Vision-Amplified Semantic Entropy for Hallucination Detection in Medical Visual Question Answering

Zehui Liao[1,2⋆], Shishuai Hu[1,2∗], Ke Zou[3], Huazhu Fu[2], Liangli Zhen[2(🖂)], and Yong Xia[1,4,5(🖂)]

[1] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China
[2] Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore 138632, Singapore
[3] National University of Singapore, Singapore 119077, Singapore.
[4] Ningbo Institute of Northwestern Polytechnical University, Ningbo 315048, China
[5] Research & Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen 518057, China
zhenll@ihpc.a-star.edu.sg; yxia@nwpu.edu.cn

Abstract. Multimodal large language models (MLLMs) have demonstrated significant potential in medical Visual Question Answering (VQA). Yet, they remain prone to hallucinations—incorrect responses that contradict input images, posing substantial risks in clinical decision-making. Detecting these hallucinations is essential for establishing trust in MLLMs among clinicians and patients, thereby enabling their real-world adoption. Current hallucination detection methods, especially semantic entropy (SE), have demonstrated promising hallucination detection capacity for LLMs. However, adapting SE to medical MLLMs by incorporating visual perturbations presents a dilemma. Weak perturbations preserve image content and ensure clinical validity, but may be overlooked by medical MLLMs, which tend to over-rely on language priors. In contrast, strong perturbations can distort essential diagnostic features, compromising clinical interpretation. To address this issue, we propose Vision Amplified Semantic Entropy (VASE), which incorporates weak image transformations and amplifies the impact of visual input, to improve hallucination detection in medical VQA. We first estimate the semantic predictive distribution under weak visual transformations to preserve clinical validity, and then amplify visual influence by contrasting this distribution with that derived from a distorted image. The entropy of the resulting distribution is estimated as VASE. Experiments on two medical open-ended VQA datasets demonstrate that VASE consistently outperforms existing hallucination detection methods. The code will be available at https://github.com/Merrical/VASE.

Keywords: Hallucination detection · Medical VQA · Vision-amplified semantic entropy.

[⋆] Z. Liao and S. Hu contributed equally. Corresponding authors: L. Zhen and Y. Xia.

1 Introduction

Medical Visual Question Answering (VQA), a task powered by multimodal large language models (MLLMs), enables the interpretation of medical images in response to natural language queries, supporting disease diagnosis, abnormality detection, and complex image analysis across domains such as radiology, pathology, and ophthalmology [27,18]. However, MLLMs in medical VQA are susceptible to hallucinations—errors where generated responses misinterpret or contradict the input image—posing substantial risks[12,19], including misdiagnoses, inappropriate treatments, and diminished trust in AI-driven medical tools. Despite the efforts to mitigate hallucinations via data optimization [30], training [13], and decoding refinements [26], hallucinations remain a persistent challenge, underscoring the critical need for effective hallucination detection methods.

Recent efforts in hallucination detection for large language models (LLMs) and MLLMs can be broadly classified into five categories. **Uncertainty estimation** methods [17,6,1] infer hallucinations via predictive uncertainty, with higher uncertainty indicating a greater likelihood of hallucinations. **Hallucination detector**-based approaches [8,28,2,9] fine-tune LLMs- and MLLMs-based detectors using additional annotated hallucination data. **Cross-check** methods [30,5] verify the consistency of responses across multiple LLMs/MLLMs. **External fact retrieval** techniques [3,21] cross-check generated responses against external knowledge bases to detect hallucinations. Within MLLMs, **visual evidence verification** methods [24,29] assess hallucinations by leveraging expert vision models. Among these approaches, uncertainty estimation stands out due to its effectiveness and simplicity [19,12,32], as it allows for hallucination detection using only the model itself, without relying on additional models, annotated hallucination data, or external knowledge. A notable advancement in uncertainty estimation for LLMs is **semantic entropy** (**SE**) [6]. SE quantifies uncertainty by calculating the entropy of the probabilistic distribution of responses at the semantic level, rather than the response level, thereby outperforming other uncertainty estimation techniques. A natural adaptation of SE to MLLMs involves incorporating perturbations in visual input during entropy estimation [31]. However, when applied to medical images, strong visual perturbations can distort critical diagnostic features and thus compromise clinical interpretation [20], while weak transformations may be overlooked by medical MLLMs, which tend to neglect visual input and rely excessively on language priors [7]. This creates a dilemma: while preserving image content with weak transformations ensures clinical validity, it may lead to inaccurate entropy estimation as medical MLLMs often overlook visual evidence.

To address this challenge, we propose **V**ision **A**mplified **S**emantic **E**ntropy (**VASE**), a novel method that incorporates weak visual transformations and amplifies the visual influence to improve hallucination detection in medical VQA. Specifically, we first estimate the semantic probability distribution of responses sampled from the medical MLLM, incorporating weak visual transformations. We then amplify the impact of visual input by comparing this distribution with that derived from a distorted image paired with the original text input. Finally,

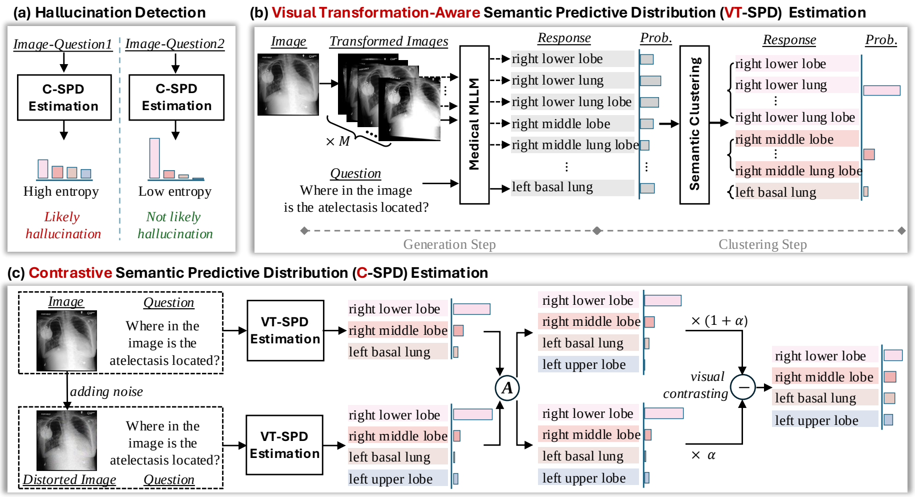

Fig. 1: Overview of VASE estimation for hallucination detection in medical VQA. VASE refers to the entropy of the estimated C-SPD. $\textcircled{A}$ means semantic equivalence classes alignment. 'Prob.' means the probability.

the entropy of the contrasted semantic predictive distribution is defined as VASE. The response is considered as hallucination if the VASE score is high. We evaluate VASE on two medical VQA datasets, demonstrating its superior performance in detecting hallucinations over existing methods across two medical MLLMs. The main contributions of this work are three-fold: (1) We introduce VASE for hallucination detection in medical VQA, an area that remains underexplored in medical contexts; (2) Unlike SE, VASE integrates visual variability and enhances the influence of visual input, addressing the issue of over-reliance on language priors in medical MLLMs; and (3) Extensive experiments shows that VASE consistently outperforms existing hallucination detection methods.

2 Our Proposed Method

Given a multi-modal input pair $(\boldsymbol{x}_v, \boldsymbol{x}_q)$, where $\boldsymbol{x}_v$ represents the input image and $\boldsymbol{x}_q$ denotes the textual query, a medical MLLM f is designed to generate a response $\boldsymbol{r}$ that effectively addresses the inquiry posted in $\boldsymbol{x}_q$ while leveraging the information contained in $\boldsymbol{x}_v$. This work aims to develop a hallucination detection mechanism that determines whether $\boldsymbol{r}$ contains hallucinations—incorrect findings that contradict the visual input. As illustrated in Figure 1 (a), we propose to detect hallucinations by estimating VASE of the model's output given the input pair $(\boldsymbol{x}_v, \boldsymbol{x}_q)$. A response $\boldsymbol{r}$ is classified as a hallucination if the VASE exceeds a predefined threshold τ. The VASE estimation consists of three key stages: (1) estimating the visual transformation-aware semantic predictive distribution (see Fig. 1 (b)), (2) measuring contrastive distribution by comparing this estimated semantic predictive distribution with the one that is derived from the distorted image with original text input (see Fig. 1 (c)), and (3) calculating the entropy of the contrastive distribution for identifying hallucinations.

2.1 Visual Transformation-Aware Semantic Predictive Distribution

To estimate the semantic predictive distribution of responses generated by a medical MLLM f, we employ a two-step process: (1) Generation: Sampling multiple responses from the model's predictive distribution for the input pair $(\boldsymbol{x}_v, \boldsymbol{x}_q)$. (2) Clustering: Grouping the sampled responses into semantically equivalent clusters using bidirectional entailment [6]. The probabilities of responses within the same cluster are then aggregated to estimate semantic probability distribution.

Specifically, to approximate the predictive distribution, we introduce ***image transformations*** to enhance response diversity while preserving clinically relevant visual features. These transformations include random cropping, rotation, translation, brightness adjustment, and contrast adjustment, each applied to $\boldsymbol{x}_v$ before it is processed by the model f. We perform M inference runs, applying a different random transformation to $\boldsymbol{x}_v$ before each run. The model f then generates M output sequences $\{\boldsymbol{s}^{(i)}\}_{i=1}^{M}$ with high-temperature sampling from the predictive distribution, ensuring a diverse set of responses. For each generated sequence, we record its probability as

$$P(\boldsymbol{s}^{(i)}|\boldsymbol{x}_v, \boldsymbol{x}_q) = \prod_j P(s_j^{(i)}|\boldsymbol{s}_{<j}^{(i)}, \boldsymbol{x}_v, \boldsymbol{x}_q), \tag{1}$$

where $s_j^{(i)}$ is the j-th token of $\mathbf{s}^{(i)}$ and $\mathbf{s}_{<j}^{(i)}$ refers to the set of preceding tokens.

Next, to estimate the probability distribution at the semantic level, the generated sequences are clustered into semantic equivalence groups. This is achieved using entailment relationships assessed by the DeBERTa-Large-MNLI model [10]. Two sequences $\boldsymbol{s}$ and $\boldsymbol{s}'$ are considered semantically equivalent if they entail each other. The clustering process on M responses $\{\boldsymbol{s}^{(i)}\}_{i=1}^{M}$ proceeds as follows: The first response $\boldsymbol{s}^{(1)}$ establishes the initial semantic equivalence class. Each subsequent response $\boldsymbol{s}^{(i)}$ is compared to the existing classes in order. If $\boldsymbol{s}^{(i)}$ is semantically equivalent to an existing class, it is included in that class; otherwise, a new semantic equivalence class is created. This process yields N_1 semantic equivalence classes $[c_i]_{i=1}^{N_1}$. The probability of class c_i is calculated by summing the probabilities of all sequences within that class:

$$P(c_i|\boldsymbol{x}_v, \boldsymbol{x}_q) = \textstyle\sum_{\boldsymbol{s}\in c_i} P(\boldsymbol{s}|\boldsymbol{x}_v, \boldsymbol{x}_q). \tag{2}$$

The resultant semantic probability distribution is given by $\{P(c_i|\boldsymbol{x}_v, \boldsymbol{x}_q)\}_{i=1}^{N_1}$.

2.2 Contrastive Semantic Predictive Distribution

To amplify the impact of visual input, we compare the semantic probability distribution derived from the original image-text input with the one that from a distorted version of the image with the original text input. Specifically, we estimate the semantic predictive distribution $\{P(c_i|\boldsymbol{x}_v, \boldsymbol{x}_q)\}_{i=1}^{N_1}$ based on $(\boldsymbol{x}_v, \boldsymbol{x}_q)$. Next, we add noise to the input image, resulting in a distorted version, denoted as $\boldsymbol{x}'_v$. The semantic predictive distribution $\{P(c_i|\boldsymbol{x}'_v, \boldsymbol{x}_q)\}_{i=1}^{N_2}$ based on $(\boldsymbol{x}'_v, \boldsymbol{x}_q)$ can be generated, where N_2 is the number of its semantic equivalence classes. Since the semantic equivalence classes derived from these two distributions may not naturally align, we utilize DeBERTa-Large-MNLI to establish alignment and ensure consistency. The aligned distributions are then redefined as $\{P(c_i|\boldsymbol{x}_v, \boldsymbol{x}_q)\}_{i=1}^{N}$ and $\{P(c_i|\boldsymbol{x}'_v, \boldsymbol{x}_q)\}_{i=1}^{N}$, where N is the number of semantic equivalence classes across both distributions.

Using a contrastive semantic analysis strategy, we compute the contrastive semantic predictive distribution as:

$$P_{con}(\boldsymbol{c}|\boldsymbol{x}_v, \boldsymbol{x}_q) = \sigma((1+\alpha)P(\boldsymbol{c}|\boldsymbol{x}_v, \boldsymbol{x}_q) - \alpha P(\boldsymbol{c}|\boldsymbol{x}'_v, \boldsymbol{x}_q)), \tag{3}$$

where $\sigma(\cdot)$ is the softmax function, $\boldsymbol{c} = [c_i]_{i=1}^{N}$, and α is the vision-amplified ratio parameter that modulates the relative contributions of the two semantic predictive distributions. Fig. 1 (c) illustrates a scenario where both the original and distorted images yield similar probabilities for the responses “right lower lobe” and “right middle lobe”. This indicates that these responses are heavily influenced by language priors, resulting in an inaccurate distribution estimation. By contrasting the distributions of the original and distorted images, the impact of language priors on responses is reduced, resulting in a contrastive semantic predictive distribution that is more grounded in visual context.

2.3 Detecting Hallucination via VASE

VASE is the entropy of the contrastive semantic predictive distribution:

$$VASE(\boldsymbol{x}_v, \boldsymbol{x}_q) = -\textstyle\sum_{i=1}^{N} P_{con}(c_i|\boldsymbol{x}_v, \boldsymbol{x}_q) log(P_{con}(c_i|\boldsymbol{x}_v, \boldsymbol{x}_q)). \tag{4}$$

A higher VASE indicates larger model uncertainty, suggesting an increased likelihood of hallucinations. If $VASE(\boldsymbol{x}_v, \boldsymbol{x}_q) > \tau$, where $\tau \in \mathcal{R}^+$ is a threshold, the response $\boldsymbol{r}$ is classified as a likely hallucination. The threshold τ can be determined using the validation set following previous research [23].

3 Experimental Study

3.1 Datasets and Experimental Setup

Datasets. We evaluated VASE on two VQA benchmarking datasets: MIMIC-Diff-VQA [11] and VQA-RAD [15]. **MIMIC-Diff-VQA** [11] is a chest X-ray VQA dataset with 700,703 question-answer pairs across seven categories: abnormality, difference, level, location, presence, type, and view. To prevent data leakage, we followed the MIMIC-CXR training set split [14] instead of the official split, as CheXagent [4] was trained on MIMIC-CXR’s official training set,

Table 1: Hallucination detection performance (AUC(%) and AUG(%)) of VASE and seven peer methods. The performance on both open-ended test samples and all test samples is reported. The best results are highlighted in **bold**.

Method	MedDiffVQA				VQA-RAD			
	Open-Ended		All		Open-Ended		All	
	AUC	AUG	AUC	AUG	AUC	AUG	AUC	AUG
CheXagent [4]								
AvgProb [17]	36.20	33.18	37.26	36.18	48.50	27.48	44.67	40.36
AvgEnt [17]	63.72	47.83	62.64	50.07	51.43	21.49	55.09	47.67
MaxProb [17]	35.65	31.73	36.53	34.08	48.14	27.32	44.20	39.31
MaxEnt [17]	64.43	47.98	63.55	50.27	52.40	19.90	55.70	47.46
Cross-Checking [30]	68.45	48.98	66.23	50.30	70.51	38.99	68.63	55.74
RadFlag [25]	85.77	61.01	85.35	62.19	74.83	38.52	77.08	61.62
SE [6]	86.51	59.82	86.06	62.58	73.01	38.36	77.54	62.40
Ours	**87.79**	**62.06**	**87.01**	**63.88**	**76.54**	**39.93**	**77.99**	**62.91**
LLaVa-Med [16]								
AvgProb [17]	45.81	36.86	44.86	38.11	51.53	36.40	51.02	45.19
AvgEnt [17]	55.07	42.54	56.02	44.88	52.14	32.10	49.61	45.16
MaxProb [17]	42.37	31.82	41.65	33.29	51.15	38.04	50.21	45.61
MaxEnt [17]	56.98	43.77	58.12	45.77	51.48	31.65	50.91	44.88
Cross-Checking [30]	60.31	44.23	57.80	44.58	59.56	38.99	61.77	51.35
RadFlag [25]	73.62	54.23	72.55	53.02	61.21	36.33	64.64	54.12
SE [6]	74.76	53.66	74.21	53.18	63.55	38.88	64.71	54.10
Ours	**75.86**	**55.58**	**74.98**	**54.25**	**66.72**	**40.47**	**66.43**	**55.28**

resulting in 13,121 test samples. To assess hallucination detection performance in open-ended VQA, "view" questions with fixed answers (*i.e.*, "AP view" or "PA view") were excluded, resulting in an open-ended test set of 12,143 samples. **VQA-RAD** [15] consists of 2,244 question-answer pairs linked to 314 radiology images, covering both open-ended and binary ("yes/no") questions. The official test set contains 451 samples, with 200 of those being open-ended questions. Hallucination detection performance on both open-ended test samples and all test samples are reported. The GREEN model [22] was used to generate ground truth labels (hallucination or not) for test samples according to the low-temperature (T=0.1) sampled responses from medical MLLM [6] and reference answers.

Evaluation Metrics. AUC and Area Under GREEN Curve (AUG) are used as metrics to evaluate VASE and other hallucination detection methods, following [6]. AUC can be interpreted as the probability that a randomly chosen correct answer has been assigned a higher confidence / lower uncertainty score than a randomly chosen hallucinated answer. For AUG, we compute the "mean GREEN score at X%", which is the mean GREEN score of the model on the most-confident X% of samples identified by the respective uncertainty method. The GREEN score is derived from the GREEN model and quantifies the correctness of the generated answer. It is calculated as $\frac{\#matched\ findings}{\#matched\ findings+\#errors}$,

Table 2: Performance (AUC(%), AUG(%)) of VASE and its four variants on the MIMIC-Diff-VQA dataset, evaluated using the CheXagent model. The best results are highlighted in **bold**.

Components of VASE			AUC	AUG
Visual Contrasting	Weak Visual Transformation	Strong Visual Transformation		
×	×	×	86.51	59.82
×	×	✓	85.70$_{\downarrow 0.81}$	59.50$_{\downarrow 0.32}$
×	✓	×	86.28$_{\downarrow 0.23}$	60.10$_{\uparrow 0.28}$
✓	×	×	87.35$_{\uparrow 0.84}$	61.18$_{\uparrow 1.36}$
✓	✓	×	**87.79**$_{\uparrow 1.28}$	**62.06**$_{\uparrow 2.24}$

where # denotes the count of matched findings/errors. To summarize the "mean GREEN score at X%" from 1% to 100% (interval=1%), we compute the AUG - the total area enclosed by the mean GREEN score at all cut-off percentage X%. Higher AUC and AUG values indicate better hallucination detection.

Implementation Details. The sampling number M is set to 10. The sampling temperature during high-temperature sampling is set to 1.0. The vision-amplified ratio α is set to 1.0. The weak transformations used include: random cropping (selecting 90% to 100% of the original area), random rotation (choosing an angle between -10° and 10°), random translation (displacing up to 10% of the image's width or height), and brightness/contrast adjustments (uniformly selected between 0.8 and 1.2). Gaussian and Poisson noise are added to distort the image. Gaussian noise has a zero mean and a standard deviation of 0.07. Poisson noise is applied by scaling the image intensity, performing Poisson sampling, and normalizing back to [0,1], with a scaling factor of 70 controlling its intensity. All experiments were performed on two NVIDIA GeForce RTX 3090 GPUs.

3.2 Experimental Results and Analysis

Comparative Experiments. We compared VASE against seven recent hallucination detection methods, including: (1) AvgProb and MaxProb [17] estimate model confidence by calculating the average and maximum token probabilities in the generated answer, with lower confidence indicating a higher likelihood of hallucinations; (2) AvgEnt and MaxEnt [17] measure uncertainty by computing the average and maximum entropy of the token probability distribution, where higher entropy indicates a higher likelihood of hallucinations; (3) SE [6] estimates semantic-level uncertainty by sampling multiple responses, constructing a probability distribution over semantically equivalent answers, and computing entropy; (4) RadFlag [25] generates multiple responses and calculates the proportion of answers agreeing with the original response. A lower agreement rate suggests a higher hallucination likelihood; and (5) Cross-Checking [30] generates additional responses using other VQA for the same input and computes the mean consistency score between the original response and these alternative outputs. Table 1

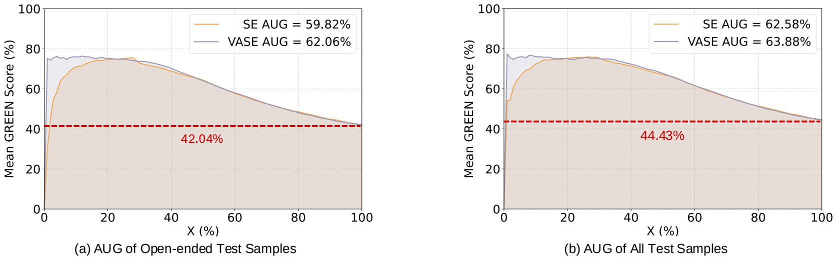

Fig. 2: The AUG curves of SE and VASE on the (a) open-ended test samples and (b) all test samples of MIMIC-Diff-VQA. Each point on the curve represents the mean GREEN score of the model on the most-confident X% of samples identified by SE/VASE. The red dash line refers to the mean GREEN score of all samples under setting (a) and (b).

reports the performance of these hallucination detection methods on VQA-RAD and MIMIC-Diff-VQA datasets using CheXagent [4] and LLaVA-Med [16]. The results indicate that VASE consistently outperforms other competing methods, especially on open-ended test samples, achieving the highest AUC and AUG scores across multiple datasets and medical MLLMs.

Ablation Analysis. We conducted an ablation study to assess the contributions of VASE's two key components: visual transformation and visual contrasting. The baseline method detects hallucinations using the entropy of the semantic predictive distribution from the original image-text pair. Visual transformation extends this by estimating semantic entropy from randomly augmented images while keeping the text unchanged. Visual contrasting enhances the baseline by comparing the semantic predictive distributions of the original and distorted images, using the entropy of the contrastive distribution for hallucination detection. VASE refers to integrating both visual transformation and visual contrasting into the baseline. As shown in Table 2, adding strong visual transformations to the baseline reduces performance, while weak transformations yield comparable AUC and AUG, indicating medical MLLMs' over-reliance on language priors. Nonetheless, when we amplifying the impact of visual input by introducing visual contrasting to baseline, and then adding weak visual transformation, the detection performance improves progressively. Additionally, as observed in Fig. 2, VASE outperforms baseline (*i.e.*, SE) on the AUG curve, particularly in the range of the most-confident 0%-20% of samples. The mean GREEN score for these samples, identified by VASE, is consistently higher than that of SE. This suggests that VASE is more effective at connecting high-confidence samples to lower entropy, underscoring its ability to accurately identify reliable samples.

4 Conclusion

This paper introduced VASE, a novel approach for hallucination detection in medical VQA. By incorporating weak visual transformations and strengthening the influence of visual inputs, VASE enhances predictive uncertainty estimation and achieves superior performance compared to existing hallucination detection methods across two medical VQA datasets. An ablation study highlights the critical role of weak visual transformations and vision contrasting in improving detection efficacy. Future work will focus on extending VASE to mitigate hallucinations in MLLMs for medical VQA, further enhancing the reliability of AI-driven medical image interpretation.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grants 62171377 and 92470101, the “Pioneer” and “Leading Goose” R&D Program of Zhejiang, China, under Grant 2025C01201(SD2), the Shenzhen Science and Technology Program under Grant JCYJ20220530161616036, the Ningbo Clinical Research Center for Medical Imaging under Grant 2021L003 (Open Project 2022LYKFZD06), the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University under Grants CX2022056 and CX2023016, and the National Research Foundation Singapore under the National Cybersecurity R&D Programme (Grand Challenge Award: CRPO-GC1-NTU-002).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chen, C., Liu, K., Chen, Z., Gu, Y., Wu, Y., Tao, M., Fu, Z., Ye, J.: INSIDE: LLMs’ internal states retain the power of hallucination detection. In: ICLR (2024)
2. Chen, J., Yang, D., Wu, T., Jiang, Y., Hou, X., Li, M., Wang, S., Xiao, D., Li, K., Zhang, L.: Detecting and evaluating medical hallucinations in large vision language models. arXiv preprint arXiv:2406.10185 (2024)
3. Chen, J., Kim, G., Sriram, A., Durrett, G., Choi, E.: Complex claim verification with evidence retrieved in the wild. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 3569–3587 (2024)
4. Chen, Z., Varma, M., Delbrouck, J.B., Paschali, M., Blankemeier, L., Van Veen, D., Valanarasu, J.M.J., Youssef, A., Cohen, J.P., Reis, E.P., et al.: Chexagent: Towards a foundation model for chest x-ray interpretation. In: AAAI 2024 Spring Symposium on Clinical Foundation Models (2024)
5. Cohen, R., Hamri, M., Geva, M., Globerson, A.: Lm vs lm: Detecting factual errors via cross examination. In: EMNLP. pp. 12621–12640 (2023)
6. Farquhar, S., Kossen, J., Kuhn, L., Gal, Y.: Detecting hallucinations in large language models using semantic entropy. Nature **630**(8017), 625–630 (2024)
7. Favero, A., Zancato, L., Trager, M., Choudhary, S., Perera, P., Achille, A., Swaminathan, A., Soatto, S.: Multi-modal hallucination control by visual information grounding. In: CVPR. pp. 14303–14312 (2024)
8. Gunjal, A., Yin, J., Bas, E.: Detecting and preventing hallucinations in large vision language models. In: AAAI. vol. 38, pp. 18135–18143 (2024)
9. Hardy, R., Kim, S.E., Rajpurkar, P.: Rextrust: A model for fine-grained hallucination detection in ai-generated radiology reports. arXiv preprint arXiv:2412.15264 (2024)

10. He, P., Liu, X., Gao, J., Chen, W.: Deberta: Decoding-enhanced bert with disentangled attention. In: ICLR (2020)
11. Hu, X., Gu, L., An, Q., Zhang, M., Liu, L., Kobayashi, K., Harada, T., Summers, R.M., Zhu, Y.: Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 4156–4165 (2023)
12. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems (2023)
13. Jiang, C., Xu, H., Dong, M., Chen, J., Ye, W., Yan, M., Ye, Q., Zhang, J., Huang, F., Zhang, S.: Hallucination augmented contrastive learning for multimodal large language model. In: CVPR. pp. 27036–27046 (2024)
14. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
15. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 1–10 (2018)
16. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. NeurIPS **36** (2024)
17. Li, Q., Geng, J., Lyu, C., Zhu, D., Panov, M., Karray, F.: Reference-free hallucination detection for large vision-language models. In: EMNLP. pp. 4542–4551 (2024)
18. Liu, B., Zou, K., Zhan, L., Lu, Z., Dong, X., Chen, Y., Xie, C., Cao, J., Wu, X.M., Fu, H.: GEMeX: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis. arXiv preprint arXiv:2411.16778 (2024)
19. Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., Peng, W.: A survey on hallucination in large vision-language models. arXiv preprint arXiv:2402.00253 (2024)
20. Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., Lu, F.: Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recognition **110**, 107332 (2021)
21. Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.t., Koh, P., Iyyer, M., Zettlemoyer, L., Hajishirzi, H.: Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In: EMNLP. pp. 12076–12100 (2023)
22. Ostmeier, S., Xu, J., Chen, Z., Varma, M., Blankemeier, L., Bluethgen, C., Michalson, A.E., Moseley, M., Langlotz, C., Chaudhari, A.S., et al.: Green: Generative radiology report evaluation and error notation. arXiv preprint arXiv:2405.03595 (2024)
23. Peng, Y., Lin, A., Wang, M., Lin, T., Liu, L., Wu, J., Zou, K., Shi, T., Feng, L., Liang, Z., et al.: Enhancing ai reliability: A foundation model with uncertainty estimation for optical coherence tomography-based retinal disease diagnosis. Cell Reports Medicine **6**(1) (2025)
24. Sahu, P., Sikka, K., Divakaran, A.: Pelican: Correcting hallucination in vision-llms via claim decomposition and program of thought verification. In: EMNLP. pp. 8228–8248 (2024)

25. Sambara, S., Zhang, S., Banerjee, O., Acosta, J., Fahrner, J., Rajpurkar, P.: Radflag: A black-box hallucination detection method for medical vision language models. arXiv preprint arXiv:2411.00299 (2024)
26. Wang, X., Pan, J., Ding, L., Biemann, C.: Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In: ACL. pp. 15840–15853 (2024)
27. Xiao, H., Zhou, F., Liu, X., Liu, T., Li, Z., Liu, X., Huang, X.: A comprehensive survey of large language models and multimodal large language models in medicine. arXiv preprint arXiv:2405.08603 (2024)
28. Xiao, W., Huang, Z., Gan, L., He, W., Li, H., Yu, Z., Jiang, H., Wu, F., Zhu, L.: Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. arXiv preprint arXiv:2404.14233 (2024)
29. Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X., Chen, E.: Woodpecker: Hallucination correction for multimodal large language models. Science China Information Sciences **67**(12), 220105 (2024)
30. Yu, Q., Li, J., Wei, L., Pang, L., Ye, W., Qin, B., Tang, S., Tian, Q., Zhuang, Y.: Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In: CVPR. pp. 12944–12953 (2024)
31. Zhang, R., Zhang, H., Zheng, Z.: Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. arXiv preprint arXiv:2411.11919 (2024)
32. Zou, K., Chen, Z., Yuan, X., Shen, X., Wang, M., Fu, H.: A review of uncertainty estimation and its application in medical imaging. Meta-Radiology p. 100003 (2023)