Deterministic Hallucination Detection in Medical VQA via Confidence-Evidence Bayesian Gain

Mohammad Asadi[1], Tahoura Nedaee[2], Jack W. O'Sullivan[3,4], Euan Ashley[3,4], and Ehsan Adeli[4,5,6]

[1] Department of Electrical Engineering, Stanford University, CA, USA
masadi@stanford.edu
[2] Department of Biology, Stanford University, CA, USA
[3] Division of Cardiology, Department of Medicine, Stanford University, CA, USA
[4] Department of Biomedical Data Science, Stanford University, CA, USA
[5] Department of Computer Science, Stanford University, CA, USA
[6] Department of Psychiatry and Behavioral Sciences, Stanford University, CA, USA

Abstract. Multimodal large language models (MLLMs) have shown strong potential for medical Visual Question Answering (VQA), yet they remain prone to hallucinations, defined as generating responses that contradict the input image, posing serious risks in clinical settings. Current hallucination detection methods, such as Semantic Entropy (SE) and Vision-Amplified Semantic Entropy (VASE), require 10 to 20 stochastic generations per sample together with an external natural language inference model for semantic clustering, making them computationally expensive and difficult to deploy in practice. We observe that hallucinated responses exhibit a distinctive signature directly in the model's own log-probabilities: inconsistent token-level confidence and weak sensitivity to visual evidence. Based on this observation, we propose **C**onfidence-**E**vidence **Ba**yesian **G**ain (CEBaG), a deterministic answer-level hallucination detection method for short-form medical VQA that requires no stochastic sampling, no external models, and no task-specific hyperparameters. CEBaG combines two complementary signals: *token-level predictive variance*, which captures inconsistent confidence across response tokens, and *evidence magnitude*, which measures how much the image shifts per-token predictions relative to text-only inference. Evaluated across four medical MLLMs and three VQA benchmarks (16 experimental settings), CEBaG achieves the highest AUC in 13 of 16 settings and improves over VASE by 8 AUC points on average, while being fully deterministic and self-contained. https://github.com/masadi-99/CEBaG.

Keywords: Hallucination detection · Medical VQA · Multimodal LLMs.

1 Introduction

Medical Visual Question Answering (VQA), powered by multimodal large language models (MLLMs), enables the interpretation of medical images in response to natural language queries, supporting tasks from abnormality detection to differential diagnosis [24, 18]. However, MLLMs are susceptible to *hallucinations*,

defined as responses that misinterpret or contradict the input image, posing substantial risks including misdiagnoses, inappropriate treatments, and diminished clinician trust [8, 19]. Despite efforts in data optimization [26], training [9], and decoding refinements [23], hallucinations remain a persistent challenge, highlighting the need for effective detection methods.

Recent approaches to hallucination detection include detector fine-tuning [5], cross-checking [26, 3], visual evidence verification [25], and uncertainty estimation [10, 1, 4]. Among these, uncertainty estimation stands out for its simplicity, as it uses only the model itself without additional annotated data or external knowledge bases [8, 19]. Semantic Entropy (SE) [4] quantifies uncertainty by computing entropy over semantically clustered responses. Vision-Amplified Semantic Entropy (VASE) [16] extends SE by contrasting semantic distributions from original and augmented images, amplifying the influence of visual input. Other contrastive methods such as VCD [12] and LCD [13] leverage output differences under perturbed inputs, while Semantic Entropy Probes [11] train a probe on internal activations, requiring task-specific data unlike our training-free approach. A common thread in these methods is their reliance on stochastic sampling, external models, or additional training, which raises the question of whether the uncertainty signal they seek is accessible through simpler means.

These methods also face a concrete efficiency bottleneck. SE requires M stochastic generations (typically M=10) with high-temperature sampling, followed by pairwise semantic equivalence checks via an external NLI model [6]. VASE doubles this to $2M$ generations plus $O(M^2)$ NLI comparisons per input, making it prohibitive for clinical deployment where real-time feedback is essential. We propose **Confidence-Evidence Bayesian Gain (CEBaG)**, which replaces stochastic sampling with deterministic log-probability analysis. CEBaG combines two complementary signals: *token-level predictive variance* (the standard deviation of per-token log-probabilities, elevated when the model fabricates content) and *evidence magnitude* (the absolute per-token log-probability shift when the image is present versus absent, capturing the strength of visual grounding). The resulting score, $\sigma \cdot (1 + |G|/L)$, is parameter-free, requires only one generation and two scoring passes, uses no external models, and is fully deterministic (Fig. 1). We position CEBaG specifically for answer-level hallucination detection in short-form medical VQA. Its two signals, the sequence-level evidence gain G and the global standard deviation of token log-probabilities, operate over the entire response and thus capture whether an answer, taken as a whole, is grounded in the image. This makes CEBaG well-suited to flagging holistic answer–image consistency in concise medical VQA.

Our contributions are: (1) We propose CEBaG, a deterministic answer-level hallucination detection method requiring three forward passes and no external models, compared to 20+ stochastic generations and an NLI model for SE/VASE. (2) We conduct an extensive evaluation on four MLLMs (MedGemma-4b, MedGemma-1.5-4b, HuatuoGPT-Vision-7B, LLaVA-Med-v1.5-7B) and three datasets (VQA-RAD, SLAKE, PathVQA), 16 settings in total. (3) CEBaG con-

sistently outperforms SE, VASE, and RadFlag [22], achieving the best AUC in 13 of 16 settings and improving over VASE by 8 AUC points on average.

2 Method

Given a medical image x_v, a textual question x_q, and a medical MLLM f, the model generates a response $r = f(x_v, x_q)$ via greedy or low-temperature decoding. Our goal is to determine whether r is a hallucination (a response that contradicts the visual evidence in x_v) without access to ground-truth answers. For evaluation, we follow prior work [16, 4] and use the GREEN model [20] to produce reference-based ground-truth labels.

2.1 Evidence Gain as Pointwise Mutual Information

The key quantity in CEBaG is the *evidence gain*: how much the model's belief in its own answer changes when it observes the image. Formally, we interpret the text-only probability $P(r \mid x_q)$ as the model's **language prior**, i.e., its belief about the answer based solely on the question and its pre-trained knowledge. The multimodal probability $P(r \mid x_v, x_q)$ represents the **posterior** belief after updating with the visual evidence x_v. We use "prior" and "posterior" in the information-theoretic sense: $P(r \mid x_q)$ represents the model's belief before observing visual evidence, and $P(r \mid x_v, x_q)$ its updated belief after conditioning on the image. This is distinct from Bayesian inference over model parameters; rather, G captures how the image updates the model's predictive distribution for a fixed set of parameters. The sequence log-probabilities are defined as:

$$\log P(r \mid x_v, x_q) = \sum\nolimits_{j=1}^{L} \log P(r_j \mid r_{<j},\, x_v,\, x_q), \tag{1}$$

$$\log P(r \mid x_q) = \sum\nolimits_{j=1}^{L} \log P(r_j \mid r_{<j},\, x_q). \tag{2}$$

The **evidence gain** G is the difference between the posterior and the prior:

$$G = \log P(r \mid x_v, x_q) - \log P(r \mid x_q). \tag{3}$$

By Bayes' rule, $P(r \mid x_v, x_q) = \frac{P(x_v \mid r, x_q) P(r \mid x_q)}{P(x_v \mid x_q)}$. Substituting this into Eq. (3):

$$G = \log \left(\frac{P(x_v \mid r, x_q) P(r \mid x_q)}{P(x_v \mid x_q)} \right) - \log P(r \mid x_q) = \log P(x_v \mid r, x_q) - \log P(x_v \mid x_q). \tag{4}$$

Since $\log P(x_v \mid x_q)$ is constant with respect to the response r, G is directly proportional to $\log P(x_v \mid r, x_q)$, the log-likelihood of the image given the response. Thus, G is the **Pointwise Mutual Information (PMI)** between the response and the image (conditioned on the question). A large positive G implies that the response makes the observed image highly probable (strong grounding), while a negative G suggests the response contradicts the visual evidence.

The **evidence magnitude** is a length-normalized visual influence measure:

$$E = \frac{|G|}{L}, \tag{5}$$

which captures the average absolute shift in belief per token. We use the magnitude $|G|$ because a strong visual influence (increasing or decreasing token probabilities) indicates attention to the image, whereas a lack of influence ($G \approx 0$) suggests the model is relying entirely on its language prior.

2.2 Confidence-Evidence Bayesian Gain

We propose CEBaG, a scoring function that combines the model's internal uncertainty with its sensitivity to visual evidence. We define the **token-level predictive variance**: $\sigma = \mathrm{std}_{j=1}^{L}\big[\log P(r_j \mid r_{<j}, x_v, x_q)\big]$, which measures the consistency of the model's confidence across the generated sequence. High σ indicates "confidence spikes" on function words interspersed with low-confidence content tokens, a pattern characteristic of hallucination. The CEBaG score is defined as:

$$\mathrm{CEBaG} = \sigma \cdot \big(1 + E\big). \tag{6}$$

This formulation can be interpreted as **posterior uncertainty weighted by visual information gain**. Hallucinations typically arise in two scenarios: (1) The model is uncertain about the content (σ is high); or (2) The visual evidence causes a large shift in belief (E is high) but fails to resolve the uncertainty (or contradicts the language prior).

The multiplicative term $(1 + E)$ amplifies the base uncertainty σ when the response is highly sensitive to the image, flagging cases where the model is "struggling" to reconcile visual evidence with its language prior. This amplification is realized as a multiplicative *gain* on σ rather than an independent additive term. The gain is monotone in both signals and preserves σ as its base, such that the score reduces to σ when the image leaves the prediction unchanged (E=0), so visual evidence can only sharpen, never override, the uncertainty estimate. Because it multiplies rather than adds, it also requires no coefficient to reconcile the differing scales of σ and E.

Importantly, this formulation is **hyperparameter-free**, requiring no hyperparameter tuning or normalization, making it robust across different models and datasets. Unlike SE and VASE, which require configuring the number of generations M, sampling temperature T, and (for VASE) the amplification ratio α, CEBaG applies the same fixed formula across all models and datasets.

3 Experiments

Datasets: We evaluate on three medical VQA benchmarks: *VQA-RAD* [14] (451 test samples, including 200 open-ended), a radiology VQA dataset covering chest X-rays, CT, and MRI; *SLAKE* [17], a bilingual medical VQA dataset with

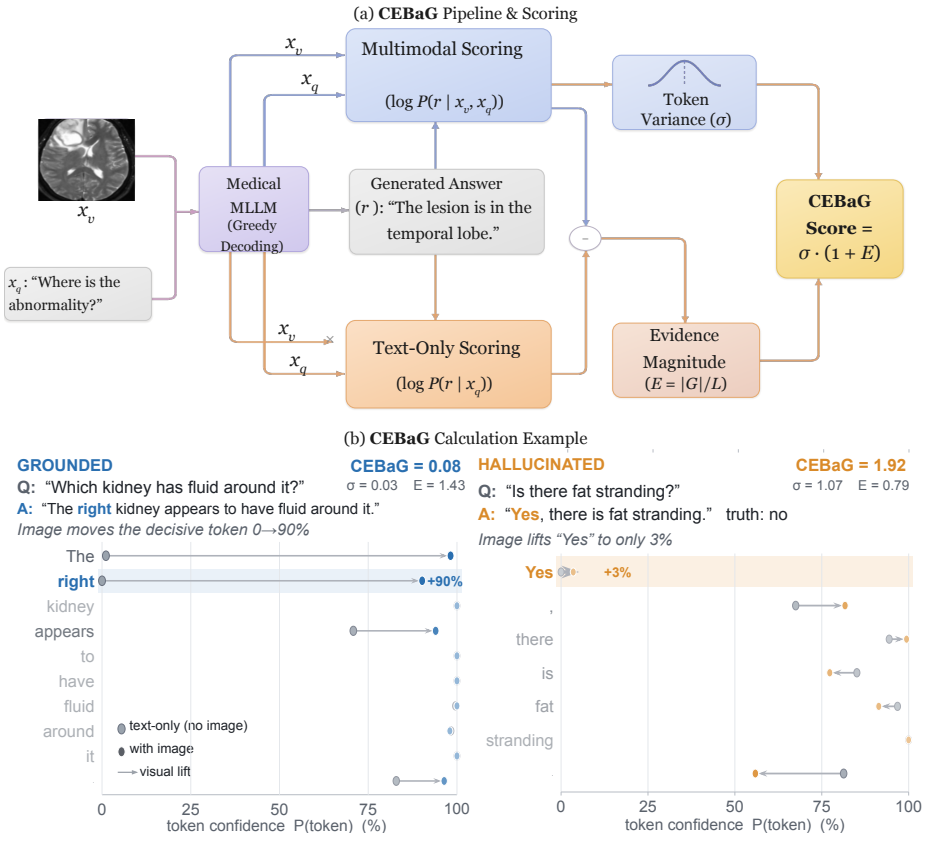

Fig. 1. Overview of CEBaG for hallucination detection in medical VQA. (a) The CEBaG pipeline: the model generates answer r, scored with and without the image to yield two signals: token-level predictive variance σ and evidence magnitude $E{=}|G|/L$. CEBaG combines these as $\sigma \cdot (1{+}E)$. (b) Per-token probabilities ($100\,P(\text{token})$) with vs. without the image, and CEBaG scores for a grounded and a hallucinated example.

semantically labeled knowledge (1061 test samples); and *PathVQA* [7] (1000 test samples), covering pathology images. For VQA-RAD, we report performance on both open-ended questions and all questions (including yes/no), following [16].

Models: Four medical MLLMs spanning two architecture families: *MedGemma-4b-it* and *MedGemma-1.5-4b-it* [21] (encoder–decoder), and *HuatuoGPT-Vision-7B* [2] and *LLaVA-Med-v1.5-Mistral-7B* [15] (LLaVA-style decoder-only).

Ground truth and metrics: Following the evaluation protocol established by VASE [16], we use the GREEN model [20] to generate reference-based ground-truth labels: a response is labeled as hallucinated if its GREEN score falls below 1.0. We adopt this protocol to ensure direct comparability with prior work and use the same metrics as [16]: the Area Under the ROC Curve (AUC) and the Area Under the GREEN Curve (AUG).

Baselines. We compare CEBaG against four state-of-the-art hallucination detection methods: (1) *AvgProb* [10], which uses the average token probability in the sequence; (2) *Semantic Entropy (SE)* [4], which measures uncertainty over semantic clusters of generated answers; (3) *Vision-Amplified Semantic Entropy (VASE)* [16], which enhances SE by generating additional answers conditioned on visually augmented images; and (4) *RadFlag* [22], which checks for consistency between the generated answer and the findings in the radiology report. We adhere to the official implementations and hyperparameters for all baselines. For SE and VASE, we use 10 generations (plus 10 augmented for VASE) and a temperature of 1.0, as specified in their respective papers.

Implementation details. For CEBaG, the generated answer r is produced via greedy decoding (temperature T=0.1). The two scoring passes (Eqs. 1–2) use teacher-forced forward passes; σ is computed from the same forward pass as $\log P(r \mid x_v, x_q)$ at no additional cost. For the text-only pass, visual input is removed at the interface level: for MedGemma, no image is included in the chat message and no pixel values are passed to the model; for LLaVA-style models (LLaVA-Med, HuatuoGPT), the `<image>` token is removed from the prompt template and the processor is called without image input. This ensures a clean text-only baseline across all architectures. All experiments were conducted on 8 NVIDIA H100 GPUs. Following the evaluation protocol of SE [4] and VASE [16], we report point estimates of AUC and AUG. Being deterministic, CEBaG avoids the baselines' sampling variance, giving exact per-answer scores.

3.1 Main Results

Table 1 presents the hallucination detection performance across all 16 experimental settings. CEBaG achieves the best AUC in **13 of 16** settings and ranks in the top two in 15 of 16. On average, CEBaG attains 67.9% AUC, outperforming VASE (59.6%), SE (57.4%), and RadFlag (57.8%) by substantial margins, with an improvement of +8.2 AUC points over the previous state-of-the-art VASE. Significantly, CEBaG achieves this without any hyperparameter tuning.

The gains are particularly pronounced on larger datasets: on PathVQA, CEBaG outperforms VASE by +10.5 (MedGemma), +6.5 (MedGemma-1.5), +28.8 (HuatuoGPT), and +17.2 (LLaVA-Med) AUC points. On SLAKE, the improvements are similarly large, with CEBaG surpassing VASE by +9.3 (MedGemma) and +7.1 (HuatuoGPT). On the smaller VQA-RAD datasets, the improvements remain consistent across most models, with CEBaG achieving the best AUC in 6 of 8 VQA-RAD settings. Furthermore, our method maintained stable detection performance across GREEN thresholds between 0.4 and 0.8, confirming its insensitivity to specific ground-truth boundaries.

3.2 Efficiency Analysis

Table 2 compares the computational requirements and resulting average performance for each method. With the standard setting M=10, SE requires 10 autoregressive generations plus $O(M^2)$=$O(100)$ pairwise entailment comparisons

Table 1. Hallucination detection performance: AUC (%)↑ and AUG (%)↑ across four medical MLLMs and three VQA datasets. Best and second-best among the detection methods (SE, VASE, RadFlag, CEBaG) are in **bold** and underlined, respectively. CEBaG achieves the best AUC in 13 of 16 settings.

	AvgProb [10]		SE [4]		VASE [16]		RadFlag [22]		**CEBaG (Ours)**	
Dataset	AUC	AUG	AUC	AUG	AUC	AUG	AUC	AUG	AUC	AUG
MedGemma-4b-it [21]										
VQA-RAD (Open)	39.9	48.0	**61.9**	**65.3**	60.8	63.3	58.4	63.6	61.0	64.8
VQA-RAD (All)	42.9	54.6	52.2	58.0	57.6	59.9	54.7	58.9	**60.8**	**64.6**
SLAKE	35.3	52.9	54.9	64.2	62.6	70.1	57.0	66.2	**71.9**	**75.9**
PathVQA	46.4	45.1	58.0	48.9	61.5	49.6	56.8	47.4	**71.9**	**59.0**
MedGemma-1.5-4b-it [21]										
VQA-RAD (Open)	69.1	47.4	53.2	37.2	60.1	40.3	54.5	39.3	**72.9**	**47.2**
VQA-RAD (All)	63.6	56.2	53.4	48.6	54.8	51.5	54.3	46.3	**66.9**	**58.2**
SLAKE	57.4	52.0	57.7	51.9	**59.5**	52.9	57.2	**53.0**	57.5	52.0
PathVQA	65.9	48.9	56.8	40.9	60.1	45.7	56.3	41.6	**66.6**	**48.8**
HuatuoGPT-Vision-7B [2]										
VQA-RAD (Open)	43.7	48.1	64.1	62.8	61.7	59.6	59.9	58.4	**70.4**	**63.1**
VQA-RAD (All)	39.3	52.2	58.3	65.1	58.9	66.1	60.0	63.5	**67.7**	**67.6**
SLAKE	30.5	38.4	57.7	54.9	59.9	58.0	60.6	54.2	**67.0**	**64.2**
PathVQA	30.5	28.7	53.6	38.2	53.2	39.5	55.5	38.9	**82.0**	**55.6**
LLaVA-Med-v1.5-Mistral-7B [15]										
VQA-RAD (Open)	62.9	38.6	63.8	**40.4**	**68.4**	40.2	65.8	40.1	65.8	39.1
VQA-RAD (All)	58.8	48.7	58.1	**50.7**	57.6	49.3	56.8	48.1	**62.1**	50.0
SLAKE	59.9	50.0	57.9	50.2	57.4	50.8	58.7	50.7	**63.8**	**52.3**
PathVQA	75.1	52.0	57.4	41.3	60.4	41.7	58.8	39.7	**77.6**	**53.1**
Average	51.3	47.6	57.4	51.2	59.6	52.4	57.8	50.6	**67.9**	**57.2**

per sample. VASE doubles the generation cost to 20 passes plus the same entailment overhead. RadFlag also requires M generations and entailment checks. All three methods additionally require loading and running a DeBERTa-v2-xlarge-MNLI model alongside the target MLLM, and all involve hyperparameters (M, sampling temperature T, and for VASE, the vision-amplified ratio α).

In contrast, CEBaG requires exactly **three forward passes**: one autoregressive generation and two teacher-forced scoring passes (with and without the image). No external model is needed, and no hyperparameters require tuning. The scoring passes are substantially cheaper than generation as they process fixed token sequences without autoregressive decoding. Despite its simplicity and efficiency, CEBaG achieves the highest average AUC (67.9%) and AUG (57.2%) across all settings, outperforming the computationally expensive baselines.

3.3 Ablation Study

We compare four variants: σ only (token-level variance); E only (evidence magnitude $|G|/L$); CEBaG, the parameter-free $\sigma \cdot (1 + E)$; and CEBaG$_\lambda$, an oracle that, per model-dataset pair, min-max normalizes G and a penalty term and

Table 2. Computational cost and performance. Despite an order of magnitude fewer forward passes, CEBaG achieves the highest average AUC and AUG.

Method	Forward passes	External model	Deterministic	Hyperparams	$\overline{AUC}$	$\overline{AUG}$
AvgProb [10]	1 gen.	None	Yes	None	51.3	47.6
SE [4]	M gen. + NLI	DeBERTa-MNLI	No	M, T	57.4	51.2
VASE [16]	$2M$ gen. + NLI	DeBERTa-MNLI	No	M, T, α	59.6	52.4
RadFlag [22]	M gen. + NLI	DeBERTa-MNLI	No	M, T	57.8	50.6
CEBaG (Ours)	1 gen. + 2 evals	None	Yes	None	**67.9**	**57.2**

Table 3. Ablation study: AUC (%)↑ of CEBaG components. σ only uses token-level variance alone; E only uses absolute per-token gain alone; CEBaG combines both (parameter-free); CEBaG$_\lambda$ adds a tunable weight (upper bound).

	AUC (%)				AUG (%)			
Setting	σ	E	CEBaG	CEBaG$_\lambda$	σ	E	CEBaG	CEBaG$_\lambda$
MedGemma-4b-it								
SLAKE	71.1	51.6	71.9	**73.5**	74.8	64.9	**75.9**	74.6
PathVQA	67.8	63.4	71.9	**77.5**	55.6	57.0	**59.0**	61.7
MedGemma-1.5-4b-it								
VQA-RAD (Open)	68.2	63.9	72.9	**75.7**	44.7	40.8	47.2	**47.0**
PathVQA	57.9	70.8	66.6	**77.8**	42.6	52.2	48.8	**56.3**
HuatuoGPT-Vision-7B								
PathVQA	80.6	67.3	**82.0**	82.5	55.3	48.3	**55.6**	55.5
VQA-RAD (All)	68.0	53.8	67.7	**68.6**	**67.5**	59.6	67.6	66.9
LLaVA-Med-v1.5-Mistral-7B								
VQA-RAD (All)	62.9	52.0	62.1	**66.2**	50.8	45.6	50.0	**54.0**
SLAKE	65.2	51.8	63.8	**68.8**	52.7	46.8	52.3	**55.3**
Average (all 16)	67.0	59.0	67.9	**71.9**	53.8	50.5	57.2	**58.8**

selects the gain sign, penalty term, and weight $\lambda \in [-5, 5]$ maximizing AUC, an in-sample configuration unattainable without test labels.

Table 3 reports the ablation; the bottom row averages all 16 settings. 1- Token-level variance dominates: σ-only achieves 67.0% average AUC, above all baselines (SE 57.4%, VASE 59.6%, RadFlag 57.8%). 2- Evidence magnitude is complementary: while E only (59.0%) trails σ only, CEBaG (67.9%) outperforms it in more than half of the settings, with notable gains on MedGemma-1.5 VQA-RAD Open (+4.7) and MedGemma PathVQA (+4.1). In the rest, $(1+E)$ lowers AUC, usually marginally (-0.3 on HuatuoGPT VQA-RAD All, -1.4 on LLaVA-Med SLAKE). Because σ and E combine without normalization, their relative influence varies across models; CEBaG keeps this factor because it raises the average and needs no tuning. 3- The tuned upper bound does not transfer. CEBaG$_\lambda$ reaches 71.9% average AUC, 4.0 points above CEBaG, entirely from per-setting λ selection: the optimal λ is unstable, ranging from -5 to $+2.75$. The best single λ recovers under one of these four points, and the tuned value is unattainable without test-label access.

4 Conclusion

We introduced CEBaG, a deterministic answer-level hallucination detection method for medical VQA combining token-level predictive variance and evidence magnitude. Unlike sampling-based approaches (SE, VASE) that need 10–20 stochastic generations and an external NLI model, CEBaG uses only three forward passes. Across four medical MLLMs and 16 settings, it achieves the best AUC in 13 and surpasses state-of-the-art VASE by 8 AUC points on average at much lower cost. Its efficiency, simplicity, and zero-configuration design make it practical for real-time clinical deployment; its main limitation is requiring white-box access to log-probabilities. Future work can extend CEBaG to black-box settings via output perturbation and to multi-turn clinical dialogue. Because both signals aggregate over the full response, CEBaG operates at the whole-answer level; extending the per-token gain into a span-level signal for long-form outputs like report generation is a natural next direction. GREEN-generated labels, used for comparability with prior work, also make validation on a clinician-annotated subset an important next step.

Acknowledgment: This work was partially supported by the NIH Grants AG089169, AG084471, Stanford HAI Hoffman-Yee Award & GCP Credits, and UIT. M. Asadi is supported by the Amazon AI PhD and the Stanford HAI Graduate Fellowships.

Disclosure of Interests. The authors have no competing interests to declare.

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., Gui, C., Ouyang, R., Gao, A., Chen, S., Chen, G.H., Wang, X., Zhang, R., Cai, Z., Ji, K., Yu, G., Wan, X., Wang, B.: HuatuoGPT-Vision, towards injecting medical visual knowledge into multimodal LLMs at scale. arXiv:2406.19280 (2024)
3. Cohen, R., Hamri, M., Geva, M., Globerson, A.: LM vs LM: Detecting factual errors via cross examination. In: EMNLP. pp. 12621–12640 (2023)
4. Farquhar, S., Kossen, J., Kuhn, L., Gal, Y.: Detecting hallucinations in large language models using semantic entropy. Nature **630**(8017), 625–630 (2024)
5. Gunjal, A., Yin, J., Bas, E.: Detecting and preventing hallucinations in large vision language models. In: AAAI. vol. 38, pp. 18135–18143 (2024)
6. He, P., Liu, X., Gao, J., Chen, W.: DeBERTa: Decoding-enhanced BERT with disentangled attention. In: ICLR (2021)
7. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: PathVQA: 30000+ questions for medical visual question answering. arXiv:2003.10286 (2020)
8. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., Liu, T.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Trans. Inf. Syst. (2025)
9. 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)

10. Li, Q., Geng, J., Lyu, C., Zhu, D., Panov, M., Karray, F.: Reference-free hallucination detection for large vision-language models. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 4542–4551 (2024)
11. Kossen, J., Han, J., Razzak, M., Schut, L., Malik, S., Gal, Y.: Semantic entropy probes: Robust and cheap hallucination detection in LLMs. arXiv:2406.15927 (2024)
12. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In: CVPR (2024)
13. Manevich, A., Tsarfaty, R.: Mitigating hallucinations in large vision-language models via language-contrastive decoding. In: ACL Findings. pp. 6008–6022 (2024)
14. 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), 180251 (2018)
15. 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** (2023)
16. Liao, Z., Hu, S., Zou, K., Fu, H., Zhen, L., Xia, Y.: Vision-amplified semantic entropy for hallucination detection in medical visual question answering. In: MICCAI (2025)
17. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: ISBI. pp. 1650–1654 (2021)
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: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:2402.00253 (2024)
20. Ostmeier, S., Xu, J., Chen, Z., Varma, M., Blankemeier, L., Bluethgen, C., Michalson, A.E., Moseley, M., Langlotz, C., Chaudhari, A.S., Delbrouck, J-B.: GREEN: Generative radiology report evaluation and error notation. arXiv:2405.03595 (2024)
21. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: MedGemma technical report. arXiv:2507.05201 (2025)
22. Zhang, S., Sambara, S., Banerjee, O., Acosta, J., Fahrner, J., Rajpurkar, P.: RadFlag: A black-box hallucination detection method for medical vision language models. arXiv:2411.00299 (2024)
23. 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)
24. 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:2405.08603 (2024)
25. 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 Inf. Sci. **67**(12), 220105 (2024)
26. 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)