because this effect was previously unreported in prior evaluations and affects results so significantly, we group our findings according to the number of options. We reproduce the accuracy of the LLaVA-7b model (Liu et al., 2024); however, we find that MCQA accuracy varies significantly between the number of question options, ranging from 69% (2 options) to 7% (5).

Table 2 shows our main result, a comparison of calibration of traditional probability-based UQ and generative UQ. The expected calibration error (ECE) is first shown, which gives an indication of the calibration of class probabilities (drawn from the model's softmax probability given to the answer token in the multiple-choice classification setting) with respect to classification accuracy. ECE was calculated between the binary accuracy indicator and the class token probability using 15 bins. Calibration with respect to the MCQA classification accuracy — in other words, whether we can predict the model's correctness on the original MCQA task based on the respective UQ metric (class token probability or CE), conveyed using the Area Under the Receiver Operating Characteristic (AUROC) curve, shows a clear discrepancy. When comparing generative calibration (AUROC_{prob}) with classification calibration (AUROC_{CE}), there is a clear trend, dropping significantly overall and severely when there are two or three options versus four.

5 Conclusion

UQ for MLLMs has recently begun to be studied with greater attention and rigor, with works such as Kostumov et al. (2024) and Chandu et al. (2024) representing important progress in this direction. We add to this growing literature by extending this work to black-box MLLMs on generative tasks. Our findings support the notion that MLLM evaluation should go beyond solely benchmarking accuracy on MCQA problems, instead looking at both performance and calibration in open-ended generation tasks. Multiple choice evaluation frameworks possibly hide model biases (e.g., with respect to choosing option answer tokens, or performance discrepancies across the number of options in the original question), and unintentionally mask overfitting and poor generalization across tasks. We find that the accuracy of MLLMs on public benchmark datasets does not necessarily align with their capability to serve as oracles of visual understanding and reasoning. Furthermore, our findings cast doubt on the direct translation of benchmark results on vision QA datasets to in-practice performance of models on real applications, such as conversational visual agents. While benchmarks are a highly useful tool for comparing MLLMs across a wide array of subject areas and types of visual reasoning, accounting for model performance and calibration in the open-ended setting is essential to contextualizing models' success on such benchmarks.

While our experiments shed light on the ways in which MLLMs still struggle to perform and the areas in which they are poorly calibrated, our findings are limited by the scope of models, data, and UQ methods that we use. In particular, future work should investigate the dynamics of uncertainty across generative and classification tasks using recent proposed works in LLM UQ, such as Park et al. (2026), which proposed technique that adopts an efficient approach to approximating output variance via disagreement among "draft models," which are light-weight models that propose candidate tokens. Overall, future work should broaden and extend uncertainty-focused, in-depth evaluations in the multi-modal space.

6 Acknowledgements

This work was supported by the NNSA Office of Defense Nuclear Nonproliferation Research and Development, U.S. Department of Energy, and Pacific North-

west National Laboratory, which is operated by Battelle Memorial Institute for the U.S. Department of Energy under Contract DEAC05–76RLO1830. This article has been cleared by PNNL for public release as PNNL-SA-220468.

References

Aichberger, L., Schweighofer, K., Ielanskyi, M., and Hochreiter, S. (2024). Semantically diverse language generation for uncertainty estimation in language models. *arXiv preprint arXiv:2406.04306*.

Bewersdorff, A., Hartmann, C., Hornberger, M., Seßler, K., Bannert, M., Kasneci, E., Kasneci, G., Zhai, X., and Nerdel, C. (2024). Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education. *arXiv preprint arXiv:2401.00832*.

Birhane, A., Kasirzadeh, A., Leslie, D., and Wachter, S. (2023). Science in the age of large language models. *Nature Reviews Physics*, 5(5):277–280.

Chandu, K. R., Li, L., Awadalla, A., Lu, X., Park, J. S., Hessel, J., Wang, L., and Choi, Y. (2024). Certainly uncertain: A benchmark and metric for multimodal epistemic and aleatoric awareness. *arXiv preprint arXiv:2407.01942*.

Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Gao, X., Zhang, J., Mouatadid, L., and Das, K. (2024). Spuq: Perturbation-based uncertainty quantification for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2336–2346.

Groot, T. and Valdenegro-Toro, M. (2024). Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In *Proceedings of TrustNLP Workshop@ NAACL 2024*.

Gui, Y., Jin, Y., and Ren, Z. (2024). Conformal alignment: Knowing when to trust foundation models with guarantees. *Advances in Neural Information Processing Systems*, 37:73884–73919.

Horawalavithana, S., Munikoti, S., Stewart, I., and Kvinge, H. (2024). Scitune: Aligning large language models with scientific multimodal instructions. In *EMNLP 2024, The first workshop on Natural Language Processing for Science*.

Huang, H., Zheng, O., Wang, D., Yin, J., Wang, Z., Ding, S., Yin, H., Xu, C., Yang, R., Zheng, Q., et al. (2023). Chatgpt for shaping the future of dentistry: the potential of multi-modal large language model. *International Journal of Oral Science*, 15(1):29.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., and Kasneci, G. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Kostumov, V., Nutfullin, B., Pilipenko, O., and Ilyushin, E. (2024). Uncertainty-aware evaluation for vision-language models. *arXiv preprint arXiv:2402.14418*.

Kuhn, L., Gal, Y., and Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Liu, H., Li, C., Li, Y., and Lee, Y. J. (2024). Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. (2022). Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Neal, R. (1992). Bayesian learning via stochastic dynamics. *Advances in neural information processing systems*, 5.

Panagoulias, D. P., Virvou, M., and Tsihrintzis, G. A. (2024). Evaluating llm–generated multimodal diagnosis from medical images and symptom analysis. *arXiv preprint arXiv:2402.01730*.

Park, S., Yeom, J., Sok, J., Park, J., Kim, H., and Kim, T. (2026). Efficient epistemic uncertainty estimation for large language models via knowledge distillation. *arXiv preprint arXiv:2602.01956*.

Pezeshkpour, P. and Hruschka, E. (2024). Large language models sensitivity to the order of options in multiple-choice questions. In Duh, K., Gomez, H., and Bethard, S., editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

Quach, V., Fisch, A., Schuster, T., Yala, A., Sohn, J. H., Jaakkola, T. S., and Barzilay, R. (2024). Con-

formal language modeling. In *The Twelfth International Conference on Learning Representations*.

Ren, A. Z., Dixit, A., Bodrova, A., Singh, S., Tu, S., Brown, N., Xu, P., Takayama, L. T., Xia, F., Varley, J., et al. (2023). Robots that ask for help: Uncertainty alignment for large language model planners. In *Conference on Robot Learning (CoRL)*. Proceedings of the Conference on Robot Learning (CoRL).

Saikh, T., Ghosal, T., Mittal, A., Ekbal, A., and Bhattacharyya, P. (2022). Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3).

Tian, J., Zhao, J., Wang, Z., and Ding, Z. (2024). Mmrec: Llm based multi-modal recommender system. *arXiv preprint arXiv:2408.04211*.

Ulmer, D., Zerva, C., and Martins, A. F. (2024). Non-exchangeable conformal language generation with nearest neighbors. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1909–1929.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*, volume 29. Springer.

Xia, Z., Xu, J., Zhang, Y., and Liu, H. (2025). A survey of uncertainty estimation methods on large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21381–21396.

Zheng, C., Zhou, H., Meng, F., Zhou, J., and Huang, M. (2023). Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

A Appendix

A.1 Dataset Processing

From the original 4,241 validation questions in ScienceQA, we removed 2,144 questions that do not include images, and also removed questions that were only answerable with their options included (i.e., those that included the text "Which of the following …" or "Which better answers the question …"), leaving a total of 1,980 questions. We run all experiments with the 7 billion parameter LLaVA-1.5 model (Liu et al., 2024).

A.2 Details on Computing Semantic Entropy

The original SE used a natural language inference (NLI) pre-trained model to establish if two responses from a model are semantically equivalent. We experimented with the NLI model reported in both Kuhn et al. (2023) and Farquhar et al. (2024), "Deberta-mnli," but found that the short outputs for our model and task, often one or two words, were often incorrectly judged to be "not entailed," and vice versa with longer texts. We found that using *lexical* matching methods, where n-gram overlap and "longest common substring" metrics are used to judge whether two texts are roughly equivalent produced better clusters. There is a trade-off in using lexical methods versus model-based semantic entailment, in particular in the case of detecting synonyms of semantically similar nouns such as "dog" and "puppy." However, in practice we found that these cases are relatively uncommon. We generally observed that calibration — i.e., whether the entropy over clustered model responses is aligned with correctness — is not sensitive to the particular method used for equivalence. While the clusters themselves vary according to the equivalence method used, the trend of CE values in relation to model correctness does not.

For computing CE on this task, we use ROUGE-L (the longest common subsequence). Two texts t_1 and t_2 are assigned the same cluster if $\text{ROUGE}_L(t_1, t_2) >= 0.6$. The threshold of 0.6 was selected in order to correctly capture both near-exact matches with a few words or less (e.g., "both samples" and "using both samples"), while also capturing overlap between larger texts (e.g., "Does the amount of light affect the size of radishes grown in a greenhouse?" and "Does the amount of light significantly affect the size of radish plants grown in a greenhouse?"). We find that model calibration with respect to this score is not sensitive to the threshold used and that, in general, CE requires only a coarse metric of textual relatedness in order to yield calibrated uncertainty scores.