Exploring Geometric Concentration for Quantifying Uncertainty in Scientific Image Caption Generation

Souradeep Chattopadhyay[1] **Brendan Kennedy**[2] **Sai Munikoti**[2]

Soumik Sarkar[1] **Karl Pazdernik**[2,3]

[1] Department of Mechanical Engineering, Iowa State University
[2] Pacific Northwest National Laboratory
[3] Department of Statistics, North Carolina State University

Abstract

Uncertainty Quantification (UQ) methods for Large Language Models (LLMs) have primarily been evaluated on question-answering benchmarks, where outputs are short, structured, and comparisons between generations are relatively well-defined. In contrast, many practical generative tasks involve open-ended, complex outputs, motivating evaluation of current state-of-the-art UQ beyond simple question-answering settings. In this work, we explore the challenging task of UQ for scientific image captioning. Using a subset of the ArxivCap dataset and two popular multimodal LLMs, we compare *Directional Concentration Uncertainty* (DCU), a geometric UQ measure proposed by Chattopadhyay et al. (2026), against semantic entropy (Kuhn et al., 2023), a leading method for UQ on structured question-answering. Our results indicate that DCU clearly outperforms SE, motivating further research into applications of DCU to other, complex tasks.

1 Introduction

Uncertainty quantification (UQ) is a fundamental requirement for making generative models trustworthy and robust, particularly in scientific and sensitive domains requiring critical decision-making. By exposing aspects of a model's underlying confidence — or lack thereof — UQ enables users to better understand model outputs and to make informed downstream decisions based on them (Huang et al., 2024; Zablotskaia et al., 2023; Kendall and Gal, 2017; Yu et al., 2024).

UQ has a well established history in conventional statistical modeling and in the development of deep learning systems like Bayesian neural networks, network uncertainty approaches (Neal, 1992; Gal and Ghahramani, 2016) and conformal prediction methods. (Vovk et al., 2005; Shafer and Vovk, 2008). More recently, UQ for large language models (LLMs) has attracted attention, as their scale and discrete generation make many classical UQ methods inapplicable (Liu et al., 2025), motivating black box approaches based on model outputs. One notable family of methods in this direction are sampling based methods, which estimate uncertainty by drawing multiple stochastic generations from a model and characterizing variability across the resulting outputs (Quach et al., 2024; Aichberger et al., 2024; Ulmer et al., 2024; Stengel-Eskin et al., 2024).

Among sampling-based approaches, semantic entropy (SE) has achieved strong performance in estimating model correctness on question-answering benchmarks (Kuhn et al., 2023) and has also been applied as a black-box approach for hallucination detection (Farquhar et al., 2024). SE operates by first grouping sampled model outputs according to semantic equivalence, using the notion of bidirectional-entailment relations, and then computes the entropy of the resulting cluster distribution. Existing evaluations of SE and related sampling based approaches have primarily focused on simple question-answering (QA) datasets, such as TriviaQA (Joshi et al., 2017). However, due to the narrowly defined notion of semantic equivalence, methods such as SE are ill-suited to tasks beyond the tried-and-true QA tasks used in LLM evaluations. In-

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

deed, the notion of semantic equivalence is typically defined only for phrases or sentences with identical meaning, whereas complex tasks such as summarization or image captioning lack a clear or empirically measurable notion of equivalence.

To overcome the limitations of SE and transcend the reliance on narrow notions of textual semantics, Chattopadhyay et al. (2026) proposed the Directional Concentration Uncertainty (DCU) as a geometry-based approach to task-agnostic, modality-agnostic uncertainty quantification for generative models. DCU estimates uncertainty by measuring the directional concentration of continuous embedding representations from multiple sampled model outputs using a von Mises–Fisher (vMF) distribution. By operating directly on the geometry of the embedding space rather than relying on pairwise semantic judgments, DCU characterizes uncertainty through the dispersion of generated outputs, making it broadly applicable across generative settings and independent of task specific structure. In experiments, Chattopadhyay et al. (2026) affirmed the viability of the method when compared directly to SE on QA tasks, including visual question-answering, and suggested further work exploring its suitability for complex tasks.

In this work, we evaluate Directional DCU as a uncertainty quantification method for scientific image caption generation. We empirically compare DCU and semantic entropy on open-ended, image-conditioned generation using state-of-the-art vision–language models, assessing their effectiveness beyond structured QA settings in identifying unreliable captions. Our experiments show that DCU and SE perform similarly under lenient evaluation settings, while under stricter evaluation criteria DCU performs significantly better, achieving AUROC improvements of over 0.12.

2 Related Work

UQ methods for large language models can be broadly divided into model based and sampling based approaches. A comprehensive survey by Xiao et al. (2022) reviewed UQ methods across multiple tasks, including text classification, generation and question answering, and provided task dependent recommendations along with a discussion of their limitations.

Within the class of sampling-based approaches, a number of methods have been developed that aim to capture uncertainty at a semantic level. Notably, semantic entropy (Kuhn et al., 2023) measures uncertainty by grouping multiple model generations according to semantic equivalence and computing entropy over the resulting clusters. Building on this general direction, Qiu and Miikkulainen (2024) proposed semantic density, which quantifies the confidence of large language model outputs by analyzing their distribution in a semantic representation space. This method evaluates response trustworthiness by measuring semantic variation across outputs and does not require any additional training or fine tuning.

Recent work has also explored alternative strategies for uncertainty estimation in large language models. For example, Gao et al. (2024) proposed a perturbation-driven framework that estimates uncertainty by analyzing model responses under controlled variations, enabling the characterization of both aleatoric and epistemic uncertainty. Their results demonstrated an average reduction of approximately 50% in expected calibration error (ECE).

3 Methodology

Here, we describe our adopted approach used for evaluating the DCU metric for image captioning proposed by Chattopadhyay et al. (2026). We describe our problem formulation, outline the major differences between DCU and SE and also provide a comprehensive overview of DCU.

3.1 Problem Formulation

For our problem we consider an image captioning task, where a vision language model receives an image input I with context x and produces a textual caption c. Due to stochastic decoding, repeated sampling for the same image I yields a set of captions $\{c_1, c_2, \ldots, c_N\}$. Our objective is to quantify the *uncertainty* associated with the model's outputs $\{c_1, c_2, \ldots, c_N\}$ for the given image I using DCU and SE and compare their performance.

3.1.1 Directional Concentration Uncertainty vs. Semantic Entropy

The SE approach proposed by Kuhn et al. (2023) computes uncertainty by first assigning sampled model outputs to discrete semantic groups and then evaluating the entropy of the resulting group distribution. Group membership is determined through pairwise semantic comparisons between outputs, typically using a bidirectional entailment model to assess whether two responses are semantically equivalent. As a result, the computation of uncertainty in SE depends on repeated pairwise comparisons and an explicit clustering step.

DCU follows a different computational procedure where each sampled output is embedded into a continuous embedding space and then uncertainty is computed using the overall spatial distribution of those

embeddings. This formulation avoids reliance on auxiliary entailment models and replaces discrete grouping with a continuous geometric characterization.

3.2 The Directional Concentration Measure

For a given image input I, a vision language model is sampled N times to obtain a set of captions $\{c_1, c_2, \ldots, c_N\}$. Each caption is mapped to a d-dimensional continuous representation using a pretrained text embedding model, and the resulting embeddings are ℓ_2-normalized to obtain vectors $\mathbf{z}_i$, $i = 1, 2, \ldots, N$ that lie on the unit hypersphere.

The distribution of the normalized embeddings $\{\mathbf{z}_i\}$ is then modeled using the vMF distribution. The vMF distribution is defined on the unit hypersphere and is parameterized by a mean direction $\boldsymbol{\mu}$ and a concentration parameter κ.

In this setting, the estimated concentration parameter κ provides a measure of consistency across the sampled captions for a given image. When generated captions are semantically similar, their embeddings concentrate around a common direction, resulting in a higher κ. Conversely, greater variation in the generated caption leads to increased dispersion in embedding space and a lower κ. The parameters of the vMF distribution are estimated from the set of normalized embeddings using maximum likelihood estimation. For more details see appendix of Chattopadhyay et al. (2026).

Finally, DCU is defined as the inverse of the estimated concentration parameter, κ^{-1}. This formulation assigns higher uncertainty to cases with greater embedding dispersion and lower uncertainty to cases where embeddings are tightly concentrated, yielding a continuous uncertainty score derived from the geometric structure of the embedding space.

4 Experimental Setup

In this section, we describe the experimental protocol used to evaluate the DCU measure and compare it against SE. Our goal here is to determine the potential of DCU and SE for scientific image captioning tasks. We compare DCU to SE directly using a popular image captioning dataset and two popular state of the art LLMs.

4.1 Description of Dataset

Our experiments were conducted using the ArxivCap dataset (Li et al., 2024), a large scale multimodal corpus constructed using scientific papers drawn from *arXiv*. The dataset consists of figures extracted from more than 500k scientific articles paired with their corresponding captions, along with metadata such as the paper title and abstract. The figures span a wide range of scientific domains and exhibit substantial variation in visual structure, including plots, diagrams and tables, making the dataset well suited for evaluating uncertainty in open ended scientific image captioning.

For our evaluation tasks we randomly selected a sample of 100 papers resulting in 512 figure-caption pairs with its corresponding metadata.

4.2 Models

For our experiments we used two open source LLMs: LLaVA 1.5 7B and Gemma 3 4B. The models were used in their pretrained forms with no additional fine-tuning.

For the implementation of DCU, an embedding model is required. For our experiments, we used the bi-modal CLIP encoder as our embedding model. Similarly to the generative models the encoder model was used in their pretrained forms with no additional fine tuning. For SE, we repeat the authors' prior implementation, using a pretrained MNLI model (Deberta-Large-MNLI)[1] to compute pairwise semantic equivalences using bidirectional entailment.

4.3 Prompting and Generation

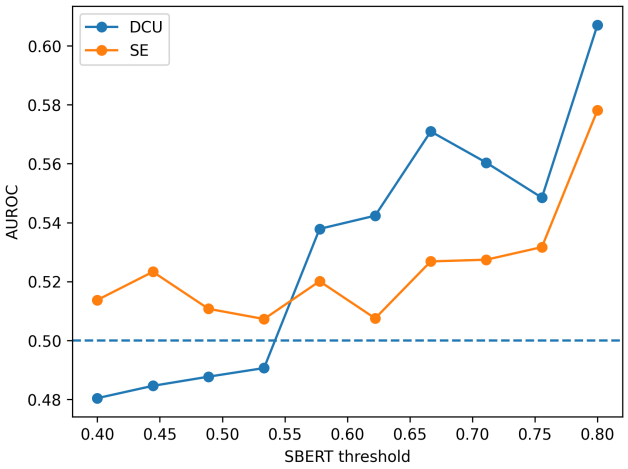
Our prompting strategies were designed to facilitate free form scientific figure captions from vision language models without imposing any restrictive output formats. For every image-caption pair, the model was provided with the figure image together with a textual prompt which instructed the LLM to produce a caption for the given figure. The prompt also included the paper title and abstract as context for better precision in the generated responses.

For our experiments, we used $N = 10$ generations with 'temperature' and 'top_P' settings similar to Kuhn et al. (2023) and Chattopadhyay et al. (2026) for effective comparison. Additionally, the prompt remained unchanged across the N generations for every model, and the hyperparameter settings were kept fixed for the two datasets for fair comparison.

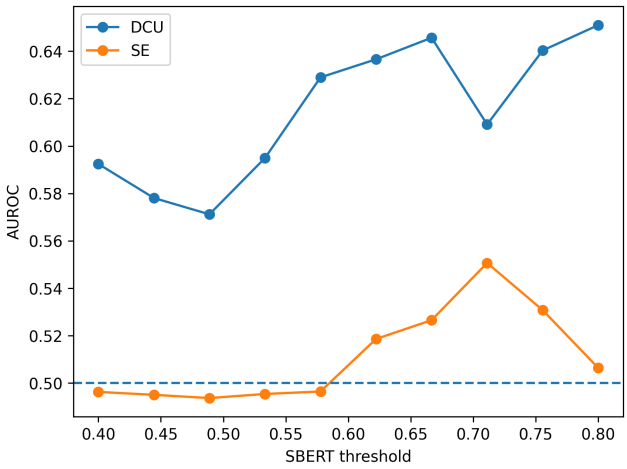
4.3.1 Performance Evaluation and Baseline

Following Kuhn et al. (2023) and Chattopadhyay et al. (2026), we evaluated DCU using both accuracy and the area under the receiver operating characteristic curve (AUROC). Accuracy was computed as the fraction of generated captions labeled as correct under this

[1] `https://huggingface.co/microsoft/deberta-large-mnli`



(a) SBERT Threshold vs AUROC for GEMMA 3 4B



(b) SBERT Threshold vs AUROC for LLaVA 1.5 7B

Figure 1: AUROC values vs SBERT score threshold for (a) GEMMA 4B and (b) LLaVA 1.5 7B models

Table 1: Accuracy and AUROC scores of DCU and SE for both models. The AUROC scores are computed using an SBERT threshold of 0.6.

Model	Acc.	AUROC_{DCU}	AUROC_{SE}
Gemma 3 4B	0.44	0.53	0.52
LLaVA 1.5 7B	0.25	0.63	0.50

criterion. AUROC measures how well the uncertainty score separates correct and incorrect model outputs across different confidence thresholds. AUROC lies between 0 and 1 with scores close to 1 indicating better predictions, and 0.5 indicating the baseline of a random prediction. The correctness labels for computing AUROC were determined by comparing the model's first generated answer to the given reference answer using a sentence-levelsimilarity score based on a pretrained sentence BERT encoder. A generation was labeled correct if its cosine similarity (referred to as SBERT score henceforth) to the reference caption exceeded a predefined threshold, set to 0.6 in our initial experiment.

5 Results and Discussion

In this section we present the empirical results of our study using our selected models and dataset. The accuracy and AUROC values using a SBERT score threshold of 0.6 for both the models are given in table 1. The accuracies for both models are relatively low, indicating the inherent complexity of the task. This is because, for tasks like scientific image captioning, multiple valid descriptions may exist for a single image, and exact semantic alignment with a specific reference is challenging. For Gemma 3 4B, the AUROC indicates that both DCU and SE perform similarly, with DCU having a slight edge over SE, and for LLaVA 1.5 7B, DCU performs significantly better than SE.

Additionally, the AUROC values for both measures at different SBERT score thresholds are given in Figure 1. The results indicate that as the threshold gets stricter, the performance of DCU increases significantly over SE. In fact, for LLaVA 1.5 7B, DCU performs better than SE consistently across all thresholds. At lower thresholds, where correctness criteria are less strict, captions with limited semantic overlap may still be labeled as correct, resulting in closer AUROC values for DCU and SE. As the correctness threshold becomes more strict, the performance gap between the methods widens, suggesting that embedding-based concentration better captures variations in caption reliability than the bidirectional entailment used by semantic entropy in image-conditioned generation tasks.

Our empirical study supports the claim made by Chattopadhyay et al. (2026) that DCU performs well on complex generation tasks such as scientific image captioning. The results also indicate that there remains scope for further improvement, for example through the use of alternative or more task specific embedding models.

6 Conclusion

In this work, we evaluated DCU on the task of scientific image captioning and compared its performance against SE, a standard UQ approach for text generation in an open-ended scientific image captioning setting. Our results show that DCU yields more reli-

able uncertainty estimates than SE for complex image-conditioned outputs, supporting its applicability beyond traditional QA benchmarks. Additionally, our results also suggest opportunities for further enhancement, such as exploring alternative embedding models or adapting the approach to other multimodal generation tasks. Overall, this study highlights the potential of embedding based uncertainty measures for improving reliability assessment in complex generative systems.

7 Acknowledgements

This work was supported by the NNSA Office of Defense Nuclear Nonproliferation Research and Development, U.S. Department of Energy, and Pacific Northwest National Laboratory, which is operated by Battelle Memorial Institute for the U.S. Department of Energy under Contract DEAC05–76RLO1830; and the AI Institute for Resilient Agriculture (USDA-NIFA #2021-647021-35329). This article has been cleared by PNNL for public release as PNNL-SA-220407.

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*.

Chattopadhyay, S., Kennedy, B., Munikoti, S., Sarkar, S., and Pazdernik, K. (2026). Directional concentration uncertainty: A representational approach to uncertainty quantification for generative models.

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.

Huang, X., Li, S., Yu, M., Sesia, M., Hassani, H., Lee, I., Bastani, O., and Dobriban, E. (2024). Uncertainty in language models: Assessment through rank-calibration. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 284–312, Miami, Florida, USA. Association for Computational Linguistics.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. (2017). Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1601–1611.

Kendall, A. and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5580–5590, Red Hook, NY, USA. Curran Associates Inc.

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*.

Li, L., Wang, Y., Xu, R., Wang, P., Feng, X., Kong, L., and Liu, Q. (2024). Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint*, abs/2403.00231.

Liu, X., Chen, T., Da, L., Chen, C., Lin, Z., and Wei, H. (2025). Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6107–6117.

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

Qiu, X. and Miikkulainen, R. (2024). Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space.

Quach, V., Fisch, A., Schuster, T., Yala, A., Sohn, J. H., Jaakkola, T. S., and Barzilay, R. (2024). Conformal language modeling. In *The Twelfth International Conference on Learning Representations*.

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

Stengel-Eskin, E., Hase, P., and Bansal, M. (2024). Lacie: Listener-aware finetuning for confidence calibration in large language models. *arXiv preprint arXiv:2405.21028*.

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.

Xiao, Y., Liang, P. P., Bhatt, U., Neiswanger, W., Salakhutdinov, R., and Morency, L.-P. (2022). Un-

certainty quantification with pre-trained language models: A large-scale empirical analysis.

Yu, L., Cao, M., Cheung, J. C., and Dong, Y. (2024). Mechanistic understanding and mitigation of language model non-factual hallucinations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7943–7956.

Zablotskaia, P., Phan, D., Maynez, J., Narayan, S., Ren, J., and Liu, J. (2023). On uncertainty calibration and selective generation in probabilistic neural summarization: A benchmark study. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2980–2992, Singapore. Association for Computational Linguistics.