Reasoning Trace Divergence: An Empirical Signal for Trustworthy Black-Box MLLMs in Histopathology Classification

Anastasiia Kazmina[1], Constantin Venhoff[2], Bryan Wong[1], Sungrae Hong[1], Yutong Xie[3], and Mun Yong Yi*[1]

[1] Korea Advanced Institute of Science and Technology, Daejeon, South Korea
{anastasiiaka, munyi}@kaist.ac.kr
[2] University of Oxford, Oxford, United Kingdom
[3] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates

Abstract. Multimodal Large Language Models (MLLMs) offer transformative potential for medical diagnostics, yet their tendency to generate plausible but ungrounded hallucinations remains a critical barrier. In black-box settings, where internal model states are inaccessible, existing uncertainty quantification methods like visual-induced agreement rely on final label consistency, which often fails to capture systematic overconfidence. In this work, we move from visual stability to reasoning stability, identifying a failure mode where models maintain consistent labels despite divergent and fabricated logic. We introduce Reasoning Trace Divergence (RTD), a novel empirical signal that exposes these latent hallucinations by probing the stability of the model's logical paths under stochastic visual perturbations. Evaluating this signal on two patch-level histopathology datasets using Gemini and MedGemma, we demonstrate that reasoning instability is a robust predictor of grounding failures. By filtering for high reasoning stability, we identify a safe operating zone that significantly improves classification reliability, providing a principled direction for trustworthy black-box medical AI. The code is available at https://github.com/anastlibr99/reasoning-trace-divergence-blackbox-mllms-histopath.

Keywords: Uncertainty Quantification · Reasoning Stability · Computational Pathology.

1 Introduction

Multimodal Large Language Models (MLLMs) enable the analysis of multidimensional data [2,12], a capability essential in medicine where diagnostic insights are distributed across heterogeneous sources such as images and text. While these models offer transformative potential, they are prone to generating plausible but

* Corresponding Author

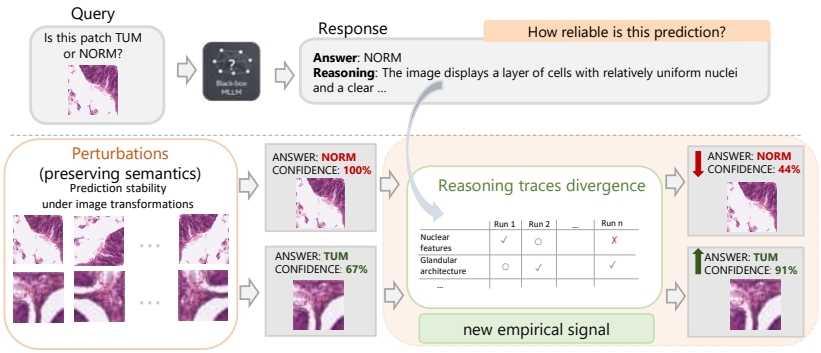

Fig. 1. Overview of the UVR Framework. UVR quantifies uncertainty by auditing reasoning equivariance across D_4 transformations. The pipeline calibrates MLLM confidence by lowering the scores for incorrect answers (Red) and increasing confidence for correct ones (Green).

wrong content [17,18]. This raises a critical question for high-stakes clinical domains: how can we evaluate the reliability of a prediction? This challenge is exacerbated in black-box settings [1,16], where massive scale or proprietary APIs preclude access to internal model states, gradients, or logits.

Hallucinations were initially formalized in text-only Large Language Models (LLMs) [8], identifying root causes such as noisy training data, knowledge gaps, or data fabrication [11]. A prominent mitigation strategy is semantic entropy [5], which quantifies uncertainty by measuring consistency across semantically equivalent responses. This concept was further extended by introducing knowledge-preserving semantic transformations to the input. Adapting these techniques to MLLMs requires a more nuanced understanding of cross-modal interactions [13]. While defining semantic-preserving transformations for multimodal inputs has shown improvement over baseline entropy [19,20], clinical applications remain complex. In medical imaging, semantic-preserving distortions differ significantly from the general domain; in histopathology specifically, where small morphological features determine malignancy, the grounding of reasoning in specific visual tokens is critical.

While recent work has demonstrated that vision-amplified semantic entropy can effectively detect hallucinations in open-ended medical Visual Question Answering (VQA) [10,21], the efficacy of these techniques in structured classification tasks remains under-explored. Although classification is often treated as a subset of closed-set VQA, the transition to histopathology introduces a unique challenge: the reliance on few-shot In-Context Learning (ICL) to adapt generalist MLLMs to specialized domains [3,7]. This shift is critical because while few-shot prompting significantly improves top-line accuracy, it often induces a calibration-accuracy gap where models produce increasingly coherent reasoning traces that mask systematic overconfidence or latent hallucinations. Consequently, there is an urgent need to analyze the uncertainty signals inherent in these generated traces. Current black-box metrics, which rely solely on final label consistency, fail to capture the subtle failures in reasoning that few-shot settings can introduce,

leaving a significant gap in our ability to detect failures in high-stakes diagnostic contexts.

To address this, we analyze the failure modes of semantic uncertainty in zero-shot and few-shot histopathology. We define a novel empirical signal that moves beyond token-level entropy to consider the divergence of the model's reasoning traces under stochastic visual perturbation (see Fig.1). We evaluate this signal across both general-purpose (Gemini 2.5 Flash Lite [16]) and domain-specific (MedGemma 1.5 4b [14]) MLLMs using two histopathology patch datasets in both zero- and few-shot settings using ICL. Our main contributions are:

- We identify **systematic overconfidence** as a primary failure mode in black-box medical MLLMs, where standard semantic entropy fails to capture decoupled reasoning.
- We introduce **Reasoning Trace Divergence (RTD)** as an empirical signal that exposes these failure patterns by measuring the stability of the model's logic under visual pressure.
- We demonstrate the utility of this signal by identifying a **safe operating zone** for clinical decision support; filtering for high reasoning stability effectively isolates correct predictions and improves the reliability of the diagnostic pipeline.

2 Methodology

2.1 Task Formulation

In this section, we define a histopathology patch classification task with a discrete label set, under both zero-shot and few-shot settings. Let $x_i \in \mathcal{X}$ denote the i-th test image, for $i = 1, \ldots, N$. $\mathcal{Y}$ is a finite set of possible labels the image may belong to. $\mathcal{Y} = \{c_1, c_2, \ldots, c_C\}$ where C is the total number of classification labels. Let $y_i \in \mathcal{Y}$ be the ground truth label for x_i.

The complete model input is encapsulated in a multimodal classification query $Q_i = (x_i, s, u_E)$ where s denotes a system prompt defining the experimental constraints and output formatting requirements, and u_E represents the user prompt. The user prompt is an ordered sequence consisting of a set of E labeled demonstrations for every class from $\mathcal{Y}$ and a task instruction u^{inst}: $u_E = \left(\{(x_e, y_e)\}_{e=1}^{E}, u^{\text{inst}}\right)$. The setting $E = 0$ corresponds to zero-shot inference, while $E > 0$ denotes few-shot learning.

A MLLM parameterized by θ produces a class label prediction and reasoning text: $\hat{y}_i, r_i = f_\theta(Q_i; \tau)$ where $Q_i = (x_i, s, u_E)$ denotes the model input, $\hat{y}_i \in \mathcal{Y}$ is the predicted class label, r_i denotes the generated reasoning, and τ is the decoding temperature.

Our primary goal is to evaluate predictive certainty within black-box settings where internal model probabilities are inaccessible. To achieve this, we first establish a deterministic reference for each test image x_i. Given the input query $Q_i = (x_i, s, u_E)$ the model produces a reference prediction $\hat{y}_i = f_\theta(Q_i; \tau = 0)$. We then assess the reliability of this prediction by generating an empirical distribution $p_i(c)$ under various conditions.

2.2 Semantic-Preserving Perturbations

To more robustly probe the model's knowledge, we move beyond simple noise and consider perturbations to the input manifold. Not all perturbations of input can lead to meaningful uncertainty results that reflect the true model uncertainty about the specific task. Following the definition of knowledge-preserving perturbations for language tasks in [6] we define semantic-preserving perturbations for a query Q_i as follows.

Definition 1 (Semantic-Preserving Perturbation). *Let Q_i denote the model query corresponding to image x_i, including the image and associated prompts. A perturbation operator δ is said to be semantic-preserving if*

$$K(\delta(Q_i)) = K(Q_i), \tag{1}$$

where $K(\cdot)$ denotes the decision-relevant knowledge required to determine the true class label y_i.

In the analysis of histopathological patches, the orientation of morphological structures is entirely stochastic. Consequently, diagnostic features such as gland architecture or serration patterns can be observed at any angle. This inherent property ensures that the D_4 transformation group [15], consisting of eight rotations and reflections, functions as a set of strictly semantic-preserving perturbations. For any patch x_i, the underlying diagnostic knowledge $K(x_i, s, u_E)$ remains invariant regardless of the spatial transformation $\delta \in D_4$. We exploit this property to define Visual Entropy (H_{vis}), which quantifies the predictive instability across these eight symmetric views. Formally, $H_{vis} = -\sum_{c \in \mathcal{Y}} p(c) \log p(c)$, where $p(c)$ is the empirical label frequency across the 8 predictions. A high H_{vis} signifies a failure in visual equivariance, indicating that the model's internal representations are sensitive to non-semantic spatial shifts.

2.3 Reasoning Trace Divergence

A distinct advantage of MLLMs is their capacity to generate natural language justifications alongside categorical predictions, providing a dual-stream signal for uncertainty quantification. We leverage this to construct a higher-fidelity metric of model instability. However, as raw justifications r_i are unstructured, we utilize a secondary LLM to perform Semantic Feature Mapping, converting free-form text into structured Reasoning Traces ($\mathcal{T}$). For a given patch x and its D_4 transformations, a meta-analytical prompt first identifies m unique morphological features $F = \{f_1, \ldots, f_m\}$ present across all justifications. Note that F emerges from the model's own justifications, not a predefined ontology; alignment with descriptors like glandular architecture is observational, not expert-verified. Each r_i is then mapped to a feature vector $\boldsymbol{v}_i \in \mathcal{S}^m$, where the state space $\mathcal{S} = \{1, 0, -1\}$ represents the presence, omission, or explicit negation of a feature, respectively. This results in a structured logical state that allows

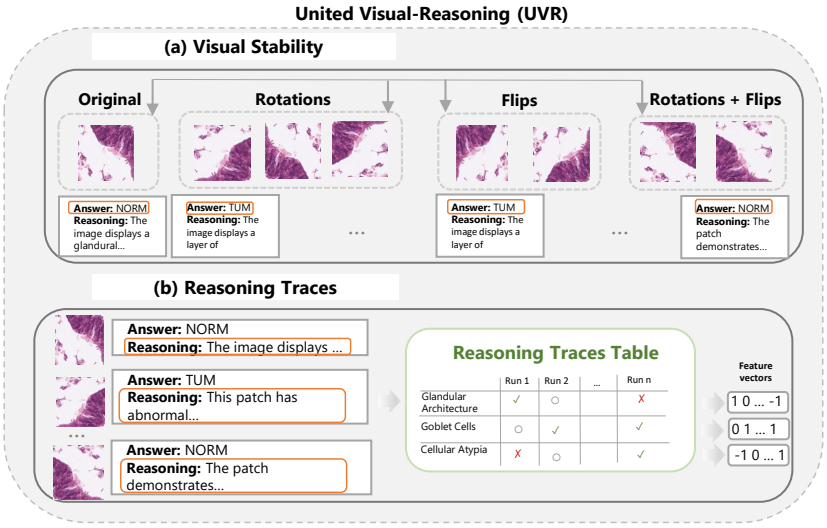

Fig. 2. UVR Framework. (a) Image-based D_4 perturbations evaluate model equivariance. (b) Multimodal reasoning traces are distilled into a Semantic Feature Matrix $(1, 0, -1)$ to compute logical uncertainty $\mathcal{H}_{RT}$.

for a precise mathematical comparison of the model's internal consistency. The Reasoning Trace is defined as the structured logical state:

$$\mathcal{T}_i = [v_{i,1}, v_{i,2}, \ldots, v_{i,m}] \tag{2}$$

We quantify logic-level instability using the Average Reasoning Entropy ($\mathcal{U}_{sem}$). For each feature f_j, we calculate the Shannon entropy $H(f_j)$ across the D_4 transformations. The average semantic uncertainty is defined as:

$$\mathcal{H}_{RT} = \frac{1}{m} \sum_{j=1}^{m} H(f_j), \tag{3}$$

where $H(f_j) = -\sum_{k \in \mathcal{S}} p_k \log p_k$. To capture the interaction between decision-level and logic-level instability, we employ a clipping operation to normalize entropy values into a strictly bounded $[0, 1]$ probability space, preventing extreme divergence from disproportionately skewing the joint metric. We propose the United Visual-Reasoning (UVR) certainty score:

$$\mathcal{C}_{UVR} = \text{clip}(1 - H_{vis}, 0, 1) \cdot \text{clip}(1 - \mathcal{H}_{RT}, 0, 1) \tag{4}$$

Fig. 2 demonstrates this two-level approach.

3 Results

3.1 Experimental Setup

Datasets and Sampling. We evaluate the proposed uncertainty estimation methods on two benchmark histopathology patch classification datasets: PatchCamelyon (PCam) [4] and NCT-CRC-100K [9]. For PCam, which targets the

Table 1. AUROC (%) for Gemini 2.5 Flash Lite and MedGemma 1.5 4B. Mean ± std over three few-shot samplings (seeds 3, 13, 42). VASE is omitted for MedGemma due to safety-induced refusals under hard perturbations.

Model	Dataset	Score	0-Shot	3-Shot	6-Shot
Gemini 2.5 [16]	CRC-100K [9]	Sampling Entropy	58.9	56.8 ± 0.2	61.0 ± 1.0
		RadFlag [21]	58.7	56.7 ± 0.4	60.9 ± 1.0
		VASE [10]	56.8	60.0 ± 4.1	65.6 ± 7.5
		Ours (UVR)	**66.1**	**67.0 ± 1.5**	**67.9 ± 6.3**
	PatchCamelyon [4]	Sampling Entropy	64.0	57.1 ± 2.8	59.0 ± 5.1
		RadFlag [21]	63.8	57.1 ± 2.8	59.0 ± 5.1
		VASE [10]	56.8	46.2 ± 0.0	52.1 ± 6.3
		Ours (UVR)	**66.0**	**64.0 ± 8.9**	**66.3 ± 9.4**
MedGemma 4B [14]	CRC-100K [9]	Sampling Entropy	52.7	46.2 ± 5.9	46.3 ± 2.5
		RadFlag [21]	55.3	46.2 ± 5.9	46.0 ± 2.1
		Ours (UVR)	**70.2**	**72.0 ± 1.6**	**74.1 ± 2.3**
	PatchCamelyon [4]	Sampling Entropy	60.0	59.6 ± 2.9	59.6 ± 2.3
		RadFlag [21]	50.4	62.2 ± 3.4	58.0 ± 8.6
		Ours (UVR)	**60.1**	**63.8 ± 1.6**	53.0 ± 3.1

detection of metastatic breast cancer in lymph node patches, we utilize a binary classification task consisting of two classes: Metastasis and No Metastasis. We randomly sample 150 patches per class from the official test set for evaluation, while few-shot demonstration examples are drawn from the validation split. For NCT-CRC-100K, which comprises 100,000 H&E patches across nine tissue types, we constrain the task to a binary classification between tumor epithelium (TUM) and normal colon mucosa (NORM). We randomly sample 150 patches per class from the validation split for our evaluation cohort and draw few-shot demonstration examples from the training split. This sampling strategy ensures that the in-context examples and evaluation cases remain strictly separated across both datasets. To analyze the impact of context on reasoning stability, we evaluate the model in zero-shot and few-shot settings using $E \in \{0, 3, 6\}$ demonstration examples per class. Few-shot demonstration sets are sampled independently three times using random seeds {7, 13, 42} to account for variability introduced by in-context example selection. All reported few-shot results are averaged across these three samplings.

Model Selection and Prompting Strategy. To evaluate generalizability across architectures, we use two state-of-the-art MLLMs: the proprietary Gemini 2.5 Flash Lite (representing high-scale commercial APIs) and the open-weights MedGemma 1.5 4B (representing domain-specific medical models) [14]. For the generation and analysis of reasoning traces in both cases, we utilize Gemini 2.5 Flash Lite [16] by prompting it to extract and process the reasoning from the corresponding models. This approach allows us to consistently capture the logical divergence of the models' explanations for our reasoning stability metrics.

Evaluation Methods. We assess the reliability of a reference prediction $\hat{y}_i = f_\theta(Q_i; \tau = 0)$ using the Area Under the Receiver Operating Characteristic (AUROC) curve. In this framework, AUROC measures the probability that a correctly classified sample is assigned a higher certainty score than an incorrect one

Table 2. Top-10% Accuracy (%) for Gemini 2.5 Flash Lite and MedGemma 1.5 4B. Mean ± std over three few-shot samplings (seeds 7, 13, 42).

Model	Dataset	Score	0-Shot	3-Shot	6-Shot
Gemini 2.5 [16]	CRC-100K [9]	Sampling	76.7	67.8 ± 5.1	68.9 ± 6.9
		RadFlag [21]	63.3	71.1 ± 5.1	74.4 ± 6.9
		VASE [10]	63.3	81.1 ± 1.9	92.2 ± 13.5
		Ours (UVR)	**86.2**	**89.7 ± 3.5**	90.8 ± 7.2
	PatchCamelyon [4]	Sampling	64.0	57.1 ± 2.8	59.0 ± 5.1
		RadFlag [21]	68.4	56.9 ± 3.7	60.9 ± 4.9
		VASE [10]	56.8	46.2 ± 0.0	52.1 ± 6.3
		Ours (UVR)	66.0	**64.0 ± 8.9**	**66.3 ± 9.4**
MedGemma 4B [14]	CRC-100K [9]	Sampling	52.7	46.2 ± 5.9	46.3 ± 2.5
		RadFlag [21]	55.3	46.2 ± 5.9	46.0 ± 2.1
		Ours (UVR)	**70.2**	**72.0 ± 1.6**	**74.1 ± 2.3**
	PatchCamelyon [4]	Sampling	90.0	78.9 ± 9.6	70.0 ± 14.1
		RadFlag [21]	90.0	78.9 ± 9.6	70.0 ± 14.1
		Ours (UVR)	**93.1**	73.6 ± 7.2	68.0 ± 5.7

Table 3. Ablation study of visual and reasoning stability components on MedGemma 4B (CRC-100K, 6-Shot). Our final method (bottom row) combines both to achieve peak performance.

Visual Stability	Reasoning Stability	AUROC (%)	Acc@10% (%)
✗	✗	46.31	70.00
✓	✗	71.91	80.00
✗	✓	64.31	87.93
✓	✓	**74.07**	**93.10**

(hallucination). To ensure comparability, all entropy measures are normalized into certainty signals $\mathcal{C} \in [0, 1]$ as $\mathcal{C} = 1 - \text{clip}(H, 0, 1)$. Beyond global performance, we evaluate the utility of these signals for **reliable decision support**. Specifically, we define a safe operating zone by filtering for the **top 10% most confident predictions** according to each metric. We then calculate the classification accuracy within this high-certainty subset to determine which signal most effectively isolates the model's most trustworthy outputs. Crucially, this requires no ground-truth labels, so the safe operating zone applies directly to unseen samples. AUROC is threshold-independent; Top-10% accuracy reflects a clinically relevant operating point. Broader calibration curves are left to future work.

3.2 Experimental Results

We evaluate our proposed UVR framework against several established uncertainty and hallucination detection strategies, including: (1) Sampling Entropy, based on the principles of semantic entropy [5], which performs $N = 8$ independent forward passes at temperature $\tau = 1.0$ to measure label-level uncertainty through the diversity of predicted classes; (2) RadFlag [21], which estimates internal consensus by generating multiple responses and calculating the proportion

of answers that align with the original diagnostic output, where lower agreement suggests a higher likelihood of error; and (3) Vision-Amplified Semantic Entropy (VASE) [10]: We adapted the VASE framework for medical classification by constructing a perturbation distribution that includes our weak D_4 transformations alongside three hard stressors: (i) high-sigma Gaussian blur ($\sigma = 15$) to preserve stain while destroying morphology, (ii) global mean pixel constant, and (iii) high-density salt-and-pepper noise. While this combined distribution exposes ungrounded logic in Gemini 2.5 Flash Lite, MedGemma 4B 1.5 frequently refused to answer when presented with these hard transformations. We therefore omit this metric for MedGemma to ensure a fair and consistent evaluation across the remaining benchmarks.

Comparison Analysis. As shown in Table 1, the proposed UVR metric demonstrates consistent improvement across all experimental settings, significantly outperforming standard sampling and existing baselines. The method reaches a peak AUROC of approximately 74.1% for the MedGemma 4B model on CRC-100K, marking a level of performance that is clinically relevant for computer-aided diagnosis. In Table 2, the Top-10% Accuracy results further confirm the metric's effectiveness as a reliable trust filter for medical decision-making. However, performance in MedGemma remains less stable in specific settings due to the model's inherent conservatism in histopathology tasks. This behavior results in minimal or repetitive reasoning chains that cause a performance drop compared to more expressive models.

Ablation Study. Ablation results in Table 3 confirm that while individual components provide significant benefits, combining visual and reasoning stability is essential for peak performance. Specifically, visual stability alone provides the largest single jump in AUROC (+25.60), while reasoning stability contributes an additional +18.00 in AUROC and +17.93 in Top-10% Accuracy. The two signals are thus complementary, capturing prediction and reasoning consistency respectively

4 Conclusion

In this work, we introduced a novel empirical signal for uncertainty quantification in black-box medical MLLMs based on RTD. By auditing the entropy of reasoning traces under visual, semantic-preserving perturbations, our framework moves beyond simple label consistency to a more robust representation of model certainty. Across two histopathology datasets, integrating reasoning traces improves both uncertainty estimation and classification accuracy for general-purpose and domain-specific models alike. Furthermore, our ablation study underscores the critical role of reasoning stability in identifying latent hallucinations that visual-only metrics often fail to capture.

Limitations and Future Work. While our framework enhances reliability, it introduces specific challenges. The extraction and analysis of reasoning traces by the MLLM itself can introduce a secondary layer of uncertainty and increased response variation. Expert validation of extracted concepts and broader calibration analysis remain important future directions. Future work will formalize RTD

to better decouple diagnostic logic from linguistic stochasticity. Additionally, we aim to investigate the scaling of this signal across more diverse clinical modalities to further establish a safe operating zone for black-box AI in medicine.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) under grant number RS-2022-NR068758. We are also deeply grateful for the generous support provided by the Seegene Medical Foundation in South Korea. We thank Dr. Chris Abishek S, MBBS, currently pursuing residency in General Surgery (Chettinad Hospital and Research Institute), for helpful discussions on histopathological concepts and terminology.

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

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. AlSaad, R., Abd-Alrazaq, A., Boughorbel, S., Ahmed, A., Renault, M.A., Damseh, R., Sheikh, J.: Multimodal large language models in health care: applications, challenges, and future outlook. Journal of medical Internet research **26**, e59505 (2024)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
4. Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., consortium, C., Hermsen, M., Manson, Q.F., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama **318**(22), 2199–2210 (2017)
5. Farquhar, S., Kossen, J., Kuhn, L., Gal, Y.: Detecting hallucinations in large language models using semantic entropy. Nature **630**(8017), 625–630 (2024)
6. Feng, Y., Htut, P.M., Qi, Z., Xiao, W., Mager, M., Pappas, N., Halder, K., Li, Y., Benajiba, Y., Roth, D.: Rethinking llm uncertainty: A multi-agent approach to estimating black-box model uncertainty. arXiv preprint arXiv:2412.09572 (2024)
7. Ferber, D., Wölflein, G., Wiest, I.C., Ligero, M., Sainath, S., Ghaffari Laleh, N., El Nahhas, O.S., Müller-Franzes, G., Jäger, D., Truhn, D., et al.: In-context learning enables multimodal large language models to classify cancer pathology images. Nature Communications **15**(1), 10104 (2024)
8. 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 **43**(2), 1–55 (2025)
9. Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., et al.: Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS medicine **16**(1), e1002730 (2019)

10. 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: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 669–679. Springer (2025)
11. Ravichander, A., Ghela, S., Wadden, D., Choi, Y.: Halogen: Fantastic llm hallucinations and where to find them. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1402–1425 (2025)
12. Rösler, W., Altenbuchinger, M., Baeßler, B., Beissbarth, T., Beutel, G., Bock, R., von Bubnoff, N., Eckardt, J.N., Foersch, S., Loeffler, C.M., et al.: An overview and a roadmap for artificial intelligence in hematology and oncology. Journal of cancer research and clinical oncology **149**(10), 7997–8006 (2023)
13. Sahoo, P., Meharia, P., Ghosh, A., Saha, S., Jain, V., Chadha, A.: A comprehensive survey of hallucination in large language, image, video and audio foundation models. Findings of the Association for Computational Linguistics: EMNLP 2024 pp. 11709–11724 (2024)
14. 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)
15. Sharma, P.: Dihedral group d4—a new feature extraction algorithm. Symmetry **12**(4), 548 (2020)
16. 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)
17. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. National Science Review **11**(12), nwae403 (2024)
18. 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)
19. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12944–12953 (2024)
20. Zhang, R., Zhang, H., Zheng, Z.: Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. arXiv preprint arXiv:2411.11919 (2024)
21. Zhang, S., Sambara, S., Banerjee, O., Acosta, J., Fahrner, L.J., Rajpurkar, P.: Radflag: A black-box hallucination detection method for medical vision language models. arXiv preprint arXiv:2411.00299 (2024)