



# Understanding the robustness of vision-language models to medical image artefacts



Zijie Cheng<sup>1</sup> , Ariel Yuhan Ong<sup>2,3</sup>, Siegfried K. Wagner<sup>2,3</sup>, David A. Merle<sup>2,3</sup>, Lie Ju<sup>2,3</sup>, Hanyuan Zhang<sup>1,4</sup>, Ruinian Chen<sup>5</sup>, Linze Pang<sup>6</sup>, Boxuan Li<sup>1</sup>, Tiantian He<sup>4,5</sup>, Anran Ran<sup>7</sup>, Hongyang Jiang<sup>7</sup>, Dawei Gabriel YANG<sup>7,8</sup>, Ke Zou<sup>9</sup>, Jocelyn Hui Lin Goh<sup>8,9,10</sup>, Sahana Srinivasan<sup>8</sup>, Andre Altmann<sup>1,4</sup>, Daniel C. Alexander<sup>4,5</sup>, Carol Y. Cheung<sup>7</sup>, Yih Chung Tham<sup>8,9,10</sup>, Pearse A. Keane<sup>2,3</sup> & Yukun Zhou<sup>2,3,4,5</sup>

Vision-language models (VLMs) show promise for answering clinically relevant questions, but their robustness to medical image artefacts remains unclear. We evaluated VLMs' robustness through their performance on images with and without weak artefacts across five artefact categories, as well as their ability to detect images with strong artefacts. We built evaluation benchmarks using brain MRI scans, Chest X-ray, and retinal images, involving four real-world medical datasets. VLMs achieved moderate accuracy on original unaltered images (0.645, 0.602 and 0.604 for MRI, OCT, and X-ray applications, respectively). Accuracy declined with weak artefacts ( $-3.34\%$ ,  $-9.06\%$  and  $-10.46\%$ ), while strong artefacts were detected at low rates (0.194, 0.128 and 0.115). Our findings indicated that VLMs are not yet capable of performing tasks on medical images with artefacts, underscoring the need to establish uniform benchmark thoroughly examining model robustness to image artefacts, and explicitly incorporate artefact-aware method design and robustness tests into VLM development.

Vision-language models (VLMs) are trained using vast and diverse images paired with textual descriptions (prompts), allowing them to deeply understand visual content and generate accurate responses<sup>1</sup>. Powerful examples, such as GPT-4o<sup>2</sup> and Claude 3.5 Sonnet<sup>3</sup>, enable diverse applications such as image captioning<sup>4</sup>. The advances have been extended to the clinical task, leading to specialised medical VLMs such as MedDR<sup>5</sup>, LLaVAMed<sup>6</sup> and BiomedCLIP<sup>7</sup>, which contribute to disease detection in multiple medical imaging modalities. However, in real-world medical scenarios, medical image artefacts commonly exist and potentially bias the performance of VLMs, which has not been systematically studied.

Image artefacts in the medical domain are distortions caused by patient movement, equipment limitations, varied technicians' skills and environmental factors<sup>8–12</sup>. They are commonly seen in many medical imaging modalities such as magnetic resonance imaging (MRI)<sup>13</sup>, optical coherence tomography (OCT)<sup>14</sup> and X-ray<sup>15</sup>. For instance, random motion in OCT imaging refers to movements of the subject, eye drift or microsaccades

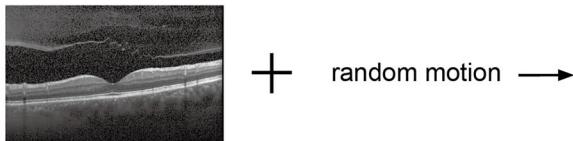
during scan acquisition<sup>14,16,17</sup>. These image artefacts can lead to errors in disease detection, including false positives, where normal cases are incorrectly identified as diseased, and false negatives, where diseased cases are mistakenly classified as normal<sup>18–21</sup>. These errors occur when random image artefacts generate patterns similar to pathological features or obscure actual abnormalities. Previous literature has studied the impact of medical image artefacts on traditional Artificial Intelligence (AI) models and have observed remarkable performance drops<sup>22–24</sup>.

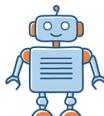
Recent studies have explored the use of VLMs in medical visual question answering (VQA) to evaluate their capabilities in disease detection<sup>25–29</sup>. Their results show that some powerful VLMs (e.g. GPT-4o) could achieve performance comparable to junior doctors<sup>30–34</sup> in certain tasks. However, few studies have explored the robustness and detection capabilities of VLMs when medical image artefacts are present. Understanding this critical knowledge from the early stages is essential for advancing trustworthy medical AI applications in clinical settings, especially

<sup>1</sup>Department of Medical Physics & Biomedical Engineering, University College London, London, UK. <sup>2</sup>Institute of Ophthalmology, University College London, London, UK. <sup>3</sup>NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust, London, UK. <sup>4</sup>UCL Hawkes Institute, University College London, London, UK. <sup>5</sup>Department of Computer Science, University College London, London, UK. <sup>6</sup>School of Computer Science and Technology, Beijing Jiaotong University, Beijing, China. <sup>7</sup>Department of Ophthalmology and Visual Sciences, The Chinese University of Hong Kong, Hong Kong Special Administrative Region, Hong Kong, China. <sup>8</sup>Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore. <sup>9</sup>Centre for Innovation and Precision Eye Health; and Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. <sup>10</sup>Ophthalmology and Visual Science Academic Clinical Program, Duke-NUS Medical School, Singapore, Singapore. e-mail: [zijie.cheng.23@ucl.ac.uk](mailto:zijie.cheng.23@ucl.ac.uk); [yukun.zhou.19@ucl.ac.uk](mailto:yukun.zhou.19@ucl.ac.uk)

## (a) false positive

 ...As an educator, conclude your answer in "retinal diseases" or "normal". Describe your reasoning in steps



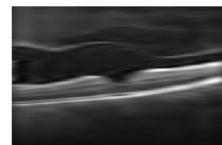
 ...  
\*\*Analysis of the Image:\*\*  
...

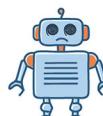
There are no irregular hyper-reflective areas beneath the RPE, suggesting the absence of CNV.



...  
\*\*Conclusion: Normal\*\*

 ...As an educator, conclude your answer in "retinal diseases" or "normal". Describe your reasoning in steps



 ...  
\*\*Analysis of the Image:\*\*  
...

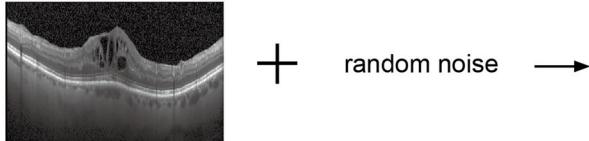
The image shows a significant elevation beneath the retina, which could indicate subretinal fluid or neovascularization. The presence of irregular elevation suggests possible CNV.

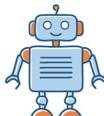


...  
\*\*Conclusion: Retinal Diseases\*\*

## (b) false negative

 ...As an educator, conclude your answer in "retinal diseases" or "normal". Describe your reasoning in steps



 ...  
\*\*Analysis of the Image:\*\*  
...

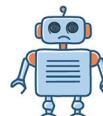
The image shows a significant elevation in the central retina with cystic spaces, indicating fluid accumulation. The presence of these features suggests a condition more aligned with CNV or DME rather than early AMD.



...  
\*\*Conclusion: Retinal Diseases\*\*

 ...As an educator, conclude your answer in "retinal diseases" or "normal". Describe your reasoning in steps



 ...  
\*\*Analysis of the Image:\*\*  
...

The image shows a smooth contour of the retina without significant thickening or cystoid spaces, suggesting the absence of DME. There are no irregular hyperreflective areas beneath the RPE or subretinal fluid, indicating no signs of CNV.



...  
\*\*Normal.\*\*

**Fig. 1 | False positive and False negative examples after adding artefacts to images.** **a** An example of a false positive case. Vision-Language models (VLMs) correctly classify normal cases when analysing original unaltered images but misclassify the same cases as diseased when random motion is introduced. **b** An

example of a false negative case. VLMs correctly identify abnormal cases in original unaltered images but misclassify them as normal after random noise is introduced. More misdetection examples with full VLMs' responses shown in Supplementary Data 1.

as VLMs are increasingly seen as promising solutions due to their strong capabilities in reasoning and user interaction. Figure 1 illustrates failure cases in disease detection by VLMs, caused by the presence of medical image artefacts.

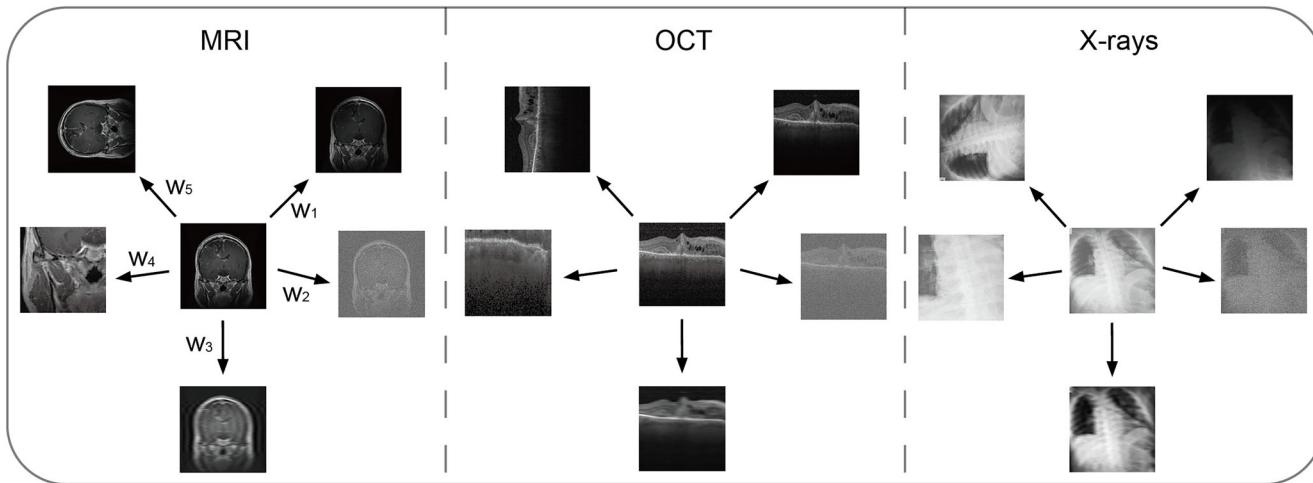
In this study, we evaluated the robustness of VLM to medical image artefacts in disease detection, with three particular contributions: (1) we constructed an evaluation benchmark for VLM robustness, spanning three medical imaging modalities, including five categories of image artefacts, in weak and strong scales; (2) we proposed to evaluate the robustness in three tiers of metrics to assess VLMs' performance, the typical classification performance (i.e. accuracy and sensitivity), performance percentage change, which describe the proportional change in VLMs' performance when adding weak artefacts to images, and the strong artefact detection rate which measures the VLMs' ability to identify images with strong artefacts; (3) we included most competitive VLMs, including GPT-4o, Claude 3.5 Sonnet, Llama 3.2, BiomedCLIP and MedGemma and assessed their robustness in both datasets with added artefacts and real-world artefacts. We revealed the insufficient robustness of current VLMs to medical image artefacts and underscore the urgent need to highlight robustness tests in VLM

benchmarking and to build more robust medical VLMs for healthcare applications.

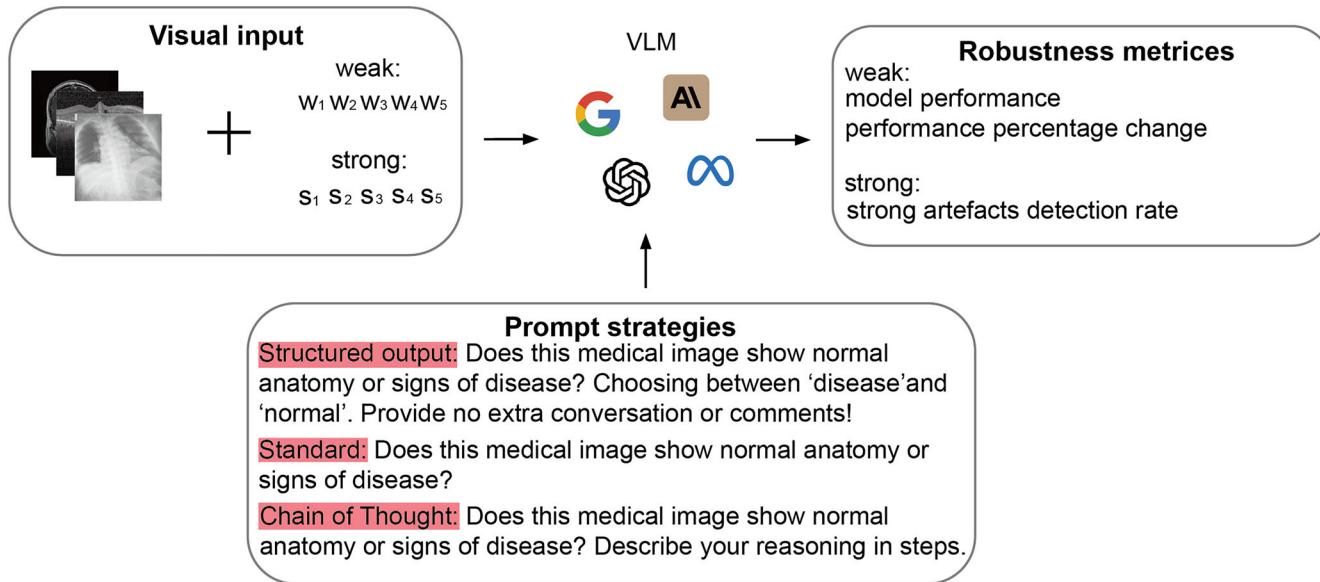
## Results

Figure 2 shows the overview of the project, including the construction of benchmarks and evaluation of VLMs' robustness to medical image artefacts in disease detection tasks, spanning over three medical imaging modalities. For each medical imaging modality, we randomly selected 200 images, 100 for normal and 100 for disease cases. We built the benchmark by introducing common artefacts such as intensity artefacts (bias field, motion and noise) and spatial artefacts (cropping and rotation) to the original unaltered images at weak and strong scales. Weak artefact datasets consist of images that are partially obscured yet still interpretable, while strong artefact datasets contain ungradable images that are severely distorted by artefacts, making it impossible to reliably classify them as diseased or normal. More details are provided in the 'Benchmark' subsection of the Methods, with example images shown in Supplementary Figs. 1–3. The benchmark structure, along with the hyperparameters used in the TorchIO library to generate artefacts at weak and strong scales, is presented in Supplementary

## A Benchmark Dataset



## B VLMs for disease diagnosis



**Fig. 2 | Overview of building benchmarks and the pipeline for evaluating robustness of Vision-Language models (VLMs) on disease detection tasks.**

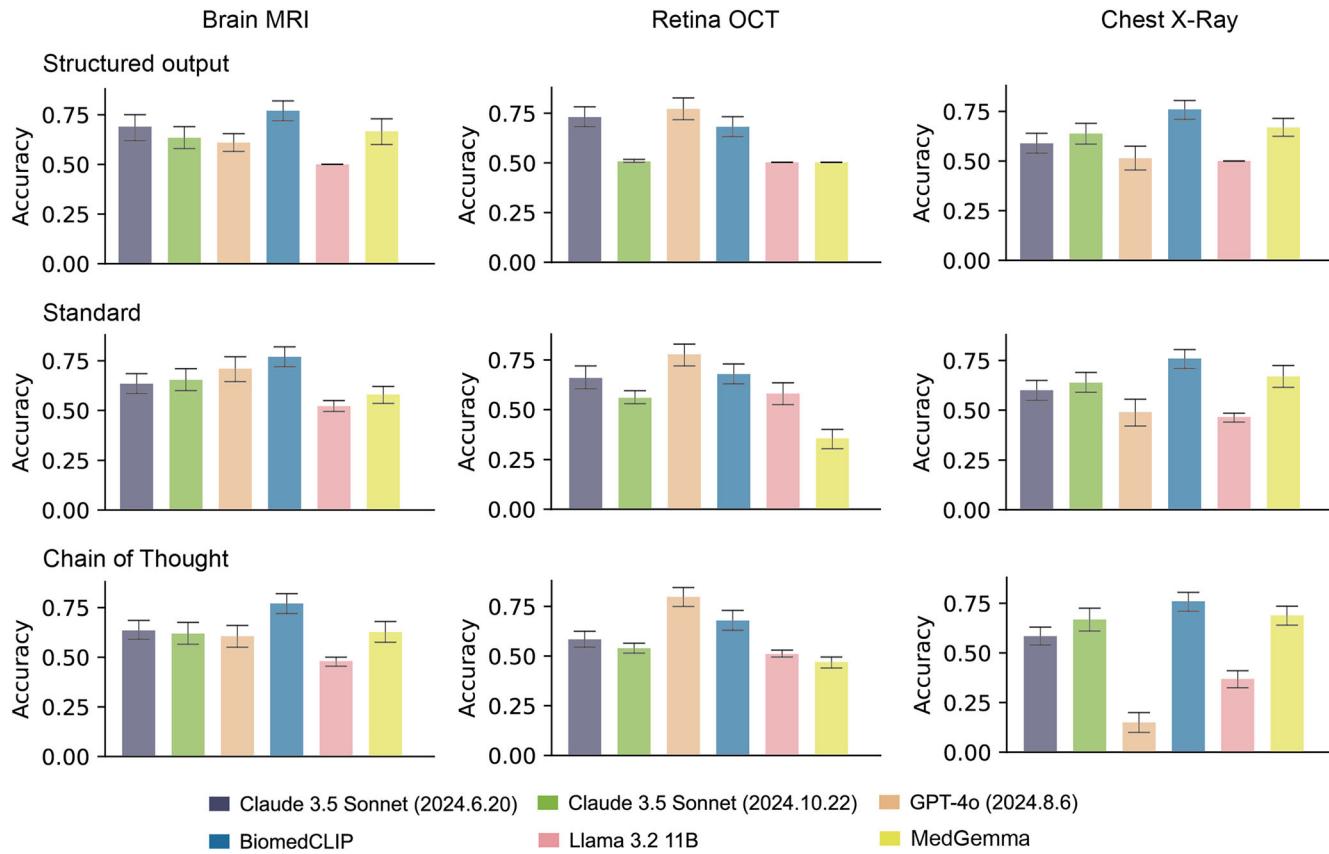
A presents robustness evaluation benchmarks covering disease detection tasks across different medical imaging modalities. Three intensity artefacts (random bias field, noise and motion) with weak ( $w_1-w_3$ ) and strong ( $s_1-s_3$ ) scales, and two spatial artefacts (random cropping and rotation) with weak ( $w_4-w_5$ ) and strong ( $s_4-s_5$ ) scales are introduced to original unaltered images. All hyperparameters used to generate artefacts at weak and strong scales are provided in Supplementary

Data 3. B illustrates the project pipeline, demonstrating the evaluation process used to assess the robustness of the VLMs. When adding weak artefacts, we evaluated model performance (e.g. accuracy) and the performance drop from its performance on the original unaltered images. When images were severely distorted by strong artefacts, we assessed VLMs' ability to detect poor image quality. All experiments were repeated through different prompt strategies from restricting reasoning through structured output to encouraging reasoning step by step through Chain of Thought.

Data 2 and 3, respectively. This study included two proprietary models (GPT-4o<sup>2</sup> and Claude 3.5 Sonnet<sup>3</sup>) and three open-source models (Llama 3.2<sup>35</sup>, BiomedCLIP<sup>7</sup> and MedGemma<sup>36</sup>), with BiomedCLIP only output probability for pre-defined categories. We evaluated the typical classification performance of VLMs using original unaltered images and those containing weak artefacts. We also examined the performance percentage change after adding weak artefacts to images, as well as VLMs' ability to detect strong artefacts (see the 'Metrics for Model Robustness' and 'Models Evaluated' subsections of the Methods section for further details). Additionally, we evaluated three prompting strategies: structured output, standard output and chain of thought, and examined their effect on performance.  $p$  values were calculated using a two-sided  $t$ -test to assess whether significant differences exist.

### Model performance on original unaltered images

We assessed the performance of various VLMs on original unaltered images and observed moderate performance (Fig. 3). For brain tumour (MRI) and COVID-19/pneumonia detection (chest X-ray), BiomedCLIP achieved the highest accuracy, with scores of 0.770 (95% Confidence Interval (CI): 0.720–0.820) and 0.760 (95% CI: 0.710–0.805) respectively, using the standard prompt. This advantage also extended to sensitivity, particularly in MRI applications, where it reached the highest value of 0.950 (95% CI: 0.910–0.990) (Supplementary Fig. 4). In macular disease detection (OCT), GPT-4o achieved the highest accuracy of 0.778 (95% CI: 0.720–0.830) and sensitivity of 0.656 (95% CI: 0.560–0.740) using the standard prompt. Llama 3.2 11B classified most cases as diseased, resulting in poor accuracy below 0.6 across all tasks (Fig. 3). In contrast, MedGemma classified most cases as



**Fig. 3 | The accuracy of Vision-Language models (VLMs) in detecting disease on original unaltered images.** Each column represents the models' accuracy in disease detection tasks across different medical imaging modalities, while each row illustrates their accuracy using various prompt strategies. Quantitative results are

detailed in Supplementary Data 6. For each task, accuracy measurements derive from 1000 stratified bootstrap samples. Results show mean accuracy with 95% confidence intervals (error bars).

normal, leading to similarly poor accuracy, high specificity, but low sensitivity (Fig. 3 and Supplementary Figs. 4, 5).

#### Model robustness to weak image artefacts

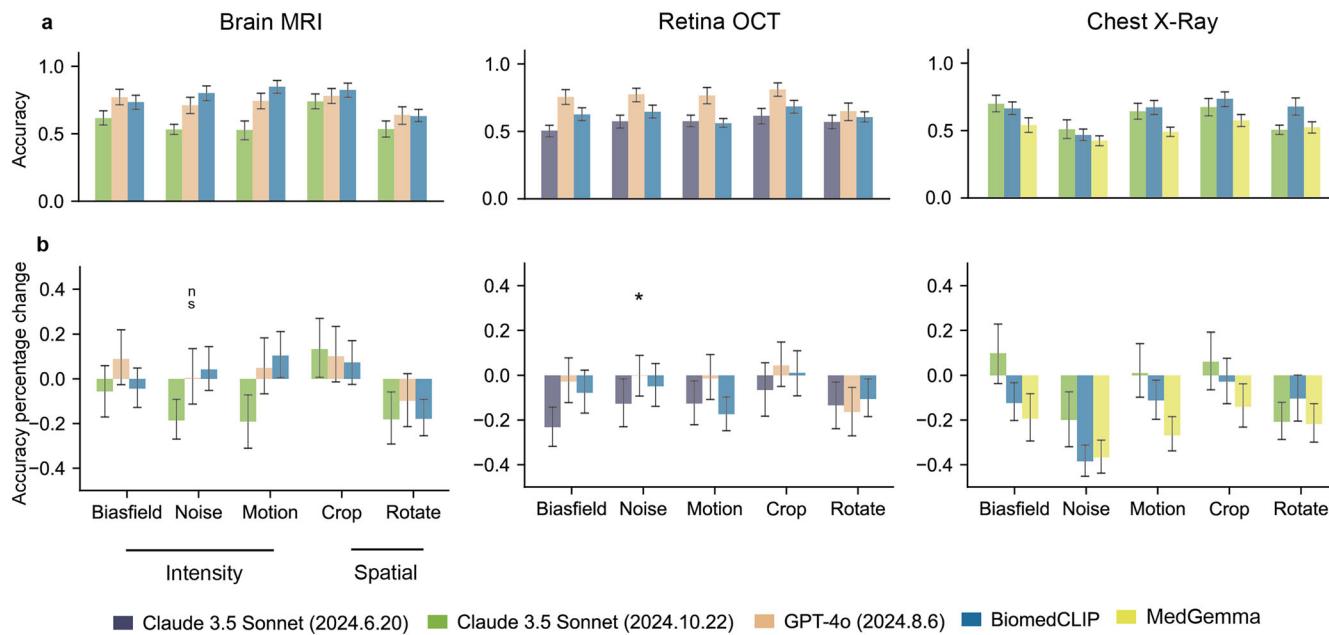
We evaluated the robustness of VLMs on images with weak artefacts in disease detection (Fig. 4 and Supplementary Figs. 6–17). Figure 4a shows the top three models per type of disease detection task, selected based on their accuracy on original unaltered images (Fig. 3). Among the models evaluated, BiomedCLIP demonstrated the highest accuracy and sensitivity in MRI applications. For instance, it achieved an accuracy of 0.801 (95% CI: 0.745–0.855) and sensitivity of 0.782 (95% CI: 0.700–0.860) when detecting MRI images with weak random noise using the standard prompt (Supplementary Fig. 6). While in OCT applications, GPT-4o demonstrated the strongest performance (Supplementary Fig. 7). In chest X-ray applications, MedGemma consistently showed the lowest accuracy in Fig. 4a and the lowest sensitivity compared to all other models (Supplementary Fig. 8). We observed that MedGemma's reasoning process is inconsistent with the conclusion in some cases. The reasoning process refers to the intermediate information provided by VLMs in analysing the provided image, while the conclusion summarises the VLM final decision, either 'Normal' or 'Diseased' in our study. We quantified the rate of such conflicts in Supplementary Fig. 19. For example, in chest X-ray applications using CoT prompting, MedGemma exhibited a conflict rate of 0.204 (95% CI: 0.150–0.260) on original images, which increased to 0.427 (95% CI: 0.360–0.490) after introducing weak random motion. Figure 4b shows the models' performance percentage change relative to their original accuracy. Random noise resulted in more significant performance degradation, especially in X-ray applications. When random noise was added to chest X-rays, BiomedCLIP's accuracy in detecting lung diseases decreased to 0.466

(95% CI: 0.425–0.510), representing a 38.5% drop ( $p < 0.001$ ) compared to its performance on original unaltered chest X-rays.

We found that adding specific image artefacts improved the disease detection accuracy of certain VLMs, particularly in MRI applications. Specifically, applying weak random cropping to MRI images improved model performance. For instance, Claude 3.5 Sonnet (2024.10.22) achieved an accuracy of 0.739 (95% CI: 0.685–0.795) in detecting brain tumours, representing a 13.3% ( $p < 0.001$ ) increase compared to the accuracy on original unaltered MRI images (Supplementary Data 8). Furthermore, introducing other image artefacts, such as the bias field artefact, to brain MRI images led to an 8.9% ( $p < 0.001$ ) increase in brain tumour detection accuracy for GPT-4o.

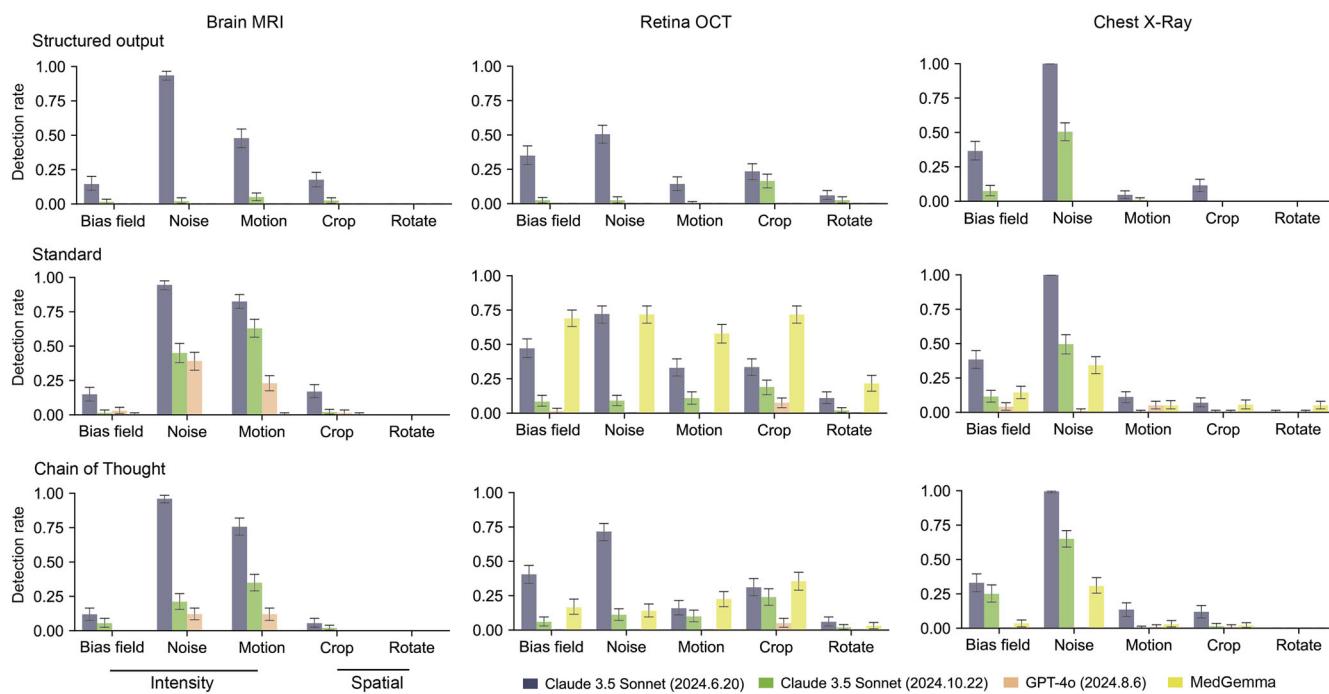
#### Strong artefacts detection rate

We assessed the VLMs' ability to detect strong artefacts in severely distorted images (Fig. 5). We excluded BiomedCLIP, as it provides only category probabilities, i.e. disease or normal. Llama 3.2 11B was not included as it consistently classified most cases as diseased (Fig. 3 and Supplementary Figs. 4, 5). Claude 3.5 Sonnet (2024.6.20) with the standard prompt achieved the highest strong artefacts detection rate in MRI and X-ray applications. For example, on brain MRI images with strong random motion, Claude 3.5 Sonnet (2024.6.20) achieved a detection rate of 0.825 (95% CI: 0.775–0.875), outperforming both its upgraded model (0.629; 95% CI: 0.565–0.695) and GPT-4o (0.230; 95% CI: 0.175–0.285). MedGemma showed relatively good capability in detecting strong artefacts in OCT applications compared to other models. Specifically, it achieved a detection rate of 0.690 (95% CI: 0.630–0.750) on OCT images containing random bias fields, while Claude 3.5 Sonnet (2024.6.20) reached 0.471 (95% CI: 0.405–0.540). Intensity artefacts were generally more easily detected by VLMs. For instance, Claude



**Fig. 4 | Accuracy of Vision-Language models (VLMs) detecting diseases in images with weak artefacts using standard prompts, and their performance percentage change compared to the accuracy obtained from original unaltered images.** We present only the top three models for lesion detection on original unaltered images because the remaining models performed poorly, with accuracies around 0.5 (Fig. 3), and their performance on distorted images provides limited insight into model robustness. **a** Shows the models' accuracy (y-axis) after adding weak artefacts to the images. Full models performance through various prompt strategies shown in Supplementary Figs. 6–8. **b** Illustrates the percentage change in models' accuracy (y-axis) after introducing weak artefacts. Positive values indicate increased

performance, whereas negative values represent decreased performance. *P* values comparing original model performance with performance after adding weak artefacts are shown above each bar. Unless otherwise specified,  $p < 0.001$ . \*\* indicates  $0.001 < p < 0.01$ , \* indicates  $0.01 < p < 0.05$  and ns indicates  $p \geq 0.05$ . Model performance percentage change through various prompt strategies shown in Supplementary Figs. 9–17. The complete quantitative results are presented in Supplementary Data 7, while all *p* values can be found in Supplementary Data 9. For each disease detection task, we performed 1000 iterations of stratified bootstrapping to calculate accuracy and sensitivity. Results show mean performance with 95% confidence intervals (error bars).



**Fig. 5 | Vision-Language models' (VLMs) detection rate for strong artefacts.** The y-axis shows the rate at which VLMs flag poor image quality in images distorted by strong artefacts, with higher values indicating greater effectiveness in artefact identification. Quantitative results are presented in Supplementary Data 10. For each

disease detection task, we performed 1000 iterations of stratified bootstrapping to calculate strong artefacts detection rate. Results show mean detection rates with 95% confidence intervals (error bars).

3.5 Sonnet (2024.6.20) achieved a 0.945 (95% CI: 0.910–0.975) detection rate when analysing MRI images with strong random noise, while showing no detecting capability for images with strong rotation.

### Role of prompt engineering

We analysed the impact of prompt engineering on VLMs' performance. When switching prompt strategies from requiring structured output to standard output, models became more capable of flagging poor quality images. For instance, in MRI applications, Claude 3.5 Sonnet (2024.10.22) performed better with the standard prompt for detecting images with random noise, with rate of 0.629 (95% CI: 0.565–0.695), compared to 0.051 (95% CI: 0.025–0.080) with the structured output prompt, as shown in Fig. 5. We also observed some unexpected findings. When detecting chest X-rays with weak artefacts, GPT-4o showed decreased accuracy and sensitivity after switching prompt strategies from requiring structured output to standard and CoT output. For example, its sensitivity decreased from 0.809 (95% CI: 0.720–0.880) with the structured output prompt to 0.331 (95% CI: 0.240–0.420) with the CoT prompt when detecting chest X-rays containing random motion artefacts, as shown in Supplementary Fig. 4.

### VLM robustness to real-world artefacts in applications with colour fundus photographs

We assessed VLM robustness in detecting diabetic retinopathy from colour fundus photographs containing real-world artefacts of varying scales (Supplementary Fig. 18). Weak artefact datasets contain images that are partially obscured yet still interpretable, whereas strong artefact datasets contain ungradable images that are images severely distorted and without clearly visible lesions. The first two subsets each include two categories, normal and diabetic retinopathy, with 50 images per category. The third subset contains severely distorted and ungradable images. When detecting high quality images, GPT-4o had the highest accuracy 0.731 (95% CI: 0.660–0.800) and highest specificity 0.960 (95% CI: 0.900–1.000). While Llama 3.2 11B had the lowest accuracy 0.470 (95% CI: 0.430–0.500), as it classified most cases as diabetic retinopathy. When detecting images with weak artefacts, most models had a drop in accuracy. Specifically, MedGemma demonstrated the weakest robustness to image artefacts, exhibiting the largest drops in performance: a 41% decrease in accuracy and a 77.5% decrease in specificity compared to its performance on images without artefacts. Similar to the findings observed in X-ray applications, Claude series models exhibited increased sensitivity after introducing weak artefacts; however, this did not result in improved accuracy. In detecting ungradable images, MedGemma achieved the highest detection rate for poor quality images, at 0.780 (95% CI: 0.700–0.860).

## Discussion

This study presents a robustness analysis of the VLMs to medical image artefacts in disease detection. We developed an evaluation benchmark and proposed to evaluate model robustness in three tier metrics. The results indicate that all VLMs perform moderately on original unaltered images, and significantly worse when weak artefacts were added. Few VLMs were able to recognise the poor quality images, showing low strong artefact detection rate. We observed a similar VLM robustness on a dataset with real-world artefacts, further revealing the limited capability of current VLMs in handling images with suboptimal quality.

VLMs do not perform well on disease detection tasks and perform even worse when artefacts are added to the original unaltered images. Prior work reports higher performance for GPT-4o on selected disease detection tasks<sup>25,27,28,37</sup>. In this study, BiomedCLIP in certain applications, such as disease detection with chest X-ray (Fig. 3) performed well, surpassing competitive and multifold larger VLMs. Previous research suggests that models pre-trained on medical data benefits certain domain-specific tasks<sup>38,39</sup>, which likely explains the good performance of BiomedCLIP of certain medical applications. This highlights the value of domain-specific pre-training for downstream clinical applications like disease detection. This is further evidenced by the poor performance of MedGemma in OCT

applications compared to applications in other medical imaging modalities, as it has not been specifically fine-tuned for this modality<sup>36</sup>. Part of our observations are consistent with past literature. Xia et al. investigated the robustness of VLMs to random noise and revealed that VLMs perform poorly under such condition<sup>27</sup>. More specifically, VLMs show poorer robustness against intensity artefacts, particularly random noise. This may be because spatial artefacts preserve more features compared to intensity artefacts (Supplementary Figs. 1–3). Additionally, despite image rotation only altering its orientation, this leads to performance degradation across all applications in our study, probably due to these models mostly trained on single-orientation images. In applications involving X-rays and colour fundus photographs with weak artefacts, Claude series models showed reduced accuracy but increased sensitivity, likely due to misinterpreting artefacts as lesions.

Few models are capable of flagging strong image artefacts in disease detection tasks. In recent studies, images in disease detection tasks are usually taken from public datasets with good image quality, without evaluating VLMs' performance in detecting the poor image quality<sup>13,15,16,40,41</sup>. In this study, we find that Claude 3.5 Sonnet (2024.6.20) demonstrates superior performance in detecting strong artefacts compared to other VLMs. Understanding the rationale behind this is important but challenging, as these are proprietary models with undisclosed training data and strategies, which further emphasises the need for data transparency in VLM research<sup>42,43</sup>. Nevertheless, our evaluation pipeline provides a comprehensive comparison of their robustness to medical image artefacts. With the advancement of more powerful open-source VLMs, we plan to evaluate their robustness to image artefacts and provide deeper insights into the factors driving model performance. We also observe that VLMs are better at flagging intensity artefacts. This is probably because intensity artefacts largely change the intensity distribution compared to spatial artefacts (e.g. rotation). In the intensity artefacts, we find that VLMs are better at identifying severely distorted images caused by random noise (Fig. 5), likely due to their ability to recognise Gaussian distributions in image intensity distribution patterns. In contrast, other intensity artefacts such as the random bias field, which do not follow Gaussian distributions, present greater detection challenges for VLMs and require enhanced processing methods.

The variation in VLMs' disease detection performance after adding weak image artefacts likely involves multiple factors. As shown in Fig. 4, the brain tumour detection performance for some VLMs was enhanced by introducing random cropping to MRI images. This improvement is likely due to cropping enabling the models to better focus on the anatomical tissue and lesion by removing extraneous blank background, with an example illustrated in Supplementary Fig. 21. We also observed that adding bias fields to brain MRI images improved the brain tumour detection accuracy for GPT-4o. This may be caused by varied selections in answered questions due to ethical safeguards within VLMs. When weak artefacts were added, some VLMs answered more questions than on the original images, while this didn't extend to Chest X-ray and retinal applications. Future work will investigate the reasons behind the variation in VLMs' performance after introducing certain types of image artefacts.

In certain cases, MedGemma's final output diverges from the reasoning it presents, with an example of this discrepancy shown in Supplementary Data 13. This conflict was not observed in other models, possibly because MedGemma has a smaller size, with limited response generation capability. We also observed that as the scale of artefacts increased, the conflict rate of MedGemma rose correspondingly (Supplementary Fig. 19). Notably, MedGemma exhibited a higher rate of conflict with intensity artefacts compared to spatial artefacts. This may suggest that images with greater deviations from the normal intensity distribution cause more confusion for the model, indicating its limited robustness to medical image artefacts.

VLMs show inconsistent disease detection performance across medical imaging modalities. In OCT applications, VLMs show higher specificity but lower sensitivity compared to other medical imaging modalities. This may be attributed to the smaller size of lesions in OCT images, which are less

distinguishable from normal tissue, leading to lower sensitivity. When detecting lesions on images with weak artefacts, VLMs show reduced sensitivity in MRI and OCT applications but increased sensitivity in X-rays. This paradox may stem from opacity being a key marker for pulmonary lesions—since healthy lungs appear with low opacity, artefacts that increase opacity may bias VLMs toward diseased predictions (Supplementary Fig. 3), raising sensitivity. In MRI applications, VLM performance typically improves with random cropping, whereas this effect is less evident in applications in other medical imaging modalities. This is likely because brain lesions are generally larger and centrally located, allowing cropping to help the model focus on the lesion (Supplementary Fig. 1d).

Allowing the reasoning process improves VLMs' performance in recognising poor image quality, although it brings limited benefit to disease detection performance in general. Previous works report that CoT prompting can improve the VLMs' performance on general tasks, particular tasks relying on logical thinking<sup>44–48</sup>. Whether this advantage extends to the medical domain, specifically in disease detection, has not been fully tested. In this study, we found that models with standard and CoT prompts outperformed structured output prompts in recognising poor image quality, possibly because they encourage models to focus more on the image itself. However, switching the prompt from structured output to standard or CoT formats could deteriorate the performance of some VLMs, particularly GPT-4o in X-ray applications (Fig. 3 and Supplementary Figs. 4, 5). This decline is caused by ethical safeguards of GPT-4o, refusing to answer more frequently with CoT prompts compared to structured output prompts. The ambiguous or refusal responses were categorised as incorrect answers in our study, as they provide limited valuable information to disease detection in clinical deployment. This choice has been widely used in other AI healthcare studies<sup>25,27,49,50</sup>. However, we acknowledge that this approach may underestimate the performance of certain VLMs in disease detection, particularly for models strongly constrained by ethical safeguards (e.g. GPT-4o on chest X-ray applications). The refusal or ambiguous outputs may reflect responsible model behaviour rather than an inability to perform the task. In the future, a benchmark considering both model performance and ethical safeguards should be designed for more comprehensive VLM evaluation.

A comprehensive evaluation of VLMs' robustness is essential for advancing their development and explainability. Current models fall short in identifying poor-quality images. Future VLMs could benefit from integrating a quality assessment module prior to diagnosis. Moreover, our results may encourage researchers to investigate the underlying causes of VLMs' suboptimal performance. For example, future work could visualise the attention maps from vision encoders to assess whether meaningful pathological features are captured across different medical images, particularly those with small lesion sizes (e.g. retinal OCT images in our study). Although the advantages of CoT reasoning may not be immediately apparent in current applications, prompt engineering remains a promising approach for enhancing disease detection performance, as shown in previous studies. Ferber et al. showed that in-context learning improves VLMs' classification of cancer pathology images<sup>50</sup>. More broadly, systematic exploration of advanced prompting strategies may lead to further performance gains.

By building the VLMs' robustness evaluation framework, we hope to accelerate the progress to better understanding and developing robust medical VLMs. Several limitations of this study warrant consideration. First, we introduced five categories of image artefacts commonly encountered in clinical scenarios. However, these may not fully capture the complexity of artefacts found in real-world medical imaging, where artefacts often appear in combination rather than isolation. It is worth exploring model robustness to combinations of image artefacts. Second, the intensity levels for weak and strong image artefacts were preset hyperparameters. Future research should provide a quantified threshold for weak and strong artefacts (e.g. quantifying the difference between images with weak artefacts and their original unaltered counterparts), therefore a more fine-grained scale of image artefacts will be explored. Third, disease detection tasks are generally simplified into a two-category format (normal vs. diseased), which is relevant to the

moderate performance of VLMs even in these simplified cases (Fig. 3). Future work will further refine the VLM robustness evaluation framework by simulating more realistic image artefacts, exploring fine-grained artefact scales and incorporating multi-class disease classification tasks. Additionally, a detailed quantitative analysis of the effects of certain image artefacts that produced varied performance shifts will be undertaken to better understand their impact on VLM performance.

In this study, we build a benchmarking framework to evaluate VLMs' robustness to image artefacts in three disease detection tasks across multiple medical imaging modalities. By pinpointing vulnerabilities tied to specific medical imaging modality and artefact types, this work provides insights for developing robust medical VLMs capable of functioning reliably in real-world clinical settings.

## Methods

### Benchmark dataset

The Brain Tumour MRI Dataset<sup>51</sup> includes images of three tumour types—gliomas, meningiomas and pituitary tumours—along with healthy scans. It combines three sources: Figshare<sup>52</sup>, the SARTAJ dataset<sup>53</sup>, and Br35H<sup>54</sup>. Images in Figshare were acquired post-Gd-DTPA injection at Nanfang Hospital and Tianjin Medical University Hospital (2005.9–2010.10); the other sources lack detailed provenance. Patient age and sex are not provided. Each MRI scan yields multiple images, labelled by experienced radiologists. The study sample comprised 100 diseased and 100 normal images.

The Retinal OCT Images dataset<sup>55</sup> includes three macular diseases—choroidal neovascularisation, diabetic macular oedema and drusen—alongside normal images. Images were collected from retrospective adult cohorts at five global institutions between July 1, 2013, and March 1, 2017. Patient age and sex are not provided. Multiple images per scan were labelled through a tiered grading process involving medical students, ophthalmologists and senior retinal specialists with over 20 years of experience. The study sample comprised 100 diseased and 100 normal images.

The COVID-19 Detection X-ray Dataset<sup>56</sup> includes three lung diseases—bacterial pneumonia, COVID-19 and viral pneumonia—alongside normal chest X-rays. X-rays with COVID-19 were sourced from public datasets<sup>54</sup>, while pneumonia and normal were from paediatric patients (ages 1–5) at Guangzhou Women and Children's Medical Centre<sup>57</sup>. Patient age, sex and number of scans per patient are not specified. Diagnoses were reviewed by two expert physicians. The study sample comprised 100 diseased and 100 normal images.

The benchmark consists of original unaltered images and those with artificially introduced artefacts at weak and strong scales (Supplementary Data 2). Weak artefact datasets consist of images that are partially obscured yet still interpretable, while strong artefact datasets contain ungradable images that are severely distorted by artefacts, making it impossible to reliably classify them as diseased or normal. Intensity artefacts (random noise, motion, bias fields) simulate issues from hardware noise and operational factors, applied using the TorchIO library<sup>58</sup>. Spatial artefacts (rotation, cropping) mimic misalignment and limited field of view during acquisition (Supplementary Figs. 1–3). The configuration and hyperparameter settings for all types of image artefacts at both weak and strong scales are provided in Supplementary Data 3. Questions were generated from ground truth labels to convert it into VQA tasks. This study did not involve the recruitment of human participants. All analyses were performed on publicly available, de-identified datasets (e.g. The Retinal OCT Images dataset<sup>55</sup>), with their details and links introduced in the 'Data availability' section. Additional ethical approval and informed consent were not required for this study.

To better assess the robustness of VLMs to compound artefacts existing in real-world data, we included DDR dataset, which contains diabetic retinopathy images of varying quality. Examples of image with compound artefacts are shown in Supplementary Fig. 20<sup>59</sup>. Collected from 147 hospitals across 23 provinces in China, the dataset lacks details on patient age, sex and scan count. Images were annotated by ophthalmologists and classified into six DR severity levels: none ('0'), mild non-proliferative ('1'), moderate non-proliferative ('2'), severe non-proliferative ('3'), proliferative ('4') and

ungradable ('5'). Ungradable images are those that are so severely distorted by artefacts that they cannot be reliably classified as diseased or normal. Due to the subtlety of mild DR, we grouped mild DR and non-DR ('0'–'1') as normal, '2'–'4' as diseased and '5' as strongly artefactual. We selected 50 high-quality images from both the normal and diseased classes. In addition, we selected 50 low-quality images from each of the normal and diseased classes to construct the weak artefact dataset, and 100 images from class '5' to represent the strong artefact dataset.

### Metrics for model robustness

We manually categorised VLM responses into four classes: '0' for those indicating normal findings without lesions, '1' for responses indicating image abnormalities, '2' for any cases where VLMs flag poor image quality and '3' for instances where VLMs refused to answer due to ethical concerns or suggested consulting a specialist. All results and manual annotations are provided in Supplementary Data 4. We used three metrics to evaluate VLM robustness.

**Model performance.** The accuracy and sensitivity of VLMs in disease detection. Some responses from VLM did not provide a specific answer between disease and normal (with answers '2' and '3'). We processed them as incorrect answers in calculating accuracy and sensitivity.

**Performance percentage change.** The performance difference between detecting diseases on images with weak artefacts and on original unaltered images, divided by original performance. We use relative performance change rather than the absolute difference to normalise performance changes against the baseline and make results comparable across tasks with different baseline accuracies. For example, suppose GPT-4o's accuracy in detecting brain tumours on MRI scans before and after adding weak noise are 0.7 and 0.5, respectively, the performance percentage change is:  $(0.5-0.7)/0.7 = -28.6\%$ .

**Strong artefacts detection rate.** The proportion of responses that flagged poor image quality in images with strong artefacts (with answer '2'), compared to the total number of responses.

### Models evaluated

This included three proprietary models—GPT-4o (2024.8.6)<sup>2</sup>, Claude 3.5 Sonnet (2024.6.20)<sup>3</sup> and the upgraded version Claude 3.5 Sonnet (2024.10.22). Claude 3.5 Sonnet was upgraded during the experiment; we hence tested the performance of both versions. Additionally, we tested three open-source models, Llama 3.2 with 11 billion parameters<sup>35</sup>, BiomedCLIP<sup>36</sup> and MedGemma<sup>36</sup>. BiomedCLIP and MedGemma are both designed for medical applications; however, BiomedCLIP provides only probability scores for each category in disease detection tasks. The chosen VLMs were selected for their popularity and strong performance across benchmarks based on previous literature<sup>36,60,61</sup>. Additionally, they have good practicality for clinicians and researchers to readily deploy and evaluate without extensive setting up and resource requirement. To maintain reproducibility, avoiding different responses for the same disease detection task, we set the temperature parameter as 0 for GPT-4o and both versions of Claude 3.5 Sonnet. For Llama 3.2, the temperature was configured to 0.01 to avoid the error that occurs when it is set to 0. BiomedCLIP and MedGemma do not define temperature parameters. The results and analysis in this study are fully reproducible. Although some proprietary VLMs didn't release the details about their training data and architectures, our open-source code to generate synthetic images and assess the robustness of VLMs is available at: [https://github.com/zijiecheng/VLM\\_robustness](https://github.com/zijiecheng/VLM_robustness). Additionally, all the benchmark data used for analysis in this study can be accessed at: <https://drive.google.com/drive/folders/1M7EldoSvxEMZjA9wJs52H1-4G2zTU8C>. As our study interprets VLM robustness to image artefacts based on VLM outputs, the study and its findings are reproducible and interpretable.

We tested three classical prompt strategies given that prompt substantially impacts VLMs' performance: (1) structured output

prompt, restricting the model to output only the categorical answer (e.g. 'diseased' or 'normal'); (2) standard prompt, a default prompt style allowing open answering without special restrictions; and (3) Chain of Thought (CoT) prompt<sup>44</sup>, including instructions such as 'Describe your reasoning step by step' to encourage detailed, sequential reasoning. The details of these prompt strategies are put in the Supplementary Data 5. The three prompt strategies fit into diverse real-world healthcare scenarios. The structured output prompt facilitates efficient community screening where VLMs provide quick decisions and triage flags for large-scale cohorts. By contrast, the standard prompt and chain of thoughts can provide more evidence to support clinical decision-making. In particular, the chain of thought offers step-wide reasoning which may give descriptions of abnormal patterns such as lesion on medical images for disease detection.

### Statistical analysis

We employed stratified bootstrapping with 1000 iterations to maintain class distributions. We calculated the mean and standard deviation of the performance over the 1000 iterations and the standard error by (standard deviation/5). We obtain the 95% CI by means of  $1.96 \times$  standard error. To assess the statistical significance of performance changes due to adding artefacts, we conducted the two-tailed *t*-test to calculate *p* values and put the result in Supplementary Data 9.

### Data availability

The benchmark dataset, which includes three imaging modalities with introduced artefacts, is derived from publicly available sources and can be accessed via the following links: MRI (<https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset>), OCT (<https://www.kaggle.com/datasets/paultimothymooney/kermany2018>) and X-ray (<https://www.kaggle.com/datasets/darshan1504/covid19-detection-xray-dataset>). We have reorganised, randomly sampled and added artefacts to create this benchmark, which is available at: [https://github.com/zijiecheng/VLM\\_robustness](https://github.com/zijiecheng/VLM_robustness). Additionally, the colour fundus dataset, containing real-world artefacts, is available at: <https://github.com/nkicsl/DDR-dataset>. We have reorganised and sampled it according to image distortion levels, and it can be found alongside the benchmark dataset.

### Code availability

The code used to introduce artefacts and evaluate the performance of VLMs from Z.C. is available at: [https://github.com/zijiecheng/VLM\\_robustness](https://github.com/zijiecheng/VLM_robustness). The code for introducing image artefacts uses TorchIO as the reference: <https://github.com/TorchIO-project/torchio>. The reference code for VLM evaluation is available for the following models: GPT-4o (<https://platform.openai.com/docs/quickstart>), Claude 3.5 Sonnet (<https://github.com/anthropic/anthropic-cookbook>), BiomedCLIP ([https://huggingface.co/microsoft/BiomedCLIP-PubMedBERT\\_256-vit\\_base\\_patch16\\_224](https://huggingface.co/microsoft/BiomedCLIP-PubMedBERT_256-vit_base_patch16_224)), Llama 3.2 (<https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct>) and MedGemma (<https://huggingface.co/google/medgemma-4b-pt>). Results were further analysed and visualised using Python v.3.9.18, NumPy v.1.26.2, Matplotlib v.3.8.0, SciPy v.1.11.3 and pandas v.1.5.0.

Received: 12 July 2025; Accepted: 22 October 2025;

Published online: 27 November 2025

### References

1. Bordes, F. et al. An introduction to vision-language modeling. *arXiv [cs.LG]* <https://doi.org/10.48550/arXiv.2405.17247> (2024).
2. OpenAI et al. GPT-4 Technical Report. *arXiv [cs.CL]* <https://doi.org/10.48550/arXiv.2303.08774> (2023).
3. Anthropic et al. The Claude 3 Model Family: Opus, Sonnet, Haiku. <https://www.anthropic.com> (2024).
4. Yang, X., Wu, Y., Yang, M., Chen, H. & Geng, X. Exploring diverse in-context configurations for image captioning. *Adv. Neural Inf. Process. Syst.* **36**, 40924–40943 (2023).

5. He, S. et al. Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. *CoRR* <http://arxiv.org/abs/2404.15127> (2024).
6. Li, C. et al. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. *Adv. Neural Inf. Process. Syst.* **36**, 28541–28564 (2023).
7. Zhang, S. et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv [cs.CV]* <https://doi.org/10.48550/arXiv.2303.00915> (2023).
8. Murray, R. & Werpy, N. Image interpretation and artefacts. in *Equine MRI* 101–145 (Wiley Online Library, 2010).
9. Goceri, E. Medical image data augmentation: techniques, comparisons and interpretations. *Artif. Intell. Rev.* **56**, 12561–12605 (2023).
10. Hindi, A., Peterson, C. & Barr, R. G. Artifacts in diagnostic ultrasound. *Rep. Med. Imaging* **6**, 29–48 (2013).
11. Arabi, H. & Zaidi, H. Deep learning-based metal artefact reduction in PET/CT imaging. *Eur. Radiol.* **31**, 6384–6396 (2021).
12. Garcea, F., Serra, A., Lamberti, F. & Morra, L. Data augmentation for medical imaging: a systematic literature review. *Comput. Biol. Med.* **152**, 106391 (2023).
13. Eggers, G. et al. Artefacts in magnetic resonance imaging caused by dental material. *MAGMA* **18**, 103–111 (2005).
14. Ray, R., Stinnett, S. S. & Jaffe, G. J. Evaluation of image artifact produced by optical coherence tomography of retinal pathology. *Am. J. Ophthalmol.* **139**, 18–29 (2005).
15. Boas, F. E. & Fleischmann, D. CT artifacts: causes and reduction techniques. *Imaging Med.* **4**, 229–240 (2012).
16. Bazvand, F. & Ghassemi, F. Artifacts in macular optical coherence tomography. *J. Curr. Ophthalmol.* **32**, 123–131 (2020).
17. Spaide, R. F., Fujimoto, J. G. & Waheed, N. K. Image artifacts in optical coherence tomography angiography. *Retina* **35**, 2163–2180 (2015).
18. Long, N. M. & Smith, C. S. Causes and imaging features of false positives and false negatives on F-PET/CT in oncologic imaging. *Insights Imaging* **2**, 679–698 (2011).
19. Gelbman, B. D. et al. Radiographic and clinical characterization of false negative results from CT-guided needle biopsies of lung nodules. *J. Thorac. Oncol.* **7**, 815–820 (2012).
20. Howarth, D. M., Forstrom, L. A., O'Connor, M. K., Thomas, P. A. & Cardew, A. P. Patient-related pitfalls and artifacts in nuclear medicine imaging. *Semin. Nucl. Med.* **26**, 295–307 (1996).
21. Millet, I. et al. Pearls and pitfalls in breast MRI. *Br. J. Radiol.* **85**, 197–207 (2012).
22. Ali, S. et al. An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. *Sci. Rep.* **10**, 2748 (2020).
23. Howell, L., Ingram, N., Lapham, R., Morrell, A. & McLaughlan, J. R. Deep learning for real-time multi-class segmentation of artefacts in lung ultrasound. *Ultrasonics* **140**, 107251 (2024).
24. Elyounssi, S. et al. Addressing artifactual bias in large, automated MRI analyses of brain development. *Nat. Neurosci.* **1**, 10 (2025).
25. Ye, J. et al. Gmai-MMBench: A comprehensive multimodal evaluation benchmark towards general medical AI. *Adv. Neural Inf. Process. Syst.* **37**, 94327–94427 (2024).
26. Gu, Z., Chen, J., Liu, F., Yin, C. & Zhang, P. MedVH: Toward systematic evaluation of hallucination for large vision language models in the medical context. *Adv. Intell. Syst.* **7**, 2500255 (2025).
27. Xia, P. et al. CARES: A comprehensive benchmark of trustworthiness in medical vision language models. *Adv. Neural Inf. Process. Syst.* **37**, 140334–140365 (2024).
28. Royer, C., Menze, B. & Sekuboyina, A. Multimedeval: A benchmark and a toolkit for evaluating medical vision-language models. *MIDL* *arXiv:2402.09262* (2024).
29. Jiang, Y. et al. Evaluating general vision-language models for clinical medicine. *medRxiv* <https://doi.org/10.1101/2024.04.12.24305744> (2024).
30. Yildirim, N. et al. Multimodal healthcare AI: identifying and designing clinically relevant vision-language applications for radiology. *CHI Conf. Hum. Factors Comput. Syst.* 1–22 (2024).
31. van Diessen, E., van Amerongen, R. A., Zijlmans, M. & Otte, W. M. Potential merits and flaws of large language models in epilepsy care: a critical review. *Epilepsia* **65**, 873–886 (2024).
32. Nasarian, E., Alizadehsani, R., Acharya, U. R. & Tsui, K.-L. Designing interpretable ML system to enhance trust in healthcare: a systematic review to proposed responsible clinician-AI-collaboration framework. *Inf. Fusion* **108**, 102412 (2024).
33. Tanno, R. et al. Collaboration between clinicians and vision-language models in radiology report generation. *Nat. Med.* **31**, 599–608 (2025).
34. Zhang, J. et al. A comparative study of GPT-4o and human ophthalmologists in glaucoma diagnosis. *Sci. Rep.* **14**, 30385 (2024).
35. Grattafiori, A. et al. The Llama 3 herd of models. *arXiv [cs.AI]* <https://doi.org/10.48550/arXiv.2407.21783> (2024).
36. Sellergren, A. et al. MedGemma Technical Report. *arXiv [cs.AI]* <https://doi.org/10.48550/arXiv.2507.05201> (2025).
37. VVan, M.-H., Verma, P. & Wu, X. On large visual language models for medical imaging analysis: An empirical study. In *2024 IEEE/ACM Conf. Connected Health: Appl. Syst. Eng. Technol.* 172–176 (2024).
38. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
39. Xie, Q. et al. Medical foundation large language models for comprehensive text analysis and beyond. *NPJ Digit. Med.* **8**, 141 (2025).
40. van Dalen, J. A., Vogel, W. V., Corstens, F. H. M. & Oyen, W. J. G. Multi-modality nuclear medicine imaging: artefacts, pitfalls and recommendations. *Cancer Imaging* **7**, 77–83 (2007).
41. Krupa, K. & Beksińska-Figatowska, M. Artifacts in magnetic resonance imaging. *Pol. J. Radiol.* **80**, 93–106 (2015).
42. Alderman, J. E. et al. Tackling algorithmic bias and promoting transparency in health datasets: The STANDING Together consensus recommendations. *Lancet Digit. Health* **7**, e64–e88 (2025).
43. Transparency (in training data) is what we want. *Nat. Mach. Intell.* **7**, 329–329 (2025).
44. Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **35**, 24824–24837 (2022).
45. Feng, G. et al. Towards revealing the mystery behind chain of thought: a theoretical perspective. *Adv. Neural Inf. Process. Syst.* **36**, 70757–70798 (2023).
46. Zhang, F., Ling, Y., Yang, J., Zhang, P. & Zhang, G. COVID-19 knowledge mining based on large language models and chain-of-thought reasoning. *BIBM*, 6574–6581 (2024).
47. Li, M., Zhou, H., Yang, H. & Zhang, R. RT: a Retrieving and Chain-of-Thought framework for few-shot medical named entity recognition. *J. Am. Med. Inform. Assoc.* **31**, 1929–1938 (2024).
48. Zhang, Z., Zhang, A., Li, M. & Smola, A. Automatic chain of thought prompting in large language models. *ICLR* *arXiv:2210.03493* (2022).
49. Jeong, D. P., Garg, S., Lipton, Z. C. & Oberst, M. Medical adaptation of large language and vision-language models: Are we making progress? *EMNLP* *arXiv:2411.04118* (2024).
50. Ferber, D. et al. In-context learning enables multimodal large language models to classify cancer pathology images. *Nat. Commun.* **15**, 10104 (2024).
51. Nickparvar, M. Brain Tumor MRI Dataset. *Kaggle*. <https://doi.org/10.34740/kaggle/dsv/2645886> (2021).
52. Cheng, J. Brain tumor dataset. *figshare*. <https://doi.org/10.6084/M9-FIGSHARE.1512427.V1> (2015).
53. Bhuvaji, S., Kadam, A., Bhumkar, P., Dedge, S. & Kanchan, S. Brain Tumor Classification (MRI). *Kaggle*. <https://doi.org/10.34740/kaggle/dsv/12745533> (2020).
54. Hamada, A. H. Br35H: Brain Tumor Detection 2020. *Kaggle*. <https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection> (2021).

55. Kermany, D. Labeled optical coherence tomography (OCT) and Chest X-Ray images for classification. *Mendeley*. <https://doi.org/10.17632/rscbjbr9sj.2> (2018).

56. Deshpande, D. COVID-19 Detection X-Ray Dataset. *Kaggle*. <https://www.kaggle.com/datasets/darshan1504/covid19-detection-xray-dataset> (2020).

57. Kermany, D. S. et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131.e9 (2018).

58. Pérez-García, F., Sparks, R. & Ourselin, S. TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Comput. Methods Programs Biomed.* **208**, 106236 (2021).

59. Li, T. et al. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Inf. Sci.* **501**, 511–522 (2019).

60. OpenAI et al. GPT-4o System Card. *arXiv [cs.CL]* <https://doi.org/10.48550/arXiv.2410.21276> (2024).

61. Ma, J. et al. Large language model evaluation in autoimmune disease clinical questions comparing ChatGPT 4o, Claude 3.5 Sonnet and Gemini 1.5 pro. *Sci. Rep.* **15**, 17635 (2025).

D.A.M., L.J., H.Z., R.C., L.P., B.L., T.H., A.R.R., H.J., D.Y., K.Z., J.H.L.G., S.S., A.A., D.C.A., C.Y.C., Y.C.T., P.A.K. and Y.Z. contributed to the drafting and revising of the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41746-025-02108-w>.

**Correspondence** and requests for materials should be addressed to Zijie Cheng or Yukun Zhou.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025

## Acknowledgements

Y.Z. is supported by a Wellcome Trust Early-Career Award (grant no. 318987/Z/24/Z). P.A.K. is supported by a Moorfields Eye Charity Career Development Award (grant no. R190028A) and a UK Research & Innovation Future Leaders Fellowship (grant no. MR/T019050/1). For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising. We acknowledge the computational resources supported by the UCL Computer Science Cluster and the Advanced Research Computing UAI platform.

## Author contributions

Z.C., B.L., L.J., T.H., C.Y.C. and Y.Z. contributed to the conception and design of the work. Z.C. and Y.Z. contributed to the data organisation, technical implementation and analysis framework. Z.C., A.Y.O., S.K.W.,