Hallucination-Aware Multimodal Benchmark for Gastrointestinal Image Analysis with Large Vision-Language Models

Bidur Khanal[1*], Sandesh Pokhrel[2*], Sanjay Bhandari[2*], Ramesh Rana[3], Nikesh Shrestha[3], Ram B. Gurung[3], Cristian Linte[1], Angus Watson[5], Yash R. Shrestha[4], and Binod Bhattarai[5]

[1] Rochester Institute of Technology, Rochester, NY, USA
[2] Nepal Applied Mathematics and Informatics Institute for Research (NAAMII), Lalitpur, Nepal
[3] Kathmandu University, Dhulikhel, Nepal
[4] University of Lausanne, Lausanne, Switzerland
[5] University of Aberdeen, Aberdeen, UK
binod.bhattarai@abdn.ac.uk

Abstract. Vision-Language Models (VLMs) are becoming increasingly popular in the medical domain, bridging the gap between medical images and clinical language. Existing VLMs demonstrate an impressive ability to comprehend medical images and text queries to generate detailed, descriptive diagnostic medical reports. However, hallucination–the tendency to generate descriptions that are inconsistent with the visual content–remains a significant issue in VLMs, with particularly severe implications in the medical field. To facilitate VLM research on gastrointestinal (GI) image analysis and study hallucination, we curate a multimodal image-text GI dataset: Gut-VLM. This dataset is created using a two-stage pipeline: first, descriptive medical reports of Kvasir-v2 images are generated using ChatGPT, which introduces some hallucinated or incorrect texts. In the second stage, medical experts systematically review these reports, and identify and correct potential inaccuracies to ensure high-quality, clinically reliable annotations. Unlike traditional datasets that contain only descriptive texts, our dataset also features tags identifying hallucinated sentences and their corresponding corrections. A common approach to reducing hallucination in VLM is to finetune the model on a small-scale, problem-specific dataset. However, we take a different strategy using our dataset. Instead of finetuning the VLM solely for generating textual reports, we finetune it to detect and correct hallucinations, an approach we call hallucination-aware finetuning. Our results show that this approach is better than simply finetuning for descriptive report generation. Additionally, we conduct an extensive evaluation of state-of-the-art VLMs across several metrics, establishing a benchmark. Dataset and code available: bhattarailab/Hallucination-Aware-VLM.

* These authors contributed equally to this work. B. Khanal was associated with Multimodal Learning Lab, UoA, while working on this paper.

Keywords: Multimodal data · Gastrointestinal image analysis · Vision-Language Model (VLM) · Hallucination · Hallucination-aware finetuning

1 Introduction

Gastrointestinal (GI) diseases affect millions of people globally, making an accurate and timely diagnosis crucial for effective patient care [2]. GI endoscopy is the gold standard tool for diagnosing gastrointestinal diseases and is widely adopted in clinical settings. In recent years, Artificial Intelligence (AI) has shown significant potential in assisting clinicians with disease understanding and decision-making by detecting conditions [24], classifying anatomical landmarks [7], and identifying anomalies [28, 29]. While AI models primarily rely on endoscopic images, integrating descriptive text enhances expressiveness and interpretability, providing a richer clinical context that enables informed clinical decision making, diagnostic support, quality assurance, communication, medical record documentation, and more [17, 23, 30]. Nevertheless, despite the importance of textual information, its practical usage remains limited due to the lack of such multimodal datasets containing both GI endoscopic images and descriptive texts.

Several medical image-text datasets exist for chest X-rays images, histopathology images, and other radiographs [32], enabling the development of VLMs for clinical applications [33]. Unlike other image-only AI tools, VLMs are inherently good at absorbing complex information, reasoning, and generating explanations that are comprehensive for both clinicians and patients. Although several GI image and video datasets exist [5, 8, 14, 27, 38], to the best of our knowledge, Kvasir-VQA [10] is the only existing text-image multimodal dataset for GI image analysis, but it has notable limitations. Its short textual responses limit the depth of expert analysis and do not fully accommodate specialized medical vocabulary. Moreover, this dataset lacks comprehensive validation of all samples by certified experts, attributed to time constraints.

To mitigate these shortcomings, we create a new multimodal dataset out of Kvasir-v2 images, describing the underlying conditions in a short descriptive report format. Instead of having experts manually annotate images from the start, we first leverage an existing commercial large VLM (ChatGPT-4 Omni [1]) to generate image descriptions by prompting it with expert-crafted questions, which serves as a cost-effective and time-efficient alternative. Studies show that ChatGPT, trained on vast internet data, demonstrates a reasonable understanding of medical prompts and performs decently on some evaluations [16, 31]. However, like any other language models, VLMs are prone to hallucinations and cannot be relied upon without certified expert supervision.

Hallucination, in this context, refers to instances where the model produces information that appears plausible but is factually incorrect or fabricated [3, 11]. VLMs generate outputs in an auto-regressive manner, predicting the next plausible token based on statistical patterns. However, this process can introduce biases, as predictions tend to favor patterns that are most frequently observed in the training data rather than being grounded in factual knowledge. A recent

study indicates that the state-of-the-art (SOTA) large VLMs exhibit a hallucination rate up to 30% when describing natural images [11]. Hallucination can be more severe in medical data, as demonstrated by our dataset (Fig. 1), where only 30.39% of the VLM-generated responses are fully correct. To address this issue, in the second stage, we employ expert gastroenterologists to analyze the images and responses generated by ChatGPT, identify potential hallucinations, and correct any inaccuracies, ensuring that the final response is accurate and reliable. As a result, we obtain expert feedback on where the VLM hallucination has occurred at the sentence level and what the corrected version is. This additional information in the dataset can be leveraged to develop hallucination detectors or create hallucination-aware models.

Some studies in educational psychology suggest that students learn effectively when actively engaged in correcting errors and self-reflecting rather than through passive learning [25]. We argue that training VLM to identify and correct hallucinations fosters learning through correction, similar to how humans learn. Our hallucination-aware strategy for finetuning VLMs is rooted in this motivation.

While several datasets and studies on hallucination in natural image-text scenarios have recently emerged [3], there have been very few attempts to explore this phenomenon within the medical domain [6, 15], and none specifically in GI analysis. Therefore, in this work, we propose a hallucination-aware multimodal dataset, Gut-VLM, for GI image analysis using Kvasir-v2 images, one of the most widely used GI image datasets, to facilitate research on the clinical applications of VLMs and study hallucinations. Our key contributions are:

1. We create a novel multimodal image-text dataset for GI image analysis, consisting of VLM-generated descriptive diagnostic reports, expert-labeled tags identifying hallucinated sentences, and their corresponding corrections.
2. We provide an extensive evaluation benchmark on four SOTA VLMs across various settings, using both existing and our proposed LLM-assisted evaluation metrics, alongside a clinical expert evaluation.
3. We demonstrate that our innovative hallucination-aware finetuning approach, trained to detect and correct hallucinations, improves test performance compared to finetuning only on corrected ground-truth responses.

2 Methodology

In this section, we introduce the Gut-VLM dataset, detailing its overall composition and outlining the annotation pipeline, and present some key dataset statistics, particularly focusing on VLM-induced hallucinations. Finally, we also present a hallucination-aware VLM finetuning strategy.

2.1 Dataset Composition

The images in the Gut-VLM dataset are sourced from Kvasir-v2 [27], covering a diverse set of normal and abnormal gastrointestinal conditions. Normal findings

include images of healthy Cecum (321), Pylorus (298), and Z-line anatomy (418), while abnormal findings include images of Polyps (185), Esophagitis (70), Ulcerative Colitis (180), Dyed-resection-margins (172), and Dyed-lifted-polyps (172). In total, we annotated 1,816 images representing these conditions, splitting them into 1,450 for training and 366 for testing. Instead of randomly splitting the entire dataset into training and test sets, we ensured proportional representation of each condition in the test set by allocating 20% of samples from each category to the test set.

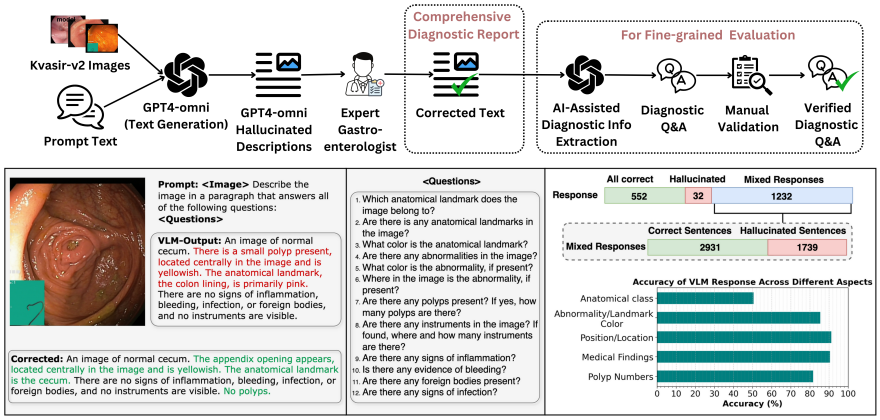

Fig. 1. Top: Overview of the data annotation pipeline. Bottom: [Left] A sample from the Gut-VLM, describing the underlying conditions based on the questions asked in the prompt, showing hallucinated (red) and corrected texts (green). [Right] Statistics of ChatGPT-4 Omni-generated responses in the dataset.

2.2 Dataset Annotation Pipeline

In this section, we outline the multistep pipeline used to generate the final annotated dataset, as illustrated in Fig. 1. This innovative, cost-effective approach involves using a large VLM to generate descriptive diagnostic reports for images, followed by expert corrections to produce verified reports. The report is further parsed to extract diagnostic Q&A for fine-grained evaluation.

VLM Diagnostic Report Generation: To generate a descriptive report for each image in our dataset, we queried ChatGPT-4 Omni [1] to describe each image using a prompt designed to elicit a detailed response covering the contents of 12 diagnostic reference questions for gastrointestinal image assessment. These questions, part of the MedVQA-GI Challenge [12], address aspects such as anatomical class, color and position of landmarks and abnormalities, polyp count, and the presence of inflammation, bleeding, foreign bodies, infection, or instruments. Since the generated responses contained some hallucinated texts, we proceeded to the next step to identify and correct these hallucinations.

The VLM-generated responses are reviewed by expert gastroenterologists to identify hallucinations. Following the M-HalDetect framework [11], each sentence in the response was labeled to indicate whether it contained hallucinated text. A sentence is marked as <non-hallucinated> if it accurately describes the content in the image, while any sentence containing inconsistent information is marked as <hallucinated>. Fig. 1 presents the corresponding hallucination statistics for the overall dataset: only 30.39% of the responses were fully correct, 1.7% contained hallucinations in all sentences, and 67.84% consisted of mixed responses with correct and hallucinated sentences. Finally, experts corrected incorrect sentences to ensure the clinical accuracy of descriptions. These annotations were collected using our in-house developed annotation tool and involved four experts.

Diagnostic Q&A Extraction for Fine-Grained Evaluation: While diagnostic reports are more comprehensive and context-rich, to objectively evaluate performance at a fine-grained level, we must assess whether the descriptions accurately address the 12 diagnostic questions. To achieve this, we prompted ChatGPT to extract information from the generated descriptive responses into a structured Q&A format. This method was also applied to extract diagnostic Q&A for other VLMs during testing. For the ground-truth responses, we manually verified the extracted diagnostic Q&A to ensure accuracy. We observed a hallucination rate of 4.29% during the information extraction process from the corresponding descriptive responses. As the same extractor was applied to all models during evaluation, the hallucination impact should be consistent across models. Additionally, since the results are available in a structured Q&A format, this data can also be utilized for Visual Question Answering (VQA) experiments.

2.3 Hallucination-aware VLM Finetuning

A standard VLM finetuning approach for generating descriptions would be to train the model to output the ground-truth texts. However, since we have the VLM-generated response, sentence-level hallucination tags, and the final corrected response, we instead finetune the model to identify hallucinated sentences and correct the response, as shown in Fig. 2. In Step 1, the model is trained to identify hallucinations and label the output as either <hallucinated> or <non-hallucinated>. In Step 2, it learns to revise hallucinated sentences and generate a corrected response. This approach leverages the existing VLM responses and makes the model aware of potential hallucination patterns.

3 Experiments

Using our proposed dataset, we experimented with four state-of-the-art (SOTA) VLM models: LLaVA-1.6-7B [22], Qwen2-7B [36], mPLUG-Owl-2B [37], and DeepSeek-7B-VL [19]. We first generated descriptive reports by prompting the pretrained models to describe images with a focus on 12 diagnostic questions. Next, we applied supervised finetuning using Rank-8 LoRA adaptation [13] with two strategies. In the first strategy, we finetuned the model to generate corrected

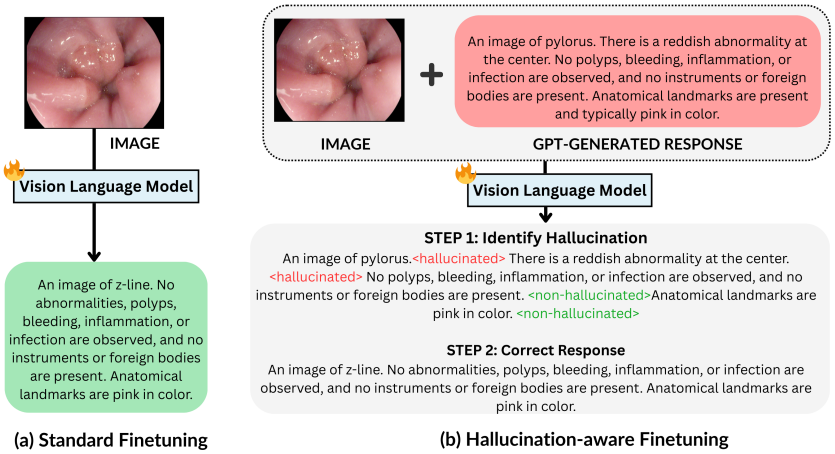

Fig. 2. Comparing standard finetuning with hallucination-aware finetuning

responses directly. In the second strategy, we finetuned the model to learn to detect hallucinated sentences and then correct them—the approach we refer to as hallucination-aware (H) fine-tuning.

As outlined in Section 2.2, since the dataset was already structured in a VQA format for fine-grained evaluation, we explicitly experimented with one of the models, LLaVA-1.6-7B, on a VQA task. We first tested the pretrained model, then applied standard VQA finetuning in a multi-conversational format.

All models were finetuned using an A100 GPU for five epochs until the loss stabilized, with a batch size of 8 and a learning rate of 1e-4, and evaluated on the test set. We used the ms-swift framework [§] to run all VLM experiments.

4 Evaluation

We initially assess VLM-generated reports against corrected ground-truth descriptions using classical metrics such as ROUGE-L [18], BLEU [26], and METEOR [4], which measure sequence overlap, n-gram matches, and semantic similarity, respectively. However, these metrics are limited by context-length dependence, insensitivity to subtle semantic differences, and inability to assess factual accuracy. Here, we propose two LLM-assisted evaluation metrics.

The first, R-Sim, rates coarse-level semantic similarity between the ground-truth texts and VLM-generated texts on a scale from 1 to 5 (worst to best), using ChatGPT. We provide both texts and prompt ChatGPT to assess their similarity by focusing on 12 GI diagnostic questions from the MedVQA-GI Challenge as a reference for judgment. A score of 5 indicates a high degree of semantic alignment with the ground-truth, while a score of 1 reflects significant divergence in meaning. Such LLM-assisted evaluations have been reported in recent

[§] https://github.com/modelscope/ms-swift

studies [20]. The second metric, Question Answering Accuracy Score (QAAS), quantitatively evaluates how accurately VLM-generated responses answer the 12 diagnostic questions. We begin by extracting the answers to these questions from the VLM responses using ChatGPT, following the procedure as in Section 2.2. These extracted answers are then compared to the corresponding ground-truth answers in Q&A pairs, with ChatGPT handling variations in wording, such as synonyms or paraphrasing. If ChatGPT determines that the answer matches the ground-truth, it is marked as correct; otherwise, it is marked as incorrect. The overall QAAS is simply calculated as the ratio of the total number of correct answers to the total number of questions.

We had an expert rate LLaVa's descriptive responses for clinical evaluation. Due to budget constraints, we randomly sampled 30 responses per model and asked the expert to provide two ratings (1 to 5): similarity, assessing the clinical resemblance of the response to the ground truth, and quality, evaluating its clinical significance independently. We averaged the two ratings to compute a final per-response score. Additionally, a new sample was randomly inserted among the 30 responses five times, unknown to the expert, to measure rating consistency.

5 Results

Descriptive Format Diagnostic Report: Table 1 presents the quantitative evaluation of descriptive reports generated by VLMs for GI images. All pretrained models exhibited significantly lower performance in terms of ROUGE-L, BLEU, METEOR, R-Sim, and QAAS scores, demonstrating a limited ability to generate clinically relevant descriptions—except for ChatGPT-4 Omni. Since ChatGPT-4 is a large-parameter model and the ground-truth descriptions are corrected versions of the original ChatGPT responses, higher scores are expected due to potential bias toward the original structure.

There was a notable improvement across all metrics when finetuning VLMs with ground-truth texts. However, the hallucination-aware finetuning (finetuned^{H}) outperformed the standard finetuning, suggesting that training the model to detect and correct hallucinations leads to improved performance and produces more reliable, context-aware models. For instance, the LLaVA-1.6-7B finetuned^{H} achieved a QAAS score of 90.89%, compared to 83.07% for its standard finetuned counterpart and only 50.89% for pretrained version. Similarly, R-Sim improved from 1.36 (pretrained) to 3.71 (standard finetuning), with further improvement to 3.96 through hallucination-aware finetuning, demonstrating a more accurate semantic alignment with expert-generated descriptions. We could not finetune ChatGPT-4 Omni as it is not an open-source model. We also evaluated QAAS across different aspects as depicted in Fig. 3, which shows that hallucination-aware training consistently outperforms standard training across most aspects.

Expert Evaluation: *For LLaVa-1.6-7B, the average scores are* 1.90 *(pretrained),* 3.24 *(finetuned), and* 3.36 *(hallucination-aware finetuned), with the latter scoring*

Table 1. Quantitative evaluation of various VLMs in descriptive report generation and VQA tasks. Finetuned$^{\mathbf{H}}$ indicates hallucination-aware finetuning.

Model	ROUGE-L↑	BLEU↑	METEOR↑	R-Sim↑	QAAS(%)↑
ChatGPT-4 *Omni*	0.87	0.80	0.89	2.97	85.99
LLaVa-1.6-7b *pretrained*	0.26	0.10	0.47	1.36	50.89
LLaVa-1.6-7b *finetuned*	0.54	0.35	0.63	3.71	83.07
LLaVa-1.6-7b *finetuned*$^{\mathbf{H}}$	**0.89**	**0.82**	**0.90**	**3.96**	**90.89**
DeepSeek7b *pretrained*	0.29	0.11	0.39	1.65	65.20
DeepSeek7b *finetuned*	0.55	0.37	0.65	**3.76**	83.73
DeepSeek7b *finetuned*$^{\mathbf{H}}$	**0.88**	**0.81**	**0.90**	3.63	**88.77**
Qwen7b *pretrained*	0.32	0.12	0.48	1.74	67.57
Qwen7b *finetuned*	0.54	0.37	0.64	3.78	83.27
Qwen7b *finetuned*$^{\mathbf{H}}$	**0.88**	**0.82**	**0.90**	**4.04**	**90.53**
MPlugOwl2b *pretrained*	0.26	0.09	0.44	1.34	55.29
MPlugOwl2b *finetuned*	0.50	0.32	0.60	3.68	82.90
MPlugOwl2b *finetuned*$^{\mathbf{H}}$	**0.85**	**0.77**	**0.87**	**3.72**	**88.40**
VQA LLaVa-1.6-7b *pretrained*	–	–	–	–	49.26
VQA LLaVa-1.6-7b *finetuned*	–	–	–	–	87.91

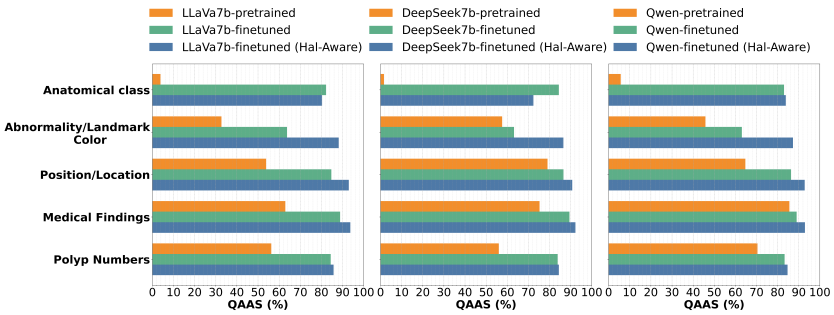

Fig. 3. Comparison of VLM responses evaluated across different aspects

the highest. We also computed the rater's average coefficient of variation [35] using the stand-out sample and found it to be 11.41%*, indicating fair consistency.*

Visual Question Answering (VQA) Evaluation: Table 1 also summarizes the performance of LLaVA-1.6-7B in VQA task. We focus on QAAS, as other metrics are mainly applicable only for description comparison. The pretrained models struggled to answer expert-designed questions effectively, achieving significantly lower QAAS, while finetuning substantially improved the performance.

6 Discussion and Conclusion

Here, we introduced Gut-VLM, a multimodal dataset for GI image analysis that includes hallucination-aware annotations to advance research on reliable and

trustworthy VLMs. Our annotation process of using VLM-generated descriptive diagnosis reports, followed by expert corrections, not only reduces annotation costs by avoiding the need to hire experts for routine annotation from start but also enables the creation of a dataset with tags identifying potential hallucination patterns. We argue that by formulating VLM finetuning tasks as hallucination detection and correction rather than just diagnostic report generation, we can elicit reasoning in VLMs, similar to how engagement via error identification and correction enhances learning in humans compared to passive learning.

This work also has some limitations. Budget constraints led to corrections being limited to responses from a single VLM (ChatGPT), potentially introducing bias toward its response structure. Additionally, our dataset offers sentence-level hallucination tags, which may limit granularity. We will expand the dataset by incorporating diverse VLM responses, annotating segment-level hallucination tags, and extending the dataset to include diverse demographics.

Future research could explore alternative hallucination detection and mitigation strategies, such as uncertainty estimation [9], reinforcement learning [11], architectural modifications [21], and feature fusion [34]. We also highlight the need for standardized benchmarks in the medical domain to accurately assess hallucinations and ensure reliable model evaluation, as well as ways to test the statistical significance of performance and its relevance to diagnosis.

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., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Arnold, M., Abnet, C.C., Neale, R.E., Vignat, J., Giovannucci, E.L., McGlynn, K.A., Bray, F.: Global burden of 5 major types of gastrointestinal cancer. Gastroenterology **159**(1), 335–349 (2020)
3. Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., et al.: Hallucination of multimodal large language models: A survey. arXiv preprint arXiv:2404.18930 (2024)
4. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: ACL workshops. pp. 65–72 (2005)
5. Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., et al.: Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific data **7**(1), 283 (2020)
6. 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)
7. Cogan, T., Cogan, M., Tamil, L.: MAPGI: Accurate identification of anatomical landmarks and diseased tissue in gastrointestinal tract using deep learning. Computers in Biology and Medicine **111**, 103351 (2019)
8. Du, H., Dong, Z., Wu, L., Li, Y., Liu, J., Luo, C., Zeng, X., Deng, Y., Cheng, D., et al.: A deep-learning based system using multi-modal data for diagnosing gastric neoplasms in real-time (with video). Gastric Cancer **26**(2), 275–285 (2023)

9. Farquhar, S., Kossen, J., Kuhn, L., Gal, Y.: Detecting hallucinations in large language models using semantic entropy. Nature **630**(8017), 625–630 (2024)
10. Gautam, S., Storås, A.M., Midoglu, C., Hicks, S.A., Thambawita, V., Halvorsen, P., Riegler, M.A.: Kvasir-VQA: A text-image pair gi tract dataset. In: Proceedings of the VLM4Bio Workshop. p. 3–12. ACM (2024)
11. Gunjal, A., Yin, J., Bas, E.: Detecting and preventing hallucinations in large vision language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 18135–18143 (2024)
12. Hicks, S., Storås, A.M., Halvorsen, P., de Lange, T., Riegler, M., Thambawita, V.L.: Overview of ImageCLEFmedical 2023 - medical visual question answering for gastrointestinal tract. In: CLEF. pp. 1316–1327 (2023)
13. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., et al.: LoRA: Low-rank adaptation of large language models. In: ICLR. vol. 1, p. 3 (2022)
14. Jha, D., Sharma, V., Dasu, N., Tomar, N.K., Hicks, S., Bhuyan, M.K., Das, P.K., Riegler, M.A., Halvorsen, P., Bagci, U., et al.: Gastrovision: A multi-class endoscopy image dataset for computer aided gastrointestinal disease detection. In: ICML ML4MHD Workshop. pp. 125–140. Springer (2023)
15. Jiang, Y., Chen, J., Yang, D., Li, M., Wang, S., Wu, T., Li, K., Zhang, L.: CoMT: Chain-of-medical-thought reduces hallucination in medical report generation. In: ICASSP. pp. 1–5. IEEE (2025)
16. Keshavarz, P., Bagherieh, S., Nabipoorashrafi, S.A., Chalian, H., Rahsepar, A.A., Kim, G.H.J., et al.: ChatGPT in radiology: A systematic review of performance, pitfalls, and future perspectives. Diagnostic and interventional imaging (2024)
17. Li, J., Hu, S., Shi, C., Dong, Z., Pan, J., Ai, Y., Liu, J., Zhou, W., Deng, Y., Li, Y., Yuan, J., Zeng, Z., et al.: A deep learning and natural language processing-based system for automatic identification and surveillance of high-risk patients undergoing upper endoscopy: a multicenter study. EClinicalMedicine **53** (2022)
18. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
19. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., et al.: DeepSeek-V3 technical report. arXiv preprint arXiv:2412.19437 (2024)
20. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Mitigating hallucination in large multi-modal models via robust instruction tuning. In: ICLR (2024)
21. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: CVPR. pp. 26296–26306 (2024)
22. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS. vol. 36, pp. 34892–34916 (2023)
23. Marques, S., Bispo, M., Pimentel-Nunes, P., Chagas, C., Dinis-Ribeiro, M.: Image documentation in gastrointestinal endoscopy: review of recommendations. GE-Portuguese Journal of Gastroenterology **24**(6), 269–274 (2017)
24. Melaku Bitew Haile, Ayodeji Olalekan Salau, B.E., Belay, A.J.: Detection and classification of gastrointestinal disease using convolutional neural network and svm. Cogent Engineering **9**(1), 2084878 (2022)
25. Metcalfe, J.: Learning from errors. Annual review of psychology **68**(1), 465–489 (2017)
26. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the ACL. pp. 311–318 (2002)
27. Pogorelov, K., Randel, K.R., Griwodz, C., Eskeland, S.L., de Lange, T., Johansen, D., Spampinato, C., Dang-Nguyen, D.T., Lux, M., Schmidt, P.T., Riegler, M., Halvorsen, P.: Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In: Proceedings of the ACM MMSys. pp. 164–169 (2017)

28. Pokhrel, S., Bhandari, S., Ali, S., Lambrou, T., Nguyen, A., Shrestha, Y.R., Watson, A., et al.: NCDD: Nearest centroid distance deficit for out-of-distribution detection in gastrointestinal vision. arXiv preprint arXiv:2412.01590 (2024)
29. Pokhrel, S., Bhandari, S., Vazquez, E., Lambrou, T., Gyawali, P., Bhattarai, B.: TTA-OOD: Test-time augmentation for improving out-of-distribution detection in gastrointestinal vision. In: MICCAI DEMI Workshop. pp. 33–42. Springer (2024)
30. Selivanov, A., Rogov, O.Y., Chesakov, D., Shelmanov, A., Fedulova, I., Dylov, D.V.: Medical image captioning via generative pretrained transformers. Scientific Reports **13**(1), 4171 (2023)
31. Shieh, A., Tran, B., He, G., Kumar, M., Freed, J.A., Majety, P.: Assessing ChatGPT 4.0's test performance and clinical diagnostic accuracy on USMLE STEP 2 CK and clinical case reports. Scientific Reports **14**(1), 9330 (2024)
32. Shrestha, P., Amgain, S., Khanal, B., Linte, C.A., Bhattarai, B.: Medical vision language pretraining: A survey. arXiv preprint arXiv:2312.06224 (2023)
33. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S.R., Cole-Lewis, H., et al.: Toward expert-level medical question answering with large language models. Nature Medicine pp. 1–8 (2025)
34. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: CVPR. pp. 9568–9578 (2024)
35. Wikipedia contributors: Coefficient of variation (2025), https://en.wikipedia.org/wiki/Coefficient_of_variation, accessed: 2025-02-27
36. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
37. Ye, J., Xu, H., Liu, H., Hu, A., Yan, M., Qian, Q., Zhang, J., Huang, F., Zhou, J.: mPLUG-Owl3: Towards long image-sequence understanding in multi-modal large language models. ICLR (2025)
38. Zhu, S., Gao, J., Liu, L., Yin, M., Lin, J., Xu, C., Xu, C., Zhu, J.: Public imaging datasets of gastrointestinal endoscopy for artificial intelligence: a review. Journal of Digital Imaging **36**(6), 2578–2601 (2023)