Detecting Clinical Hallucinations in LVLMs via Counterfactual Visual Grounding Uncertainty

Xiao Song[1], Haonan Qin[1], Zhaoxu Zhang[1], Jiong Zhang[3], Yuqi Fang[1,2(✉)], and Caifeng Shan[1(✉)]

[1] School of Intelligent Science and Technology, Nanjing University, Suzhou, China
{yqfang, cfshan}@nju.edu.cn
[2] National Institute of Healthcare Data Science at Nanjing University, Nanjing University, Nanjing, China
[3] School of Biomedical Engineering, Nanjing University, Suzhou, China

Abstract. Large vision-language models (LVLMs) are increasingly used for clinical image understanding, yet they remain vulnerable to *hallucinations*–producing textual findings or attributes not supported by the image. We present a vision-traceable hallucination detection framework that audits arbitrary LVLM responses via visual evidence grounding, requiring neither modification nor internal access to the hidden states of LVLMs. Given an LVLM response, we extract visually verifiable entities and use a medical-domain-adapted Qwen-VL grounding verifier to localize each entity on the input image. To enhance the robustness of our detection method, we introduce a counterfactual entity perturbation method and estimate visual evidence uncertainty by contrasting factual and counterfactual grounding results. Specifically, we compute an entity-level uncertainty score from the positive confidence, counterfactual confidence, and their grounding overlap for binary hallucination decision-making. Experiments on multiple medical imaging modalities and LVLM backbones demonstrate that our method consistently improves hallucination detection performance over recent baselines, while providing interpretable localization evidence and strong cross-model transferability. Code and dataset are available at `https://github.com/Agentic-CliniAI/CounterVHD`.

Keywords: Hallucination Detection · Large Vision-Language Models · Uncertainty Estimation · Counterfactual · Visual Grounding .

1 Introduction

Medical large vision–language models (LVLMs) have demonstrated impressive capability in interpreting images and generating free-form responses for a wide range of clinical tasks, such as visual question answering, report generation,

X. Song and H. Qin—Contributed Equally.

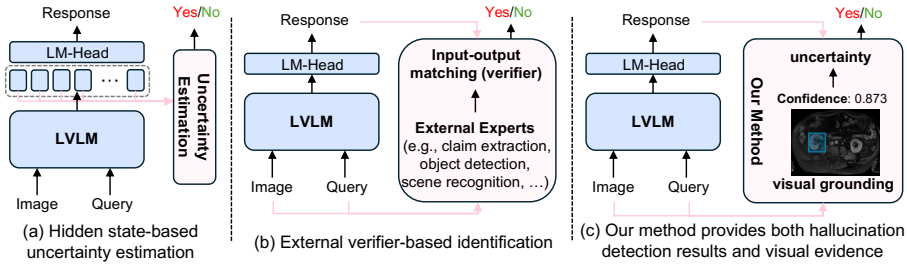

Fig. 1. Comparison of hallucination detection paradigms. (a) **Hidden state-based methods**: access LVLM's internal hidden states. (b) **External verifier-based methods**: rely on external expert models. (c) **Ours**: identifies hallucinations by aligning responses to visual regions, improving interpretability via visual evidence.

and decision support. Despite rapid progress, LVLMs remain prone to *hallucinations* [7,9,2,17,13]: they may describe findings, anatomy, or devices that are not supported by the input image or assign incorrect attributes (e.g., laterality, location, severity). In safety-critical medical scenarios, such errors can mislead clinicians and undermine trust, motivating reliable auditing tools.

Recent research on hallucination detection in LVLMs can be grouped into two paradigms [9], as depicted in Fig. 1 (a) and (b). (a) **Hidden state-based methods** [5,8,18] apply LVLM's internal hidden state for predictive uncertainty as a proxy for hallucination rate, where higher uncertainty signals higher likelihood of erroneous generation. While effective, their applicability is often constrained by the black-box nature of most commercial models, which do not grant users access to their internal states. (b) **External verifier-based methods** [2,6,15,10] deploy well-performed LVLMs or specifically trained expert models to verify whether the outputs align faithfully with both the query and visual evidence. However, beyond binary detection conclusions, explicitly localizing the corresponding visual evidence is essential for validating findings and ensuring diagnostic tracing, which is a critical capability that existing methods neglected.

To tackle this problem, in this paper, we propose a vision-aware hallucination detection method via Counterfactual-driven Visual Grounding Uncertainty Estimation (see Fig. 1 (c) and Fig. 2). Specifically, we extract entity mentions from responses and localize them using a Qwen-VL *visual grounding* verifier fine-tuned specifically for the medical domain. Furthermore, to enhance robustness of our method, we introduce a *counterfactual entity perturbation branch* that prompts the verifier to ground a hard negative counterpart that is semantically related yet spatially non-overlapping from the original factual entity. Subsequently, we derive a vision-aware uncertainty score by comparing the IoU-reweighted grounding confidences of factual and counterfactual bounding boxes, thereby enabling robust and accurate hallucination detection. Extensive experiments across diverse medical imaging modalities and LVLM backbones demonstrate that our

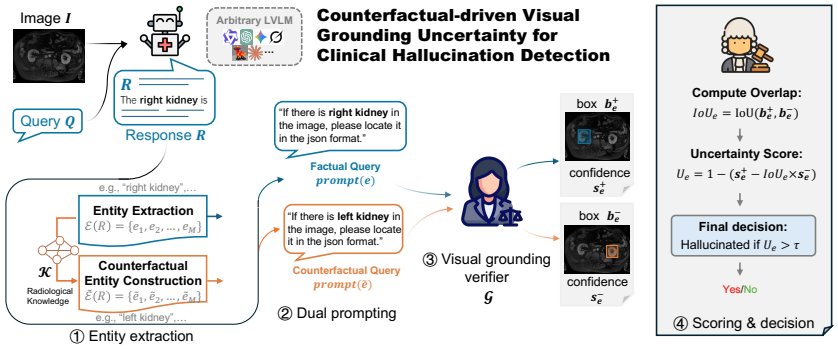

Fig. 2. Pipeline of the proposed Counterfactual-driven Visual Grounding Uncertainty Estimation method. ① Given a response R from an arbitrary LVLM, we extract entities $\mathcal{E}$ and construct counterfactual entities $\tilde{\mathcal{E}}$ using radiological knowledge. ② Constructing the factual and counterfactual queries. ③ A trained grounding verifier predicts bounding boxes (b_e^+, b_e^-) and confidence scores (s_e^+, s_e^-). ④ Uncertainty estimation and the final decision is made by comparing the aggregated score against threshold.

method achieves superior performance with robust generalizability over recent baselines, while our vision-aware design simultaneously provides interpretable visual evidence for clinical auditing. Our main contributions are threefold:

- We introduce a vision-aware hallucination detection method for LVLMs in clinical scenarios. By leveraging visual grounding, the method provides clinicians with both hallucination decisions and precise entity-level localization.
- We propose a counterfactual-driven uncertainty estimation method that identifies spurious evidence by contrasting factual and counterfactual grounding, ensuring robust hallucination identification.
- Experiments across diverse medical modalities and LVLM backbones demonstrate that our method achieves superior performance while ensuring robust generalizability and visual interpretability.

2 Methodology

2.1 Problem Setup and Overview

In this paper, we study plug-and-play detection of hallucinations in the responses of arbitrary LVLMs in clinical scenarios. Given an image I and a black-box LVLM $\mathcal{M}$, the model generates a response $R = \mathcal{M}(I, Q)$ to a user query Q. Our goal is to decide whether R contains *visual-grounded claims* that contradict the visual facts in I. To this end, we introduce an external visual grounding verifier $\mathcal{G}$ (instantiated by a fine-tuned Qwen3-VL-2B [1] model), which is decoupled from $\mathcal{M}$ and thus generalizes across arbitrary LVLMs.

The key idea is to convert hallucination detection into *visual evidence alignment*: As illustrated in Fig. 2, we extract entity mentions from R, ground them onto I to obtain factual evidence, and additionally construct *counterfactual* entity descriptions that should not visually overlap with the original entity. We then estimate a test-time uncertainty by contrasting the factual and counterfactual grounding results, yielding a hallucination probability.

2.2 Entity Extraction from LVLM Responses

We first extract a set of entity mentions from R corresponding to visually verifiable concepts in medical images (e.g., anatomy, lesions):

$$\mathcal{E}(R) = \{e_{\mathrm{m}}\}_{\mathrm{m}=1}^{\mathrm{M}}, \tag{1}$$

where M is the number of extracted entities. In practice, we prompt the GPT-4.1-mini model via API as a sophisticated text parser.

2.3 Visual Grounding Verifier

For each entity $e \in \mathcal{E}(R)$, we prompt the grounding model $\mathcal{G}$ to return a bounding box and a confidence score:

$$(b_e^+, s_e^+) = \mathcal{G}(I, \mathrm{prompt}(e)), \tag{2}$$

where $b_e^+ \in \mathbb{R}^4$ denotes the predicted box and $s_e^+ \in [0, 1]$ indicates grounding confidence. Specifically, the grounding prediction b_e^+ is obtained from a single deterministic inference of $\mathcal{G}$. To compute s_e^+, we consider two alternative algorithms: (a) **Monte-Carlo Sampling**: Under elevated temperature and stochastic sampling, $\mathcal{G}$ generates N diverse candidate boxes $\{b_e^{+(\mathrm{n})}\}_{\mathrm{n}=1}^{\mathrm{N}}$; s_e^+ is defined as their pairwise mIoU:

$$s_e^+ = \frac{2}{\mathrm{N}(\mathrm{N}-1)} \sum_{1 \le i < j \le \mathrm{N}} \mathrm{IoU}(b_e^{+(i)}, b_e^{+(j)}). \tag{3}$$

(b) **Internal Logit Probability**: Compute s_e^+ by averaging the softmax probabilities of the predicted box logits via the hidden state of $\mathcal{G}$. Note that different from hidden state-based methods that require internal LVLM access, our method leverages the internal states of $\mathcal{G}$ while remaining model-agnostic, thereby ensuring generalizability across various LVLMs.

2.4 Counterfactual Entity Construction

Grounding confidence alone is insufficient. Obstacles, including poor image quality or insufficient grounding capability of verifier, may result in misidentification of hallucinations. To enhance the robustness of our method, we therefore

introduce a counterfactual branch. Specifically, for each entity e, we generate a counterfactual description $\tilde{e}$ that is *intended not to be visually co-located* with e:

$$\tilde{e} = \mathcal{C}(e, \mathcal{K}), \tag{4}$$

where $\mathcal{C}$ is a counterfactual perturbation function supported by the radiological knowledge $\mathcal{K}$. We design $\mathcal{C}$ to satisfy the constraint of **non-overlapping**: $\tilde{e}$ refers to a mutually exclusive perturbation (e.g., contralateral side, different organ class, or negated attribute) that should not share the same visual evidence region with e. For instance, "left kidney" versus "right kidney", as shown in Fig. 2. This non-overlap guarantee is specifically built into $\mathcal{K}$'s construction: we apply a strict spatial exclusion rule that discards any entity whose bounding box could overlap with others on identical imaging slices. We then ground the counterfactual as:

$$(b_e^-, s_e^-) = \mathcal{G}(I, \text{prompt}(\tilde{e})). \tag{5}$$

2.5 Counterfactual–driven Uncertainty Estimation Algorithm

We first quantify the overlap between factual and counterfactual grounding:

$$\text{IoU}_e = \text{IoU}(b_e^+, b_e^-). \tag{6}$$

Intuitively, if the response is correct and the entity e is truly present, the factual grounding should be confident and localized, while the counterfactual should either have low confidence or localize elsewhere. Conversely, if the LVLM model hallucinates e or the grounding verifier $\mathcal{G}$ is confused by the aforementioned obstacles in Sec. 2.4, both prompts might locate the same spatial region, yielding high overlap that confuses the hallucination decision-making. To address this issue, we novelly combine these signals into an **IoU-reweighted uncertainty score** U_e, which directly measures the likelihood of hallucination:

$$U_e = 1 - (s_e^+ - \text{IoU}_e \cdot s_e^-). \tag{7}$$

This algorithm ensures U_e remains low when factual grounding is confident and spatially distinct from the counterfactual. However, if the LVLM hallucinates e or $\mathcal{G}$ is confused, it may confidently localize to spurious regions, risking missed detection. Here, the counterfactual branch acts as a safeguard: significant overlap penalizes s_e^+ via the $\text{IoU}_e \cdot s_e^-$ term, thereby increasing U_e to flag the hallucination. We then classify e as hallucinated if $U_e > \tau$, where τ is a tunable threshold.

3 Experiments

3.1 Experimental Setup

Datasets. We construct both training and testing datasets sourced from IMIS-Bench [4], a collection of 80 medical image segmentation datasets covering diverse anatomical regions, where this broad coverage significantly ensures

robust generalization. For **verifier training**, we extract bounding boxes from segmentation masks and form grounding queries in Qwen3-VL format, resulting in 33,266 pairs. For **hallucination detection testing**, we collect 1,904 image-report pairs (958 supporting, 946 hallucinated) generated by four LVLMs: MedGemma-27B [12], GPT-5.1 [11], Gemini-3-Flash [14], and Grok-4-Fast [16], where the hallucinations are induced via mutually exclusive counterfactual labels. This test set spans 1,129 CT and 775 MRI scans.

Baselines and Metrics. As external verifier-based detection is still in its nascent stage, competing methods are limited. Many existing approaches require specialized training on curated datasets or rely on closed-source parameters, rendering direct reproduction infeasible. Consequently, we compare our approach with leading hallucination detection methods, including a GPT-5.1-based prompting method, UniHD [3] and Faithscore [6], where the former two methods are pure API-based and Faithscore relies on trained external expert. We report the hallucination detection rate (HDR/recall) and precision in Tab. 1.

Implementation Details. We fine-tune a Qwen3-VL-2B model to act as the verifier with LoRA. Specifically, the model is optimized with the AdamW optimizer, using a learning rate of 0.0001 and a weight decay of 0.0001. A CosineLR scheduler is adopted, with the warmup phase accounting for 3% of the total training steps. The per-device batch size is 16, and gradients are accumulated over 8 steps. Training is conducted for 5 epochs on 2× NVIDIA A100 GPUs.

3.2 Main Experimental Results

To evaluate hallucination detection performance and plug-and-play reusability, we benchmark our detector across multiple LVLM backbones (spanning both general and medical domains) and diverse imaging modalities (CT and MRI), as detailed in Tab. 1. Our method demonstrates robust detection capabilities and consistently outperforms recent baselines across the evaluated backbones and modalities. Notably, on MRI data, our approach achieves superior results on both detection rate and precision. While UniHD reports a higher detection rate on the CT modality, it exhibits the lowest precision among all compared methods, indicating a high rate of false positives (i.e., misidentifying non-hallucinated samples). In contrast, our method maintains the highest precision across all modalities, providing a more reliable balance between sensitivity and accuracy. Therefore, the consistent performance across multiple LVLMs validates the plug-and-play capability of our method without requiring access to internal states, which significantly enhances the applicability in real-world clinical settings. Moreover, our method is more lightweight than FaithScore (2B vs. 13B), demonstrating superior computational efficiency.

3.3 Ablation Studies

Effect of Different Confidence Scoring Mechanisms. We compare the proposed two confidence scoring mechanisms to evaluate their effectiveness in hallucination detection: **MCS** and **ILP**. As shown in Fig. 3 (a), both methods

Table 1. Main experimental results: hallucination detection rate (HDR) and precision (Prec.) across LVLM backbones and imaging modalities (%).

Methods	Param.	MedGemma		GPT-5.1		Gemini-3		Grok-4	
		HDR	Prec.	HDR	Prec.	HDR	Prec.	HDR	Prec.
CT									
GPT-5.1-based	-	67.84	61.57	71.07	63.43	71.13	64.71	56.74	53.49
UniHD [3]	-	94.33	56.84	82.01	58.16	78.87	58.95	87.86	52.56
Faithscore [6]	13B	58.16	61.19	48.92	64.76	52.11	64.35	55.00	65.81
Ours	2B	63.12	64.03	61.15	68.00	62.68	60.14	60.00	59.57
MRI									
GPT-5.1-based	-	65.64	59.33	80.20	75.00	77.49	69.77	68.75	62.00
UniHD [3]	-	91.58	48.33	65.66	45.46	79.79	45.18	86.46	47.43
Faithscore [6]	13B	46.32	67.69	44.44	65.67	34.04	66.67	41.67	63.49
Ours	2B	95.79	91.92	91.93	92.00	96.81	90.10	95.83	93.88

achieve high performance, with AUC values of 0.8811 and 0.8763. The MCS method slightly outperforms the ILP, suggesting that evaluating the semantic consistency of multiple outputs is more robust for detecting hallucinations than relying solely on the model's internal logit probability based confidence scores.

Effect of Counterfactual–driven Uncertainty Algorithm (CUA). To investigate the contribution of each component within our proposed CUA framework, we conduct an ablation study as illustrated in Fig. 3 (b). The results demonstrate that the inclusion of counterfactual (CF) branch is critical for both scoring mechanisms. Specifically, removing the CF component leads to a severe performance degradation and a notable score-direction inversion for the MCS: the raw AUC drops to 0.3572, which falls below the random guess level. This below-random performance indicates an anti-correlation between the predicted scores and the ground-truth labels. However, when applying a sign flip to correct the prediction direction, the MCS-only baseline achieves a reasonable standalone AUC of 0.6428. This mathematically confirms that while the MCS inherently possesses discriminative capability, the CF branch is indispensable for rectifying the prediction direction and substantially boosting the overall AUC from 0.6428 to 0.8811. Meanwhile, for the ILP, removing the CF component results in a direct performance decrease from 0.8763 to 0.8404. This suggests that CF samples effectively expose the model's underlying uncertainty by creating challenging decision boundaries. Furthermore, we evaluate the role of IoU-based method that reweights the confidence of CF by computing the overlapping of two branches. Removing this filtering step (w/o IoU) results in a moderate decrease in AUC (to 0.8595 for MCS and 0.8539 for ILP), which confirms that filtering out spatially irrelevant samples helps refine the reliability of the uncertainty scores.

Sensitiveness to Threshold τ. We further analyze how the selection of the threshold τ impacts the hallucination detection performance. As shown in Fig. 3 (c), as τ increases from 0.1 to 0.9, the hallucination detection rate for both MCS and ILP exhibits a steady decline. Conversely, the precision of both methods shows a consistent upward trend, reaching its peak at higher τ values. This

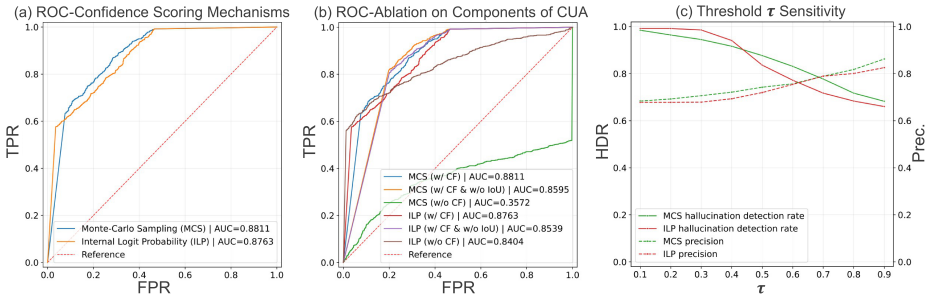

Fig. 3. Ablation studies on hallucination detection. (a) Performance comparison of two confidence scoring mechanisms: Monte-Carlo Sampling (MCS) and Internal Logit Probability (ILP). (b) Ablation analysis of the components in Counterfactual-driven Uncertainty Estimation Algorithm (CUA), highlighting the impact of the counterfactual branch (w/CF vs. w/o CF) and IoU-based reweighting (w/CF vs. w/CF & w/o IoU). (c) Sensitivity analysis of the threshold τ with respect to HDR and Prec.

reflects a typical trade-off in anomaly detection: a lower threshold allows for the identification of more potential hallucinations, while a higher threshold ensures that the identified instances are more likely to be genuine hallucinations. The intersection of the detection rate and precision curves around $\tau = 0.6$ suggests an optimal balance point for general applications.

3.4 Qualitative Analysis

Fig. 4 shows two hallucination detection cases. Specifically, for the non-hallucinated response "a tumor in the brain" in Fig. 4(a), our method performs factual and counterfactual grounding. The verifier accurately localizes the tumor (blue box) with high confidence (0.839), while the counterfactual query "a normal brain" yields no detection, consistent with pathological fact. This contrast produces a low uncertainty score (0.161), confirming factual accuracy. In Fig. 4(b), for "a normal right kidney", the verifier fails confident localization (0.536), whereas the counterfactual "thoracic cavity" yields high confidence (0.954). This discrepancy gives $U_e = 0.815$, indicating hallucination. Beyond binary detection, our method provides interpretable visual evidence, justifying detection and facilitating targeted rectification for precise response generation.

4 Conclusion

We propose a plug-and-play, vision-traceable clinical hallucination detection method for LVLMs via visual grounding experts. To enhance robustness, we introduce a counterfactual entity perturbation branch for uncertainty estimation.

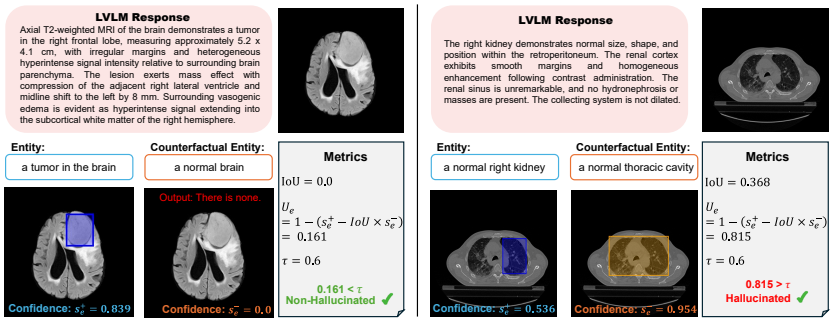

Fig. 4. Visualization of hallucination detection. (a) Non-hallucination case where factual branch detects grounded visual evidence and makes correct decision. (b) Hallucination case missed by factual branch but successfully detected by our algorithm.

This enables robust uncertainty estimation, effectively distinguishing hallucinations by suppressing spurious visual alignments from a single factual branch. Experiments across medical imaging modalities and LVLM backbones demonstrate superior detection performance and provide clinicians with interpretable, entity-level visual localization evidence. Future work will integrate this into hallucination mitigation, extend to complex spatiotemporal relationships in medical imaging, and overcome static, spatially exclusive priors by developing dynamic knowledge bases for visually entangled concepts. Ultimately, this research explores ways for reliable, safety-critical multimodal AI in the clinical domain.

Acknowledgments. This study was funded by the New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123502).

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

References

1. Bai, S., Cai, Y., Chen, R., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
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, X., Wang, C., Xue, Y., Zhang, N., Yang, X., Li, Q., Shen, Y., Liang, L., Gu, J., Chen, H.: Unified hallucination detection for multimodal large language models. In: ACL. pp. 3235–3252 (2024)
4. Cheng, J., Fu, B., Ye, J., Wang, G., Li, T., Wang, H., Li, R., Yao, H., Cheng, J., Li, J., et al.: Interactive medical image segmentation: A benchmark dataset and baseline. In: CVPR. pp. 20841–20851 (2025)

5. Hardy, R., Kim, S.E., Rajpurkar, P., et al.: Rextrust: A model for fine-grained hallucination detection in ai-generated radiology reports. In: AAAI Bridge Program on AI for Medicine and Healthcare. pp. 173–182 (2025)
6. Jing, L., Li, R., Chen, Y., Du, X.: Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. In: Findings of EMNLP. pp. 5042–5063 (2024)
7. Khanal, B., Pokhrel, S., Bhandari, S., Rana, R., Shrestha, N., Gurung, R.B., Linte, C., Watson, A., Shrestha, Y.R., Bhattarai, B.: Hallucination-aware multimodal benchmark for gastrointestinal image analysis with large vision-language models. In: MICCAI. pp. 235–245. Springer (2025)
8. Li, Q., Geng, J., Lyu, C., Zhu, D., Panov, M., Karray, F.: Reference-free hallucination detection for large vision-language models. In: Findings of EMNLP. pp. 4542–4551 (2024)
9. 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. pp. 669–679. Springer (2025)
10. Liao, Z., Hu, S., Zou, K., Jin, M., Zhang, Y., Fu, H., Zhen, L., Xia, Y.: Univrse: Unified vision-conditioned response semantic entropy for hallucination detection in medical vision-language models. arXiv preprint arXiv:2503.20504 (2025)
11. OpenAI: Introducing GPT-5 (Aug 2025), `https://openai.com/zh-Hans-CN/index/introducing-gpt-5`
12. 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 preprint arXiv:2507.05201 (2025)
13. Song, X., Liu, J., Liu, Y., Li, Y., Lei, W., Wang, R.: Rethinking radiology report generation via causal inspired counterfactual augmentation. In: ACM BCB. pp. 1–10 (2024)
14. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
15. Wang, J., Wang, Y., Xu, G., Zhang, J., Gu, Y., Jia, H., Wang, J., Xu, H., Yan, M., Zhang, J., et al.: Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. arXiv preprint arXiv:2311.07397 (2023)
16. xAI: Grok 4 (Jul 2025), `https://x.ai/news/grok-4`
17. Xiao, W., Huang, Z., Gan, L., He, W., Li, H., Yu, Z., Shu, F., Jiang, H., Zhu, L.: Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. In: AAAI. vol. 39, pp. 25543–25551 (2025)
18. Zou, K., Bai, Y., Liu, B., Chen, Y., Chen, Z., Zhou, Y., Yuan, X., Wang, M., Shen, X., Cao, X., et al.: Uncertainty-aware medical diagnostic phrase identification and grounding. IEEE Trans. Pattern Anal. Mach. Intell. (2025)