University of New Haven

Ph.D. Dissertation

Paraphrase Sensitivity in Medical Vision-Language Models: Measurement, Mechanisms, Mitigation, and Deployment Safety

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY IN ENGINEERING AND APPLIED SCIENCE

BY

Binesh Sadanandan

UNIVERSITY OF NEW HAVEN
West Haven, Connecticut

August 2026

Paraphrase Sensitivity in Medical Vision-Language Models: Measurement, Mechanisms, Mitigation, and Deployment Safety

SUBMITTED BY:

Binesh Sadanandan
Ph.D. Candidate, Author

APPROVED BY:

Vahid Behzadan, Ph.D.
Committee Chairperson and Advisor; Associate Professor, University of New Haven

Khaled Sayed, Ph.D.
Committee Member; Assistant Professor, Electrical & Computer Engineering and Computer Science, University of New Haven

Muhammad Aminul Islam, Ph.D.
Committee Member; Assistant Professor, Electrical & Computer Engineering and Computer Science, University of New Haven

Xenophon Papademetris, Ph.D.
External Committee Member; Professor of Biomedical Informatics & Data Science, Yale School of Medicine

Shue Wang, Ph.D.
Coordinator, Doctoral Program, University of New Haven

Ronald Harichandran, Ph.D.
Dean, Tagliatela College of Engineering, University of New Haven

Nancy Ortins Savage, Ph.D.
Provost and Senior Vice President for Academic Affairs, University of New Haven

Acknowledgments

I am deeply grateful to my advisor, Dr. Vahid Behzadan, for his guidance, patience, and mentorship throughout this work. His questions shaped the direction of this dissertation and pushed me to make its claims sharper and its evidence stronger.

I thank the members of my dissertation committee for their careful reading and constructive feedback, which improved both the rigor and the clarity of this research. I am grateful to the clinician collaborators who lent their radiological and clinical expertise to the safety evaluation: Dr. Lekshmy Jayan (Department of Radio-Diagnosis, Mount Zion Medical College Hospital, Adoor, Kerala, India) and Dr. Arun Gopinatha Kurup (Badr Al Samaa Hospitals, Nizwa, Muscat, Oman), whose clinical judgment grounded the technical findings. I also thank Dr. Shue Wang, coordinator of the doctoral program, for guiding me through its requirements and milestones.

To my colleagues in the Secure and Assured Intelligent Learning (SAIL) Lab at the University of New Haven, thank you for the discussions, the shared compute, and the day-to-day companionship that made this work possible.

I am also grateful to my colleagues and mentors at Medtronic, in particular Jason Williams, Andrew Miesse, Gerald Hodgkinson, Alfonso Aranda Hernandez, Jeffrey Mulreed, and Adam Haas, whose encouragement and belief in this work kept me going through the long stretch of balancing research with a full-time engineering career. This degree was made possible in part by Medtronic's education reimbursement program, and I am thankful for the company's investment in my growth.

To my father and mother, thank you for your constant motivation and for always being there with a helping hand, through this journey and every one before it. None of this happens without you.

I remember my late father-in-law, Somanpillai, and mother-in-law, Sujatha, with love and gratitude. Their kindness and support touched our family deeply, and I wish they could have seen this finished.

Finally, to my wife Anju and our children, Arwin and Nitara: this dissertation and the research behind it were done with the hours you gave me and sacrificed, on evenings and weekends that belonged to you. Anju carried far more than her share so I could do this work, and Arwin and Nitara reminded me daily why any of it matters. This is as much yours as it is mine.

Dedication

To Anju, Arwin, and Nitara

Ethical Statement

This dissertation studies the reliability and safety of medical Vision-Language Models. Its central finding is that some models achieve apparent consistency while their answers change little when the medical image is removed or replaced with one showing the opposite finding, so answers that look trustworthy can add little to what the question text alone would support. Disclosing this failure mode, and the evaluation tools that expose it, is intended to make the deployment of medical AI safer, not to enable misuse.

The research uses only de-identified third-party chest radiograph datasets. NIH ChestX-ray14 was obtained from the openly released NIH Clinical Center set. MIMIC-CXR and VinDr-CXR were obtained through credentialed PhysioNet access under their data use terms, while PadChest, including the PadChest-GR subset used by the probe, was obtained from the Medical Imaging Databank of the Valencia Region (BIMCV) under its research use agreement. No new patient data was collected and no protected health information was accessed. The clinical validation described in this work is a consultation with clinician collaborators and is not a human-subjects study.

The methods and benchmarks developed here are diagnostic: they measure when a model should not be trusted. We release them so that model developers and clinical adopters can audit systems before deployment, and we caution explicitly against interpreting a low paraphrase-flip rate as evidence of safety. The overarching goal of this dissertation is to support the responsible, evidence-based use of medical Vision-Language Models in ways that benefit patients while minimizing the risk of confident errors that the image did little to inform.

Use of large language models. Large language models played two distinct roles in this dissertation. Inside the research, they are components of the systems under study: the PSF-Med paraphrase bank was generated and filtered by an LLM, a second model family provided cross-family equivalence checks, and the medical Vision-Language Models under evaluation are themselves built on language models. These uses are documented in the chapters where they arise. Outside the research, Anthropic's Claude served as an editing aid, used for copy editing prose, formatting LaTeX and figures, and checking numbers and cross-references in the text against the underlying experimental artifacts. No result, analysis, or conclusion was produced by a language model: the research questions, experimental design, experiments, and interpretation are the author's own, and every AI-assisted passage was reviewed and verified by the author before inclusion.

Contents

List of Tables

List of Figures

List of Publications

The following publications form the basis of this dissertation. Chapter mappings are indicated in brackets.

Peer-reviewed / accepted

1. **PSF-Med: A Paraphrase Sensitivity Failure Benchmark for Medical Vision-Language Models.** Accepted, MMFM-BIOMED Workshop, CVPR 2026. *[Chapter 3]*
2. **Mechanistically-Guided Targeted LoRA for Paraphrase Consistency in Medical VLMs.** Accepted, Conference on Health, Inference, and Learning (CHIL) 2026. *[Chapter 5]*
3. **Consistent but Dangerous: When Paraphrase Consistency Hides Ungrounded Medical VLMs.** Accepted, CVPR 2026 Workshop (MedReasoner). *[Chapter 6]*
4. **Chain-of-Thought Backfires: Reasoning-Induced Failures in Language Models** (companion, text-only). Accepted, 2AI Conference 2026.
5. **Attention Without Grounding: Causal Evaluation of Visual Explanations in Medical VLMs.** Accepted, iMIMIC 2026 (Interpretability of Machine Intelligence in Medical Image Computing), 9th edition, a satellite workshop of MICCAI 2026 (archival, Springer LNCS). *[Chapter 6]*
6. **Predictive Entropy as a Joint Screen for Error and Paraphrase Instability in Medical VLMs.** Accepted, UNSURE 2026 (Uncertainty for Safe Utilization of Machine Learning in Medical Imaging), 8th edition, a satellite workshop of MICCAI 2026 (archival, Springer LNCS). *[Chapter 7]*
7. **VSF-Med.** Early pilot abstract, ISBI 2026; established the initial paraphrase-vulnerability and attention-stability setup that matured into PSF-Med.

Under review / in revision

8. **A Scoping Review of Trustworthiness Evaluation for Medical Vision-Language Models.** Under review, JMIR AI. *[Chapter 2]*

List of Abbreviations

AUGRC	Area Under the Generalized Risk-Coverage curve
AUROC	Area Under the Receiver Operating Characteristic curve
AURC	Area Under the Risk-Coverage curve
CXR	Chest X-Ray
ECE	Expected Calibration Error
FVU	Fraction of Variance Unexplained
LLM	Large Language Model
LoRA	Low-Rank Adaptation
MC Dropout	Monte Carlo Dropout
MI	Mutual Information
NLP	Natural Language Processing
OOD	Out-of-Distribution
PSF	Paraphrase Sensitivity Failure
ROI	Region of Interest
SAE	Sparse Autoencoder
UQ	Uncertainty Quantification
VLM	Vision-Language Model
VQA	Visual Question Answering

Abstract

Medical vision-language models are judged reliable when they answer the same clinical question the same way. They often fail even this basic test: across six models and three clinical populations, the answer changes on 6.4% to 54.7% of paraphrase pairs. The natural response is to train for consistency, and that is where the harder problem appears. A model can be consistent because it reads the chest radiograph the same way each time, or because its answer is driven by the language of the question and changes little when the radiograph is taken away. When we look, the second case is common: across ten model-dataset settings, a mean of 81% of each model's consistent answers are unchanged when the image is removed. In this thesis we study paraphrase sensitivity in medical vision-language models along both of these layers, and we show that neither inconsistency nor consistency, on its own, tells us whether a model is using the image.

We approach the problem in five stages: measure the failure, diagnose where it comes from, reduce it, evaluate whether the reduction is safe, and test what a deployment-time monitor can and cannot catch. We build PSF-Med, a benchmark of 26,850 chest X-ray questions and 92,856 evaluation pairs across three clinical populations, and show that six medical vision-language models flip on 6.4% to 54.7% of paraphrase pairs. We separate consistency from image dependence with controlled image-removal and image-swap experiments, and find that across ten model-dataset settings a mean of 81% of each model's consistent predictions are unchanged when the image is removed. We apply sparse autoencoders to locate where the failure is expressed, and use that account to design a targeted intervention that reduces the flip rate by about 59% on patients and images never seen in training while modifying 0.1% of the parameters. We then re-evaluate that intervention with the same controls used to indict the base models, and characterize a single-forward-pass uncertainty signal for deployment. To check that this failure begins on the language side rather than in the image, we train a scaled-down, domain-adapted replica of MedGemma's architecture, a compact Gemma-3 decoder over a frozen MedSigLIP-448 encoder, so that every paraphrase of a question is answered from identical image features and any answer change is language-side by construction, and we train it on two hospitals while holding out two others. Broadening the training-phrasing distribution lowers its flip rate from 67.1% to 4.8%, and to 26.6% when the test wording is withheld from training, while a rank-1 edit at its early layers restores the flipped answer. The yes-minus-no margin, the scalar a binary answer discards, catches a paraphrase flip in one forward pass at an area under the curve of 0.923 to 0.974, whereas an image-unreliant answer carries no single-pass signal.

The methods and the benchmark have been reviewed outside the university, with six papers drawn from this work accepted at peer-reviewed venues. The central conclusion is that safety claims for medical vision-language models require joint evaluation of semantic invariance, correctness, image dependence, and calibration, reported on margins rather than on binary answers alone, because a model optimized for consistency alone may answer the same way whether or not the patient's radiograph shows the finding.

Chapter 1

Introduction

1.1 Motivation

A medical Vision-Language Model (VLM) reads a chest radiograph and answers a clinical question about it, such as whether a pneumothorax is present. In practice the same question reaches the model in many forms. One clinician writes "Is there a pneumothorax?", another writes "Do you see air in the pleural space?", and a reporting template writes "Evaluate for pneumothorax." These are the same question, and a reliable model should give the same answer to all of them. The first finding of this thesis is that current models do not. Across six medical vision-language models and three clinical populations, the answer changes on 6.4% to 54.7% of paraphrase pairs. In a setting where the wording of a question is not standardized, a model that agrees with itself only most of the time is already a problem.

The natural response is to train the model to be consistent, and consistency can indeed be improved. This is where the harder problem appears. A model can hold its answer fixed across phrasings because it reads the radiograph the same way each time, which is what we want, or because its answer comes from a prior over how such questions are usually answered and does not change when the radiograph is taken away. These two are indistinguishable on any metric that only checks whether answers agree. And when we look, the second case is common. Across ten model-dataset settings, a mean of 81% of each model's consistent answers are unchanged when the image is removed entirely, and the models that flip the least are often the ones relying least on the image.

So paraphrase sensitivity in medical vision-language models has two layers. A model may be inconsistent, giving different answers to the same question, or it may be consistent for the wrong reason, giving a stable answer that does not depend on the image. In this thesis we study both, and we formulate the problem as follows. Given a clinical question q about an image I, and a set of clinically equivalent paraphrases of q, we ask two questions of the model. Does its answer stay the same across the paraphrases, and does its answer depend on the image? The field measures only the first. The two properties are distinct, and the gap between them is where unsafe behavior hides.

1.2 Why Models Fail This Way

It helps to ask why a medical vision-language model would answer from the question rather than the image in the first place. The thesis investigates this in its mechanistic chapter, so we state here the candidate explanations it works from, ordered from the ones that are already measurable to the ones that remain hypotheses. They are not exclusive, and together they describe a model that has the option to ignore the image, no reason not to, and a mechanism by which a rephrasing flips the answer.

The first explanation is the simplest: for many clinical questions, the words are enough. Chest X-ray findings are not evenly distributed. In the datasets studied here, PadChest questions are positive

83% of the time, and VinDr questions are positive by construction, so a model that answers from the base rate of a finding, without consulting the image, is right most of the time. Standard accuracy rewards this. A model has little pressure to consult the image when the prior alone gives the expected answer, and paraphrase consistency is then satisfied trivially by a model that reads only the words. This explanation accounts for consistency without grounding; it does not by itself explain why a rephrasing flips an answer.

The second explanation is a matter of how these models are trained and scored. Medical vision-language models are typically fine-tuned and evaluated on one fixed wording per question, with no signal that rewards the same answer across phrasings and no signal that penalizes reaching an answer without the image. Nothing in the objective asks the model to be invariant to phrasing or dependent on the image, so there is little reason to expect it to be either. Chapter 5 constructs this failure deliberately to size it. A model trained on phrasings whose register is correlated with the answer flips on 88.4% of paraphrase pairs, against 67.1% for a model trained on one fixed phrasing and 4.8% for one trained on all of them, with accuracy falling in the same order (58.1%, 66.8%, and 75.3%). The two narrow regimes separate from each other (Cliff's $\delta = 0.97$, $p = 3.1 \times 10^{-4}$), so a training distribution in which the wording predicts the answer is a distinct and larger harm than one that is merely narrow.

The third explanation is mechanistic, and it is the one this thesis tests rather than assumes. If the answer depends on the surface form of the question, some internal representation must carry that surface form into the decision. In Chapter 5 we identify a candidate: a sparse feature at layer 17 that responds to the register and operator of a question, separating a presence framing such as "Is there X?" from an exclusion framing such as "Can X be ruled out?". A paraphrase that shifts register moves this feature, and the shift propagates toward the answer. We present this as a candidate account rather than a settled mechanism, because a clean causal test with matched controls is still required, but it gives a concrete place to look for the origin of the failure.

A fourth explanation concerns the pathway rather than either end of it. The models studied here join a vision encoder to a language model through a learned projection, and the evidence suggests the strength of that join matters more than the quality of the encoder. The most text-reliant model in our study pairs a biomedical vision encoder with its language model, yet leans on the question more than a model with a general encoder does. This points to the pathway that connects vision and language, rather than the vision features themselves, as the place where image evidence is lost.

An earlier and smaller version of the probe model of Chapter 5 tests this directly, holding the encoder, its features, and the decoder architecture fixed and varying only how the encoder's output reaches the decoder. That earlier probe was trained on a different pair of hospitals and evaluated without further training on NIH ChestX-ray14. Reading the projected patch tokens in the usual way, its decoder is at chance there, with a mean area under the receiver operating characteristic curve of 0.500 over five seeds; a supervised probe fitted to its internal representation reaches only 0.518, so the visual evidence is absent from the representation rather than merely unexpressed. Supplying the same encoder's own attention-pooled embedding as one additional token raises the mean to 0.604, with wide seed-to-seed spread, and the gain disappears when that token is shuffled across images (0.502), so it is specific to the radiograph in question. The evidence was available in the encoder throughout: a linear probe on the encoder's own pooling head separates the same findings at 0.812, against 0.514 for the same probe on the pooled patch tokens the decoder is handed. What changed was whether the join delivered it. The scaled probe of Chapter 5 carries that grounding token and is not blind: it answers at 74.8% accuracy with the image and at exactly 50.0% without it. The mechanism is that the encoder concentrates its finding signal in a trained pooling head, while the interface hands the decoder the unpooled patch tokens and leaves it to relearn an equivalent readout, which a small decoder does not do when a phrasing shortcut is available. This supports the hypothesis without settling it for the

deployed models, whose decoders are large enough to learn a readout the probe's cannot.

A fifth explanation attributes the failure to capacity rather than to training. Modern vision-language models are heavily overparameterized, and in a network of this depth a small shift in the embedding of a rephrased question can propagate and grow across layers until it crosses a decision boundary [1, 2]. On this account, paraphrase sensitivity is a generic property of large deep networks, not something specific to medical training data. Two results in this thesis bound how much of the failure this account can carry on its own. First, scale within the MedGemma family buys no stable consistency advantage: the 27B model is the most consistent on MIMIC but not on PadChest (Chapter 3). Second, the probe of Chapter 5 holds the architecture and parameter count fixed and varies only the training phrasing distribution, and the flip rate moves from 4.8% to 67.1% to 88.4%. Capacity may set the stage, but the training distribution decides what plays out on it.

The thesis measures the first explanation, argues the second, investigates the third, provides a first controlled test of the fourth, and bounds the fifth.

1.3 Background

Three lines of work set up the problem. Each has made real progress, and each leaves the specific gap this thesis addresses. Chapter 2 reviews the trustworthiness-evaluation literature in full; here we state only the three advances that lead into our challenges.

1. **Accuracy of medical vision-language models**. A series of models, including MedGemma, LLaVA-Rad, and CheXagent, now reach strong accuracy on fixed chest X-ray question answering. This progress is measured almost entirely on one fixed wording per question, so it says nothing about behavior under rephrasing, and nothing about whether a correct answer came from the image or from the question.
2. **Robustness evaluation**. Work in general language and vision has long shown that models are sensitive to input perturbation, including paraphrasing. In the medical setting this evaluation is rare, and where it exists it seldom controls whether a paraphrase truly preserves clinical meaning, and almost never separates a stable answer that uses the image from a stable answer that ignores it.
3. **Mechanistic interpretability**. Sparse autoencoders and attention analysis now allow feature-level inspection of what a model represents. These tools have been applied mostly to general models and to localization, not to diagnosing a specific clinical failure or to guiding an intervention against it.

1.4 Challenges

Although each of these lines has been successful on its own terms, a few critical challenges remain for anyone who wants to know whether a medical vision-language model is reliable under rephrasing.

- There is no benchmark that measures paraphrase sensitivity in medical vision-language models at scale, with control over whether a paraphrase preserves clinical meaning, and across more than one clinical population.
- Consistency and image dependence are evaluated separately, never together. It is therefore unclear whether a low flip rate is evidence that a model uses the image, or only evidence that it is stable for some reason.
- The metric that would settle this is itself blind to it. Binary accuracy on a balanced set is read off a single decision threshold per finding, so a model given no image evidence, a model given another patient's, and a model given its own can all score near 50% while what they have to work with differs completely. Whether a model uses the image is therefore not only unmeasured in current practice, it is hard to see through the measurement that practice reports.

- There is no account of where in the model paraphrase sensitivity is expressed, and therefore no principled place to intervene.
- It is unknown whether paraphrase sensitivity can be reduced, and if it can, whether the reduction improves the model's use of the image or merely trades one shortcut for another.
- At deployment time there is no cheap signal that flags a prediction as risky, and no audit that catches a model that is consistent because it is reading the text rather than the image.

1.5 Contributions

This thesis addresses these challenges with a single framework and five studies that instantiate it.

The framework treats a prediction as an object with two properties rather than one. For a question q about an image I, let $a(q, I)$ be the model's answer, let $P(q)$ be a set of clinically equivalent paraphrases of q, and let $I_\varnothing$ be the same input with the image removed. We define

$$c(q, I) = \mathbb{1}[\, a(p, I) = a(q, I) \text{ for all } p \in P(q) \,], \tag{1.1}$$

$$g(q, I) = \mathbb{1}[\, a(q, I) \neq a(q, I_\varnothing) \,], \tag{1.2}$$

where c records whether the answer is *consistent* across paraphrases and g records whether the answer *depends on the image*. The field measures c and reads it as safety. This thesis measures the pair (c, g). The prediction of concern is the one that is consistent but not image-dependent, $c = 1$ and $g = 0$: a stable answer that would be the same whether or not the radiograph showed the finding. Correctness is a third, separate axis, measured against the reference label.

Each thrust answers one of the challenges above, and a controlled probe isolates the cause of the failure itself.

- **Thrust 1 (Measurement, Chapter 3).** We build PSF-Med, a benchmark of 26,850 chest X-ray questions and 92,856 evaluation pairs across MIMIC-CXR in the United States, PadChest in Spain, and VinDr-CXR in Vietnam. Paraphrases are filtered by a rubric-based equivalence audit over 122,778 pairs and spot-checked against a 1,200-pair clinician adjudication drawn before the second PadChest iteration. On this benchmark, six medical vision-language models flip on 6.4% to 54.7% of paraphrase pairs once two degenerate single-class cells are set aside. This answers the first challenge: paraphrase sensitivity is common, and it varies by model and population.
- **Thrust 2 (Diagnosis, Chapter 4).** We separate c from g with controlled image-removal and image-swap experiments. A text-reliant model is stable because it leans on the question, while a more image-dependent model is less stable but reacts to the image. This answers the second challenge: a low flip rate is not evidence of image use.
- **Thrust 3 (Mechanism and mitigation, Chapter 5).** We apply sparse autoencoders to locate where the failure is expressed, identifying a candidate two-stage account through Feature 3818 at layer 17 and Feature 12139 at layer 29. We use this account to design a targeted low-rank adaptation on layers 15 to 19, which reduces the flip rate from 8.5% to 3.5% over five seeds, about 59%, on a split that shares no patient and no image with training, while modifying 0.1% of the parameters and showing no observed accuracy reduction. This answers the fourth and fifth challenges.
- **Thrust 4 (Safety evaluation, Chapter 6).** We turn the same controls on the field's assumptions and on our own intervention. A four-quadrant screen built on (c, g) shows that across ten model-dataset settings a mean of 81% of consistent predictions are image-invariant. Re-evaluated with these controls, the Thrust 3 intervention reduces paraphrase sensitivity without improving grounding, and partly at its expense (text-only agreement rises from 53.5% to about 77% on the clean re-audit), which we report as a property of the intervention rather than set aside.
- **Thrust 5 (Deployment gating, Chapter 7).** We characterize a single-forward-pass uncertainty signal that predicts both errors and flips across the evaluated model families, and we show what a

selective-prediction gate does and does not catch. This answers the sixth challenge: a cheap signal exists, but the admitted set must still be audited for image use.

- **The controlled probe (Chapter 5).** We build a scaled-down, domain-adapted replica of MedGemma's architecture, a compact Gemma-3 decoder over a frozen MedSigLIP-448 encoder, so that every paraphrase of a question is answered from identical image features and the training-phrasing distribution becomes a manipulable input rather than a frozen prior. This speaks to the third and fourth challenges: it identifies the training-phrasing distribution as a cause of paraphrase sensitivity, and it shows that binary accuracy cannot tell a blind model from a seeing one. Its design requirements form a validated training framework for a stable, image-using model in this setting, and they read as training and evaluation practices for dependable medical VLMs: balance the training distribution per finding so that the question text alone predicts nothing, give the decoder an explicit grounding signal, train on all phrasings rather than one fixed wording, and read the answer margin rather than the binary answer alone.

The probe's results are the evidence for this account. Trained on 86,288 binary presence questions from NIH ChestX-ray14 and PadChest with MIMIC-CXR and VinDr-CXR held out entirely, every paraphrase of a question is answered from identical image features. The training-phrasing distribution alone sets the flip rate: 67.1% when training on one phrasing per question, 4.8% when training on all paraphrases, and 88.4% when the training phrasings are allowed to predict the answer; scored only on wording withheld from training, the flip rate is 26.6%. A rank-1 edit at the probe's early language layers restores the flipped answer, which is why a targeted low-rank adaptation at the early-to-middle language layers is the efficient parametric fix, while a full-model LoRA over all 34 layers, rather than five, reaches the same query-level consistency. Only the qualitative account transfers: the probe's absolute layer index is not comparable to the deployed model's layer 16, and the deployed adapter is the one on layers 15 to 19 described above. The same probe turns the readout into a monitor: the absolute margin catches a paraphrase flip in one forward pass, at an area under the receiver operating characteristic curve of 0.923 in-distribution and 0.974 on each held-out hospital, whereas an image-unreliant answer has no single-pass signal and per-case image reliance is visible only at the population level and not per prediction.

The methods and the benchmark have been reviewed outside the university. Six papers drawn from this work have been accepted at peer-reviewed venues, one more is under review, and PSF-Med is released for public use. Overall, the thesis moves the target for evaluating a medical vision-language model from a single question, "Is the model consistent?", to a joint one: is the model consistent, correct, dependent on the image, and appropriately uncertain?

1.6 Research Questions

The five thrusts are organized around the following research questions, which Chapter 8 revisits, twelve of them in a summary table and the rest in the chapters that produced them.

1.6.1 Thrust 1: Measurement

- **RQ1.1**: What are the flip rates of current medical VLMs on a large-scale, multi-dataset paraphrase benchmark?
- **RQ1.2**: Do flip rates vary systematically by paraphrase transformation type (lexical, syntactic, negation)?
- **RQ1.3**: Does model scale within a family correlate with paraphrase consistency?
- **RQ1.4**: Does distribution shift (different clinical population, different imaging equipment) amplify paraphrase sensitivity?

1.6.2 Thrust 2: Diagnosis

- **RQ2.1**: To what extent do models rely on text priors versus image content when answering clinical questions?
- **RQ2.2**: Is there a trade-off between paraphrase consistency and visual grounding?
- **RQ2.3**: Do models that flip their answers attend less to the relevant anatomical region?

1.6.3 Thrust 3: Mechanism and Mitigation

- **RQ3.1**: Can Sparse Autoencoder (SAE) features identify interpretable mechanisms underlying paraphrase sensitivity?
- **RQ3.2**: Can targeted Low-Rank Adaptation (LoRA) fine-tuning reduce flip rates while preserving accuracy and visual grounding?
- **RQ3.3**: Does the optimal intervention location match the mechanistic diagnosis?

1.6.4 Thrust 4: Safety Evaluation

- **RQ4.1**: Does reducing flip rate improve model safety, or can it create a false sense of safety?
- **RQ4.2**: Do attention heatmaps reliably indicate visual grounding?
- **RQ4.3**: Are demographic disparities (sex, age) larger for the models with the worst overall calibration?

1.6.5 Thrust 5: Calibrated Uncertainty and Deployment Gating

- **RQ5.1**: Can predictive entropy from a single forward pass predict paraphrase flips, not only errors?
- **RQ5.2**: Do calibration and selective-prediction guarantees generalize under image-quality corruption and cross-site shift?
- **RQ5.3**: Do multi-signal deployment gates admit genuinely image-grounded cases, or do they raise admitted-case accuracy by selecting text-answerable cases?
- **RQ5.4**: Does any internal monitoring signal (e.g., readout cosine, transport probes) transfer across architecture families?

1.7 Roadmap

The remainder of this thesis follows the five thrusts. Chapter 2 reviews the trustworthiness-evaluation literature and proposes a minimum reporting checklist. Chapter 3 presents PSF-Med, the equivalence audit, the clinician adjudication, and the flip-rate evaluation. Chapter 4 presents the image-removal and image-swap diagnosis. Chapter 5 presents the mechanism and the targeted mitigation with its patient- and image-disjoint evaluation. Chapter 6 presents the four-quadrant safety screen, the grounding controls, and the fairness analysis; the clean-adapter re-audit of the intervention is reported in Chapter 5 (Section 5.7.5). Chapter 7 presents the uncertainty signal and the deployment gate. Chapter 8 concludes with the central finding and its consequences for how medical vision-language models should be evaluated before deployment.

Chapter 2

A Scoping Review of Trustworthiness Evaluation in Medical Vision-Language Models

This chapter provides the background that motivates the rest of the dissertation. Before measuring, diagnosing, and mitigating specific failure modes in the thrusts that follow, it is useful to ask a prior question: across the published literature, which trustworthiness properties of medical Vision-Language Models (VLMs) are actually evaluated, with what methods, and where are the gaps? To answer this, the chapter presents a scoping review conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR). The review maps eight trustworthiness dimensions across the medical VLM literature, catalogs the datasets, metrics, and controls used to evaluate them, and distills the findings into an eight-item Minimum Trustworthiness Evaluation Checklist (MiTEC). The review shows that evaluation practice lags well behind modeling capability, with the largest gaps in visual grounding controls, calibration, fairness analysis, and deployment safeguards. Those gaps are exactly the targets of the five thrusts, and the closing section maps each gap to the chapter that addresses it.

A medical VLM is a system that jointly processes a clinical image and natural-language text to answer a diagnostic question, generate a radiology report, or classify a finding. The term is used interchangeably in the literature with multimodal large language model (MLLM); this review treats both under the same population definition.

2.1 Medical Vision-Language Models and the Trustworthiness Gap

2.1.1 Capability Outpacing Evaluation

Medical VLMs have advanced quickly since 2022. General-purpose systems such as GPT-4V [3] and Gemini [4] have been applied to medical imaging with minimal adaptation, while domain-specific models including LLaVA-Med [5], Med-Flamingo [6], PubMedCLIP [7], BiomedCLIP [8], PLIP [9], BiomedGPT [10], and MedGemma [11] have been built for clinical imaging workflows [12, 13, 14]. The references in this paragraph identify the models themselves and illustrate the field's development; they are background rather than members of the evidence synthesis, and one of them [4] also appears in the reasoned-exclusion list, where it was excluded for reporting an accuracy comparison with no trustworthiness evaluation. Performance benchmarks have grown in parallel: medical visual question answering (VQA) datasets such as VQA-RAD, SLAKE, and PathVQA, together with report-generation benchmarks built on MIMIC-CXR and CheXpert, now provide standardized accuracy comparisons across model families [15].

Accuracy on these benchmarks captures only one facet of fitness for clinical use. The problem of text-prior shortcuts, where a model answers from question statistics rather than image content, was first shown in general-domain VQA [16, 17] and has proven persistent. A medical VLM may score highly on a VQA benchmark by exploiting these shortcuts rather than interpreting the image [18], may produce confident but hallucinated findings that contradict the visual evidence [19], or may fail unpredictably when a clinician rephrases a routine question [20].

2.1.2 The Trustworthiness Gap

Trustworthiness in medical artificial intelligence spans several dimensions beyond task accuracy: robustness to input variation, calibration of expressed confidence, grounding in visual evidence, fairness across patient demographics, and transparency of the decision process [21]. Each carries direct clinical implications. On the clinician side, Ketabi et al. [22], in a preprint posted after this review's search cutoff and therefore outside the included sample, surveyed 46 radiologists on their requirements for explainable machine learning in image analysis and found that 42 of 46 (91%) considered explainability a prerequisite for trust, with the most cited reason being confidence in using the model.

A radiology VLM that flips its answer from "pneumothorax present" to "pneumothorax absent" when a clinician rephrases "Is there a pneumothorax?" as "Do you see evidence of air in the pleural space?" introduces a diagnostic inconsistency that could compromise care. This paraphrase-sensitivity failure has been documented in medical VQA systems [20, 23] but remains largely uncharacterized across clinical VLM deployments. A subtler problem also arises: a model can be highly consistent across paraphrased queries while its answer stays invariant when the image is withheld, which indicates that the image contributes little to the model's decision and that learned associations between question text and answer distributions carry much of the weight. Such a model looks reliable by standard consistency metrics yet poses a consistent-but-dangerous failure mode in which reliability masks the absence of any demonstrated link between the answer and the visual evidence. This is not hypothetical. Asadi et al. [24] showed that frontier multimodal models generate detailed image descriptions and reasoning traces even when no image is provided, a phenomenon they call "mirage reasoning"; a 3-billion-parameter text-only model trained on the ReXVQA benchmark [25] outperformed all frontier multimodal systems on a chest X-ray VQA task without access to any image. Text-only baselines, in which the model receives the question without the image, are the simplest control for this failure [26], yet this review finds them reported in only 2 of the 33 included studies [18, 27].

2.1.3 Prior Reviews and What They Miss

Several reviews have surveyed parts of this space, but none has systematically mapped how trustworthiness is evaluated. Hartsock and Rasool [15] reviewed 16 medical VLMs and 18 datasets for report generation and VQA, focusing on architectures and natural language generation (NLG) metrics rather than trustworthiness. Buess et al. [28] conducted a PRISMA-ScR review of 145 generative artificial intelligence studies in medicine spanning all applications rather than evaluation rigor. Ryu et al. [29] offered a taxonomy of medical VLM architectures by imaging modality, and Schouten et al. [30] mapped technical challenges and clinical applications of multimodal artificial intelligence without cataloging evaluation practice. On the trustworthiness side, Schwabe et al. [21] developed the METRIC framework for data quality rather than model evaluation, Tang et al. [31] reviewed pre-VLM medical artificial intelligence ethics, and Aljohani et al. [32] surveyed trustworthiness of text-only large language models (LLMs) in healthcare rather than VLMs. Hallucination benchmarks such as MedVH [19] and MedHEval [33] address a single dimension, while CARES [34] spans five dimensions but is a benchmark rather than a synthesis of evaluation practice. No existing review synthesizes the full spectrum of trustworthiness evaluation for medical VLMs.

2.1.4 Review Objectives

This review maps how trustworthiness is evaluated for medical VLMs and MLLMs applied to clinical imaging. It addresses six research questions (Table 2.1).

Table 2.1: The six research questions guiding the scoping review.

RQ	Question
RQ1	Which medical VLM/MLLM families, clinical specialties, imaging modalities, and tasks are studied?
RQ2	Which trustworthiness dimensions are evaluated: robustness, hallucination, visual grounding, calibration and uncertainty, fairness, interpretability, distribution shift, and safety?
RQ3	Which datasets, perturbation protocols, and metrics are used to evaluate those dimensions?
RQ4	Which controls test whether models truly use visual evidence: text-only baselines, image-swap tests, region-of-interest (ROI) ablations, counterfactuals, saliency sanity checks, causal localization?
RQ5	Which mitigations and deployment safeguards are reported: parameter-efficient fine-tuning (PEFT) or Low-Rank Adaptation (LoRA), prompting, selective prediction, abstention, human oversight, calibration, guardrails?
RQ6	What reporting gaps are most common: missing external validation, no subgroup analysis, no leakage checks, no grounding controls, no reproducibility artifacts?

2.2 Review Methodology

The review followed PRISMA-ScR [35]. The protocol was not prospectively registered; eligibility criteria, search strategy, and charting form were fixed before screening began.

2.2.1 Eligibility Criteria (Population, Concept, Context)

The review adopted the Population, Concept, Context (PCC) framework for scoping reviews [35]. The *population* was medical VLMs or MLLMs that process clinical images with natural-language text, or that generate text from medical images. The *concept* was trustworthiness evaluation or mitigation, defined as assessment of at least one of eight dimensions: robustness to prompt or image perturbations, hallucination, visual grounding or image reliance, calibration and uncertainty, fairness or demographic bias, interpretability or explainability, distribution-shift generalization, and safety (adversarial robustness, jailbreak resistance, harmful-output suppression). The *context* was clinical imaging across radiology, pathology, ophthalmology, dermatology, ultrasound, endoscopy, and related specialties.

Inclusion required a peer-reviewed journal article or full conference or workshop paper, published between January 1, 2022, and March 25, 2026, in the medical or clinical imaging domain, using a multimodal model with images and natural language, reporting at least one trustworthiness-related evaluation or safeguard, with English-language full text. Exclusions covered text-only LLM studies, pure computer-vision studies, non-medical multimodal studies, accuracy-only studies with no trustworthiness evaluation, and non-research document types (editorials, viewpoints, letters, tutorials, protocols, theses, patents). Studies available only as preprints were tracked in a separate horizon scan but excluded from the main synthesis.

2.2.2 Information Sources and Search Strategy

Five electronic databases were searched: PubMed/MEDLINE, Embase (via Ovid), Scopus, Web of Science Core Collection, and IEEE Xplore. The search ran on March 15, 2026, and was updated on March 25, 2026. Backward and forward citation chasing was performed on all included studies and on six recent reviews identified during protocol development [15, 28, 29, 21, 36, 30]. The search strategy combined three concept blocks with Boolean operators: model terms (vision-language model, multimodal large language model, LLM with vision or image terms), medical imaging context (medical, clinical, radiology, pathology, ophthalmology, dermatology, ultrasound, endoscopy, chest X-ray), and

trustworthiness evaluation terms (robustness, safety, trustworthiness, hallucination, uncertainty, calibration, grounding, faithfulness, reliability, fairness, bias, explainability, interpretability, distribution shift, generalizability). The complete string submitted to each of the five databases is reproduced in Section 2.2.3.

2.2.3 Database-Specific Search Strings

The five strings below are given as recorded at execution on March 15, 2026, and at the update on March 25, 2026. Two of the five cannot be presented as verbatim reproducible: the Embase (Ovid) search history was not retained, and the IEEE Xplore string exceeded the platform's documented term capacity. The strings, and the counts that follow, are therefore best read as a limited mapping snapshot of the literature as preserved at screening rather than as a fully reproducible PRISMA-ScR search history; the known defects and the count reconciliation are stated at the end of this subsection. Each combines the three concept blocks with the Boolean operator AND, and combines terms within a block with OR. The date limit was January 1, 2022 to March 25, 2026, applied to publication date; two platforms (Ovid and Scopus) restrict by publication year rather than by exact date, so on those platforms the limit was set to 2022 through 2026 and records published after March 25, 2026 were removed at the title and abstract stage. No full-text searching was performed: every string searches title, abstract, and, where the platform provides it, author or indexed keywords. No language filter and no publication-type filter was applied at any database; English-language full text and document type were eligibility criteria applied at screening.

PubMed/MEDLINE. Field tag `[tiab]` (title, abstract, and author keywords); date tag `[dp]`, exact dates.

```
(("vision-language model*"[tiab] OR "vision language model*"[tiab]
 OR "large vision-language model*"[tiab]
 OR "visual large language model*"[tiab]
 OR "multimodal large language model*"[tiab]
 OR (("large language model*"[tiab] OR LLM[tiab])
     AND (vision[tiab] OR visual[tiab]
          OR multimodal[tiab] OR image*[tiab])))
AND
(medical[tiab] OR clinical[tiab] OR medicine[tiab]
 OR radiolog*[tiab] OR pathology[tiab]
 OR ophthalmolog*[tiab] OR dermatolog*[tiab]
 OR ultrasound[tiab] OR endoscop*[tiab]
 OR "medical imaging"[tiab] OR "chest x-ray"[tiab])
AND
(robust*[tiab] OR safety[tiab] OR trustworth*[tiab]
 OR hallucinat*[tiab] OR uncertainty[tiab]
 OR calibrat*[tiab] OR grounding[tiab]
 OR grounded[tiab] OR faithfulness[tiab]
 OR reliab*[tiab] OR fairness[tiab]
 OR bias[tiab] OR explainab*[tiab]
 OR interpretab*[tiab] OR "distribution shift"[tiab]
 OR generaliz*[tiab]))
AND ("2022/01/01"[dp] : "2026/03/25"[dp])
```

Embase via Ovid. Field labels `.ti,ab.` (title and abstract); Ovid `limit` command for the date restriction, at annual granularity.

```
1  ("vision-language model*" OR "vision language model*"
    OR "large vision-language model*"
    OR "visual large language model*"
    OR "multimodal large language model*").ti,ab.
2  (("large language model*" OR LLM) AND (vision OR visual
    OR multimodal OR image*)).ti,ab.
3  1 OR 2
4  (medical OR clinical OR medicine OR radiolog*
    OR pathology OR ophthalmolog* OR dermatolog*
    OR ultrasound OR endoscop* OR "medical imaging"
    OR "chest x-ray").ti,ab.
5  (robust* OR safety OR trustworth* OR hallucinat*
    OR uncertainty OR calibrat* OR grounding OR grounded
    OR faithfulness OR reliab* OR fairness OR bias
    OR explainab* OR interpretab*
    OR "distribution shift" OR generaliz*).ti,ab.
6  3 AND 4 AND 5
7  limit 6 to yr="2022-2026"
```

Scopus. Field code `TITLE-ABS-KEY` (title, abstract, author keywords, indexed keywords); `PUBYEAR` date limit, at annual granularity.

```
TITLE-ABS-KEY(
  ("vision-language model*" OR "vision language model*"
   OR "large vision-language model*"
   OR "visual large language model*"
   OR "multimodal large language model*"
   OR (("large language model*" OR "LLM")
       AND (vision OR visual OR multimodal OR image*)))
  AND
  (medical OR clinical OR medicine OR radiolog*
   OR pathology OR ophthalmolog* OR dermatolog*
   OR ultrasound OR endoscop* OR "medical imaging"
   OR "chest x-ray")
  AND
  (robust* OR safety OR trustworth* OR hallucinat*
   OR uncertainty OR calibrat* OR grounding
   OR grounded OR faithfulness OR reliab*
   OR fairness OR bias OR explainab*
   OR interpretab* OR "distribution shift"
   OR generaliz*))
AND PUBYEAR > 2021 AND PUBYEAR < 2027
```

Web of Science Core Collection. Field tag `TS=` (Topic: title, abstract, author keywords, Keywords Plus); publication date range set in the interface.

```
TS=(("vision-language model*" OR "vision language model*"
     OR "large vision-language model*"
     OR "visual large language model*"
     OR "multimodal large language model*"
     OR (("large language model*" OR "LLM")
         AND (vision OR visual OR multimodal OR image*)))
    AND
    (medical OR clinical OR medicine OR radiolog*
     OR pathology OR ophthalmolog* OR dermatolog*
     OR ultrasound OR endoscop* OR "medical imaging"
     OR "chest x-ray")
    AND
    (robust* OR safety OR trustworth* OR hallucinat*
     OR uncertainty OR calibrat* OR grounding
     OR grounded OR faithfulness OR reliab*
     OR fairness OR bias OR explainab*
     OR interpretab* OR "distribution shift"
     OR generaliz*))

Timespan: publication date 2022-01-01 to 2026-03-25
Editions: SCI-EXPANDED, SSCI, CPCI-S, ESCI
```

IEEE Xplore. Field specifier `"All Metadata"` (title, abstract, index terms, and other record metadata; full text is not searched); year range set in the interface.

```
("All Metadata":"vision-language model" OR
 "All Metadata":"multimodal large language model" OR
 "All Metadata":"visual large language model" OR
 ("All Metadata":"large language model" AND
  ("All Metadata":"vision" OR
   "All Metadata":"visual" OR
   "All Metadata":"multimodal" OR
   "All Metadata":"medical image")))
AND
("All Metadata":"medical" OR
 "All Metadata":"clinical" OR
 "All Metadata":"radiology" OR
 "All Metadata":"pathology" OR
 "All Metadata":"ophthalmology" OR
 "All Metadata":"dermatology" OR
 "All Metadata":"chest x-ray")
AND
("All Metadata":"robust" OR
 "All Metadata":"safety" OR
 "All Metadata":"trustworth" OR
 "All Metadata":"hallucination" OR
 "All Metadata":"uncertainty" OR
 "All Metadata":"calibration" OR
```

```
 "All Metadata":"grounding" OR
 "All Metadata":"faithfulness" OR
 "All Metadata":"reliability" OR
 "All Metadata":"fairness" OR
 "All Metadata":"bias" OR
 "All Metadata":"explainability" OR
 "All Metadata":"interpretability" OR
 "All Metadata":"distribution shift" OR
 "All Metadata":"generalization")

Filters: Year 2022-2026, Conferences and Journals
```

Platform behavior and known limitations of the executed search. Phrase and truncation syntax differ across the five platforms, and the differences change what each string retrieved (Table 2.2). Two properties of the executed search should be stated plainly. First, on PubMed an asterisk inside a quoted phrase truncates rather than matching literally, so every plural and inflected form of the six quoted model phrases was retrieved. Re-executing the string above against the PubMed E-utilities interface in the original date window on July 29, 2026 returned 848 records, with the wildcards preserved in the returned query translation and no phrase-not-found warning. One genuine omission exists: `LLM[tiab]` matched only the singular acronym, because a field-tagged term is not stemmed and `LLM*` is invalid below four characters. Adding `LLMs[tiab]` retrieves 6 further records, 0.7% of the PubMed yield. Second, the IEEE Xplore strategy under-retrieved. Quotation marks on that platform disable stemming, so the quoted stems `"robust"` and `"trustworth"` matched only those literal strings and did not retrieve *robustness* or *trustworthiness*, and no wildcards were used to recover them; the printed query also exceeds the platform's documented capacity of 40 terms with at most 15 per clause. The Embase strategy cannot be verified as reported, because the executed Ovid search history was not retained and two versions of the string were recorded during protocol development. Per-database record yields were not recorded at the time of searching; the earliest preserved figure is the combined post-deduplication import of 506 records, to which citation chasing and targeted update searches added 10, for the 516 records identified in Section 2.3. Only the PubMed yield could be recovered, by the re-execution described above. The 848-record PubMed re-execution against the preserved 506-record combined import requires reconciliation, and three causes combine here. PubMed indexes new records daily, so a re-execution four months later in the same date window retrieves records indexed after March 25; the multi-database deduplication discarded duplicates whose magnitude was not recorded per database; and the export and deduplication history itself was not fully retained, so the 506 cannot be decomposed into per-database contributions. Which share each cause contributes cannot be established from the preserved records. The review is therefore presented as a limited mapping snapshot: the screening, charting, and counts in this chapter describe the preserved 516-record set exactly, but the search that produced it should not be read as exactly reproducible. Platform documentation consulted: PubMed User Guide (https://pubmed.ncbi.nlm.nih.gov/help/); NLM Technical Bulletin 2024 May-Jun;(458):e5; Scopus Support Center; Ovid Advanced Search Techniques; and IEEE Xplore Advanced Search Tips.

2.2.4 Selection and Data Charting

Records were exported to Zotero for deduplication and imported into Rayyan for screening [37]. Two reviewers independently screened titles and abstracts, then assessed full texts, after a calibration exercise on the first 50 records. Disagreements were resolved by discussion, with a third reviewer as adjudicator when consensus could not be reached. Inter-rater agreement at the title and abstract stage was almost perfect (Cohen's $\kappa = 0.91$), computed over the 461 of 506 database records that carried a

Table 2.2: Phrase and truncation behavior on each search platform, and the constraint that shaped the corresponding strategy.

Platform	**Phrase syntax**	**Truncation inside a phrase**	**Constraint**
PubMed/ MEDLINE	Double quotes, or a search tag appended to the phrase	Supported; the asterisk truncates and is not literal	At least four characters must precede an asterisk, so `LLM*` does not expand; an unindexed quoted phrase falls back to automatic term mapping
Embase (Ovid)	Double quotes	Guidance is inconsistent and the executed form was not retained	The date limit is annual, so line 7 retrieves the whole of 2026
Scopus	Double quotes give a loose phrase; braces give an exact phrase	Supported inside double quotes; an asterisk inside braces is literal	Braces are never used; `PUBYEAR` is annual
Web of Science	Double quotes give an exact phrase and disable lemmatization	Supported	A wildcard must be preceded by at least three characters in Topic and Title
IEEE Xplore	Double quotes give an exact phrase and disable stemming	Supported, for example `"health inform*"`	Maximum 40 terms and 15 terms per clause, with words inside a phrase counted separately

recorded decision from both reviewers; the remaining 45 records were decided by a single reviewer with the second concurring. A standardized charting form, piloted on 10 studies, captured study metadata, clinical context, model characteristics, the eight trustworthiness dimensions (binary per dimension), evaluation methods and metrics, controls for image reliance, external and out-of-distribution testing, uncertainty and calibration methods, fairness analyses, mitigation strategies, and reproducibility indicators (code, data, and model release). Exclusion-reason tags in Rayyan were optional at the title and abstract stage, so 290 of the 436 title and abstract exclusions (66.5%) carry no recorded reason label, and the tagged counts reported in Section 2.3 describe the labeled subset only. Every full-text exclusion was recorded individually.

2.2.5 Synthesis

Given the heterogeneity of study designs and evaluation methods, the review used a descriptive synthesis consistent with scoping-review methodology [35]; no meta-analysis was performed. Results are organized by research question and presented as frequency tables and narrative summaries, and the MiTEC checklist was developed inductively from the identified reporting gaps.

2.3 Selection Flow

The database search identified 506 records. Citation chasing on key reviews and targeted update searches added 10 more, for 516 records identified in total. The 506 database records were screened at title and abstract, and 436 were excluded. Exclusion-reason tags were recorded for 146 of them (33.5%): non-medical scope ($n = 87$), review or survey article ($n = 31$), pure computer vision with no language component ($n = 12$), text-only large language model ($n = 8$), editorial or tutorial ($n = 7$), and wrong outcome ($n = 1$). A further 9 records were excluded after discussion or third-reviewer adjudication of a screening disagreement, with no reason tag recorded for those decisions. The remaining 281 records (64.4%) were excluded as clearly not meeting the Population, Concept, and Context criteria, again with no reason tag recorded. The unit of accounting is the record excluded at title and abstract screening, and each such record is counted exactly once and assigned to exactly one of the three components, which sum to the total: $146 + 9 + 281 = 436$. Reason tags were applied selectively rather than mandatorily at the title and abstract stage, so the tagged counts characterize the labeled subset and not the full distribution of exclusion reasons. Eighty records advanced to full-text assessment (70 from screening plus 10 from citation chasing and targeted update searches); 8 were removed as duplicates indexed in multiple databases, leaving 72 unique full-text reports. Full-text screening redirected 29 preprints to a separate horizon scan per the preprint exclusion criterion and excluded 10 records that on full reading reported no evaluated trustworthiness dimension: 9 that evaluated only task accuracy, and EviVLM, whose evidential-uncertainty component is a cross-modal fusion mechanism rather than an evaluated trustworthiness property. Thirty-three peer-reviewed studies (14 journal, 19 conference) met all inclusion criteria and entered the main synthesis; the 29 preprints were tracked separately. Figure 2.1 shows the PRISMA-ScR flow.

Thirty-three peer-reviewed studies entered the main synthesis. Unless otherwise stated, all study-level frequencies and percentages in Sections 2.4 through 2.7 use this 33-study denominator. Records with an undetermined value for a specific field remain in the denominator and are identified as undetermined.

2.4 Findings

2.4.1 Study Characteristics (RQ1)

The included studies showed a sharp upward trajectory (Table 2.3): 3 in 2022, 2 in 2023, 7 in 2024, 19 in 2025, and 2 in the first quarter of 2026. Studies are charted by the year of the peer-reviewed

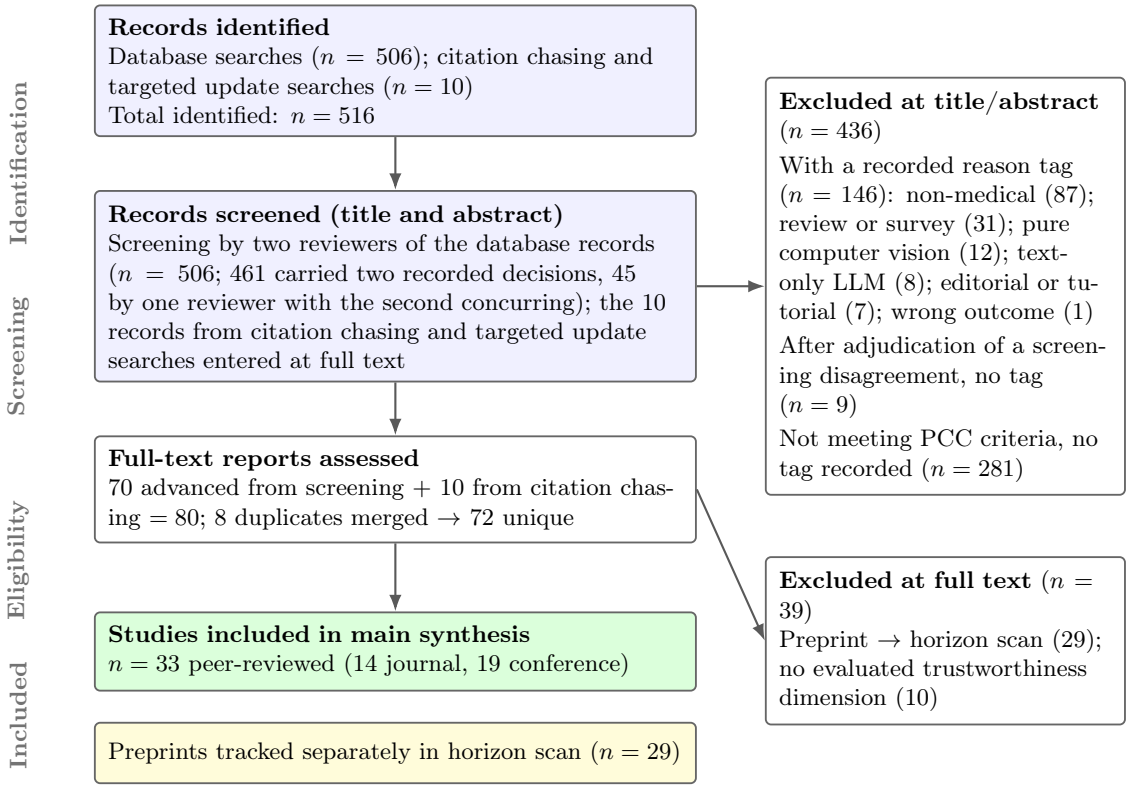


Figure 2.1: PRISMA-ScR flow diagram. Of the 516 records identified, the 506 database records were screened at title and abstract: 461 carried a recorded decision from both reviewers ($\kappa = 0.91$ over those 461) and 45 were decided by a single reviewer with the second concurring; 436 were excluded and 70 advanced. The 10 records from citation chasing and targeted update searches entered at full text, giving 80 full-text reports, of which 8 duplicates were merged, leaving 72 unique reports. The 436 title and abstract exclusions decompose as 146 records carrying a reviewer-entered reason tag, 9 excluded after adjudication of a screening disagreement, and 281 excluded as clearly not meeting the Population, Concept, and Context criteria with no tag recorded; tags were applied selectively rather than mandatorily. After full-text screening, 29 preprints were redirected to a separate horizon scan and 10 records were excluded for reporting no evaluated trustworthiness dimension, yielding 33 peer-reviewed studies in the main synthesis.

publication that made them eligible, not by the year of any earlier preprint, so SURE-VQA is charted to 2025 for its publication in Transactions on Machine Learning Research rather than to its 2024 preprint. This growth parallels the release of major model families, with a marked inflection after the public availability of GPT-4V in September 2023.

Table 2.3: Publication year of the 33 included peer-reviewed studies. The 2026 row covers January through March 25 only.

Year	Studies	Cumulative %
2022	3	9.1%
2023	2	15.2%
2024	7	36.4%
2025	19	93.9%
2026	2	100.0%

Most studies were charted as spanning several specialties and modalities rather than one. On the specialty axis, 24 of 33 studies (72.7%) were charted as multi-specialty, radiology was named in 8 (24.2%), and pathology in 1 (3.0%). On the modality axis, 25 studies (75.8%) were charted as multi-modality, chest X-ray was named in 7 (21.2%), and histology in 1 (3.0%). Radiology and chest X-ray were therefore the most frequently named single specialty and single modality, appearing in studies built on MIMIC-CXR [38], CheXpert [39], and PadChest. This concentration likely reflects the public availability of large annotated chest X-ray datasets rather than any intrinsic suitability of chest imaging for trustworthiness research. Pathology, ophthalmology, and dermatology were named far less often. Cross-domain coverage cannot be read off the charted record: only Cheng et al. [40] enumerated three or more imaging types explicitly (chest X-ray, fundus, and magnetic resonance imaging), while the 25 studies charted as multi-modality did not list the modalities they covered, so it remains unclear whether findings in chest X-ray transfer to retinal imaging or histopathology.

The charting form recorded model family at the level of *multiple*, *custom*, or a single named system, and on that record no single named family dominates. Twenty-one of 33 studies (63.6%) evaluated multiple model families, and the same 21 covered both general-purpose and medical-specific systems; 8 studies (24.2%) evaluated a custom or purpose-built architecture. Only 4 studies (12.1%) were charted against a single named family: GPT-4V [41], Med-PaLM M [42], BiomedCLIP [43], and the CheXzero and BioViL pair [44]. Eleven studies (33.3%) evaluated medical-specific models only and exactly 1 (3.0%) evaluated a general-purpose model only. Claims about which named family the literature evaluates most are therefore not supportable from this sample: Gemini was charted in no included study, GPT-4V in exactly 1, and LLaVA-Med, MedGemma, RadFM, and CheXagent in none. Those systems appear earlier in this chapter as background citations to the modeling literature, not as evaluated models in the included studies. A recurring finding was that medical domain adaptation did not consistently improve trustworthiness even when it improved task accuracy [19]. Visual question answering was the most common task format, charted in 15 of 33 studies (45.5%), followed by classification (7, 21.2%) and report generation (3, 9.1%) on MIMIC-CXR and IU-Xray, where clinical factuality was scored with CheXbert F1 or RadGraph F1 [45, 18]; no included study evaluated trustworthiness for open-ended, multi-turn clinical consultation.

2.4.2 Trustworthiness Dimensions Evaluated (RQ2)

A pronounced imbalance emerged across the eight prespecified dimensions (Table 2.4). Robustness was the most evaluated dimension, followed by hallucination, while uncertainty was addressed in 2 studies and safety in 1.

Table 2.4: Trustworthiness dimensions evaluated across the 33 included studies. Studies could evaluate more than one dimension. The uncertainty row counts studies that evaluated uncertainty or confidence reliability in any form; no included study reported a formal calibration metric, which is recorded separately in Table 2.7.

Dimension	Studies	%	Key references
Robustness	14	42.4%	[40, 43, 46, 41, 34, 27]
Hallucination	9	27.3%	[19, 47, 45, 18, 34, 48]
Distribution shift	6	18.2%	[40, 49, 50, 42, 27]
Visual grounding	5	15.2%	[51, 52, 53]
Interpretability	4	12.1%	[54, 55, 56]
Fairness / bias	4	12.1%	[44, 34]
Uncertainty / confidence	2	6.1%	[47, 34]
Safety	1	3.0%	[34]

Robustness. Robustness spanned image-side, language-side, and adversarial perturbations. On the image side, Cheng et al. [40] evaluated VLMs against five artefact categories (dust, blur, contrast changes, motion, noise) across brain magnetic resonance imaging, chest X-ray, and retinal images, finding modality-dependent degradation; Mozhegova et al. [46] and Han et al. [43] probed and defended against adversarial perturbations. On the language side, 5 of the 33 studies (15.2%) applied a perturbation to the question or instruction text and measured its effect: Mozhegova et al. [46] (synonym substitution and word deletion), Cheng et al. [40] (instruction format), Gu et al. [19] (option rewording), Huppertz et al. [41] (added context), and Han et al. [43] (adversarial text). None of the five reported an item-level agreement rate across semantics-preserving rephrasings of the same question, so none is counted as a paraphrase-consistency test in Table 2.7. Tascon-Morales et al. [57] introduced consistency-preserving VQA, the earliest included study to impose a consistency constraint across related questions. The relation it enforces is logical implication between a reasoning question and its entailed perception sub-questions rather than rephrasing of a single question, so it is not a test of semantics-preserving paraphrase consistency and is not counted as one in Table 2.7. He et al. [58] advanced reliable medical VQA through robustness evaluation; its full text could not be obtained through any route available to the review, so whether it perturbed question text is charted as undetermined rather than as absent, and the count of 5 is a lower bound. SURE-VQA [27] argued that medical VQA robustness should be assessed under real-world distribution shift, with semantic rather than token-level answer scoring and sanity baselines that test whether image input contributes at all. Xian et al. [59] argued that generic robustness tests are insufficient for biomedical foundation models, and Tu et al. [42] implicitly tested robustness through cross-task generalization of Med-PaLM M.

Hallucination. Nine studies evaluated hallucination [60]. Gu et al. [19] introduced MedVH, finding that domain-specialized VLMs were sometimes more susceptible to certain hallucination types than general-purpose models. Liao et al. [47] proposed vision-amplified semantic entropy and scored it with area under the receiver operating characteristic curve and area under the GREEN curve. Neither quantity is a calibration metric, and no included study reported one (Table 2.7); Liao et al. and CARES [34] are the only two included studies that paired hallucination evaluation with any measure of uncertainty or confidence reliability. Shah et al. [61] and Khatri et al. [62] folded hallucination into broader benchmarking frameworks, Huppertz et al. [41] quantified 258 fabricated imaging findings across 412 GPT-4V responses, and Mahdavi et al. [48] proposed Med-VCD, a sparse visual contrastive decoding method. On report generation, Delbrouck et al. [45] addressed factual correctness through semantic reward reinforcement learning, and Yu et al. [18] showed that standard NLG metrics correlate poorly with clinical accuracy. The broadest multi-dimension effort was CARES [34], which evaluated trustworthiness across five dimensions (trustfulness, fairness, safety, privacy, robustness) using roughly

41,000 question-answer pairs spanning 16 imaging modalities and 27 anatomical regions, finding that even specialized models exhibit factual inaccuracies, fairness gaps, and adversarial vulnerability.

Visual grounding. Only 5 of 33 studies (15.2%) explicitly evaluated visual grounding. Nguyen et al. [51] required localization before answering, He et al. [52] enabled grounding through PEFT, and Chen et al. [53] combined visual referring input with pixel-level grounding output. The remaining 28 studies reported accuracy or other dimensions without evaluating visual grounding as a dimension; separately, 30 of the 33 reported no control for image reliance of any kind (Table 2.5). Evidence from outside the peer-reviewed medical literature sharpens the concern: Rahmanzadehgervi et al. [63] showed that frontier VLMs fail at elementary visual tasks, and Asadi et al. [24] showed that models fabricate coherent visual reasoning, including medical diagnoses, with no image at all.

Calibration and uncertainty. Formal calibration reporting was absent: 0 of 33 studies reported a calibration metric. Two studies (6.1%) evaluated uncertainty or confidence reliability without measuring calibration. Liao et al. [47] proposed vision-amplified semantic entropy and reported area under the receiver operating characteristic curve and area under the GREEN curve, and CARES [34] reported an uncertainty-based accuracy and an over-confidence ratio; neither quantity is a calibration metric. No included study reported expected calibration error (ECE), Brier score, negative log-likelihood, a reliability diagram, or any other standard calibration metric for VLM outputs. This absence contrasts with the mature calibration toolkit for unimodal medical image classifiers, where ECE, reliability diagrams, and temperature scaling are routine [64, 65]. Three barriers help explain the gap: VLMs generate open-ended text rather than a fixed label distribution, so mapping answers to calibrated probabilities requires extra steps such as semantic clustering [66]; many evaluations use proprietary models whose logits are inaccessible; and the multitask nature of VLMs means one calibration method may not generalize across output formats. These barriers are real but not insurmountable, since semantic entropy and consistency-based approaches are model-agnostic.

Fairness, interpretability, distribution shift, and safety. Four studies (12.1%) evaluated demographic fairness, and the same four reported subgroup analyses: Yang et al. [44] documented disparities across sex, race, and age; CARES [34] included fairness among its five dimensions; and Wang et al. [67] and Xue et al. [68] proposed mitigations (a fairness-oriented mixture of experts and a logit fairness adjustment). Disparities persisted even without explicit demographic labels in training, suggesting the models encode demographic information from images themselves, consistent with unimodal findings [69]. Four studies evaluated interpretability, all using post-hoc explanation methods (cross-modal attention visualization [54], VLM-assisted concept selection [55], natural-language rationales [56, 70]); none applied mechanistic interpretability such as sparse autoencoders or circuit analysis, and none established that the explanation was faithful to the actual decision process. Six studies addressed distribution shift, framed either as performance degradation under artefacts [40] and cross-task transfer [42] or, more usefully for deployment, as out-of-distribution detection [49, 50] that asks whether the model can recognize when it is operating outside its training distribution. Only one study (3.0%) evaluated safety as a measured construct: CARES [34] tested resistance to toxic, harmful, or privacy-violating content under adversarial prompts. Mozhegova et al. [46] probed adversarial perturbations that induce diagnostic errors, which we code as robustness rather than safety, and Huppertz et al. [41] invoke patient safety in discussion but measure diagnostic accuracy, truthfulness, and fabrication rather than safety itself. No study evaluated safety guardrails, input filtering, or refusal behavior on out-of-scope queries.

2.4.3 Evaluation Methods and Metrics (RQ3)

VQA-RAD and SLAKE were the most frequently used medical VQA benchmarks; MIMIC-CXR served both VQA and report generation because it pairs images with free-text reports [45, 18]; CheXpert ap-

peared in fairness evaluations for its demographic metadata [44]; and OmniMedVQA [71] spanned 12 modalities. MedVH [19] was the only dataset designed from the ground up for trustworthiness evaluation; most studies repurposed accuracy-oriented benchmarks by adding perturbations or metrics, indicating that the field lacks purpose-built trustworthiness infrastructure. Image-side perturbations (Gaussian noise, blur, contrast, domain-specific artefacts [40], adversarial noise [43, 46]) were most common; a language-side text perturbation was applied and measured in 5 of 33 studies (15.2%) [46, 40, 19, 41, 43], with one further study charted as undetermined because its full text could not be obtained [58]; whether image and text were varied jointly within a single condition was not a charted field, so this review reports no count for combined perturbation. Accuracy, F1, and area under the receiver operating characteristic curve dominated as outcome metrics, with CheXbert F1 and RadGraph F1 preferred over lexical NLG metrics for report generation [18, 45]. No included study reported a formal calibration metric (0 of 33), and the only consistency rate reported under a language-side constraint was the logical-implication consistency of Tascon-Morales et al. [57]; no included study reported a consistency rate under semantics-preserving paraphrase.

2.4.4 Controls for Image Reliance (RQ4)

Explicit controls for visual grounding were rare (Table 2.5). Text-only baselines, the simplest and most fundamental control, appeared in only 2 of 33 studies (6.1%). Yu et al. [18] benchmarked report generators against random-retrieval baselines that ignore the image, finding the image-conditioned models ahead on the composite RadCliQ metric, though only narrowly for impression generation and not on every individual metric, and SURE-VQA [27] included a no-image sanity baseline, finding it could perform surprisingly well on medical VQA shifts. Recent preprint evidence reinforces the need: Asadi et al. [24] showed that a 3-billion-parameter text-only model trained on ReXVQA [25] with images removed topped a chest X-ray VQA task, outperforming all frontier multimodal systems, which demonstrates that benchmark scores can be driven almost entirely by text-prior statistics. Image-swap experiments were used in no included study; ROI ablation appeared only in Nguyen et al. [51]; and saliency sanity checks and counterfactual testing were charted as absent in every study. Causal localization was not a charted field and was assessed narratively; no included study reported it. In total, 90.9% of studies reported task performance or trustworthiness evaluations without establishing whether the answer depends on the image.

Table 2.5: Controls for image reliance across the 33 included studies. The first five rows correspond to charted fields. Causal localization was not a charted field; it was assessed narratively across the full texts and no included study reported it.

Control type	**Studies**	**%**	**Examples**
Text-only baseline	2	6.1%	[18, 27]
Image swap / shuffle	0	0%	none
ROI / region ablation	1	3.0%	[51]
Counterfactual testing	0	0%	none
Saliency sanity check	0	0%	none
Causal localization	0	0%	none
Any grounding control	3	9.1%	
No grounding control	30	90.9%	

2.4.5 Mitigation and Deployment Safeguards (RQ5)

Only 7 of 33 studies proposed any mitigation; the remaining 26 evaluated trustworthiness without offering a solution (Table 2.6). PEFT-based defenses were most common: Han et al. [43] used lightweight fine-tuning against adversarial noise, and He et al. [52] enabled visual grounding without full retraining.

Delbrouck et al. [45] applied semantic reward reinforcement learning to improve factual correctness, Tascon-Morales et al. [57] proposed consistency-preserving training, Wang et al. [67] and Xue et al. [68] targeted fairness, and Mahdavi et al. [48] added decoding-time hallucination mitigation. No included study evaluated conformal prediction, selective prediction, runtime guardrails, or structured human-artificial-intelligence collaboration for medical VLMs.

Table 2.6: Mitigation strategies reported in the included studies. Mitigation strategy is not a column in the released charting table; the seven studies below were identified narratively from the full texts, and the strategy counts are counts of named studies that sum to that total ($2 + 1 + 1 + 2 + 1 = 7$).

Strategy	Studies	Description
PEFT / LoRA	2	Lightweight fine-tuning against adversarial noise [43]; PEFT for visual grounding [52]
Semantic rewards	1	CheXbert/RadGraph reinforcement-learning rewards for report factuality [45]
Specialized curricula	1	Consistency-preserving training for medical VQA [57]
Fairness-specific mitigation	2	Fairness-oriented mixture of experts [67]; logit fairness adjustment [68]
Visual contrastive decoding	1	Med-VCD selects visually informative tokens at decoding [48]
Conformal / selective prediction	0	Not reported in any included study
Safety guardrails	0	Not reported in any included study
Human and artificial-intelligence collaboration	0	Not reported in any included study

2.4.6 Reporting Gaps (RQ6)

Nine reporting-quality indicators were assessed across all included studies (Table 2.7). The most critical gap was the absence of grounding controls: a model that reaches high accuracy through text-prior shortcuts will look trustworthy by every other metric (low flip rate, low hallucination rate, good calibration) while nothing in the evaluation record shows that its answers depend on the image. A related gap was the complete absence of leakage and contamination checks; no included study tested whether its benchmarks could be solved without visual input, though post-hoc cleaning methods that identify and remove text-answerable items offer a practical path forward [24]. External validation was reported in 11 of 33 studies (33.3%), which sits well below clinician-stated expectations: in the survey by Ketabi et al. [22], 78% of participants named multi-site external validation and 74% named prospective validation as prerequisites for trusting a clinical model. Semantics-preserving paraphrase consistency was tested in no included study (0 of 33), formal calibration reporting was likewise absent (0 of 33), subgroup analysis was rare (4 of 33, 12.1%) despite strong evidence that imaging models encode demographic bias [69, 44], and no study reported any deployment safeguard.

2.5 Discussion

2.5.1 Summary of Evidence

Across 33 studies published between 2022 and 2026, the central finding is a mismatch between the sophistication of model development and the rigor of trustworthiness assessment. Hallucination evaluation has drawn attention through purpose-built benchmarks, but dimensions critical for safe deployment (calibration, fairness, and visual grounding verification) remain systematically under-evaluated.

Table 2.7: Reporting-gap analysis across the 33 included studies. Every percentage uses all 33 studies as the denominator. Code release could not be determined for 10 studies (30.3%), which are counted with the 6 studies that released no code; the 51.5% figure is therefore a lower bound on release and the residual 48.5% mixes 18.2% confirmed non-release with 30.3% undetermined.

Quality indicator	**Reported**	**%**	**Implication**
Any grounding control	3	9.1%	Cannot separate image-using from text-shortcut models
Paraphrase consistency testing	0	0%	Paraphrase sensitivity undetected
External validation	11	33.3%	Unknown generalizability
Formal calibration metric	0	0%	Confidence estimates unverified
Subgroup / fairness analysis	4	12.1%	Disparities undetected
Deployment safeguards	0	0%	No safe abstention mechanism
Leakage / contamination checks	0	0%	Benchmark validity uncertain
Code released	17	51.5%	
Data released	7	21.2%	

Three findings deserve emphasis.

First, the grounding gap is the most consequential reporting failure. Only 3 of 33 studies (9.1%) reported any control for image reliance, and only 5 (15.2%) evaluated visual grounding at all. Given evidence that models can reach high benchmark scores through mirage reasoning alone [24, 72], this is the most critical methodological gap. The MIRAGE study [24], a preprint outside the peer-reviewed sample, showed that frontier models fabricate detailed visual reasoning when no image is provided, that 74% to 77% of questions on the MicroVQA, MedXpertQA-MM, and MMMU-Pro benchmarks were removed as compromised because at least one of three frontier models answered them with the image withheld, and that the fabricated reasoning is biased toward findings requiring urgent intervention. The consistent-but-dangerous failure mode, high consistency and low hallucination while the answer stays invariant when the image is withheld, would pass every evaluation that omits a grounding control.

Second, formal calibration is absent from VLM evaluation in this sample (0 of 33) despite being standard for classifiers [64, 65, 73]. VLM-expressed confidence directly influences clinician behavior: a confidently stated "pneumothorax is present" is trusted more than a hedged response, and if the model is poorly calibrated the clinician has no way to tell that the confident answer is as likely to be wrong.

Third, fairness evaluation remains sparse. Only four studies [44, 34, 67, 68] assessed or mitigated VLM performance disparities, even though demographic bias in unimodal classifiers is well documented [69, 74, 75]. The language component adds new axes along which bias can emerge (for example, differential performance on clinical jargon versus plain English), and no study tested language-side or intersectional fairness [76, 77].

2.5.2 Comparison with Prior Reviews

These findings extend prior work. Hartsock and Rasool [15] noted the inadequacy of NLG metrics; this review quantifies how few studies have moved beyond them. Buess et al. [28] reviewed generative artificial intelligence broadly; this review narrows the lens to VLMs and asks which dimensions and controls are actually covered. Ryu et al. [29] and Schouten et al. [30] organized architectures and applications, and Aljohani et al. [32] surveyed text-only LLM trustworthiness; this review adds the orthogonal axis of evaluation rigor and shows that architectural sophistication says little about how thoroughly a model has been evaluated. Schwabe et al. [21] addressed data quality; together with this review, the two cover the chain from data provenance through model evaluation.

2.5.3 A Minimum Trustworthiness Evaluation Checklist

Drawing on the identified gaps and on established practice in the broader medical artificial intelligence literature, the review proposes the Minimum Trustworthiness Evaluation Checklist (MiTEC), an eight-item minimum reporting framework rather than a complete validation protocol (Table 2.8). Items 1 through 7 map to gaps measured directly in the review; item 8 draws on the reproducibility literature. Item 7 maps to the finding that 0 of 33 studies reported any deployment safeguard. MiTEC complements broader guidance such as DECIDE-AI [78], STARD-AI [79], CONSORT-AI and SPIRIT-AI [80, 81], and TRIPOD+AI [82] by focusing on medical VLM-specific gaps: visual grounding, paraphrase consistency, calibration, subgroup analysis, and selective prediction. Items 4 and 7 are more demanding for proprietary models with inaccessible logits; for these, model-agnostic alternatives apply, including semantic entropy [66, 47], consistency-based confidence across paraphrased queries, and verbalized confidence elicitation [83].

Table 2.8: The Minimum Trustworthiness Evaluation Checklist (MiTEC) for medical vision-language models. Items 1 and 2 address grounding; 3 and 4 address reliability; 5 addresses equity; 6 addresses generalization; 7 addresses deployment readiness; 8 addresses reproducibility.

#	**Item**	**What to report**
1	Text-only baseline	Model performance with the image withheld, reported as text-only agreement: the fraction of questions for which the model gives the same answer with and without the image. High text-only agreement indicates that the image contributes little to the model's decision.
2	Image perturbation or swap test	Replace the target image with an unrelated same-modality image, or with an image carrying the opposite label. An answer that is unchanged under swap is not conditioned on the image content, which is stronger evidence than image removal alone.
3	Prompt paraphrase consistency	At least 3 clinically equivalent rephrasings of each query, with the flip rate (proportion of queries whose answer changes).
4	Calibration metrics	At least one of ECE, Brier score, or area under the generalized risk-coverage curve (AUGRC). If no confidence scores, use semantic entropy or consistency-based estimation.
5	Subgroup stratification	Results stratified by at least one demographic variable (sex, age, or race/ethnicity where available), with performance gaps alongside aggregate metrics.
6	External or out-of-distribution evaluation	At least one dataset not used in training or development, with degradation relative to the primary set.
7	Selective prediction curve	A risk-coverage curve as the model abstains on its least-confident predictions, with area under the risk-coverage curve (AURC) or AUGRC.
8	Reproducibility artifacts	Code, evaluation data, and model weights or a clear access mechanism; at minimum a data availability statement.

MiTEC is a minimum, not an exhaustive protocol. Studies targeting high-stakes settings such as emergency triage should additionally evaluate performance under time pressure and conduct reader studies with domain experts [13]. All eight items can be evaluated with existing tools: a text-only baseline runs in minutes and reveals immediately whether accuracy depends on the image, and para-

phrase testing requires only minor prompt variation. Domain-specific fine-tuning does not automatically confer trustworthiness; Gu et al. [19] found medical-domain VLMs sometimes more susceptible to hallucination than general-purpose models, so trustworthiness evaluation should be built into the development pipeline from the outset alongside targeted safeguards such as specification-tailored testing [59], conformal prediction [84], and continuous monitoring [85].

2.6 Limitations of This Review

Eight limitations bound what the counts in this chapter can support.

First, exclusion-reason tags were optional at the title and abstract stage. Of the 436 title and abstract exclusions, 146 (33.5%) carry a reviewer-entered reason label, 9 were adjudicated without a tag, and 281 (64.4%) were excluded as clearly outside the Population, Concept, and Context criteria with no tag recorded ($146+9+281 = 436$, each record counted once). The tagged distribution therefore describes the labeled subset and should not be read as the distribution of reasons across all 436.

Second, per-database record yields were not recorded at the time of searching. The earliest preserved figure is the combined post-deduplication import of 506 records. Only the PubMed yield could be recovered afterwards, by re-executing the submitted string in the original date window. A reader cannot check how much each database contributed.

Third, two of the five executed search strings have known defects. The IEEE Xplore string quoted stem forms without wildcards, which disabled the platform's stemming and lost *robustness* and *trustworthiness*, and it exceeded the platform's documented term capacity; the Embase string cannot be verified as reported, because the Ovid search history was not retained. Both are documented in Section 2.2.3. The review is therefore likely to under-represent conference literature indexed primarily in IEEE Xplore.

Fourth, the charting form recorded several fields at a granularity coarser than the claims a reader might want to make from them. Specialty, imaging modality, and model family were charted as *multiple*, *custom*, or a single named value, so a study charted as multi-modality does not disclose which modalities it covered, and the review cannot report how often any particular model family was evaluated. Claims of the form "model family X is the most studied" are outside what this charting supports.

Fifth, two dimension codings were revised after the initial charting, and the revised values are the ones reported here. The calibration dimension was split into two constructs. Evaluation of uncertainty or confidence reliability in any form stands at 2 of 33 (6.1%), and a formal calibration metric (expected calibration error, Brier score, negative log-likelihood, or a reliability diagram) at 0 of 33: the two studies that evaluate confidence reliability report area under the receiver operating characteristic curve and area under the GREEN curve [47], and an uncertainty-based accuracy and an over-confidence ratio [34], and none of those four quantities measures calibration. Membership of the dimension also changed. CARES [34] was added, and EviVLM was removed because its evidential uncertainty is a cross-modal fusion mechanism rather than an evaluated calibration property; that judgment rests on the abstract and on the complete author-released codebase, in which word-boundary searches for *ece*, *calibrat*, *brier*, *nll*, and *reliability* return no hits, because the full text sits behind a paywall. Removing that code leaves EviVLM with no evaluated trustworthiness dimension, so it fails the Concept criterion of the review and is excluded from the synthesis; the 33-study corpus and every count in this chapter reflect that exclusion. Semantics-preserving paraphrase consistency is tested by 0 of 33 rather than 1: the study originally coded for it measures logical implication between a reasoning question and its entailed sub-questions, not rephrasings of a single question. Both revisions widen the gaps the review reports, so neither flatters the argument. The charting table released with the review carries both layers, so a reader can check either: the eight original dimension flags, with the calibration flag now defined

as evaluation of uncertainty or confidence reliability, and six added columns that separate a formal calibration metric from a reported uncertainty metric, and separate a measured text perturbation from a paraphrase-consistency rate. The count of studies applying a text perturbation, 5 of 33 (15.2%), is a lower bound, because one study is charted as undetermined on that column rather than as absent [58].

Sixth, code release could not be determined for 10 of 33 studies (30.3%). Those 10 are counted with the 6 studies confirmed not to have released code, so the 51.5% release rate in Table 2.7 is a lower bound rather than a point estimate.

Seventh, one control coding was corrected after a full-text re-read. CARES [34] was initially charted as reporting a text-only baseline, but adjudication against the NeurIPS 2024 camera-ready main text, its supplemental appendix, and the released evaluation code located no no-image or image-withholding condition (every CARES condition supplies an image; its robustness dimension corrupts images or substitutes unseen-modality images), and the cell is corrected to a coding error. On the 33-study corpus the correction moves three of the review's most-quoted numbers: text-only baselines fall to 2 of 33 (6.1%), any grounding control falls to 3 of 33 (9.1%), and the share of studies with no image-reliance control rises to 90.9%. The corrected values are the ones reported throughout this chapter; the direction of the correction is to make the gap larger, not smaller.

Eighth, the review reports frequencies and does not weight studies by size, rigor, or influence, and it excludes preprints from the main synthesis. Two preprints are nonetheless cited in this chapter for evidence unavailable in the peer-reviewed sample, MIRAGE [24] and the radiologist survey of Ketabi et al. [22]; both sit outside the included set and neither contributes to any count reported here. Because the search cutoff was March 25, 2026, work published after that date is absent, and in a field moving at the observed rate (19 of 33 included studies published in 2025 alone) the counts should be read as a snapshot rather than a stable rate.

2.7 Positioning of This Dissertation

The gaps this review identifies map almost one-to-one onto the five thrusts that follow, and the dissertation's own studies are worked examples of the under-used practices the review calls for. The review found that no included study tested semantics-preserving paraphrase consistency (0 of 33), even though a rephrased clinical question is among the most plausible real-world triggers of failure. Thrust 1 (Chapter 3) addresses this directly by building PSF-Med, a paraphrase-sensitivity benchmark that measures flip rates across six medical VLMs and three chest X-ray populations, operationalizing MiTEC item 3 at scale. The review also found that only 3 of 33 studies (9.1%) reported any control for image reliance, that image-swap tests appeared in no included study, and that text-only baselines appeared in only 2 (6.1%). Thrust 2 (Chapter 4) supplies these controls, diagnosing whether apparent consistency reflects genuine visual grounding or reliance on language priors through text-only baselines and image-swap tests corresponding to MiTEC items 1 and 2.

The review noted that no included peer-reviewed study applied mechanistic interpretability, and that mitigation was reported in only 7 of 33 studies, most often through PEFT or LoRA. Thrust 3 (Chapter 5) occupies both spaces: it uses sparse-autoencoder and circuit analysis to locate the internal features responsible for paraphrase sensitivity, then applies targeted LoRA as a mitigation, moving beyond the post-hoc attention and concept explanations that dominate the interpretability studies in the sample. Calibration and deployment safeguards were the starkest gaps of all, with 0 of 33 studies reporting a formal calibration metric and 0 of 33 reporting any selective-prediction or abstention mechanism. Thrust 5 (Chapter 7) targets this gap with calibration analysis (ECE, Brier score, AURC, and AUGRC), an uncertainty bridge that carries confidence estimates across distribution shift, and deployment gating through selective prediction, implementing MiTEC items 4 and 7 that no study in the review satisfied.

Finally, the review's central warning, that consistency without visual grounding is a false sense of safety, motivates Thrust 4 (Chapter 6). Its safety evaluation formalizes the consistent-but-dangerous failure mode into a four-quadrant taxonomy that crosses answer consistency with measured image reliance, and it stratifies results by patient demographics to address the fairness gap (4 of 33 studies) that the review flags. Read together, the thrusts do not merely study medical VLM trustworthiness; they demonstrate that the MiTEC checklist is tractable, applying its grounding controls, paraphrase-consistency measurement, calibration and selective-prediction reporting, external and out-of-distribution evaluation, and subgroup analysis to real medical VLMs, and showing what each check reveals that accuracy alone conceals.

Chapter 3

Thrust 1: Measuring Paraphrase Sensitivity in Medical VLMs

Research questions. This chapter answers **RQ1.1** (the magnitude of paraphrase sensitivity across models and datasets), **RQ1.2** (whether it varies by paraphrase transformation type), **RQ1.3** (whether model scale buys consistency), and **RQ1.4** (whether distribution shift amplifies it).

Main contribution. PSF-Med is a benchmark of 26,850 chest X-ray questions and 92,856 equivalence-filtered paraphrase pairs across three clinical populations. It was constructed through a 122,778-pair Large Language Model-based rubric audit and evaluated against a stratified clinician adjudication of 1,200 pairs from the pre-regeneration audit pool. On the binary subset, six medical vision-language models flip on 6.4 to 54.7% of paraphrase pairs. The benchmark and evaluation code are publicly released.

This chapter presents PSF-Med, a benchmark for measuring Paraphrase Sensitivity Failure (PSF) in medical Vision-Language Models (VLMs). The goal is straightforward: quantify how often a model changes its answer when the same clinical question is rephrased.

The benchmark contains 26,850 questions drawn from three chest X-ray datasets covering three continents and three healthcare systems: MIMIC-CXR (United States), PadChest (Spain), and VinDr-CXR (Vietnam). After a rubric-based equivalence audit over 122,778 paraphrase pairs (GPT-5-mini judge with 91.6% Claude Haiku cross-family agreement on the recorded 500-pair cross-validation subset) and a regenerated PadChest paraphrase set that removes laterality and scope leakage, the full evaluation pool contains 8,938 MIMIC-CXR pairs, 59,573 second-iteration PadChest pairs, and 24,345 VinDr pairs; the flip rates reported in this chapter are computed on the genuine binary yes/no subset of this pool, for which the yes/no readout is well defined.

Across six medical VLMs, post-filter flip rates on the binary yes/no subset range from 6.4% (MedGemma-27B on MIMIC) to 54.7% (RadFM on VinDr) once two cells of degenerate yes-bias are set aside. The chapter also identifies the first signs of a pattern that Thrust 2 investigates in detail: models with low flip rates often show high agreement with text-only baselines, suggesting that their consistency may come from language priors rather than visual analysis. The audit confirms that this core signal, low flip rate coupled with high text-only agreement, survives after removing the strictest non-equivalent cases. Restricting to the binary yes/no subset does change the in-distribution flip ordering, however, so we do not claim a preserved model ranking: MedGemma-27B is the most consistent model on MIMIC on the binary subset, where the earlier mixed-question ordering placed it lower.

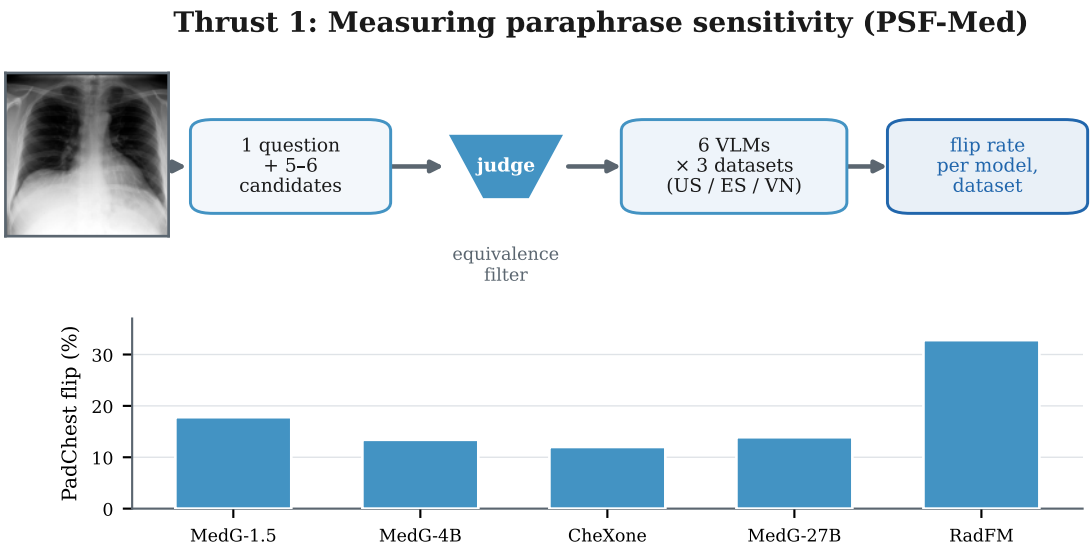


Figure 3.1: Thrust 1 at a glance. Each clinical question is expanded into five or six candidate paraphrases (mean 3.5 retained per question after equivalence filtering), a rubric-based judge keeps only the clinically equivalent ones, and six medical VLMs are scored on three chest X-ray populations. Post-filter flip rates span 6.4% to 54.7% once two degenerate yes-bias cells are set aside: LLaVA-Rad on PadChest (99.6% positive, omitted from the bar panel) and MedGemma-1.5-4B on VinDr (100% positive) reduce to a single-class prior rather than genuine consistency.

3.1 Background and Related Work

3.1.1 Medical VLMs and Their Evaluation

Chapter 2 surveys the medical VLM literature and its trustworthiness gaps in detail; here we note only what PSF-Med builds on and departs from. The models evaluated in this chapter (MedGemma [11], LLaVA-Rad [86], RadFM [87], CheXone [88], and earlier contrastive and LLaVA-Med systems [89, 90, 91, 5]) are typically assessed on fixed-question VQA benchmarks such as VQA-RAD [92] (3,515 questions, 315 images), SLAKE [93] (642 images, bilingual), and OmniMedVQA [71] (12 imaging modalities). Trustworthiness-oriented frameworks including CARES [34], VHELM [94], MedVH [19], and MedHalt [95] add hallucination and adversarial-robustness dimensions, and Med-PaLM's selective-prediction analysis used self-consistency across sampled reasoning paths as an uncertainty proxy [96]. None of these systematically tests whether a model gives the same answer when an equivalent question is rephrased.

3.1.2 Robustness and Language Priors

The tendency of vision-language models to rely on language priors rather than visual content is well documented in the general domain. VQA-CP [17] showed that VQA models exploit question-type correlations, and VQA v2 [16] attempted to reduce this bias through balanced pairs. Cycle-consistency approaches [97] test whether a model's answers are logically self-consistent. In natural language processing, CheckList [98] introduced systematic behavioral testing, distinguishing invariance tests (equivalent inputs should yield equivalent outputs) from directional and minimum-functionality tests; adversarial paraphrasing [99] showed that syntactic transformations can fool text classifiers; and ProSA [100] studied prompt sensitivity in large language models through template-level variation. PSF-Med operationalizes CheckList's invariance testing for medical VLMs and adapts ProSA's systematic-variation

insight to question-level clinical paraphrasing, where the consequences of inconsistency are clinical rather than academic. The specific gap it fills is a benchmark that generates controlled paraphrases at scale, measures flip rates across multiple models and datasets, analyzes sensitivity by transformation type, and connects consistency to visual grounding.

3.2 The PSF-Med Benchmark

3.2.1 Source Data

PSF-Med draws from three chest X-ray datasets spanning three continents and three healthcare systems:

- **MIMIC-CXR** [38]: Frontal chest radiographs from Beth Israel Deaconess Medical Center, Boston. The dataset includes structured labels derived from radiology reports via CheXpert [39]. We use the official test split.
- **PadChest** [101]: Radiographs from Hospital San Juan, Valencia, Spain. This dataset provides a distribution shift from MIMIC-CXR in terms of institution, imaging equipment, patient demographics, and language of origin (reports originally in Spanish).
- **VinDr-CXR** [102]: Radiographs from hospitals in Hanoi, Vietnam, with per-finding labels and radiologist annotations. VinDr provides a second axis of distribution shift, complementing PadChest's US-to-Europe transition with a US-to-Southeast-Asia transition that differs in equipment, patient demographics, prevalence base rates, and reporting conventions.

Using three datasets from three continents and healthcare systems lets us test whether paraphrase sensitivity appears across clinical populations or only in one. Because the three sets also differ in task composition, label balance, finding mix, and templates, differences between them are reported as cross-population differences rather than attributed to distribution shift alone. As we will see, flip rates are generally higher on PadChest and VinDr than on MIMIC, but that gap confounds institution with task composition and label balance, so it cannot be read as an isolated distribution-shift effect.

3.2.2 Question Generation

The clinical questions in PSF-Med are constructed from dataset labels using template-based generation. For MIMIC-CXR, each CheXpert label (14 pathologies $\times$ 4 certainty levels) generates a presence question: "Is there [finding]?" For PadChest, the broader label vocabulary (174 radiographic findings) generates questions for all findings with at least 10 positive instances in the evaluation set, yielding 861 unique question strings covering 23 distinct clinical findings. The 861 figure counts unique question wordings; paired against the images that satisfy each finding's positivity criterion, these strings expand to the 14,935 (image, question) instances tabulated for the second-iteration PadChest set in Table 3.1, which counts evaluation instances rather than distinct wordings.

Questions are restricted to binary (yes/no) format for two reasons. First, binary questions have unambiguous ground truth derived from dataset labels, enabling automated evaluation at scale. Second, the binary format isolates the paraphrase sensitivity problem from answer extraction challenges: there is no ambiguity about whether "Yes" and "Present" constitute the same answer. Open-ended questions ("Describe the findings") and multi-label questions ("List all abnormalities") present different consistency challenges that we leave to future work.

The question set is balanced by answer label where possible. For MIMIC-CXR, we select images ensuring approximately equal positive and negative instances for each finding. For PadChest, the natural prevalence distribution is preserved, as balancing would require discarding the majority of the dataset. This means PadChest flip rates reflect the actual clinical distribution, including the base-rate effects that Chapter 6 identifies as a driver of Dangerous predictions.

3.2.3 Paraphrase Generation

For the benchmark as evaluated, paraphrases were generated with GPT-5-family models: MIMIC-CXR and VinDr-CXR with GPT-5-mini at five paraphrases per question, and PadChest with GPT-5.4-mini at six per question, the PadChest set regenerated with judge-informed constraints after the equivalence audit of the original set. The original construction used GPT-4 with three to five paraphrases per question and is documented for provenance in Appendix A.1; the protocol described here is the one the evaluated data artifacts record. The generation prompt instructs the model to produce meaning-preserving reformulations that vary along specific linguistic dimensions while maintaining the clinical intent of the original question. We provide few-shot examples for each transformation type to guide generation quality and diversity.

The generation strategy applies five types of meaning-preserving transformations:

1. **Lexical substitution**: replacing content words with synonyms ("pleural effusion" → "fluid in the pleural space").
2. **Syntactic restructuring**: changing clause order or sentence structure while preserving meaning.
3. **Specificity modulation**: shifting how specifically the finding is described ("Is there cardiomegaly?" → "Does this patient have an enlarged heart?").
4. **Scope variation**: adjusting quantifiers ("is there evidence of" → "are there findings of"). This category is the least reliably meaning-preserving of the five. The generator was also prompted with threshold-shifting examples such as "any evidence of" → "significant evidence of", which are *not* equivalent: "significant" raises the clinical threshold, so a "no" to it is compatible with a "yes" to "any". Pairs of that kind are precisely what the equivalence audit is designed to reject, and the scope-quantification category is accordingly one of the more heavily filtered.
5. **Negation-adjacent**: rephrasing that introduces negation or exclusion vocabulary. As generated, many of these inverted the answer ("Is there pneumothorax?" → "Can pneumothorax be excluded?", whose correct answer is opposite) and were rejected by the equivalence audit; only same-polarity negation rephrasings that preserve the answer are retained.

Some of these paraphrases are genuinely meaning-preserving (a "No" to "Is there X?" and a "Yes" to "Can X be excluded?" express the same clinical judgment). Others may subtly shift the expected answer polarity. We analyze this category separately throughout the chapter and flag it as a validity threat.

3.2.4 Semantic Filtering

Not every generated paraphrase is truly meaning-preserving. We apply a two-stage filtering process to remove invalid paraphrases:

Stage 1: Embedding similarity. We encode each original question and its paraphrases using BioClinicalBERT [103] and compute cosine similarity; for MIMIC-CXR, candidates below 0.95 are excluded, the value recorded in the released data artifact. The original construction used a 0.90 threshold, selected through a 200-pair annotation study documented in Appendix A.1.

Stage 2: Semantic inversion detection. We apply rule-based checks for negation operators and polarity inversions. Paraphrases that invert the expected answer (e.g., "Is there X?" paired with "Is X absent?") are flagged and excluded from the primary evaluation, though we retain them for analysis of negation sensitivity.

After this construction-stage filtering, the original benchmark (MIMIC-CXR and the first iteration of PadChest) contains approximately 92,000 candidate paraphrases with a mean of 4.7 per question. The filtering removes approximately 8% of generated paraphrases, with the highest rejection rate in the negation-adjacent category (14%) and the lowest in lexical substitution (3%). This construction-stage candidate count should not be confused with the final judge-filtered evaluation set of 92,856 (question, paraphrase) pairs (mean 3.5 per question; Table 3.1): the two figures are numerically close

by coincidence but describe different pipeline stages, the first counting candidate paraphrases produced at construction and the second counting pairs retained for evaluation after the audit and the PadChest regeneration. A later stricter audit, described next, re-adjudicates 122,778 paraphrase pairs across all three datasets[1] and separates a smaller equivalence-validated core from clinically natural but non-equivalent reframings.

3.2.5 Quality Assurance

In total, 122,778 paraphrase pairs across MIMIC-CXR, PadChest, and VinDr were adjudicated using a structured bidirectional-entailment rubric: a pair was retained only if the original and paraphrased questions preserved the same clinically relevant yes/no truth conditions with no material change in scope, laterality, certainty, or operator.

Under this stricter audit, 61,761 pairs (50.3%) were retained in the core-equivalent set, 59,788 (48.7%) were rejected as adversarial or clinically non-equivalent reframings, and 1,229 (1.0%) were marked uncertain for human review.[2] The retained fraction varies substantially by dataset: 72.9% for the original PSF-Med / MIMIC benchmark, 35.7% for PadChest, and 78.6% for VinDr. This pattern is informative rather than alarming. The original MIMIC benchmark was already tightly filtered, whereas the later expanded sets include more scope changes, laterality drift, and near-paraphrases that are clinically natural but not strictly equivalent. The 92,856-pair evaluation set is not a subset of the 61,761-pair audited core: the audit's retained pairs from the first PadChest iteration were superseded by a regenerated, separately judge-filtered second iteration (59,573 pairs), which together with the audit-retained MIMIC (8,938) and VinDr (24,345) pairs yields the 92,856 pairs evaluated throughout this dissertation.

To test whether the LLM judge was introducing family-specific bias, we re-scored a stratified 500-pair subset with Claude Haiku. Cross-family agreement reached 91.6% on the MIMIC + PadChest subset, with disagreements concentrated in scope and quantification cases. Rejections were especially common in the highest-risk categories: negation-pattern paraphrases were rejected 93.8% of the time, while lexical substitutions were retained most often. The benchmark's central conclusion, the contrast between low flip rate and weak image reliance, remained visible in the retained core after this stricter filtering, even though the finer flip-rate ordering shifts once the population is restricted to genuine binary yes/no questions.

3.2.6 Clinician Adjudication of Equivalence

The equivalence audit above is automated, so we tested the audit procedure through blinded clinician adjudication of a stratified sample of 1,200 pairs. The sample deliberately over-represents difficult, adversarial, and uncertain cases rather than estimating the equivalence rate of the final benchmark. It was also drawn before the second iteration of the PadChest paraphrase set was produced. The study therefore provides a targeted human check of the filtering procedure, rather than representative validation of the full 92,856-pair evaluation set.

Sampling. Pairs were drawn with a fixed seed (20260501) stratified over three buckets, 600 `core_equivalent`, 400 `adversarial_reframing`, and 200 `uncertain_review`, so that the sample de-

[1] The 122,778-pair total is the rubric-based equivalence audit run added during the review of an earlier version of this benchmark, which re-scored the construction-stage paraphrase pairs across MIMIC-CXR, PadChest, and VinDr-CXR. It is the union of two GPT-5-mini batch runs: 91,783 pairs covering MIMIC-CXR (12,259), PadChest (79,378), and a 146-pair VinDr pilot, plus the full 30,995-pair VinDr run.

[2] These tallies are recomputed directly from the stored judge verdicts by `scripts/analysis/recompute_audit_retention.py`, which reproduces the 122,778-pair population exactly. They supersede a slightly earlier snapshot of the same audit (60,712 retained / 60,766 rejected / 1,300 uncertain) that was reported while the VinDr run was still completing; the population, the per-dataset ordering, and every downstream conclusion are unchanged.

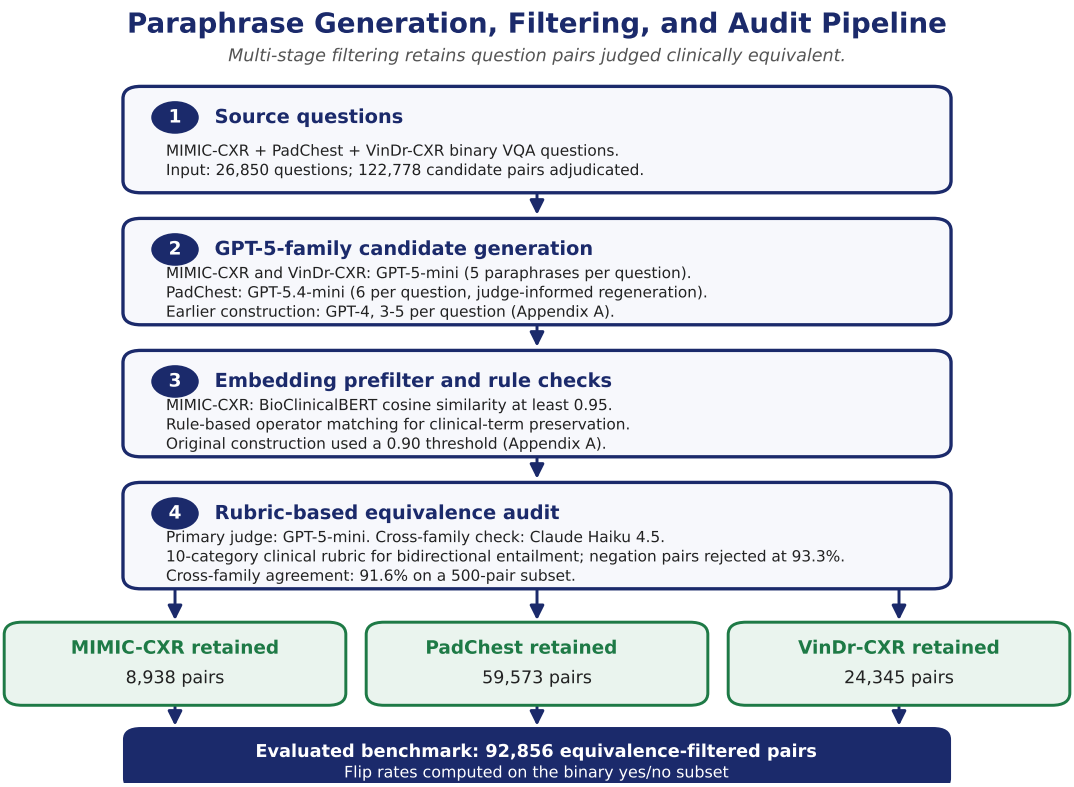


Figure 3.2: Paraphrase-validation pipeline used in the stricter post-submission audit. The original benchmark-construction filter is followed by a more conservative rubric-based equivalence audit that separates strict core-equivalent pairs from clinically natural but non-equivalent reframings.

liberately over-represents the cases where the automated judge is least reliable rather than being uniform over retained pairs. A floor of 20 pairs per dataset was enforced and operator-changing candidates were over-sampled. The delivered sample spans all four source pools (PadChest 658, PSF-Med/MIMIC 174, VinDr 325, VinDr-full 43) and all five transformation types (lexical substitution 262, syntactic restructuring 262, specificity modulation 260, scope quantification 224, negation pattern 192).

Blinding and design. Reviewers worked from a blinded sheet containing only the image, the original question, and the candidate question; the model's answer, the flip label, and the automated judge's verdict were withheld, so a reviewer could not anchor on either the model or the judge. Each reviewer recorded a verdict (equivalent, not equivalent, uncertain), a change-type tag, a confidence rating, and free-text notes. Of the 1,200 pairs, 400 were independently reviewed by all three reviewers to measure agreement, and the remaining 800 were single-reviewed for coverage.

Consensus rule. "Consensus" means two different things depending on how many reviewers saw the pair, and we state both because the label set mixes them. For the 400 triple-reviewed pairs it is the majority verdict of the three reviewers; all 400 have a majority, since no pair drew three different verdicts. For the remaining 800 single-reviewed pairs it is simply that reviewer's verdict, which carries no agreement information. The resulting labels are 689 equivalent, 409 not equivalent, and 102 uncertain over the 1,200 pairs. Any statement below that rests on the consensus label therefore rests on a single judgment for two thirds of the sample.

Agreement. Inter-reviewer agreement is estimated from the 400 triple-reviewed pairs; the other 800 pairs increase coverage but carry only one clinician judgment each. On the 400 triple-reviewed pairs, pairwise raw agreement was 98.5% (394/400) with Cohen's $\kappa = 0.97$ for each of the three reviewer pairs, and 391 of the 400 were unanimous. Agreement between the clinician consensus and the automated GPT-5-mini judge is substantially weaker: 868 of 1,200 pairs, or 72.3% raw agreement, $\kappa = 0.52$; the denominator is the full sample, mixing the majority labels of the 400 with the single-reviewer labels of the 800. Restricted to the 400 triple-reviewed pairs, where the consensus is best supported, judge agreement is 78.2% (313/400). This gap is why the adjudication was run: the LLM judge and the clinicians agree on the easy cases and diverge on the adversarial and uncertain buckets that the sample was designed to over-represent.

Effect on the reported results. This sample cannot be used to restate the benchmark's flip rates. The 1,200 pairs were drawn from the audit pool, which for PadChest is the first-iteration paraphrase set; that set was subsequently superseded by the regenerated second iteration actually evaluated in this chapter. The committed clinician-review records support the adjudication totals and agreement statistics above, but they do not include a reproducible manifest linking the reviewed sample to the final evaluated set. No overlap count or clinician-restricted flip-rate estimate is therefore reported. Because the sample is deliberately enriched for hard cases, it remains a targeted check rather than an unbiased estimate of retention on the full pool. And because two thirds of the consensus labels rest on a single reviewer, the check is weaker than the 400-pair agreement statistic alone would suggest.

3.2.7 Dataset Statistics

Table 3.1 summarizes the benchmark.

3.3 Evaluation Protocol

3.3.1 Flip Definition

A flip occurs when the model's answer to any paraphrase differs from its answer to the original question. Formally, for a model M and question q with paraphrase set $P(q)$:

$$\text{Flip}(M, q) = \mathbf{1}\left[\exists\, p \in P(q) : M(q) \neq M(p)\right] \tag{3.1}$$

Table 3.1: PSF-Med equivalence-filtered evaluation set. The benchmark spans three clinical populations on three continents, with binary clinical questions and paraphrases retained by a structured bidirectional-entailment rubric. Clinician adjudication provides a targeted check of the filtering procedure on a pre-regeneration audit sample. Pair counts are the post-filter evaluation set used throughout the dissertation, and each question-paraphrase pair is evaluated on every model.

	MIMIC-CXR	PadChest	VinDr-CXR	Total
Origin	US	Spain	Vietnam	N/A
Questions	3,266	14,935	8,649	26,850
Paraphrase pairs	8,938	59,573	24,345	92,856
Mean pairs/question	2.7	4.0	2.8	3.5

Table 3.2: Models evaluated in PSF-Med. All models are evaluated in their publicly released configurations with greedy decoding (temperature zero).

Model	Architecture	Parameters
MedGemma-27B	Gemma 3	27B
MedGemma-1.5-4B	Gemma 3	4B
MedGemma-4B	Gemma 3	4B
CheXone	Qwen2.5-VL	3B
LLaVA-Rad	LLaVA	7B
RadFM	Custom	14B

The flip rate is the mean of this indicator across all questions. We also report the pairwise disagreement rate (the fraction of individual question-paraphrase pairs that disagree) as a complementary metric, since it is less sensitive to the number of paraphrases per question.

3.3.2 Answer Extraction

All models in PSF-Med are evaluated on binary (yes/no) questions, where answer extraction is unambiguous. We parse model outputs by matching against affirmative keywords ("yes," "present," "visible," "confirmed") and negative keywords ("no," "absent," "not seen," "negative"). Outputs that match neither set are excluded from analysis. The exclusion rate ranges from 3.2% to 8.7% across the five audited models, driven primarily by hedge responses ("It's possible but..."), refusals, and off-topic outputs.

To validate the answer extraction protocol, we manually audited 500 randomly sampled outputs from the five models in the evaluation roster at the time of the audit (CheXone and MedGemma-1.5-4B joined the roster later). The parser agreed with human judgment in 97.4% of cases. Disagreements were concentrated in hedge responses where the model expressed uncertainty without committing to yes or no.

3.3.3 Models Evaluated

We evaluate six medical VLMs spanning different architectures, parameter counts, and training approaches:

All models are evaluated in their publicly released configurations with greedy decoding (temperature 0) to ensure reproducibility. We do not fine-tune any model for this evaluation; the goal is to measure sensitivity as users would encounter it. Greedy decoding eliminates sampling variance as a confound: any answer difference between the original and paraphrased question is deterministic and attributable to the input change, not to stochastic sampling. Temperature-based sampling at $T > 0$ would introduce additional variance that could either mask or amplify true paraphrase sensitivity.

Each model receives the image and question through its standard inference interface. For

MedGemma models, this uses the Hugging Face Transformers library with the model's default chat template. For LLaVA-Rad, we use the LLaVA inference pipeline with BiomedCLIP image preprocessing. For RadFM and CheXone, we use their respective inference code. All images are preprocessed to the model's expected resolution (typically 224×224 or 336×336 pixels) using the model's standard preprocessing pipeline.

The evaluation is exhaustive: every question is paired with every paraphrase, and every combination is evaluated on every model. No subsampling or stratification is applied. The full filtered pool was run across 8 NVIDIA A100 GPUs over approximately 72 hours. The flip rates reported below are computed on the genuine binary yes/no subset of that pool: questions with an unambiguous yes/no reference answer, because the yes/no readout is only defined for those. On PadChest and VinDr this filter keeps only presence questions ("Is there [finding]?"); on MIMIC it keeps all yes/no-answer questions, a mix of 956 presence, 177 view-identification ("Is this a PA view?"), and 406 abnormality questions. Location ("where is the [finding]?") and laterality ("which side shows [finding]?") questions, and the nine VinDr non-binary "other" references, are excluded because the yes/no readout is undefined for them. After restriction, the reported rates draw on 1,539 MIMIC-CXR questions (5,076 paraphrase pairs), 8,445 PadChest questions (36,244 pairs), and 2,807 VinDr-CXR questions (8,612 pairs).

3.4 Results

3.4.1 Flip Rates Across Models

Table 3.3 presents the main results on the equivalence-filtered PSF-Med run across all three datasets, restricted to the binary yes/no subset defined above. This subset is distinct from the curated 158-pair FlipBank and the 861-question PadChest flip bank used in later chapters. All rates in this section are *pairwise*: the denominator is the paraphrase pair, so the rate answers how often a single rephrasing changes the yes/no answer. The complementary *query-level* rate (the fraction of questions that flip under at least one of their roughly 3.3 paraphrases) is larger and is reported for the focal model in Table 3.4; the two must not be conflated. Excluding the two cells flagged for degenerate yes-bias (LLaVA-Rad on PadChest at 99.6% positive predictions and MedGemma-1.5-4B on VinDr at 100%), pairwise flip rates range from 6.4% (MedGemma-27B on MIMIC) to 54.7% (RadFM on VinDr), an 8.5× spread. These numbers reflect the rubric-based filtering (GPT-5-mini judge with 91.6% Claude Haiku cross-family agreement on the recorded subset) and the regenerated PadChest paraphrase set; obsolete pre-audit values such as the 42.4% on PadChest for MedGemma-4B (reported in the unfiltered earlier version of this benchmark, not a current result) were inflated by adversarial reframing pairs (broader/narrower scope, laterality drift) that did not preserve clinical meaning.

Three patterns stand out. First, scaling shows no monotonic pattern among the three tested MedGemma variants, which do not rank stably across datasets. MedGemma-27B is the most consistent of the three on MIMIC (6.4%) and on VinDr (8.1%), but MedGemma-4B edges it on PadChest (13.4% vs. 13.9%), and MedGemma-1.5-4B is the least consistent of the three on PadChest (17.5%) while collapsing into degenerate yes-bias on VinDr. No size dominates among the three tested variants on this benchmark. Second, the two yes-biased cells deserve careful interpretation rather than dismissal. LLaVA-Rad on PadChest answers "yes" on 99.6% of original questions; the resulting 0.7% flip rate looks like consistency, but the text-only controls (Chapters 4 and 6) show that almost all of these answers also match the text-only output. The defensible reading of this cell is the simplest one: a model answering "yes" at that rate has collapsed onto a single-class prior, and a near-constant answer cannot discriminate between one scan and the next, so its 0.7% flip rate measures the stability of that prior rather than stable visual grounding. The contrast with MedGemma-4B, which on the same images predicts positive on only 24.4% of questions and so does vary with the input, makes the point concrete. MedGemma-1.5-4B on VinDr exhibits the same pattern (100% yes, 0.8% flips), and both cells

Table 3.3: Pairwise paraphrase flip rates (%) on the equivalence-filtered PSF-Med run, restricted to questions with an unambiguous yes/no reference answer across three clinical populations: MIMIC-CXR (Beth Israel Deaconess, US; 1,539 questions, 5,076 paraphrase pairs), PadChest (Hospital San Juan, Spain; 8,445 questions, 36,244 pairs), and VinDr-CXR (hospitals in Hanoi, Vietnam; 2,807 questions, 8,612 pairs). This is a binary yes/no set, not a pure presence set: on PadChest and VinDr the filter keeps only presence questions ("Is there [finding]?"), but on MIMIC it keeps all yes/no-answer questions, a mix of 956 presence, 177 view-identification ("Is this a PA view?"), and 406 abnormality questions. Location and laterality questions, and the nine VinDr non-binary "other" references, are excluded because the yes/no readout is undefined for them. VinDr contains only positive presence cases, so it is a positive-case presence-question test that measures phrasing stability rather than balanced accuracy; reference prevalence and model positive-prediction rate are reported in Table 3.4. A pairwise flip is a paraphrase whose yes/no answer differs from the original's; the denominator is the paraphrase pair, not the question (see Table 3.4 for the query-level rates). Equivalence judged by GPT-5-mini with 91.6% Claude Haiku cross-family agreement on the recorded subset. Lower is more consistent. Daggers (†) flag cells where the model exhibits degenerate yes-bias (>98% positive predictions on originals): the resulting flip rate reflects a single-class prior, not paraphrase robustness, and is reported only to keep the column complete.

Model	MIMIC-CXR	PadChest	VinDr-CXR
MedGemma-1.5-4B	7.4	17.5	0.8†
MedGemma-4B	8.3	13.4	15.3
LLaVA-Rad	15.6	0.7†	11.6
MedGemma-27B	6.4	13.9	8.1
CheXone	8.2	13.0	11.1
RadFM	13.7	32.8	54.7

are flagged with a dagger in Table 3.3 for this reason. The dissertation treats both cells as the central exhibit for the consistency-without-grounding problem rather than as legitimate low-flip results. Third, RadFM has the highest flip rate on both out-of-distribution sets (32.8% PadChest, 54.7% VinDr) and the second-highest on MIMIC (13.7%, behind LLaVA-Rad's 15.6%), and it is the only model whose flip rate strictly increases with each step away from MIMIC (13.7% → 32.8% → 54.7%), suggesting a baseline instability in question processing that filtering does not relieve and that compounds with distribution shift.

Flip rates are generally higher on the two out-of-distribution populations, but these are *cross-population differences* rather than an isolated distribution-shift effect: the three datasets differ in task composition (MIMIC mixes presence, view, and abnormality questions while PadChest and VinDr are presence-only), label balance (VinDr contains only positive cases), finding mix, and question templates, as well as in institution. The trajectory is also non-monotonic for several models. Reading along the MIMIC → PadChest → VinDr direction, MedGemma-4B goes 8.3% → 13.4% → 15.3%; MedGemma-27B 6.4% → 13.9% → 8.1%; CheXone 8.2% → 13.0% → 11.1%; RadFM 13.7% → 32.8% → 54.7%. Which out-of-distribution set is harder is model-dependent: VinDr is harder than PadChest for MedGemma-4B and RadFM, while PadChest is harder for MedGemma-27B and (marginally) CheXone. Two models exhibit degenerate yes-bias on one dataset each: LLaVA-Rad on PadChest (99.6% yes, 0.7% flips) and MedGemma-1.5-4B on VinDr (100% yes, 0.8% flips). Both cases reduce to a single-class prior rather than meaningful low-flip behavior, and Chapters 4 and 6 treat them as instances of the consistency-without-grounding pattern.

Cross-model flip correlation is low (mean pairwise $r = 0.06$ on MIMIC, 0.04 on PadChest, 0.07 on VinDr), indicating that different models largely fail on different questions. This suggests that sensitivity is driven by model-specific internal representations rather than inherent difficulty of certain questions. If certain questions were inherently "hard to paraphrase" we would expect models to fail on the same questions. The near-zero correlation argues against this explanation.

Table 3.4: Denominator-explicit flip rates for MedGemma-4B on the binary yes/no subset, with reference prevalence and model positive-prediction rate reported alongside. The *pairwise* rate (used in Table 3.3 and throughout this chapter) divides flipped paraphrases by paraphrase pairs; the *query-level* rate divides questions with at least one flipped paraphrase (of the roughly 3.3 paraphrases per question) by questions. The two must never be conflated. Prevalence and positive-prediction rate are reported because the datasets are not answer-balanced: VinDr contains only positive presence cases (a positive-case presence-question test, no negatives), and PadChest is 83% positive, so these datasets measure phrasing stability but cannot support balanced-accuracy or sensitivity/specificity claims. Population and inclusion rules are fixed by the manifest `results/uai/revision/psf_binary_inclusion_manifest.jsonl`; the nine VinDr non-binary "other" references are excluded.

Dataset	Questions	Pairs	Ref. pos.	Model pos.	Pairwise flip	Query-level flip
MIMIC-CXR	1,539	5,076	51%	61%	8.3%	18.1%
PadChest	8,445	36,244	83%	24%	13.4%	32.2%
VinDr-CXR	2,807	8,612	100%	50%	15.3%	28.5%

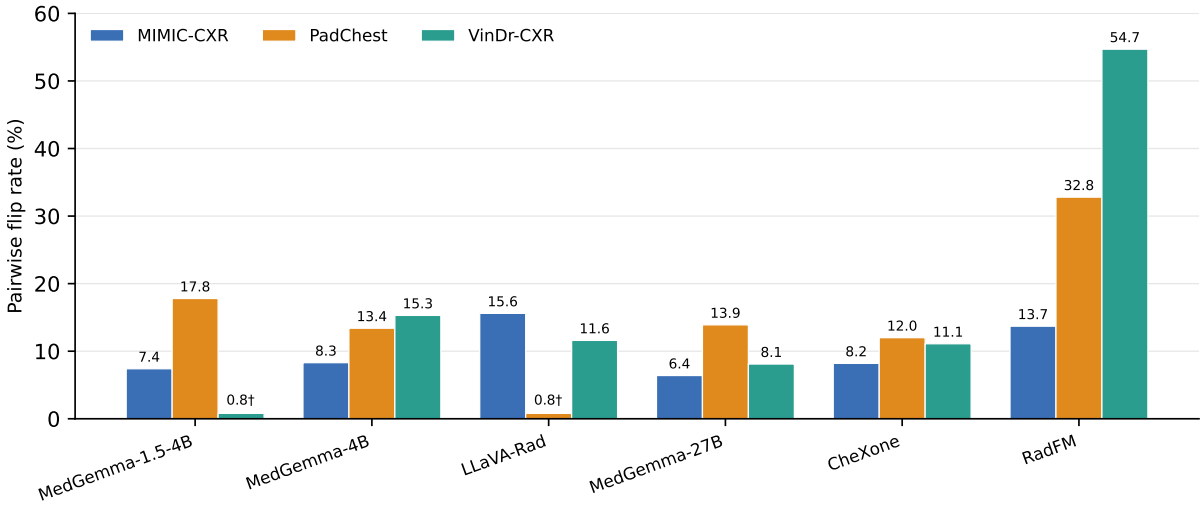


Figure 3.3: Flip rate for each of the six base medical VLMs across the three datasets, from the equivalence-filtered run (Table 3.3). Daggers mark cells with degenerate yes-bias, where the rate reflects a single-class prior rather than paraphrase robustness. No single model is best everywhere, and scaling within the MedGemma family does not reduce sensitivity.

We quantify inter-model agreement more precisely using the Jaccard index on the query-level flip sets (the set of questions each model flips on). The mean pairwise Jaccard index is 0.13 on MIMIC-CXR, 0.18 on PadChest, and 0.14 on VinDr. The overlap is modest but not negligible: it is well below the 0.5 that near-identical failure sets would produce, yet larger than on the earlier unfiltered run, because restricting to genuine binary yes/no questions concentrates the shared vocabulary of clinically hard cases. The practical implication survives: an ensemble of multiple medical VLMs could reduce flip rates by exploiting the largely disjoint remainder of their failure sets, though Chapter 6 shows that LoRA ensembles face their own challenges.

3.4.2 Per-Model Analysis

Each model's flip rate profile reveals distinct failure patterns:

MedGemma family. The scaling trend does not produce a stable ordering. Post-filter, MedGemma-1.5-4B and MedGemma-4B are close on MIMIC (7.4% vs. 8.3%, a 0.9 point gap), but on PadChest MedGemma-1.5-4B is the *less* consistent of the two (17.5% vs. 13.4%), and on VinDr

they diverge sharply because MedGemma-1.5-4B collapses into 100% yes-bias (its 0.8% flip rate is artifactual) while MedGemma-4B retains a meaningful 15.3% rate. MedGemma-27B, the largest model, is the *most* consistent of the three on MIMIC (6.4%) and on VinDr (8.1%), yet MedGemma-4B edges it on PadChest (13.4% vs. 13.9%). Parameter count therefore neither predicts consistency nor yields a fixed ranking across datasets: the model that is most consistent in distribution is not the most consistent under either shift. Chapter 4 returns to the 27B model and shows that its consistency on the curated FlipBank rests heavily on text priors rather than image content.

LLaVA-Rad. LLaVA-Rad's filtered results are bifurcated: 15.6% on MIMIC-CXR and 11.6% on VinDr-CXR (both meaningful), but only 0.7% on PadChest with 99.6% yes-bias.[3] The PadChest cell looks like a strong consistency result and is the dissertation's clearest counter-example to flip rate as a safety metric. Chapter 4 re-examines this behavior using text-only and image-swap controls and shows that most of LLaVA-Rad's PadChest answers are unchanged when the image is removed entirely; predictions on regenerated paraphrases are stable for the same reason that predictions on swapped images are stable, namely, the model leans heavily on the question text on this distribution. The four-quadrant analysis (Chapter 6) places the large majority of LLaVA-Rad's consistent PadChest predictions in the image-invariant (DANGEROUS) cell. By contrast, the 15.6% MIMIC and 11.6% VinDr rates reflect more genuine sensitivity: text-only agreement is lower, and the model's predictions are not collapsed onto a single class.

RadFM. RadFM has the highest flip rate on both out-of-distribution sets (32.8% PadChest, 54.7% VinDr) and the second-highest on MIMIC (13.7%, behind LLaVA-Rad's 15.6%), and it is the only model whose flip rate strictly increases with each step of distribution shift away from MIMIC. Filtering does not narrow the gap. The PadChest figure is computed on the 8,092 evaluable binary presence questions (a small number skipped due to image loading failures); the VinDr figure on the 2,807-question binary presence set. Rather than a distribution-shift effect alone, the RadFM pattern looks like a baseline instability in question processing that compounds with shift.

CheXone. CheXone [88] rises from 8.2% on MIMIC to 13.0% on PadChest, then eases slightly to 11.1% on VinDr; both out-of-distribution sets are harder than MIMIC, with PadChest marginally the hardest. Unlike the earlier unfiltered run, there is no dataset on which its OOD flip rate falls below its MIMIC rate, so the shift effect is monotone in aggregate for this model. CheXone replaced CheXagent in the post-rebuttal evaluation; the older CheXagent runs on disk are stale and not reported here.

3.4.3 Cross-Dataset Analysis

The rise in flip rates from MIMIC-CXR to PadChest and VinDr-CXR is informative about how paraphrase sensitivity varies across clinical populations. The three sets differ in institution, task composition, label balance, and templates simultaneously, so the ratios below are cross-population contrasts and should not be read as the effect of distribution shift alone. Post-filter, the per-model ratios (relative to MIMIC) are:

- MedGemma-4B: 8.3% → 13.4% → 15.3% (1.61× PadChest, 1.84× VinDr)
- MedGemma-1.5-4B: 7.4% → 17.5% → 0.8%† (2.36× PadChest; VinDr cell is yes-bias artifact)
- MedGemma-27B: 6.4% → 13.9% → 8.1% (2.17× PadChest, 1.27× VinDr)
- CheXone: 8.2% → 13.0% → 11.1% (1.59× PadChest, 1.35× VinDr)
- LLaVA-Rad: 15.6% → 0.7%† → 11.6% (PadChest cell is yes-bias artifact; VinDr is 0.74×)
- RadFM: 13.7% → 32.8% → 54.7% (2.39× PadChest, 3.99× VinDr)

Three patterns are visible. First, the two yes-bias cells (LLaVA-Rad on PadChest, MedGemma-1.5-

[3] On the smaller, answer-balanced LLaVA-Rad flip bank used in Chapters 6 and 7 ($n = 732$ PadChest, with an $n = 88$ MIMIC counterpart, curated to contain both answer polarities), LLaVA-Rad flips more (20.5%) and its positive rate falls to 89.6%, but it remains the most text-reliant model there; the two rates differ because the benchmark's question mix is dominated by positive findings while the flip bank is balanced.

Table 3.5: Mean pairwise flip rate (%) by paraphrase transformation type on the equivalence-filtered PSF-Med binary yes/no subset, averaged across the six base models (MedGemma-4B, MedGemma-1.5-4B, MedGemma-27B, LLaVA-Rad, CheXone, RadFM). The negation-pattern row is dataset-asymmetric: the rubric audit rejected 93.8% of generated negation pairs as polarity- or operator-changing, leaving very few on MIMIC (11 per model) and effectively none on PadChest, while the VinDr regeneration retained about 1,126 per model that meet the equivalence rubric. Per-cell sample sizes are binary-subset paraphrase pairs per model; the MIMIC negation cell is reported for completeness but rests on only 11 pairs.

Transformation Type	MIMIC-CXR	PadChest	VinDr-CXR
Lexical substitution	11.1 (n=1,511)	17.2 (n=14,662)	12.5 (n=2,816)
Syntactic restructuring	8.7 (n=1,400)	16.5 (n=8,235)	18.6 (n=1,145)
Scope quantification	9.8 (n=1,183)	17.2 (n=8,095)	14.1 (n=1,842)
Specificity modulation	10.0 (n=971)	5.3 (n=5,252)	14.3 (n=1,715)
Negation pattern	18.2 (n=11)	N/A	**37.9** (n=1,126)

4B on VinDr) sit far below the rest because the model has collapsed onto a single class on that dataset and is not making meaningful binary decisions; both cells are read as instances of the consistency-without-grounding pattern in Chapter 6. Second, when both OOD cells are meaningful, the VinDr rate is sometimes lower than the PadChest rate (MedGemma-27B, CheXone) and sometimes higher (MedGemma-4B, RadFM); LLaVA-Rad is even more consistent on VinDr than on MIMIC (0.74×). The two distribution shifts are not comparable on a single difficulty axis. Third, RadFM's 3.99× ratio on VinDr is the most severe in the table: whatever combination of institution, task composition, and label balance separates VinDr from MIMIC degrades this model far more than the others.

These patterns have practical implications for deployment. A model's paraphrase sensitivity on in-distribution data may underestimate its sensitivity when deployed in a new clinical environment, and the choice of OOD test set changes the answer: the US-to-Spain shift in PadChest produces different rankings than the US-to-Vietnam shift in VinDr. Sensitivity testing should include multiple out-of-distribution datasets, and clinical deployment should not generalize from a single OOD evaluation.

3.4.4 Analysis by Paraphrase Type

Table 3.5 breaks down flip rates by transformation type on the filtered run, averaged across the six base models. The post-filter picture is dataset-dependent in a way the unfiltered earlier version of this benchmark did not surface.

Three patterns stand out. First, on MIMIC-CXR the four well-populated categories range from 8.7% to 11.1%, a spread of about two points; transformation type explains little of the cross-pair flip variance once the audit removes adversarial reframing. Second, on PadChest the lexical, syntactic, and scope categories cluster around 16.5 to 17.2% while specificity modulation sits well below the others at 5.3%; the regenerated paraphrase set retained tighter specificity matches than the other categories. Third, and most importantly for revisiting the original claim of that earlier version, the negation-pattern row is highly dataset-dependent: the audit rejected 93.8% of generated negation pairs as polarity- or operator-changing, leaving only 11 pairs per model on MIMIC (too few to estimate reliably) and almost none on PadChest, but the VinDr regeneration retained roughly 1,126 surviving negation pairs per model and these flip at 37.9% on average across base models, the highest rate among the VinDr categories.

The directional pattern of the VinDr negation flips is striking. On five of the six base models every retained-pair negation flip is a no-to-yes transition: Base 700/700, MedGemma-27B 330/330, LLaVA-Rad 115/115, CheXone 471/471, MedGemma-1.5-4B 6/6. RadFM is the lone exception, with 936/936 going yes-to-no. The pattern is consistent with the Feature 3818 register-gate hypothesis from Chapter 5: certain framings systematically push the model in one direction rather than producing

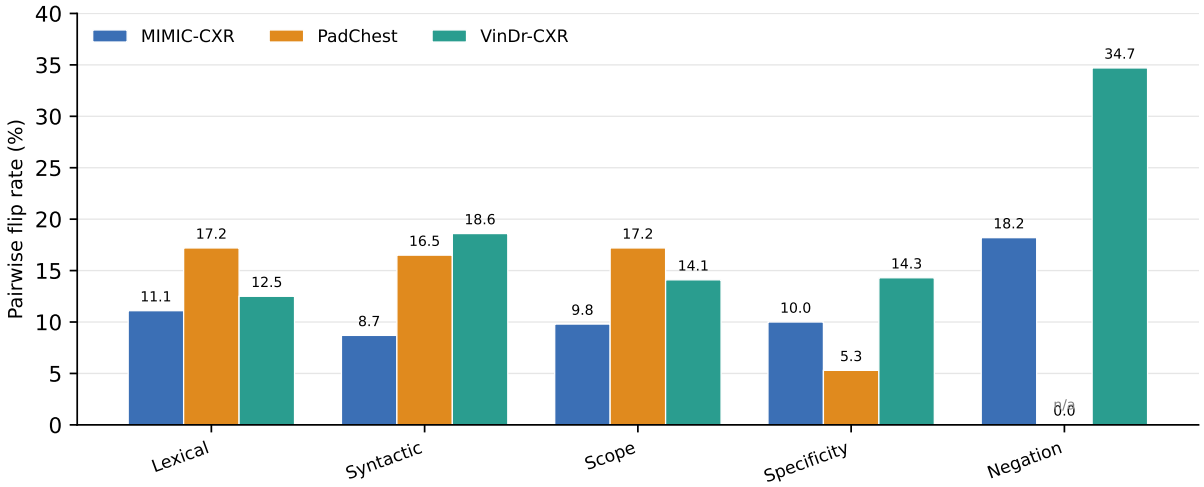


Figure 3.4: Flip rate by paraphrase transformation type on the binary yes/no subset, averaged across the six base models (Table 3.5). Negation pairs spike to 37.9% on VinDr-CXR, where about 1,126 negation paraphrases per model survived the equivalence audit. On MIMIC and PadChest almost all negation pairs were rejected as polarity-changing, so the type is sparse (11 pairs per model on MIMIC, almost none on PadChest).

balanced bidirectional fragility, and RadFM's opposite asymmetry is consistent with its inverted register sensitivity in the Sparse Autoencoder (SAE) analysis.

The VinDr finding partially rehabilitates the unfiltered claim of that earlier version that negation paraphrases are disproportionately disruptive: when the equivalence audit retains a substantial population of negation pairs, the disruption returns. But the path from the unfiltered claim to the filtered evidence is not a simple confirmation. The rebuttal-period audit demonstrated that most generated negation paraphrases on MIMIC and PadChest changed the operator and were therefore not bidirectional-entailment paraphrases at all; the original 25 to 35% headline rate on MIMIC was inflated by these adversarial reframings. The VinDr negation pairs that survive the audit are operator-preserving in the rubric sense, and their 37.9% flip rate is the cleanest evidence in the dissertation that genuinely meaning-preserving negation paraphrasing remains a hard test for medical VLMs. The "yes-to-no asymmetry" analysis on the unfiltered set is preserved in Appendix A.1 as a sensitivity analysis on adversarial reframings. The new VinDr-restricted directional asymmetry reported in the previous subsection is the cleaner post-filter version of the same phenomenon, computed on operator-preserving pairs and showing the no-to-yes direction across five base models.

The mechanistic implication is that Feature 3818, the SAE-level register gate identified in Chapter 5, is not specifically a negation handler. If it were, removing the negation pairs on MIMIC would have visibly weakened the gate's predictive power on the remaining flips, and the VinDr negation rate would track the feature directly. Chapter 5 shows that Feature 3818 still ranks as a top flip-predictor on the filtered FlipBank, and that its predictive role generalizes across paraphrase types rather than concentrating on negation. The 37.9% VinDr negation rate is therefore plausibly downstream of multiple mechanisms, only one of which the chapter identifies.

3.4.5 Embedding Distance and Flip Prediction

If paraphrase sensitivity were driven purely by semantic distance between the original and paraphrased questions, we would expect flip pairs to have lower embedding similarity than non-flip pairs. This is true, but the effect is small.

Using BioClinicalBERT embeddings, flip pairs have a mean cosine similarity of 0.966 versus 0.969 for non-flip pairs. The difference is statistically significant ($p < 10^{-24}$) but the point-biserial correlation

is only $r = -0.09$. In Euclidean distance, flip pairs average 0.258 versus 0.243 for non-flip pairs.

This weak relationship tells us something important: paraphrase sensitivity is not well predicted by how "different" the paraphrase looks in embedding space. Two questions can be nearly identical by any standard similarity metric and still produce opposite answers. The instability is driven by features of the model's internal processing, not by properties of the input alone. This observation motivates the mechanistic investigation in Chapter 5.

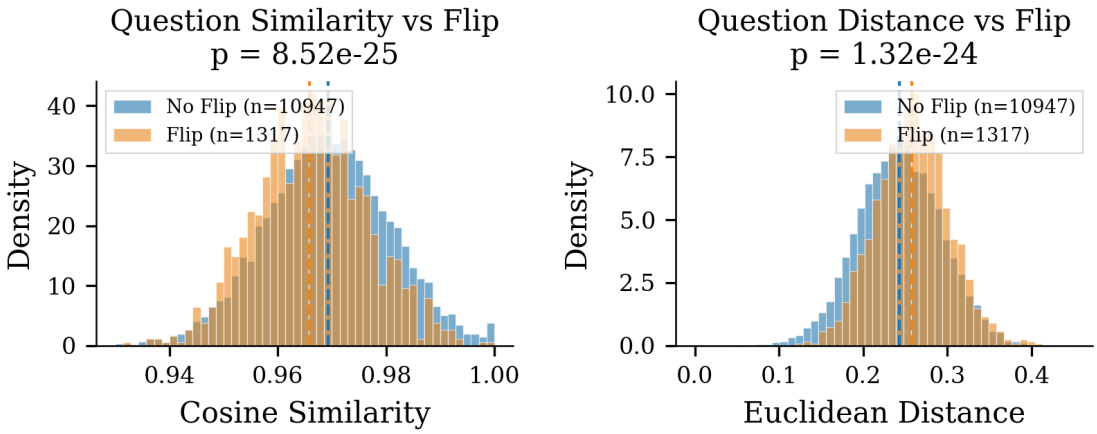


Figure 3.5: Embedding distance between original and paraphrased questions, stratified by flip outcome. Flip pairs and stable pairs overlap substantially, confirming that semantic distance is a poor predictor of sensitivity ($r = -0.09$).

3.4.6 Operator Change Correction

Not all apparent flips are genuine paraphrase sensitivity failures. Some paraphrases subtly change the *operator* of the question, from a presence framing ("Is there X?") to an exclusion framing ("Can X be ruled out?"). In these cases, the expected yes/no polarity inverts: a "Yes" to "Is there X?" and a "No" to "Can X be ruled out?" express the same clinical judgment (the finding is present). A model that gives "Yes" to the first and "No" to the second is not flipping; it is correctly tracking the operator change.

We develop a rule-based operator classifier that identifies presence, exclusion, likelihood, and neutral framings based on lexical cues. The classifier categorizes each question-paraphrase pair as same-operator (both use the same framing) or different-operator (the paraphrase changes the framing). Approximately 28% of apparent flips in the raw PSF-Med results involve operator changes. After correction (reclassifying operator-appropriate answer changes as non-flips) the true PSF rate is approximately 0.72× the raw reported rate.

This correction is conservative: we only reclassify cases where (1) the operator clearly changes and (2) the answer change is in the expected direction for that operator change. Cases where the answer changes in the *unexpected* direction (e.g., the operator changes from presence to exclusion, but the model flips from "No" to "Yes") remain classified as flips.

On the unfiltered earlier version of this benchmark, the operator correction affected all models roughly equally (27 to 30% of flips were operator-related across models), so the relative ranking of models was unchanged. Post-filter, the equivalence audit has already rejected most operator-changing pairs as "not equivalent" (the 48.7% rejected slice in the audit), so the residual operator correction on the filtered run is small. We report filtered raw rates throughout this dissertation. The unfiltered operator-correction analysis is preserved in Appendix A.1 as a sensitivity analysis on the original benchmark.

3.4.7 A Limited Frontier-Model Comparison

The six models above are open-weight systems evaluated with full logit access. A natural question is whether closed frontier models, with chain-of-thought reasoning available, behave differently. We tested this in two limited comparisons, both restricted to the binary yes/no population. First, a matched reasoning ablation re-asks the same 1,960 paraphrase pairs with reasoning disabled and then enabled within each of three API models; this comparison is reported in the workshop version of this benchmark. Enabling reasoning never reduces the flip rate, and it significantly increases the rate in two of the three models (Claude Haiku 4.5: 17.2% to 22.7%, $p = 1.4 \times 10^{-5}$; Gemini 3 Flash: 12.7% to 15.0%, $p = 0.025$; GPT-5-mini: 7.8% to 8.2%, $p = 0.57$). Second, a larger arm evaluates four frontier configurations on the 500 hardest binary questions per dataset, 24,336 pairs in total: pooled pairwise flip rates are 5.7% (Kimi K3), 8.4% (Claude Opus 5 with reasoning disabled), 9.5% (Claude Opus 5 with reasoning enabled), and 9.7% (GPT-5.6), for an 8.4% pooled rate overall.

Two qualifications bound these comparisons. The larger arm evaluates an upper-tail population, the hardest binary questions rather than a representative sample, so its rates must not be compared cell-by-cell with Table 3.3. And when the image-reliance controls of Chapter 4 are applied to these same arms, the frontier models show the same pattern as the open-weight fleet: a mean of 68.5% of each frontier model's consistent predictions are unchanged when the image is removed, against 81% for the open-weight models, and the most flip-stable frontier model (Kimi K3) is the most image-invariant one on PadChest (90.8%). Paraphrase sensitivity is therefore not an artifact of the six open-weight models studied here, and a lower flip rate on these populations is no more evidence of image use for a frontier model than it was for an open-weight one. The mechanistic and mitigation claims of this dissertation remain scoped to open-weight models, whose internals are accessible.

3.4.8 The Text-Only Signal

As a preliminary signal, we ask whether models need the image at all to answer consistently, previewing results that Chapter 4 establishes in detail. Chapter 4 measures text-only agreement, the fraction of questions answered identically with the image removed entirely, on a 107-pair diagnostic set; high text-only agreement indicates that the image contributes little to the model's decision.

The results are striking. On the 107-pair diagnostic set, text-only agreement is 85.0% for MedGemma-27B, 96.3% for LLaVA-Rad, and 66.4% for MedGemma-4B, the most image-dependent. This means that for LLaVA-Rad, nearly all of its answers can be predicted from the text alone, without looking at the chest X-ray.

High text-only agreement raises a troubling possibility. If a model returns the same answer with and without the image, its consistency across paraphrases follows from the language pathway alone, and a low flip rate is then not a sign of good visual reasoning. What the measurement establishes is narrower than it first appears: removing the image did not move the answer across the yes/no boundary, which bounds how much the image can be contributing but does not by itself show that the model never consults it. The same observation would arise from a question text that already carries the answer, from a margin that shifted without crossing the threshold, and from a base rate that makes the text prior right on most items. In the two degenerate cells the argument is cleaner, because a model answering "yes" on 99.6% or 100% of questions has collapsed onto a single-class prior and a constant answer cannot track any image.

Chapter 4 investigates this possibility in detail using controlled image swap experiments, which change what the image shows rather than removing it and therefore support a stronger reading than removal alone. The key finding, previewed briefly here, is that the correlation between low flip rate and high text-only agreement is not coincidental. It reflects a trade-off between paraphrase consistency and image reliance.

3.5 The FlipBank

For the mechanistic and mitigation work in Chapter 5, we construct a curated subset of flip cases called the FlipBank. This is a set of 158 question pairs (the canonical mechanistic FlipBank used in Chapter 5; not to be confused with the 107-pair three-backend diagnostic set of Chapter 4) collected as flip cases under the original evaluation pipeline, where MedGemma-4B changed its binary answer between the original and a paraphrase. Chapter 5's re-analysis finds that 17 of these pairs are polarity switches (where an answer change is correct behavior) and that, of the 141 operator-preserving pairs, the model flips on 76, so downstream analyses treat the bank as a mixture of flip and non-flip pairs. Three filtering criteria were applied at collection:

1. The model gives a definitive binary answer change (yes $\rightarrow$ no or no $\rightarrow$ yes).
2. The paraphrase has high semantic similarity to the original (> 0.95 cosine similarity in BioClinicalBERT).
3. Both the original and paraphrased answers parse unambiguously (no hedge, refusal, or off-topic response).

Of the 158 FlipBank cases as collected, 94 (59%) are yes-to-no answer changes and 64 (41%) are no-to-yes, drawn from 142 unique images and 23 distinct clinical findings. The FlipBank is used as the analysis set for SAE feature extraction (Chapter 5) and is kept separate from any training or evaluation splits to prevent leakage.

The FlipBank construction is deliberately conservative. The 0.95 cosine similarity threshold matches the MIMIC prefilter recorded in the evaluated data artifact and is well above the 0.90 threshold of the original construction, ensuring that FlipBank pairs are nearly indistinguishable in embedding space. This conservatism serves the mechanistic analysis: when we observe a flip in the FlipBank, we can be highly confident that the semantic content of the question has not changed, and any answer change reflects instability in the model's internal processing.

The distribution of FlipBank cases across clinical findings is non-uniform. Pleural effusion accounts for 28 cases (17.7%), cardiomegaly for 22 (13.9%), and pneumothorax for 19 (12.0%). Less common findings contribute fewer cases, with 6 findings contributing only 1 to 3 cases each. This distribution roughly tracks the overall prevalence of flip-prone questions in the benchmark, though high-prevalence findings are slightly over-represented because they have more opportunities for flip-inducing paraphrases.

The asymmetry between yes-to-no (59%) and no-to-yes (41%) flips is notable. It suggests a slight directional bias: the model is more likely to retract a positive diagnosis under paraphrasing than to introduce one. This asymmetry interacts with base rates: positive predictions for rare findings are particularly unstable, while positive predictions for common findings are more stable. Chapter 5 analyzes this asymmetry at the feature level and finds that Feature 3818 is more active for presence-style questions (which tend to elicit "yes"), partially explaining the directional bias.

3.6 Discussion

PSF-Med establishes three facts about medical VLMs. First, paraphrase sensitivity is common and varies widely across models and clinical populations. On the equivalence-filtered binary yes/no subset, the lowest meaningful flip rate is 6.4% (MedGemma-27B on MIMIC) and the highest is 54.7% (RadFM on VinDr), an 8.5$\times$ spread; two additional cells (LLaVA-Rad on PadChest and MedGemma-1.5-4B on VinDr) drop below 1% but reflect degenerate yes-bias rather than genuine consistency. Second, the problem is model-specific: different models fail on different questions (mean pairwise $r = 0.06$ on MIMIC, 0.04 on PadChest, 0.07 on VinDr), suggesting internal representational causes rather than inherent question difficulty. Third, the relationship between consistency and accuracy is not straightforward. A model can be consistent without being correct (by always answering the same wrong

answer) or correct without being consistent (by answering correctly for one phrasing and incorrectly for another).

3.6.1 Clinical Implications of Flip Rates

The flip rates measured in PSF-Med can be translated into concrete clinical scenarios, provided the denominator is stated. On MIMIC-CXR, MedGemma-4B's 8.3% pairwise rate means that roughly 1 in 12 individual rephrasings of a question receives a different answer; the more relevant per-question figure is the query-level rate of 18.1%, meaning that about 1 in 5 questions receives at least one contradictory answer across its rephrasings. In a workflow processing 100 images per day with one query per image, that is roughly 18 questions per day for which the answer depends on wording. On PadChest the same model's query-level rate rises to 32.2%, roughly 1 in 3 questions. These per-question rates are well above what a clinician would tolerate from a deterministic decision support tool, and the gap between the pairwise and query-level framings is itself a caution: a benchmark that reports only the smaller pairwise number understates the per-question exposure that a deployed clinician actually faces.

Negation framings remain clinically important even though the equivalence audit removed most negation-pattern paraphrases from the headline benchmark. Clinicians routinely use exclusion phrasings ("Is pneumothorax excluded?", "Can we rule out effusion?", "Is there no evidence of consolidation?"), and the rubric audit showed that 93.8% of GPT-generated negation paraphrases are not strict bidirectional-entailment paraphrases of presence questions: rephrasing "Is X present?" as "Can X be ruled out?" usually changes the operator and therefore the expected answer polarity. The clinical risk is real, but it is a question of operator-aware reasoning rather than of paraphrase robustness, and we treat it that way in the safety evaluation in Chapter 6. The adversarial reframing slice (the rejected 48.7% of audited pairs) is preserved and reported separately in Appendix A.1.

3.6.2 The Text-Only Concern

The most concerning finding is the text-only signal. Models that appear consistent may be achieving that consistency from the question text rather than from the image. This is not detectable by measuring flip rate alone. The correlation between low flip rate and high text-only agreement is not a coincidence. It reflects a systematic pattern where paraphrase consistency arises from text-driven behavior rather than from stable visual processing.

This finding has immediate practical implications. Any evaluation of paraphrase sensitivity in medical VLMs must include text-only baselines or controlled image swap tests to distinguish genuine visual grounding from text-driven shortcuts. Reporting flip rate without text-only agreement is incomplete and potentially misleading: a model with 2% flip rate and 98% text-only agreement appears reliable, yet 98% of its answers are unchanged when the image is removed, so its flip rate certifies nothing about whether the scan is being read. The next chapter takes up this challenge directly.

3.6.3 Comparison to General-Domain VLM Robustness

The paraphrase sensitivity rates we observe in medical VLMs can be contextualized against known robustness issues in general-domain vision-language models. In the VQA-CP benchmark [17], question-type bias causes accuracy drops of 15 to 40 percentage points when the question-answer distribution shifts between training and test. In adversarial paraphrasing of text classifiers [99], syntactically controlled paraphrases fool models on 20 to 30% of inputs. Our post-filter flip rates (meaningful range 6.4 to 54.7%) are comparable in magnitude to these adversarial baselines on the OOD datasets and smaller on MIMIC, but arise in a distinct setting: the perturbation preserves meaning (unlike adversarial attacks), and the consequence of failure is clinical rather than academic.

The medical setting also introduces unique considerations. General-domain VQA datasets have diverse question types and answer vocabularies, making consistency harder to define. Medical binary questions have a clear consistency criterion (the answer should not change under paraphrasing), making

PSF measurement unambiguous. However, the clinical stakes also mean that even low flip rates (8%) translate to frequent inconsistencies in daily clinical workflows, as discussed in Section 3.6.

3.6.4 Summary

The take-away from Thrust 1 is that paraphrase sensitivity is a real, measurable, and widespread property of how medical VLMs relate clinical queries to diagnostic images, not an artifact of evaluation or imperfect paraphrases. The text-only signal, the identification of MedGemma-4B as high-flip but image-dependent, and the FlipBank set up the diagnosis and mitigation in the chapters that follow.

Chapter 4

Thrust 2: Diagnosing the Origin of Paraphrase Sensitivity

Research questions. This chapter answers **RQ2.1** (how far models rely on text priors versus image content), **RQ2.2** (whether paraphrase consistency trades off against visual grounding), and **RQ2.3** (whether flipped answers attend less to the relevant anatomical region).

Main contribution. A controlled diagnosis that separates consistency from image dependence: on a 107-pair three-backend diagnostic set, text-only, gray-placeholder, and opposite-label-swap controls show that the model which flips least also reads the image least, so a low flip rate is not evidence of visual grounding, a tendency across the models studied rather than a universal law.

Chapter 3 established that paraphrase sensitivity is prevalent across medical Vision-Language Models, with post-filter flip rates on the binary yes/no subset ranging from 6.4% to 54.7% across six models on three clinical populations (MIMIC-CXR, PadChest, VinDr-CXR), once two cells of degenerate yes-bias are set aside. The text-only signal in Section 3.4.8 hinted at a troubling possibility: models with low flip rates may achieve their consistency not through stable visual reasoning, but by answering from language priors that the image does little to move. This chapter investigates that possibility directly.

The central question is: *when a model gives consistent answers across paraphrases, is that consistency grounded in visual evidence?* We design three controlled experiments to answer this question: text-only evaluation (removing the image entirely), controlled image swap (replacing the image with one showing the opposite finding), and attention-bounding box analysis (measuring whether the model's attention targets the correct anatomical region). The experiments are applied to three model backends that span the consistency spectrum: MedGemma-4B (high flip rate, 42.1%), MedGemma-27B (moderate flip rate, 14.0%), and LLaVA-Rad (low flip rate, 6.5%) on the 107-pair diagnostic set.

The findings reveal a robustness-grounding trade-off that is the central tension of this dissertation. MedGemma-4B exhibits high flip rates but maintains meaningful image dependence: its answers change 30.8% of the time when the image is swapped to show the opposite finding. LLaVA-Rad shows near-zero flip rate but is almost entirely text-driven: its answers match the text-only condition 96.3% of the time and change under image swap only 10.3% of the time. MedGemma-27B sits between these extremes, showing that scaling reduces flip rate but also reduces visual grounding. These results have a direct practical implication: mitigations developed in Thrust 3 should target models that are grounded but unstable (like MedGemma-4B), not models that are stable because their answers barely move with the image (like LLaVA-Rad).

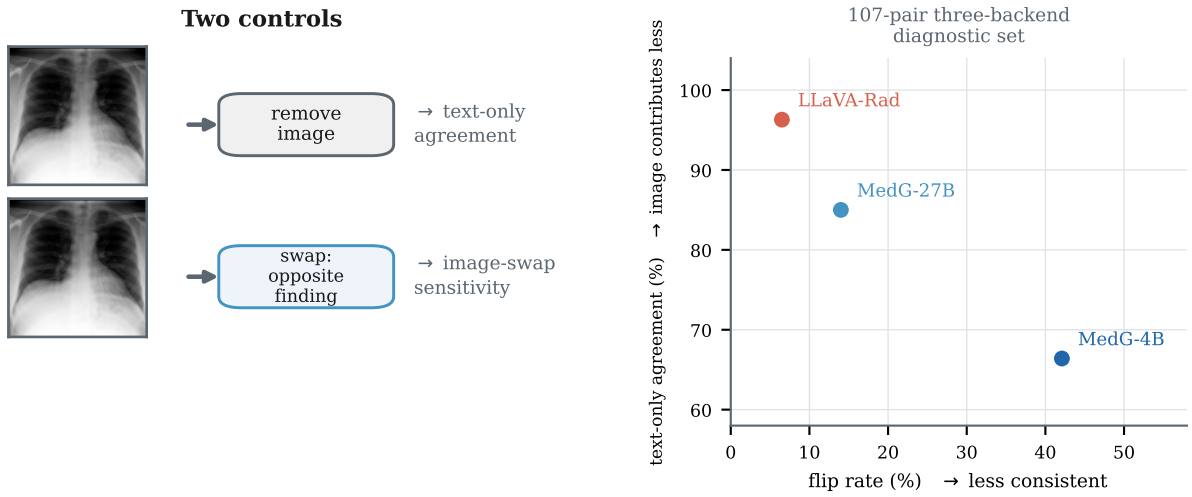


Figure 4.1: Thrust 2 at a glance. Two controls separate the ways a model can appear consistent: removing the image tests text-only agreement, and swapping in an image of the opposite finding tests image-swap sensitivity. On the resulting trade-off, LLaVA-Rad is stable because it leans on the question text, while MedGemma-4B uses the image but flips more often.

4.1 Related Work

4.1.1 Language Priors in Vision-Language Models

The tendency of vision-language models to rely on language priors rather than visual content is well documented in the general domain (Chapter 3 and [17, 16]): a model can achieve high benchmark accuracy while largely ignoring the image, a special case of the shortcut learning that Geirhos et al. [104] document across domains. In the medical setting this takes a specific form. Chest X-ray findings have strong base rates (cardiomegaly and pleural effusion are far more common than pneumothorax or pneumomediastinum), so a model that learns these rates can answer "Yes" to common findings and "No" to rare ones without analyzing the image. LLaVA-Med's training data is built from figure captions enriched for positive findings [5], a construction that favours a positive prior, and Eslami et al. [7] found that medical CLIP models show less visual-grounding benefit than in general-domain tasks, suggesting the language pathway dominates when clinical text provides sufficient signal.

The contribution of this chapter is to show that language-prior reliance is not merely a training artifact but a systematic failure mode that correlates with paraphrase consistency: the models most consistent across paraphrases are, paradoxically, the ones that rely most heavily on text priors.

4.1.2 Visual Grounding Analysis Methods

Evaluating whether a model uses visual information requires interventions that go beyond standard accuracy testing. Several approaches have been developed in the general domain.

Text-only baselines. The simplest test of visual grounding is to remove the image entirely and observe whether the model's behavior changes. A changed answer establishes that the image mattered for that prediction. An unchanged answer establishes less than it appears to: it shows that removing the image did not move the decision across the yes/no boundary, which is also what one would observe if the question text carried redundant evidence, if the margin shifted without crossing the threshold, or if the finding's base rate made the text prior right anyway. Removal therefore bounds how much the

image can be contributing rather than proving that the model did not consult it, which is why we pair it with the controlled image swap of Section 4.2, a control that changes what the image shows instead of taking the image away. This approach has been used in VQA evaluation [17, 16] and has revealed that many VQA models perform well above chance even without images. We extend this approach to medical VLMs, where it has received less systematic attention.

Image perturbation. A complementary approach replaces the correct image with an incorrect one and measures whether the model's answer changes. Shah et al. [97] enforce cycle-consistency between a question, a generated rephrasing of it, and the answer on a *fixed* image, which tests linguistic rather than visual perturbation; our controlled swap protocol is the image-side complement to that idea. Our controlled swap protocol extends this idea by specifically pairing each question with an image that shows the opposite ground-truth label for the queried finding, providing a stronger test of visual dependence than random image substitution.

Attention analysis. Attention heatmaps have been widely used to visualize which image regions a model attends to [105]. In medical imaging, attention overlap with radiologist-annotated pathology regions has been used as evidence of visual grounding. However, the interpretability of attention weights is contested. Jain and Wallace [106] showed that attention weights are frequently uncorrelated with gradient-based importance measures and that different attention distributions can produce equivalent predictions. Wiegreffe and Pinter [107] challenged this view, arguing that the existence of alternative attention distributions does not invalidate the explanatory value of learned attention. We use attention analysis as one signal among several, while acknowledging its limitations.

Causal interventions. The strongest evidence for visual grounding comes from causal interventions: if removing or altering a specific image region changes the model's prediction, that region was causally important. Adebayo et al. [108] introduced sanity checks for saliency methods, showing that many explanation methods produce similar outputs regardless of model weights. We use region of interest (ROI) ablation (removing the annotated pathology region and observing whether the answer changes) as a causal complement to attention overlap analysis.

4.1.3 The Attention Faithfulness Debate

The question of whether attention weights faithfully represent a model's reasoning process has generated substantial debate in the natural language processing (NLP) and machine learning communities. This debate is directly relevant to our work because attention heatmaps are commonly presented as evidence of visual grounding in medical VLMs.

Jain and Wallace [106] argued against attention as explanation on two grounds: learned attention weights correlate poorly with gradient-based importance, and adversarial attention distributions substantially different from the learned ones can produce nearly identical outputs. Wiegreffe and Pinter [107] countered that the existence of alternative explanations does not invalidate a given one, and that attention carries task-relevant information under appropriate diagnostic tests.

For medical VLMs, this debate has practical implications. If a model's attention heatmap highlights the correct lung region when answering a question about pneumothorax, clinicians may interpret this as evidence that the model identified the pneumothorax in the image. Our results suggest caution: we find that attention overlap with pathology regions is poor for MedGemma-4B, the model measured here, and Chapter 6 shows that attention saliency is uncorrelated with causal importance under patch occlusion. The attention debate in NLP translates directly to the medical imaging setting, and the evidence in our case supports skepticism about attention as a grounding signal.

4.1.4 Robustness Versus Reliability in Clinical AI

The distinction between robustness (consistent behavior under perturbation) and reliability (correct behavior grounded in evidence) is implicit in much of the AI safety literature but rarely made explicit.

In the clinical AI context, this distinction is critical.

A model can be consistent without being reliable: it gives the same answer regardless of input perturbation, but that answer is not based on the evidence. This is the "Clever Hans" failure mode identified in the general domain [104], where models learn superficial correlations that happen to work on standard benchmarks but fail when the correlation is broken. In medical imaging, this manifests as models that learn text-based priors ("most chest X-rays are normal, so answer No to most findings") rather than visual features.

Conversely, a model can be reliable without being consistent: it correctly identifies findings when the question is phrased in a specific way, but its internal representation of the image evidence is fragile enough that rephrasing the question shifts the decision boundary. This is the failure mode identified in Chapter 3: MedGemma-4B correctly identifies pleural effusion on a chest X-ray when asked "Is there pleural effusion?" but changes its answer when asked "Does this X-ray show fluid in the pleural space?"

The goal is a model that is both consistent and reliable: consistent under paraphrasing *because* its answers are grounded in stable visual representations. Current models achieve at best one of these properties. The experiments in this chapter characterize which property each model achieves and why.

4.2 Experimental Design

4.2.1 Research Questions

This thrust addresses three specific questions:

1. **RQ1: Text reliance.** To what extent do models rely on text priors versus image content when answering clinical questions? Can the model's answer be predicted from the question alone, without the image?
2. **RQ2: Image sensitivity.** When the image is changed to show the opposite finding, does the model change its answer? An unchanged answer means the decision did not depend on that change in image content, which is the strongest available bound on image reliance short of a causal ablation.
3. **RQ3: Spatial grounding.** When the model does appear to use the image, is its attention directed at the correct anatomical region? Does attention overlap with radiologist-annotated pathology regions differ between flip and non-flip cases?

These questions form a diagnostic hierarchy. RQ1 tests the coarsest form of image dependence (any dependence at all). RQ2 tests whether the dependence is on the *content* of the image (not just its presence). RQ3 tests whether the model attends to the *relevant* content. A model must pass all three tests to demonstrate genuine visual grounding.

4.2.2 Model Selection

We evaluate three model backends chosen to span the consistency-grounding spectrum identified in Chapter 3:

- **MedGemma-4B** [11]: A medical VLM with 4 billion parameters, based on Gemma 3 with a MedSigLIP vision encoder. It showed the highest flip rate among the MedGemma family (42.1% on the diagnostic set) and is the model targeted for mitigation in Thrust 3. Its high flip rate but potentially strong image grounding makes it a candidate for targeted improvement.
- **MedGemma-27B** [11]: The larger variant with 27 billion parameters. It achieved a lower flip rate (14.0%) on the diagnostic set, allowing us to test whether model scaling improves consistency through better visual grounding or through stronger text priors. Comparing the 4B and 27B variants isolates the effect of scale within a fixed architecture family.
- **LLaVA-Rad** [86]: A radiology-adapted VLM with 7 billion parameters, based on the LLaVA framework with BiomedCLIP as the vision encoder and Vicuna-7B as the language backbone. LLaVA-Rad was developed by Microsoft specifically for radiology report generation and showed

the lowest flip rate (6.5%) in our preliminary analysis. Its completely different architecture from MedGemma tests whether the robustness-grounding trade-off is architecture-specific or general.

All models are evaluated in deterministic mode (temperature 0, greedy decoding) to ensure reproducibility. No fine-tuning is performed in this diagnostic chapter; the goal is to characterize each model's behavior as deployed.

4.2.3 Evaluation Dataset

The diagnostic experiments use a three-backend diagnostic set of 107 semantically close question-paraphrase pairs (distinct from the 158-pair mechanistic FlipBank of Section 3.5; all three backends are evaluated on the identical 107 pairs). Each item contains an original question and a paraphrase with high semantic similarity (> 0.95 cosine similarity in BioClinicalBERT [103]), paired with a chest X-ray image. The questions are binary presence-style ("Is there [finding]?") where the model output can be parsed unambiguously as Yes or No. This curated set provides controlled conditions for the diagnostic experiments: we know the ground truth, the paraphrase relationship, and the flip status for each model. Because it is small and deliberately curated for these controlled conditions, its flip rates are not population-level estimates: MedGemma-4B flips on 42.1% of these 107 pairs, roughly five times its 8.3% headline flip rate on the full PSF-Med binary yes/no benchmark (Chapter 3). The 42.1% diagnostic-set figure should not be read as, or compared against, that benchmark number.

The diagnostic set is drawn from both MIMIC-CXR [38] and PadChest [101], though the primary analyses use the MIMIC-CXR subset where ground-truth labels are available from structured radiology reports. PadChest provides the bounding box annotations needed for the attention grounding analysis (RQ3).

4.2.4 Text-Only Evaluation Protocol

For RQ1, we evaluate each model under two conditions on the same question set:

1. **Real image condition**: The model receives the chest X-ray image and the clinical question. This is the standard evaluation setting.
2. **Text-only condition**: The model receives the same question with the image removed entirely, so no image tokens are supplied. This forces the model to answer based solely on the question text and any prior knowledge encoded in its language model weights.

We measure **text-only agreement**: the fraction of questions for which the model gives the same answer with and without the image; high text-only agreement indicates that the image contributes little to the model's decision. The answers compared are the binary yes/no readouts of the two conditions above. Formally:

$$\text{TextOnlyAgree}(M) = \frac{1}{|Q|} \sum_{q \in Q} \mathbf{1}[M(q, I) = M(q, \varnothing)] \tag{4.1}$$

where $M(q, I)$ is the model's answer to question q with image I, and $M(q, \varnothing)$ is the answer when no image is supplied.

The metric reads in one direction only. Every question on which the answer changes when the image is removed is a question the image demonstrably influenced, so $1 - \text{TextOnlyAgree}(M)$ is a *lower bound* on the fraction of predictions that depend on the image, and $\text{TextOnlyAgree}(M)$ is correspondingly an *upper bound* on the fraction that could be decided without it. The bound is not tight, and it is not an entailment. A model at 100% text-only agreement produced no answer that removal moved, which is compatible with the image contributing nothing, but equally compatible with a question text that already carries the answer, with a margin that shifted without crossing the yes/no threshold, and with a base rate that makes the text prior right on most items. A model at 50% text-only agreement (for binary questions with balanced labels) is shown to depend on the image on at least half its predictions; nothing follows about the other half, which may also have been informed by the image without the

argmax changing. Read as a bound, the metric still gives the intuition it is meant to give: neither extreme is desirable, and a useful diagnostic model should show substantial but not total text-only agreement, indicating that it draws on both text knowledge and visual evidence.

We also compute a complementary metric, the **image agnostic rate**: the fraction of questions where the model returns the same answer under all three image conditions, namely the real chest X-ray, a uniform gray placeholder image (all pixels set to RGB value 128, so the visual tokens are present but carry no diagnostic content), and a swapped image showing the opposite finding (introduced in the next subsection). Unlike text-only agreement, which removes the image entirely, this metric keeps the number of visual tokens fixed and varies only their content, isolating the model's sensitivity to *what* the image shows rather than to its mere presence.

4.2.5 Controlled Image Swap Protocol

For RQ2, we replace the real image with an **opposite-label image**: a different patient's chest X-ray that shows the opposite ground-truth label for the same clinical finding. If the original question asks "Is there pleural effusion?" and the real image is positive for pleural effusion, the swap image is selected from the dataset as a confirmed negative for pleural effusion.

We measure **swap sensitivity**: the fraction of questions where the model changes its binary answer when the image is swapped. Formally:

$$\text{SwapSens}(M) = \frac{1}{|Q|} \sum_{q \in Q} \mathbf{1}[M(q, I_{\text{real}}) \neq M(q, I_{\text{swap}})] \tag{4.2}$$

High swap sensitivity indicates that the model's answer depends on the specific image content. Low swap sensitivity indicates that the model gives the same answer for the real image and for one carrying the opposite finding, which is the signature of text-driven behavior.

The swap pool is constructed by matching each question to the nearest opposite-label image by finding type. We ensure that the swap image differs from the original only in the queried finding, not in overall image quality or patient demographics, to the extent possible. The swap pool size is reported per finding to identify potential selection bias. For findings with fewer than 5 opposite-label images available, we exclude the question from the swap analysis.

4.2.6 Attention-Bounding Box Overlap Protocol

For RQ3, we use PadChest's radiologist-annotated bounding boxes to test whether model attention targets the correct anatomical region. PadChest provides radiologist-annotated bounding boxes across multiple pathological findings. From the PadChest grounded reports we identify a large pool of question-image pairs that have at least one annotated box (16,345 flip and 33,177 no-flip pairs, 49,522 in total); the attention analysis below uses a balanced subsample of 200 of these pairs (100 flip, 100 no-flip). This 200-pair balanced subsample is the sample size for the flip versus no-flip comparison reported in this chapter (Table 4.2); the separate 637-box, all-positive PadChest grounding subset used for the controlled-baseline grounding analysis (spatial-shift and region-deletion controls) belongs to a different analysis reported in Chapter 6.

For each question with an associated bounding box, we extract the visual attention map from the decoder's self-attention, restricted to the entries where text-token queries attend to image-token keys (MedGemma is a decoder-only model that interleaves projected image tokens with text tokens in a single sequence and has no dedicated cross-attention modules). We compute three metrics:

- **Ground-truth (GT) coverage**: The fraction of attention weight that falls within the annotated bounding box. This measures whether the model attends to the pathology region.
- **Precision**: The fraction of the model's high-attention cells (grid cells whose attention exceeds a high-attention threshold) that fall inside the annotated bounding box. Whereas coverage is

Table 4.1: Backend comparison on Thrust 2 diagnostics. Flip rate measures disagreement between original and paraphrase on the real image. Pairwise accuracy requires the original and paraphrased answers to be correct. Text-only agreement and swap sensitivity quantify dependence on visual input. Higher text-only agreement and lower swap sensitivity indicate more text-driven behavior.

Metric	**MedGemma-4B**	**MedGemma-27B**	**LLaVA-Rad**
N	107	107	107
Accuracy Metrics			
Accuracy (Original)	61.7%	57.0%	47.7%
Accuracy (Paraphrase)	53.3%	50.5%	43.0%
Pairwise Accuracy	36.4%	46.7%	42.1%
Paraphrase Sensitivity			
Flip Rate	42.1%	14.0%	6.5%
Image Grounding			
Text-Only Agreement	66.4%	85.0%	96.3%
Swap Sensitivity	30.8%	19.6%	10.3%
Image Agnostic Rate	43.0%	60.7%	85.0%
Baselines			
Always Yes	43.9%	43.9%	43.9%
Always No	56.1%	56.1%	56.1%

normalized by the total attention mass over the whole image, precision is normalized by the model's own high-attention region, so it measures how much of where the model looks most strongly lands on the pathology.

- **Statistical comparison**: GT coverage for flip versus no-flip cases is compared with a two-sample t-test to test whether attention to the pathology region differs by flip status.

We compare these metrics between flip cases (questions where the model's answer changes under paraphrasing) and non-flip cases to test whether spatial grounding differs between consistent and inconsistent predictions.

4.3 Results: Text-Only Analysis

4.3.1 Main Results

Table 4.1 summarizes the diagnostic metrics across all three models. The results reveal a clear and striking pattern: as models become more stable under paraphrasing, they simultaneously become more text-driven and less grounded in the image.

MedGemma-4B shows the highest flip rate (42.1%) but also the highest swap sensitivity (30.8%) and the lowest text-only agreement (66.4%). Nearly a third of its predictions change when the image is replaced with one showing the opposite finding. This confirms that MedGemma-4B relies meaningfully on image content for its decisions, but this reliance is fragile. The same image-grounded representation that enables sensitivity to visual evidence also enables sensitivity to question phrasing. When the model attends to the image, it processes the image differently depending on how the question frames the clinical query.

MedGemma-27B shows reduced flip rate (14.0%) compared to the 4B variant, suggesting that model scaling helps with paraphrase stability. However, this improvement comes at a measurable cost to visual grounding. Text-only agreement increases to 85.0%, meaning 85% of the 27B model's answers can be predicted from the question text alone without the image. Swap sensitivity drops to

19.6%, down from 30.8% for the 4B variant. The larger model is more consistent, but it achieves this consistency partly by relying more heavily on text priors.

LLaVA-Rad presents the most extreme case. It shows the lowest flip rate (6.5%) but is almost entirely text-driven. Its answers match the text-only condition for 96.3% of questions, and the controlled swap changes its answer only 10.3% of the time. For practical purposes, on this diagnostic set LLaVA-Rad behaves close to a text-only model that happens to accept image input. Its apparent consistency under paraphrasing is not accounted for by stable visual grounding. The evidence for that is the image agnostic rate of 85.0%, which is the stronger of the two controls because it requires the answer to survive an opposite-label swap and a gray placeholder rather than only the removal of the image: for 85% of questions LLaVA-Rad returns the same answer whether or not the queried finding is present in the scan, which bounds at 15% the share of these decisions in which image content can be changing the outcome.

4.3.2 Interpreting the Baseline Rates

The "Always Yes" and "Always No" baseline rates in Table 4.1 provide context for interpreting model accuracy. In this dataset, the "Always No" baseline achieves 56.1% accuracy, reflecting the fact that most queried findings are absent in most images. A model that achieves accuracy near these baselines may be doing little more than predicting the majority class.

MedGemma-4B achieves 61.7% accuracy on original questions, only 5.6 percentage points above the "Always No" baseline. MedGemma-27B reaches 57.0%, barely above baseline. LLaVA-Rad at 47.7% actually falls below the "Always No" baseline. It answers "Yes" more often than the base rate would justify, indicating a positive bias in its text priors. These accuracy figures are modest, which is important context: these models are not achieving high accuracy and then failing on paraphrases; they are achieving marginal accuracy while simultaneously exhibiting high text reliance. This also bounds what the chapter claims. Because accuracy here sits within a few points of the majority-class rate, and below it for LLaVA-Rad, these backends are not competent on this curated set, so what follows is a statement about the relationship between consistency and image reliance and not a claim about diagnostic quality. The set was assembled from near-paraphrase pairs for diagnosis rather than sampled to estimate performance; benchmark accuracy on the full evaluation set is reported in Chapter 3.

4.3.3 Cross-Model Comparison of Text Reliance

The relationship between flip rate and text-only agreement follows a monotonic pattern across the three models:

- **MedGemma-4B**: 42.1% flip rate, 66.4% text-only agreement
- **MedGemma-27B**: 14.0% flip rate, 85.0% text-only agreement
- **LLaVA-Rad**: 6.5% flip rate, 96.3% text-only agreement

As flip rate decreases, text-only agreement increases. The Pearson correlation between these two metrics across the three models is approximately $r = -0.98$, nearly perfect. Three cautions apply. First, and most important, the three backends were deliberately chosen to span the consistency spectrum, so the coefficient is computed across models selected on one of the two variables being correlated; a near-perfect value is close to guaranteed by that design, and the number should be read as a description of the ordering these three models happen to realize rather than as an estimate of an effect size in any population. Second, this is a sample of only three models, so the correlation alone does not establish a general law. Third, the flip rates here are computed on the 107-pair diagnostic set, which was constructed before the equivalence audit and therefore still contains operator-changing pairs, so the individual flip labels are not clean. The trade-off is quantified properly on the audited benchmark and the four-quadrant analysis of Chapter 6; here the three-model pattern is illustrative and consistent with the hypothesis that current medical VLMs achieve paraphrase consistency primarily through text

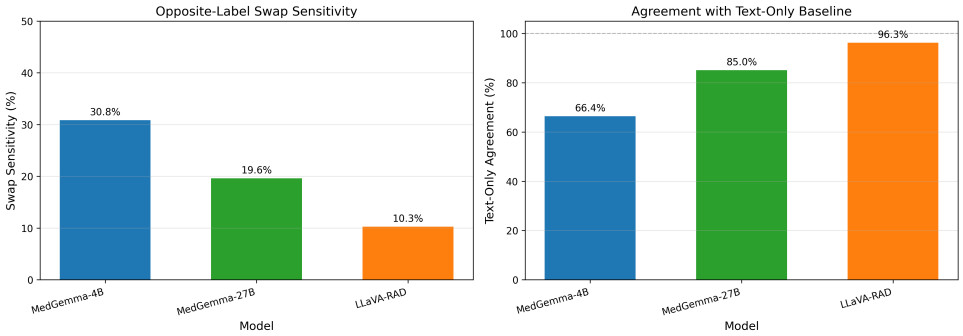


Figure 4.2: Controlled opposite-label swap sensitivity. Higher values indicate stronger dependence on visual evidence. MedGemma-4B shows the highest sensitivity, confirming it relies on the image despite its high flip rate. LLaVA-Rad shows minimal sensitivity, consistent with near-complete text reliance.

priors rather than through stable visual representations.

This pattern has been confirmed across a larger set of models and datasets in the Thrust 4 evaluation (Chapter 6), where a mean of 81% of each model's consistent predictions turn out to be image-invariant regardless of its flip rate (the $r = -0.86$ correlation between flip rate and the image-invariant fraction that one might report from these data is largely definitional and is discussed there).

4.4 Results: Image Swap Sensitivity

4.4.1 Controlled Swap Results

Figure 4.2 shows the swap sensitivity for each model. When the image is replaced with one showing the opposite finding, MedGemma-4B changes its answer 30.8% of the time, MedGemma-27B changes 19.6% of the time, and LLaVA-Rad changes only 10.3% of the time.

The swap sensitivity numbers should be interpreted in the context of the experimental design. A perfect model, one that always answers based on the image content, would show 100% swap sensitivity when the swapped image has the opposite label. The fact that even MedGemma-4B achieves only 30.8% indicates that all three models rely substantially on text priors. The difference is one of degree: the swap moves MedGemma-4B's answer on roughly a third of decisions and LLaVA-Rad's on one in ten, so image content is demonstrably doing work at least that often in each case. Because the swap can only reveal image reliance that changes the argmax, these are lower bounds on how often the image is consulted, not counts of the predictions that use it.

4.4.2 Swap Sensitivity by Clinical Finding

Swap sensitivity varies across clinical findings, reflecting differences in how models process different pathologies. Restricting to findings with at least five questions in the diagnostic set, swap sensitivity for MedGemma-4B is highest for consolidation (62.5%, $n = 8$), opacity (45.5%, $n = 11$), and pleural effusion (40.0%, $n = 15$), and lowest for atelectasis (12.9%, $n = 31$). Pneumothorax and pleural thickening sit in between (both 25.0%, $n = 8$ and $n = 12$). Findings with fewer than five questions (for example cardiomegaly, $n = 3$) are excluded because their rates are too noisy to interpret.

This finding-level variation suggests that visual grounding is not a binary property but a spectrum: the model may ground its answers in the image for some findings and rely on text priors for others. The findings with the highest swap sensitivity (consolidation, opacity, pleural effusion) all produce prominent, spatially extended radiographic changes, while the lowest swap sensitivity occurs for atelectasis, a subtler and often regional finding. Per-finding counts in the diagnostic set are small, so these rates should be read as suggestive rather than precise.

Table 4.2: Attention-bounding box analysis on PadChest ($N = 200$ samples with ground-truth boxes). Flip cases show substantially lower attention to pathology regions than non-flip cases, suggesting that spatial grounding failures contribute to paraphrase sensitivity.

Metric	Flip Cases	No-Flip Cases
GT Coverage	8.4%	14.2%
Precision	10.6%	29.0%

4.4.3 The Image Agnostic Rate

Recall the **image agnostic rate** defined in Section 4.2: the proportion of questions where a model returns the same answer under all three image conditions, the real image, the uniform gray placeholder, and the opposite-label swap. These are predictions where neither stripping the diagnostic content (the gray placeholder) nor replacing it with the opposite finding (the swap) moves the answer across the yes/no boundary. The image agnostic rates are 43.0% for MedGemma-4B, 60.7% for MedGemma-27B, and 85.0% for LLaVA-Rad. For LLaVA-Rad, 85% of its predictions are identical whether the patient has the queried finding or not. The clinical import is that for these predictions the yes/no output does not distinguish a patient who has the finding from one who does not, whatever the model computed internally.

4.5 Results: Attention Grounding Analysis

4.5.1 Attention-Pathology Overlap

Using the balanced 200-pair subsample introduced in Section 4.2 (100 flip and 100 no-flip pairs, drawn from the 49,522-pair annotated pool), we measure whether model attention targets the correct anatomical region. For each question with an associated bounding box, we extract the visual attention map and compute the ground-truth (GT) coverage: the fraction of total attention weight that falls within the annotated pathology region.

Table 4.2 presents the results for MedGemma-4B on a subset of 200 PadChest questions. Two findings stand out.

First, attention to pathology regions is low overall. Even for non-flip cases (where the model answers consistently), GT coverage is only 14.2%. The model directs roughly one-seventh of its attention to the correct region, distributing the remainder across the entire image. Precision (the concentration of within-box attention) is 29.0% for non-flip cases, meaning that when the model does attend to the pathology region, its attention is not tightly concentrated.

Second, flip cases show 41% lower GT coverage than non-flip cases (8.4% versus 14.2%, $p < 10^{-8}$, two-tailed t-test, $t = -6.13$). When the model flips its answer under paraphrasing, it was attending less to the relevant anatomical region. This correlation between spatial grounding and paraphrase consistency suggests that both phenomena share a common cause: when the model's visual processing of the relevant region is weak, it is more vulnerable to perturbation from the text side. One caveat qualifies the strength of this claim: the flip/non-flip split uses paraphrase pairs from before the equivalence audit, so some "flip" cases are operator changes rather than genuine paraphrases, and the attention gap may be partly confounded by paraphrase category. The direction of the effect is stable, but its magnitude should be read as suggestive rather than a clean paraphrase-specific estimate.

Even for non-flip cases, the GT coverage of 14.2% is low in absolute terms: the model directs only about one-seventh of its attention mass into the annotated pathology region. The gap between flip (8.4%) and non-flip (14.2%) coverage is the load-bearing comparison, indicating that flip cases attend even less to the pathology region. We did not run displaced-box or random-location control baselines for this $N = 200$ subset, so the absolute coverage values should be read as relative (flip versus no-flip)

rather than against a null spatial baseline.

4.5.2 Limitations of Attention-Based Grounding Measures

Several caveats apply to the attention grounding analysis. These limitations are important for interpreting the results and for understanding what the analysis can and cannot tell us about model behavior.

First, attention weights may not faithfully reflect model reasoning [106, 107]. The debate over attention interpretability in NLP extends to the vision-language setting. A model may process the pathology region through mechanisms not captured by attention maps, such as distributed representations in the multi-layer perceptron (MLP) layers or indirect propagation through multiple attention heads. Low attention overlap does not necessarily mean the model ignores the pathology; it may attend to it through a channel we are not measuring.

Second, the sample size of 200 is sufficient to detect the observed effect ($p < 0.05$) but limits generalization. A larger sample with more diverse findings and patient demographics would provide stronger evidence.

Third, bounding box annotations may not capture all relevant image regions. Radiologists annotate the primary pathology, but the model may attend to secondary signs (e.g., a lateral shift of the mediastinum as evidence of pleural effusion, rather than the effusion itself). Attention to diagnostically relevant but un-annotated regions would register as low GT coverage despite meaningful visual grounding.

Fourth, the analysis uses attention at a single layer. Different layers may attend to different image regions, and aggregating across layers could produce different coverage statistics. We use the final decoder layer's text-to-image self-attention, which is the most directly relevant to the model's output, but this is a design choice that affects the results.

Despite these limitations, the attention analysis provides a useful signal when combined with the text-only and swap experiments. No single metric is definitive, but the three signals point the same way for the models tested: high text-only agreement, low swap sensitivity, and (for MedGemma-4B, the only model measured here for attention) low attention overlap with the annotated pathology region. The swap result is the one that carries the most weight, because it varies image content rather than removing the image, and it is what licenses the statement that image content contributes little to these answers.

4.6 The Robustness-Grounding Trade-off

4.6.1 Quantifying the Trade-off

Figure 4.3 visualizes the central finding of this chapter. Each model is plotted with flip rate on the horizontal axis and text-only agreement on the vertical axis. The three models trace a clear trajectory from the upper-left (LLaVA-Rad: low flips, high text reliance) to the lower-right (MedGemma-4B: high flips, lower text reliance). MedGemma-27B occupies the middle.

This trade-off has a simple intuitive explanation. A model's answer depends on two sources of information: the image and the text. If the model weighs text heavily, its answer is determined primarily by the question phrasing and clinical priors. Since the question's clinical meaning is preserved under paraphrasing (by construction), the text-driven answer is stable. If the model weighs the image heavily, its answer integrates visual evidence that may be ambiguous or borderline. The same ambiguous image evidence can tip the model's decision differently depending on how the question frames the clinical query, leading to flips.

Visual grounding tends to create a vulnerability to text perturbation. A model that genuinely processes the image maintains a representation that can be sensitive to input framing, whereas a

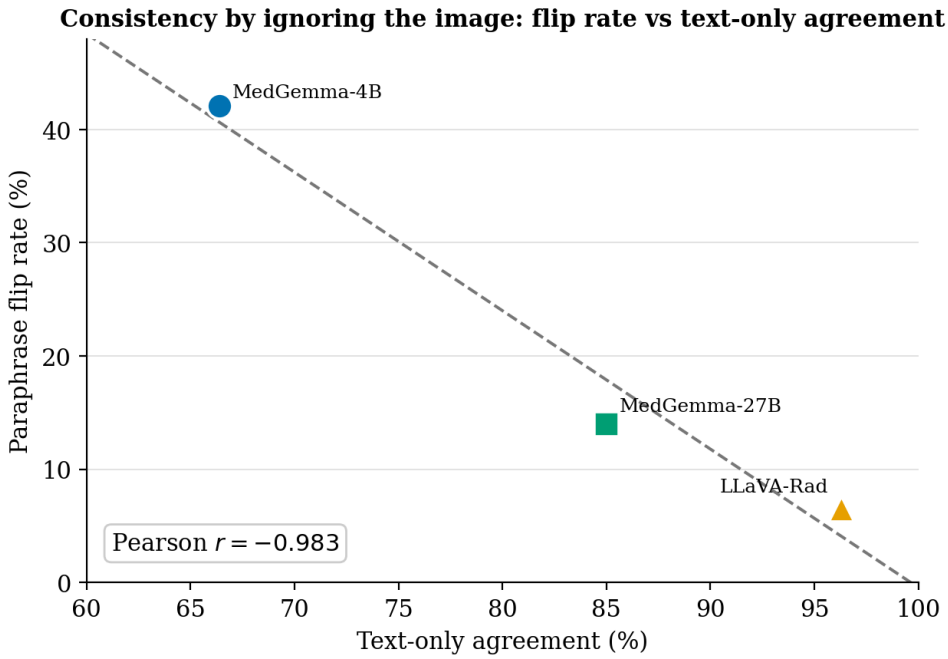


Figure 4.3: Flip rate versus text-only agreement for the three backends. A lower flip rate does not necessarily indicate better visual grounding. The strong negative correlation (approximately $r = -0.98$ across these three backends) illustrates the tendency for a lower flip rate to coincide with weaker image dependence; it is consistent with, but does not on its own establish, a general trade-off.

model that leans on the text prior has less such vulnerability. This mechanism helps explain why lower flip rates often coincide with weaker grounding among the models studied, but it is a tendency, not a deterministic law: a model can in principle be both consistent and grounded (the Ideal quadrant below).

4.6.2 The Four-Quadrant Framework

The trade-off motivates a classification framework that we develop fully in Chapter 6 but preview here. Each prediction can be classified along two axes:

1. **Consistency**: Does the model give the same answer to all paraphrases? (Consistent vs. Inconsistent)
2. **Image reliance**: Does the model's answer change when the image is removed or swapped? (Image-reliant vs. Text-reliant)

   This creates four quadrants:

The Dangerous quadrant is the most concerning from a clinical perspective. A model in this quadrant appears safe by every conventional metric: it gives the same answer every time the question is asked, and it does so with high confidence. But the answer is the same one it would have given with the image removed or replaced by its opposite, so the confidence it projects is not backed by the scan. If the image showed a clear pneumothorax and the model answers "No" because of text priors, the consistency of that answer makes the error harder to detect. An inconsistent model at least signals its unreliability through disagreement across paraphrases.

From the diagnostic data in this chapter, LLaVA-Rad places the vast majority of its predictions in the Dangerous quadrant: its answers are both highly consistent and almost entirely unchanged when the image is removed. MedGemma-4B distributes predictions across Fragile (image-reliant but inconsistent) and other quadrants. MedGemma-27B is intermediate. Chapter 6 quantifies the quadrant shares across ten model-dataset combinations.

Table 4.3: Four-quadrant classification of model predictions based on consistency and image reliance. The Ideal quadrant (consistent and image-reliant) is the target; the Dangerous quadrant (consistent but text-reliant) is the most concerning failure mode.

	Consistent	**Inconsistent**
Image-reliant	**Ideal**: Reliable and evidence-based. The model gives the same answer to all phrasings, and that answer depends on the image.	**Fragile**: Uses evidence but is wording-sensitive. The model processes the image but is unstable under rephrasing.
Text-reliant	**Dangerous**: Looks reliable, but the answer does not track the evidence. The model gives the same answer across phrasings, and that answer is unchanged when the image is removed or swapped.	**Worst**: Neither reliable nor evidence-tracking. The answer is unchanged when the image is removed or swapped, yet it still varies across phrasings.

4.6.3 Implications for Model Evaluation

The robustness-grounding trade-off has immediate implications for how medical VLMs should be evaluated. Current evaluation practices typically report accuracy on fixed question sets and sometimes add consistency metrics like flip rate. Our findings show that both metrics are insufficient:

- **Accuracy alone is insufficient**: A model can achieve accuracy at or above baseline by relying on text priors. LLaVA-Rad's accuracy does not demonstrate visual diagnostic capability.
- **Flip rate alone is misleading**: A model with low flip rate may be achieving that consistency through text-driven behavior, not through reliable visual processing. LLaVA-Rad's 6.5% flip rate is the best of the three models, but it is the worst in terms of visual grounding.
- **Joint evaluation is necessary**: Any claim of paraphrase consistency must be accompanied by evidence of visual grounding. We recommend reporting (1) accuracy, (2) flip rate, (3) text-only agreement, and (4) swap sensitivity as a minimum evaluation set for medical VLMs.

The PSF-Med benchmark (Chapter 3) provides infrastructure for all four metrics. The text-only and swap experiments require no additional data beyond the original images and a simple swap pool.

4.7 Analysis by Paraphrase Phenomenon

4.7.1 Per-Phenomenon Flip Rates

Table 4.4 breaks down flip rates by paraphrase transformation type for all three models. The patterns both confirm and extend the findings from Chapter 3.

For all three models, negation-related paraphrases produce the highest flip rates. Even MedGemma-27B and LLaVA-Rad, which achieve low overall flip rates, flip on 60% and 20% of negation paraphrases respectively. This category must be read with care, because it is confounded by operator changes. "Is there pneumothorax?" and "Can pneumothorax be ruled out?" do *not* have identical clinical meaning: on a positive image the first is correctly answered Yes and the second No, so a change in the yes/no output is correct behavior, not a failure. A share of the high negation flip rate therefore reflects the model correctly handling operator changes rather than genuine paraphrase sensitivity, which is exactly why the equivalence audit of Chapter 3 rejected 93.8% of generated negation paraphrases as polarity- or operator-changing. The genuine vulnerability is confined to same-polarity negation reformulations that preserve the answer; because this 107-pair diagnostic set predates that audit, its negation flip

Table 4.4: Flip rate by paraphrase phenomenon across the three diagnostic models (pre-audit pairs). The negation category ($n = 5$) shows the highest rates but is confounded by operator changes (see text); MedGemma-27B shows the largest scaling gain on syntactic restructuring.

Phenomenon	N	**MedGemma-4B**	**MedGemma-27B**	**LLaVA-Rad**
Negation	5	60.0%	60.0%	20.0%
Syntactic Restr.	12	50.0%	0.0%	0.0%
Lexical Subst.	40	50.0%	17.5%	10.0%
Specificity Mod.	25	48.0%	12.0%	4.0%
Scope/Quant.	25	16.0%	8.0%	4.0%

labels are not clean, and the effect is quantified properly only on the audited benchmark.

MedGemma-27B shows the largest scaling benefit on syntactic restructuring (50.0% $\rightarrow$ 0.0%) and scope/quantification (16.0% $\rightarrow$ 8.0%). These are phenomena where the 27B model's larger language model backbone provides better syntactic processing. The negation column shows no change (3 of 5 pairs for both 4B and 27B), but given the operator confound above and $n = 5$, this set cannot establish whether scaling helps on genuine same-polarity negation; the audited benchmark of Chapter 3 is the right instrument for that question.

LLaVA-Rad's low flip rates across all phenomena (except negation at 20%) are consistent with its text-driven behavior. When a model's answers move little with the image, the text processing pathway is the main remaining source of flips. LLaVA-Rad's text pathway handles most paraphrase types reliably, but negation still causes flips because negation changes the text processing path itself, not just the relationship between text and image.

4.8 Discussion

4.8.1 What Drives Paraphrase Sensitivity?

The diagnostic experiments point to two distinct mechanisms that produce paraphrase sensitivity, depending on the model.

For models that use the image (MedGemma-4B), paraphrase sensitivity arises from the interaction between ambiguous visual evidence and text-side framing. The model processes the image and forms a representation of the visual evidence. When this evidence is ambiguous (borderline findings, subtle pathology), the model's decision is close to its threshold. Different question phrasings shift the threshold slightly, pushing some borderline cases across the decision boundary. The model is not failing at image analysis; it is failing at reading a fixed visual representation the same way when the question changes. The controlled probe of Chapter 5 shows that ambiguous visual evidence is not a necessary ingredient: with the encoder frozen so that every paraphrase reads identical features, and the answer prior removed by balancing, the flip rate still ranges from 4.8% to 88.4% on the training-phrasing distribution alone.

For models whose answers move little with image content (LLaVA-Rad, at 85.0% image agnostic under the swap and placeholder controls), paraphrase sensitivity arises predominantly from the text pathway. The model applies language priors to the question text and produces an answer based on patterns in its training data. When the paraphrase changes the text sufficiently (as with negation), the text pathway itself produces a different answer. Image content does not measurably contribute to these flips: the controlled swap moves LLaVA-Rad's answer on only 10.3% of questions and 85.0% of its predictions are image agnostic (Section 4.4).

These two mechanisms have very different implications for mitigation. For MedGemma-4B, the goal is to stabilize the decision threshold without removing image dependence. Chapter 5 develops a targeted intervention (LoRA on layers 15 to 19) aimed at exactly this; as that chapter's clean re-audit

shows, the intervention improves consistency but partly at the expense of visual grounding, so the goal is only partly achieved. For LLaVA-Rad, a consistency intervention has little left to act on, because the model already achieves near-perfect consistency through text priors and reducing its flip rate further would not increase its 10.3% swap sensitivity. Improving LLaVA-Rad would require a very different approach: forcing the model to attend to image content, not making its text processing more stable.

4.8.2 The Scaling Trade-off

The comparison between MedGemma-4B and MedGemma-27B reveals a scaling trade-off that parallels the robustness-grounding trade-off. Scaling from 4B to 27B parameters improves paraphrase consistency (14.0% vs. 42.1% flip rate) but reduces visual grounding (85.0% vs. 66.4% text-only agreement). The larger model has a stronger language model that relies more heavily on text priors.

This suggests that simply scaling language models may not be the path to achieving both consistency and visual understanding in medical VLMs. The scaling trajectory, if extended, leads toward models that are very consistent but mostly text-driven, approaching LLaVA-Rad's profile at the limit. This is the opposite of what clinical deployment requires.

The implication is that mitigation should target the 4B-scale model, which has genuine image grounding that can be stabilized, rather than the 27B model, which has already sacrificed substantial grounding for consistency. Chapter 5 follows this recommendation by developing the LoRA intervention for MedGemma-4B.

4.8.3 Cross-Family Generalization of the Diagnosis

Later deployment-audit results show that the central diagnosis of this chapter generalizes beyond the original three-model comparison, but not in a uniform way. The high-level pattern is stable: lower observed flip rates often coincide with weaker image dependence and stronger text-dominant admission. What changes across families is *where* that failure appears.

Gemma-family models exhibit a late readout-misalignment signature: internal transport representations remain stable across paraphrases while the final readout state diverges. LLaVA-Rad shows a broader instability pattern in which both transport and readout can shift, consistent with a more diffuse failure than the Gemma case. Qwen2-VL is more opaque still: behavioral evidence shows strong text-dominant selection, but cosine-based internal probes are near chance. The diagnosis therefore generalizes at the behavioral level, not as a single shared internal mechanism.

Thrust 2 establishes that apparent robustness can arise from weak grounding; the later cross-family audit shows that this is not a quirk of one benchmark or one backbone. At the same time, it warns against over-generalizing from the Gemma mechanistic story. A single model family can support mechanistic diagnosis, but safe deployment across families requires behavioral audits that verify what kinds of cases each architecture actually solves.

4.8.4 Implications for Mitigation Design

The diagnostic findings impose constraints on what a good mitigation looks like:

1. **Preserve image reliance**: Any consistency improvement must maintain or increase swap sensitivity and decrease or maintain text-only agreement. An intervention that reduces flip rate while increasing text-only agreement has achieved consistency through the wrong mechanism.
2. **Target the right model**: Models that are already text-driven (LLaVA-Rad) have little to gain from a consistency intervention, because their flip rate is already near zero and reducing it further does not act on image reliance. Mitigation should target models with genuine but fragile visual grounding (MedGemma-4B).
3. **Address same-polarity negation reformulations**: the genuine negation vulnerability identified in Section 4.7 is confined to answer-preserving reformulations; any effective mitigation should

demonstrate improvement on this category specifically, evaluated on audited answer-preserving pairs.

4. **Test out-of-distribution**: The PadChest results from Chapter 3 show that flip rates are generally higher under distribution shift. Any mitigation must be evaluated on out-of-distribution data to confirm generalization.

These constraints guide the design of the Thrust 3 intervention (Chapter 5), where we develop a Targeted LoRA that reduces flip rate; the clean patient-disjoint re-audit (Section 5.7.5) shows it does not preserve image reliance, which is why any deployment must pair it with the image-reliance checks above.

4.8.5 Mild-Contrast Stability Test

As an illustrative validation experiment, we test whether minimal image perturbations affect model behavior. We apply small pixel-level edits to each image: JPEG recompression at quality 95 and a 5% global contrast adjustment. These edits are intended to be visually near-imperceptible and to carry no additional clinical information.

Under these perturbations, answer-change rates stay low across the base MedGemma-4B model and its two LoRA variants (Chapter 5), each evaluated on a 500-question PadChest sample: the answer changes under either edit on 1.2% of questions for the base model (0.6% under the contrast edit alone), 2.0% for Targeted LoRA, and 1.4% for Full LoRA. These rates are more than an order of magnitude below the base model's own paraphrase flip rate (42.1% on the diagnostic set), indicating that answer changes under paraphrasing are driven by text-side processing rather than by marginal differences in image encoding. The model's visual processing is stable under small pixel-level perturbations even when its text processing is unstable under semantic-preserving rephrasing.

This finding complements the causal hierarchy from the main experiments. The model's internal representation of the image is stable under low-level perturbation but changes when the text query frames the clinical question differently. The instability lies in the text-image integration pathway rather than either modality's representation alone. This result guides the mechanistic analysis in Chapter 5, where Feature 3818 is identified as a text-side feature that modulates how the model interprets visual evidence.

4.8.6 Relationship to General-Domain Findings

The robustness-grounding trade-off we identify is consistent with broader findings in the AI safety literature. Geirhos et al. [104] describe "shortcut learning" as a general phenomenon where neural networks exploit spurious correlations in training data. In our setting, the "shortcut" is the text prior: models learn that certain question patterns correlate with certain answers, and this correlation is a reliable (if clinically meaningless) feature for prediction. The text prior is not a bug in the training data but a genuine statistical regularity (certain findings are more common than others) that the model exploits.

The medical setting makes the shortcut particularly pernicious. In general-domain VQA, a model that answers from text priors rather than image content achieves degraded but often usable performance. In medical imaging, a model whose answers do not track the patient's image is unfit for purpose. The entire value of a medical VLM lies in its ability to analyze the specific patient's image. A model that achieves high consistency without its answers tracking the image has achieved the wrong objective.

Kim et al. [109] and Park et al. [110] developed approaches for generating natural language explanations of vision-language model decisions. In principle, such explanations could reveal whether a model's answer is based on visual evidence or text priors. However, explanations generated by text-reliant models may themselves be text-driven; the model could generate a plausible explanation that

Table 4.5: Contrasts between MedGemma-4B and LLaVA-Rad on the 107-pair diagnostic set. Brackets give 95% Wilson score intervals. The p-values come from two-sided Fisher exact tests on the per-backend counts, computed without pairing; because both backends score the identical 107 items, a paired McNemar test would be more powerful still.

Metric	MedGemma-4B	LLaVA-Rad	Fisher p
Flip rate	42.1% [33.1, 51.5]	6.5% [3.2, 12.9]	8.6×10^{-10}
Text-only agreement	66.4% [57.0, 74.6]	96.3% [90.8, 98.5]	1.1×10^{-8}
Image agnostic rate	43.0% [34.0, 52.5]	85.0% [77.1, 90.6]	1.5×10^{-10}
Swap sensitivity	30.8% [22.9, 40.1]	10.3% [5.8, 17.5]	3.1×10^{-4}

references image features it never actually processed. This creates a circularity that attention-based and causal methods attempt to break.

4.8.7 Statistical Precision of the Diagnostic Set

A natural objection to the analysis above is the size of the diagnostic set. Section 4.2 already restricts what its rates mean: the set is small and deliberately curated, and its flip rates are not population-level estimates. The curation criteria are strict. Every item pairs an original question with a paraphrase at BioClinicalBERT cosine similarity above 0.95, parses unambiguously as yes or no under all three backends, has known ground truth, and has at least five opposite-label swap candidates; each item is then scored under four conditions (real image, text-only, gray placeholder, opposite-label swap) for each of the three backends. The set functions as a diagnostic instrument rather than a survey sample, trading breadth for internal validity in the way a controlled experiment does. Population estimation belongs to Chapter 3 and its audited benchmark sets. Generalization across models and datasets belongs to Chapter 6, where a mean of 81% of consistent predictions are image-invariant across ten model-dataset settings. The causal account belongs to Chapter 5. What this chapter must establish is narrower: that consistency and image reliance can come apart, and that the controls above detect the separation when it occurs.

The contrasts the chapter relies on are statistically decisive at this sample size. Table 4.5 reports 95% Wilson score intervals and two-sided Fisher exact p-values for the four headline metrics, contrasting the two backends at the ends of the consistency-grounding spectrum, MedGemma-4B and LLaVA-Rad. The claimed effects span 20 to 42 percentage points, while $n = 107$ gives roughly ± 9 points of precision at these proportions, so each contrast is resolved well beyond chance. The Fisher tests are conservative in a specific way: they treat the two backends as independent samples even though both score the identical 107 items, and the paired alternative (McNemar's test) would only be more powerful. The interval method does not drive the conclusion either: nonparametric bootstrap intervals for all four metrics and all three backends, computed independently of the Wilson figures, agree with them (`thrust2/finalize/out/bootstrap_table.csv`).

4.8.8 Summary

The diagnostic findings here motivate the Targeted LoRA intervention of Thrust 3 and the safety screen of Thrust 4. MedGemma-4B has genuine but fragile visual grounding, and the dissociation between consistency and grounding (developed in Chapter 6, where a mean of 81% of consistent predictions are image-invariant) means a low flip rate does not certify grounding: models that achieve consistency frequently do so at the expense of image reliance, producing predictions that look reliable but do not change when the image is removed or swapped.

Chapter 5

Thrust 3: Mechanistic Diagnosis and Targeted Mitigation

Research questions. This chapter answers **RQ3.1** (whether sparse-autoencoder features can identify an interpretable mechanism underlying paraphrase sensitivity), **RQ3.2** (whether a targeted LoRA can reduce flip rates while preserving accuracy), and **RQ3.3** (whether the best intervention site matches the mechanistic diagnosis).

Main contribution. Two independent localizations of the paraphrase-flip commit at layers 16 to 17, a controlled probe that identifies the training-phrasing distribution as a cause of the sensitivity, and a targeted LoRA on layers 15 to 19 that reduces the flip rate from 8.5% to 3.5% (about 59%) on patients and images never seen in training while modifying 0.1% of parameters.

Chapter 4 reported evidence that paraphrase sensitivity in MedGemma-4B tracks fragile visual grounding more closely than reliance on the question text alone. That reading rests on the 107-pair three-backend diagnostic set, on which MedGemma-4B has the lowest text-only agreement (66.4%) and the highest image-swap sensitivity (30.8%) of the three backends; the set is small and predates the equivalence audit, so its per-pair flip labels are not clean and the reading is an indication rather than a settled result. On that evidence the image contributes materially to the model's decision on this backend, and the answer nonetheless changes when the question is rephrased. This chapter asks two questions: *why* does rephrasing change the answer, and *can we fix it without sacrificing image reliance*?

We answer the first question with Sparse Autoencoders (SAEs). We transfer pretrained SAEs from Gemma Scope 2 [111] to MedGemma-4B and identify a specific sparse feature, Feature 3818 at layer 17, that tracks question register: whether a question uses presence framing ("Is there pneumothorax?") or exclusion framing ("Can you rule out pneumothorax?"). Activation patching confirms this feature causally influences the model's yes/no margin. A distributional analysis across 20 FlipBank pairs shows Feature 3818 is the top contributor in roughly 25% of flips, with the rest distributed across many features. This FlipBank includes operator-changing pairs, however, so part of that association reflects the feature correctly distinguishing presence from exclusion framing rather than same-polarity paraphrase sensitivity; a stricter operator-preserving re-analysis (Section 5.3.1) finds Feature 3818 prominent but not a sufficient cause, so we present the 3818 $\rightarrow$ 12139 account as an exploratory hypothesis pending matched same-polarity controls. Either way the distributed remainder is consistent with superposition [112] and motivates an intervention that acts on a layer range rather than a single feature.

We answer the second question with targeted Low-Rank Adaptation (LoRA) [113]. We train LoRA adapters on layers 15 to 19 using a combined loss that balances paraphrase consistency (symmetric

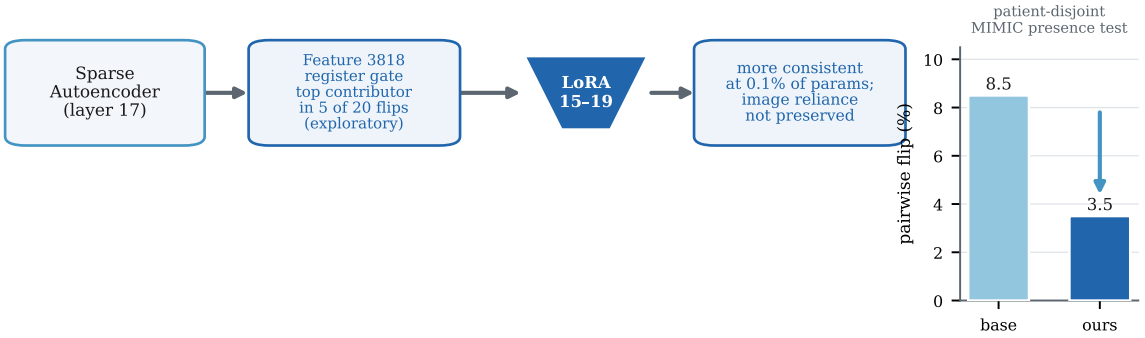


Figure 5.1: Thrust 3 at a glance. A sparse autoencoder localizes a clinical-query register gate (Feature 3818) at layer 17 associated with a substantial share of paraphrase flips (a contributing, not sole, factor; an exploratory account pending same-polarity controls). A targeted Low-Rank Adaptation on the layers around it, trained on operator-preserving paraphrases, cuts the presence-of-finding flip rate on a patient- and image-disjoint MIMIC-CXR held-out test by about 59% (pairwise 8.5% to 3.5% $\pm$ 0.6, standard deviation across five training seeds; paired McNemar $p < 0.002$ in every seed) while training 0.1% of the parameters, though a clean re-audit shows it does not preserve image dependence.

KL divergence) with answer accuracy (cross-entropy). Trained on operator-preserving paraphrases (full expanded pool, five seeds), this reduces the presence-of-finding flip rate on a patient- and image-disjoint MIMIC-CXR held-out test (pairwise 8.5% to 3.5% $\pm$ 0.6, about 59%, where the $\pm$ is a standard deviation across five training seeds; paired McNemar $p < 0.002$ in every seed) with no observed accuracy reduction (95% interval across those same five seeds, $[-1.00, +1.51]$ percentage points), while modifying only 0.1% of model parameters; the test shares no patient and no image with training. On PadChest, the out-of-distribution test set, accuracy increases by 6.4 percentage points (85.2% to 91.6%) and the margin difference narrows by 75%, although the flip rate on this balanced set is already low for the base model and changes little. The combined loss is essential: training on consistency alone causes mode collapse, where the model predicts the same answer for every question and reaches 46.4% accuracy.[1]

The answer to the second question is no, and that answer is stated here rather than left to the audit section that establishes it. A clean re-audit of these same patient-disjoint adapters (Section 5.7.5) finds that targeted adaptation does not preserve image dependence. Text-only agreement, the fraction of questions for which the model gives the same answer with and without the image, rises from 53.5% for the base model to about 77% for the targeted adapter and to about 77% for the full adapter; high text-only agreement indicates that the image contributes little to the model's decision. The two adapters are statistically indistinguishable on that measure and on image-swap sensitivity, so the earlier claim that targeted adaptation preserves grounding relative to full adaptation is not supported. Both adapters therefore buy part of their consistency from the question text. The targeted adapter remains the better choice on accuracy and on parameter cost, not on grounding, and this chapter's central

[1] Flip rates in this dissertation are reported on several distinct evaluation sets and are not directly comparable across thrusts: the equivalence-filtered PSF-Med benchmark (MIMIC n=3,266, PadChest second iteration n=14,935; Chapter 3); the patient- and image-disjoint MIMIC-CXR binary test set (n=373 questions, with the presence-of-finding primary endpoint at n=238) and the separate 158-pair mechanistic FlipBank (Chapter 5); the 107-pair three-backend diagnostic set (Chapter 4); and the curated behavioural flip banks used for the four-quadrant analysis (MIMIC n=98, PadChest n=861, with LLaVA-Rad reported on its n=732 text-only-baseline subset; Chapter 6). Each reported flip rate is qualified by its set.

mitigation result has to be read with that bound attached.

A layer ablation study reveals that early layers (0 to 10) reduce margin differences more than the mechanistically-identified middle layers (15 to 19). The SAE analysis explains *where sensitivity manifests* (layer 17), but the best intervention acts upstream, before the register-sensitive representation forms. We present the mechanistic analysis as a case study demonstrating SAE methodology for medical VLMs, separate from the empirical contribution of the combined loss approach.

5.1 Background: Sparse Autoencoders for Interpretability

Transformer models represent concepts as superposed combinations of directions in activation space [112, 114]. A Sparse Autoencoder decomposes each activation vector $\mathbf{h} \in \mathbb{R}^d$ into a sparse set of interpretable features:

$$\mathbf{f} = \text{ReLU}(\mathbf{W}_{\text{enc}}\mathbf{h} + \mathbf{b}), \quad \hat{\mathbf{h}} = \mathbf{W}_{\text{dec}}\mathbf{f} \tag{5.1}$$

where $\mathbf{f} \in \mathbb{R}^D$ is the feature activation vector ($D \gg d$), and reconstruction quality is measured by Fraction of Variance Unexplained (FVU):

$$\text{FVU} = \frac{\|\mathbf{h} - \hat{\mathbf{h}}\|^2}{\text{Var}(\mathbf{h}) \cdot d} \tag{5.2}$$

The SAE architecture uses JumpReLU activations, which produce strictly sparse representations by zeroing activations below a learned threshold. This differs from standard ReLU SAEs in that the sparsity level (L0) is explicitly controlled during training rather than emerging as a side effect of the sparsity penalty. The 16k-feature SAE used throughout this chapter produces $L0 = 70 \pm 10$ active features per input token. This extreme sparsity (more than 99% of features inactive) is what makes individual features interpretable: each active feature represents a specific, identifiable concept in the model's representation.

Gemma Scope 2 [111] provides pretrained SAEs for every layer of the Gemma 3 4B language model. These SAEs were trained on general web text, not medical data. For them to be useful on MedGemma-4B, which shares the Gemma 3 language backbone but was fine-tuned on medical data, we need to validate that the learned features still decompose MedGemma's activations faithfully. The key question is whether medical fine-tuning changes the geometry of the residual stream enough to invalidate the feature directions learned from web text. If the SAE's decoder directions no longer align with meaningful directions in MedGemma's activation space, the features would be uninterpretable.

5.2 SAE Transfer Validation

We validate SAE transfer by comparing reconstruction quality on 100 medical prompts (clinical questions about chest X-rays with images) and 100 general prompts (web-style text without images). For each prompt, we extract the residual stream at layer 17 and pass it through the Gemma Scope 2 SAE (16k features, medium L0 target).

Table 5.1: SAE transfer validation ($n = 100$ prompts per domain). Gemma Scope 2 SAEs (16k features, layer 17) reconstruct MedGemma-4B activations with >99.7% variance explained on both medical and general inputs ($H_0 : R^2 \leq 0.95$, $p < 0.001$).

Metric	**Medical**	**General**
R^2	0.9972 ± 0.0005	0.9974 ± 0.0006
FVU	0.28%	0.26%
L_0 (active features)	70 ± 10	76 ± 10

The reconstruction is nearly identical across domains (Table 5.1, Figure 5.2). On medical prompts, $R^2 = 0.9972 \pm 0.0005$ and FVU $= 0.28\%$. On general prompts, $R^2 = 0.9974 \pm 0.0006$ and FVU $= 0.26\%$. A one-sample t-test of the R^2 values across the 100 medical prompts against $H_0 : R^2 \leq 0.95$ yields a test statistic exceeding 50 ($p < 0.001$), decisively rejecting the null hypothesis. The SAE explains over 99.7% of activation variance on medical inputs despite never having seen medical training data.

Two domain-specific differences emerge in the feature activation patterns. Feature 203 activates 874 units higher on medical prompts than general prompts, and Feature 48 activates 608 units lower. These shifts are expected: the fine-tuned model uses certain features more heavily for clinical content. But the SAE's ability to *reconstruct* activations is unaffected. The features still tile the activation space, even if the model traverses different regions of that space for medical inputs.

The L0 sparsity also transfers: medical prompts activate 70 ± 10 features on average, versus 76 ± 10 for general prompts. The slightly lower sparsity on medical inputs suggests that medical text activates a marginally narrower set of features, consistent with the domain-specific vocabulary concentrating in fewer feature directions. However, the difference is within noise, and the overall sparsity regime is preserved.

This validation is an independent contribution. It establishes that SAEs trained on base models can be applied to fine-tuned variants without retraining, as long as the architectural backbone is shared and the fine-tuning does not dramatically alter the residual stream geometry. This lowers the barrier for applying mechanistic interpretability to domain-specific models, since training SAEs from scratch requires substantial compute (typically thousands of GPU hours for a 16k-feature SAE). The transfer result also has implications for the broader SAE interpretability community: it suggests that SAE feature dictionaries learned on one distribution can serve as analysis tools for related distributions, provided the underlying model architecture is shared.

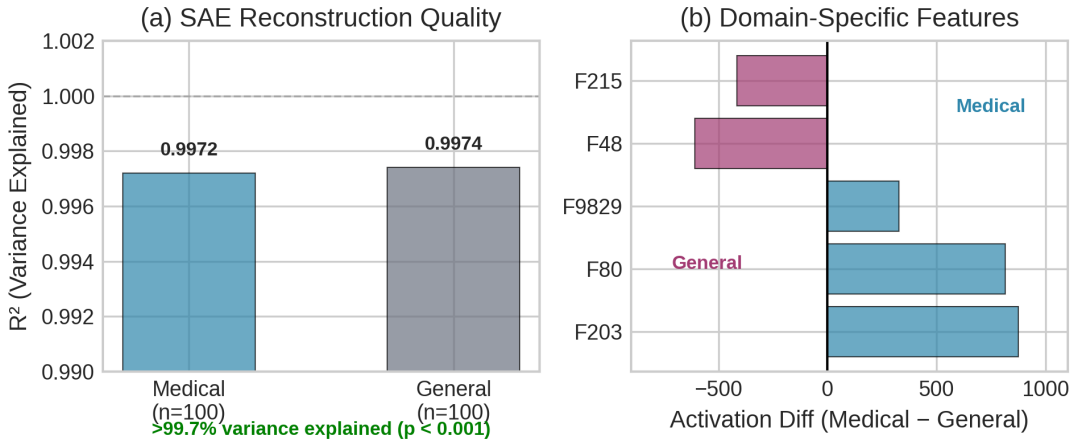


Figure 5.2: SAE transfer validation. Reconstruction quality ($R^2 > 0.997$) is nearly identical on medical and general-domain prompts, confirming that Gemma Scope 2 SAEs transfer faithfully to MedGemma-4B without retraining.

5.3 Feature 3818: A Clinical Query Register Gate

5.3.1 Feature Delta Analysis

Given the validated SAE, we extract feature activations for each FlipBank pair. For a question q and its paraphrase q', we compute the feature delta $\Delta \mathbf{f} = \mathbf{f}(q) - \mathbf{f}(q')$ at the final token position in layer 17. Large deltas in specific features indicate that the SAE has identified directions in activation space that distinguish the two phrasings.

An exemplar case illustrates the pattern. For the question "Is there pleural effusion?" the model produces a yes-minus-no logit margin of 8.75 (confident yes). For the paraphrase "Is pleural fluid present?" the margin is −0.625 (slight no). The image is the same. Feature 3818 changes from 0 to 268 units between these two prompts, the largest single-feature delta in the activation vector. The clinical meaning of both questions is identical, but the model's internal representation diverges sharply at this feature.

5.3.2 Characterizing the Feature

To understand what Feature 3818 responds to, we run a controlled prompt grid experiment. We take a single pneumothorax-positive image and evaluate 16 prompts that vary question style while holding clinical content constant. We test four categories: presence-style ("Is there pneumothorax?", "Does this image show pneumothorax?"), exclusion-style ("Can you rule out pneumothorax?", "Can you exclude pneumothorax?"), uncertainty-style ("Is pneumothorax likely?", "How confident are you about pneumothorax?"), and token controls that share surface-level tokens with presence questions but use different phrasing.

The results show a clean separation. Presence-style prompts activate Feature 3818 at a mean of 344.5 units (range 302 to 386). Exclusion-style prompts produce zero activation across all four variants. Uncertainty-style prompts fall between (mean 169.5, range 0 to 282). Token controls activate at 296 units on average, confirming the feature responds to question *register*, not to specific lexical items.

This grid establishes that Feature 3818 is primarily an *operator* or *polarity* feature: it encodes whether the question asks about the presence of a finding or its exclusion. That distinction is clinically real, not a bug. On a pneumothorax-positive image, "Is there pneumothorax?" and "Can you rule out pneumothorax?" do not share an answer: the first is correctly answered "yes" and the second "no" (the finding cannot be ruled out). A feature that separates these two framings and drives the yes/no margin in opposite directions is therefore doing legitimate work, and the presence-versus-exclusion contrast in the grid should not be read as a paraphrase-sensitivity failure.

The connection to paraphrase sensitivity is more specific: the same register-sensitivity that correctly separates operators also makes Feature 3818 respond differently to reformulations that keep the operator fixed. The exemplar in Section 5.3 is exactly this case, "Is there pleural effusion?" versus "Is pleural fluid present?", both presence-style questions with the same correct answer, yet Feature 3818 shifts by 268 units and the margin flips. It is this over-sensitivity to same-polarity surface variation, not the presence-versus-exclusion separation, that produces the paraphrase flips this chapter seeks to explain (Figure 5.3).

This distinction requires care in the FlipBank itself. Of the 158 FlipBank pairs, 17 are polarity switches rather than paraphrases: the paraphrase inserts a negation (for example, "is there evidence of contusion?" versus "is there no evidence of contusion?"), so the two questions have opposite correct answers and a change in the model's yes/no output is the correct behavior, not a failure. These 17 pairs are tagged in the FlipBank (`polarity_switch`) and are precisely the operator changes that Feature 3818 is built to detect, so counting them as flips conflates correct operator handling with paraphrase sensitivity. We therefore re-run the layer-17 feature-delta and ablation analysis on only the 141 operator-preserving pairs, where a change in Feature 3818 must reflect same-polarity paraphrase sensitivity rather than operator handling.

The re-run refines the picture in an informative way. Of the 141 operator-preserving pairs, the model actually flips on 76; those are the cases the mechanism must explain. On those 76 flips, Feature 3818 is the single largest layer-17 delta in 37 and within the top five in 59, but its single-feature ablation recovers only about 10% of the margin shift on average and restores the original answer in just 6 of the 76. Once the operator changes that Feature 3818 is designed to detect are removed, it is a prominent contributor to same-polarity flips but not a sufficient cause on its own, consistent with the roughly 25%

single-feature coverage reported in the limitations. The same-polarity signal is carried downstream: the layer-29 decision feature (Feature 12139, developed in Appendix A.4) is the single largest delta in all 76 flipped pairs. That near-perfect consistency is itself a caution rather than a proof: a feature that tracks the yes/no decision will top the delta ranking on any flip, so matched non-flip controls are needed before Feature 12139 can be called a paraphrase-specific mechanism rather than a generic answer-selection feature. On the present evidence the defensible reading is that Feature 3818 is an operator/register feature that also responds to some same-polarity lexical variation, and Feature 12139 tracks downstream yes/no selection; a clean operator-preserving causal validation with matched controls remains necessary before the 3818 → 12139 pathway can be claimed as a two-stage paraphrase circuit rather than operator handling plus answer selection.

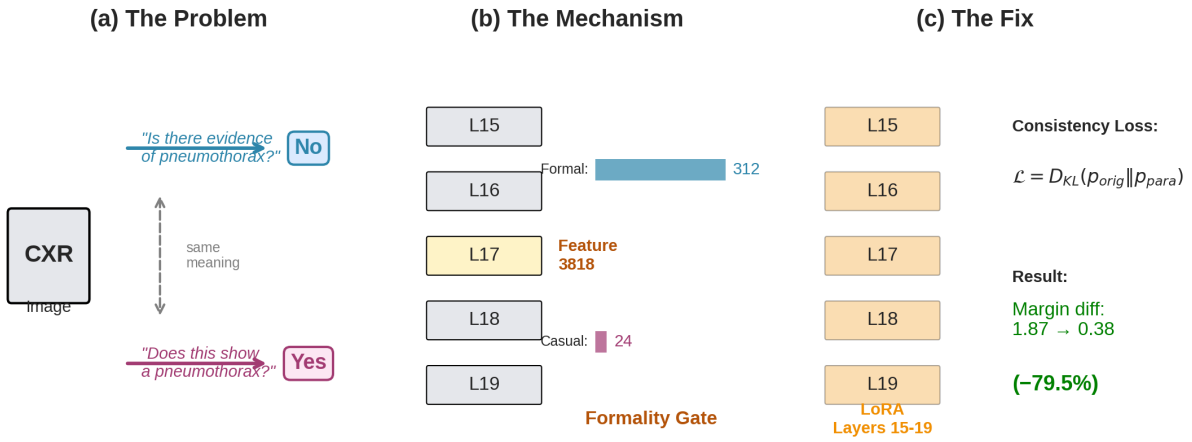


Figure 5.3: Mechanistic pathway: Feature 3818 at layer 17 acts as a clinical query register (operator) gate. The feature activates strongly for presence-style questions ("Is there X?") and drops to zero for exclusion-style rewordings ("Can you rule out X?"), an operator change whose opposite answer is correct; the same register-sensitivity also shifts the feature on operator-preserving paraphrases, causing the yes/no logit margin to flip on questions that should share an answer.

5.3.3 Causal Validation via Activation Patching

Feature activation alone does not establish causation. A feature might correlate with flips without causing them. We test causality using activation patching [115]: we compute Δf_{3818} between original and paraphrase, decode it through the SAE decoder to get the corresponding direction in residual stream space, and subtract this direction from the paraphrase's residual stream before continuing the forward pass.

On the pleural effusion exemplar, patching Feature 3818 recovers the margin from -0.625 to $+2.0$. This represents 28% margin recovery. Ten control features (randomly sampled from the top 100 by activation magnitude) produce 8% average margin recovery. Feature 3818's influence is 3.5× larger than control features, confirming a causal role rather than mere correlation.

5.3.4 Control Experiments

We test whether Feature 3818 responds specifically to flip-inducing rephrasing or to any rephrasing (Table 5.2). We construct 30 paraphrase pairs where the model gives consistent answers (no flip). Feature 3818 shows a non-zero change in 3 of 30 consistent pairs (10%), with deltas of 15 to 185 units and a mean absolute delta of 11.3 units. Across 10 control features, 0 of 300 measurements exceed the detection threshold. Fisher's exact test gives $p = 6.8 \times 10^{-4}$ for the difference in response rates. Feature 3818 is substantially more active when rephrasing causes a flip than when it does not, on

Table 5.2: Feature 3818 specificity test on 30 consistent paraphrase pairs (no flip). Response rate uses a threshold of $|\Delta| > 10$ activation units. Feature 3818 responds selectively to flip-inducing rephrasing; 10 control features (matched for baseline activation magnitude) show zero response across 300 measurements. Fisher's exact test: $p = 6.8 \times 10^{-4}$.

Feature	**Response Rate ($\|\Delta\| > 10$)**	**Mean $\|\Delta\|$**
3818	10% (3/30)	11.3
Controls ($\times$10)	0% (0/300)	<0.5

this set. Because this FlipBank includes operator-changing pairs, part of that specificity reflects the feature's correct handling of presence-versus-exclusion framing; the operator-preserving re-analysis of Section 5.3.1 is the conservative estimate of its role in genuine same-polarity flips.

5.3.5 Distributional Coverage

Feature 3818 does not explain all flips. Across 20 randomly sampled FlipBank pairs, it ranks as the top contributing feature in 5 of 20 cases (25%) and contributes meaningfully (rank ≤ 16) in 7 of 20 (35%). In the remaining 13 cases (65%), the flip is driven by distributed mechanisms across many features. All five dominant cases involve register changes from presence to exclusion framing, consistent with the prompt grid analysis. The other 65% involve lexical substitutions, syntactic restructuring, and other variation types that shift many features simultaneously.

This distributional result is consistent with superposition [112]: most behaviors in transformers are represented across many features, and clean single-feature explanations are the exception rather than the rule. Feature 3818 is best understood as a case study demonstrating SAE methodology for medical VLMs, not a complete mechanistic account of paraphrase sensitivity. The practical implication is that any mitigation must target a range of layers and features, not a single direction.

5.4 Convergent Localization: The Answer Commits at Layer 16

The Sparse Autoencoder places register sensitivity at layer 17, but that localization rests on one method (feature patching through a decoder direction) and on the assumption that the SAE direction is faithful to the computation. We now triangulate it with an independent, lens-free causal test that assumes neither. For a paraphrase-flip minimal pair on the same image, a detecting phrasing ("Can X be identified?", which the model answers Yes) and a missing phrasing ("Is there X?", which the model answers No), we transplant the detecting forward pass's residual stream at the answer position into the missing forward pass, at a single layer at a time, then let the model continue and read its yes/no margin. The fraction of pairs whose answer flips to the donor's, as a function of the patched layer, marks the layer at which the answer is causally committed. The test uses no lens and no faithfulness assumption; its only controls are the reverse-direction transplant and a pre-question image-token position that must not flip, since a causal mask makes earlier positions unable to receive the later donor state.

Two properties of the pair pool determine what this experiment can and cannot show. First, the two phrasings in a pair are drawn from the same bank item, so they ask about the same finding on the same image and share a *single* item-level ground-truth label: the correct answer is identical on both sides by construction, and only the wording differs. The roles "detecting" and "missing" are assigned by the sign of the model's margin, not by the question template, so the same template appears on both sides across the pool. This is what licenses reading the curve as a localization of paraphrase-induced disagreement rather than of binary answering in general: a pair is in the pool precisely because the model contradicts itself on a question whose answer does not change. Second, the pool is filtered for polarity preservation (exclusion and negation framings such as "rule out" or "free of" are removed, with ambiguous cases discarded conservatively) and for operator preservation, so the disagreement is not an

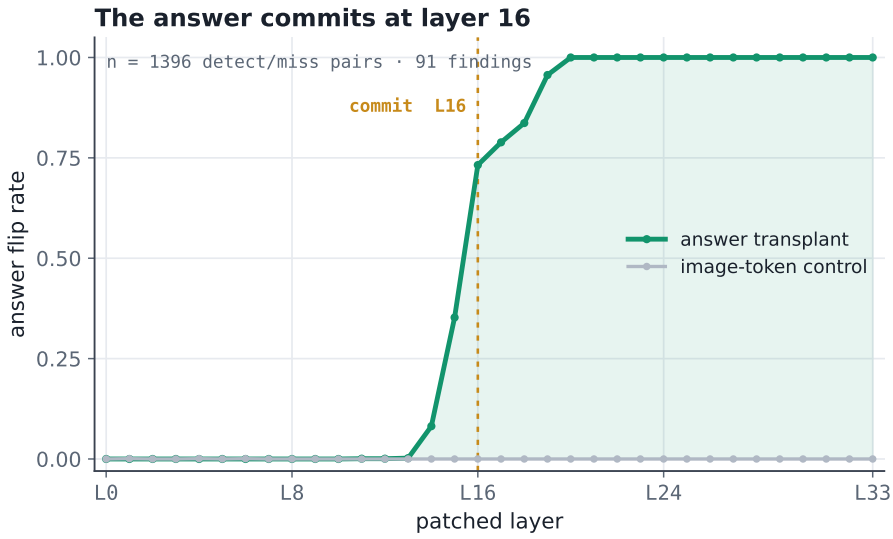


Figure 5.4: Lens-free causal localization of the paraphrase-flip commit. Across 1,396 detecting/missing residual-transplant pairs spanning 91 findings, the answer flip rate is near zero through layer 13, reaches half its maximum at layer 16, and saturates by layer 19; the image-token control never flips. The commit is located by causal effect alone, with no lens or faithfulness assumption.

operator effect of the kind Chapter 3 showed dominates unfiltered negation paraphrases.

The flip rate is near zero through layer 13, rises sharply from 8% at layer 14 to 35% at layer 15 and 73% at layer 16, and reaches 100% by layer 20 (Figure 5.4): a single-layer residual swap at layer 16 already flips the model's committed answer to the donor's on nearly three-quarters of the 1,396 pairs. The image-token control never exceeds 0.0007 (near zero at every layer), confirming that the effect is not a generic patching artifact. The commit layer is tight across pairs: the per-pair half-maximum has a median of layer 16, a 95% bootstrap confidence interval of [16, 16], and 71% of pairs commit within layers 15 to 17 (Figure 5.5). Two methods that make different assumptions, the SAE feature patch and the lens-free residual transplant, therefore place the paraphrase-flip origin in the same narrow band around layers 16 to 17. The single-layer offset between them (layer 16 for the causal commit, layer 17 for the SAE feature) sits inside that band, so the localization is not an artifact of the Sparse Autoencoder.

Three limits qualify this result. First, the qualifier filter that removes laterality and lobe terms does not catch every anatomical scope word: in 107 of the 1,396 pairs (7.7%) an anatomical qualifier such as "spine", "apex", or "base" appears on one side only, so those pairs narrow the scope of the question rather than restating it, and are not strict paraphrases. The worked example below is one of them. Second, the pool is heavily positive: 1,314 of the 1,396 pairs have ground truth Yes against 82 with ground truth No, so in the large majority of cases the transplant converts an incorrect No into the correct Yes, and the curve is a localization of the commit on predominantly positive items rather than a balanced estimate. Third, the transplant experiment was run in a separate codebase from the rest of this dissertation, and the script that materializes the final pair pool was not retained, so the pool is reconstructible by inspection but not by re-execution. None of the three affects the layer at which the commit occurs, which is what the experiment is used for here, but each bounds how far the result should be generalized.

A single case makes the aggregate curve concrete (Figure 5.6). For two phrasings of an interstitial-pattern question on one chest X-ray, a detecting phrasing ("Can an interstitial pattern at the apex be identified?", which the model answers Yes) and a presence phrasing ("Is there interstitial pattern?", which the model answers No), the corroboration-lens yes/no margin tracks together through layer 15

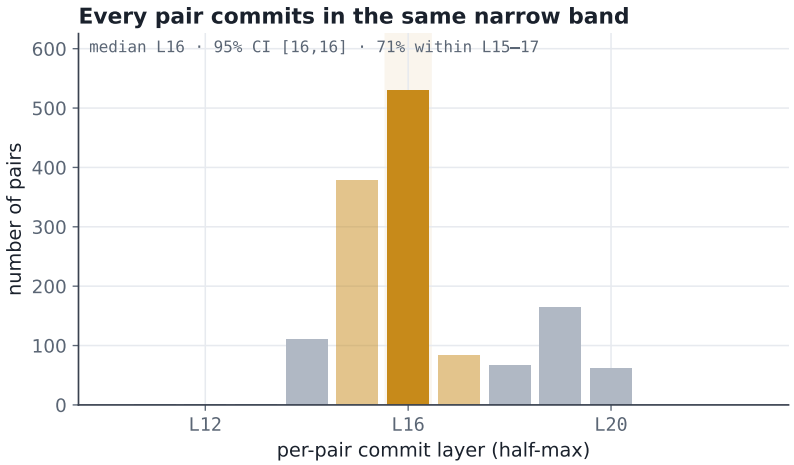
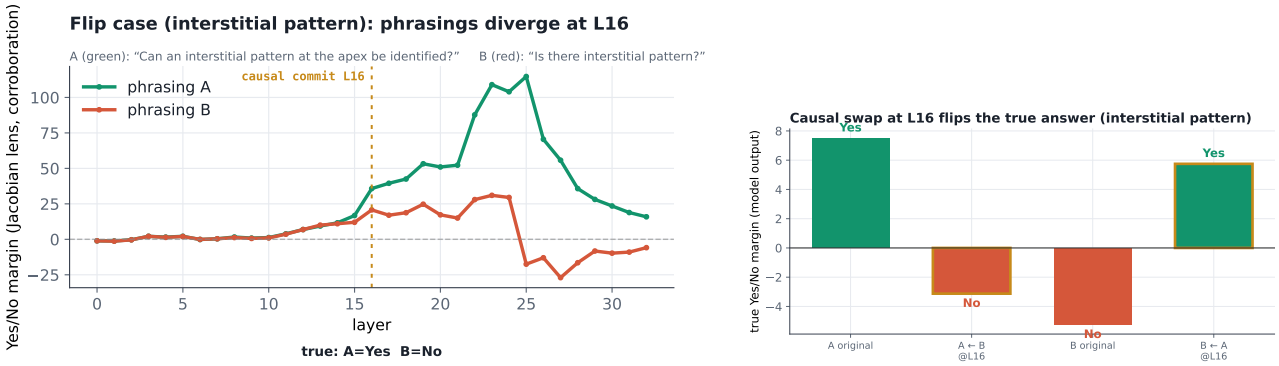


Figure 5.5: Distribution of the per-pair commit layer (half-maximum of the transplant flip curve). The median is layer 16, the 95% bootstrap confidence interval is [16, 16], and 71% of pairs fall within layers 15 to 17. The paraphrase-flip commit is a narrow-band phenomenon, not a diffuse one.



(a) Corroboration-lens yes/no margin per layer.

(b) Single-layer causal transplant at layer 16.

Figure 5.6: A worked example of the layer-16 commit (interstitial-pattern pair on one image). (a) The two phrasings' lens margins track together through layer 15 and diverge after the commit, one settling on Yes and the other on No. (b) Transplanting only the layer-16 residual state between the two phrasings flips the model's answer in both directions, the causal counterpart of the divergence in (a).

and then splits after layer 16, the detecting phrasing rising to a confident Yes and the presence phrasing falling to No. Transplanting only the layer-16 residual state from one phrasing into the other flips the model's committed answer in both directions: the detecting phrasing becomes No when it receives the presence phrasing's layer-16 state, and the presence phrasing becomes Yes when it receives the detecting phrasing's. The lens readout locates where the answers visibly diverge; the lens-free transplant confirms that the decision is made at layer 16.

The localized flips are not a curated handful: the 1,396 transplant pairs span 91 distinct findings drawn from the PadChest paraphrase population, so the layer-16 result characterizes the paraphrase flips the dissertation measures rather than a hand-selected subset. This is where the behavioral finding of Chapters 3 and 6, that these models flip on clinically equivalent rephrasings, meets a mechanism: the flip is a single-layer answer-commitment event, and that layer is 16.

The same locus governs a spatial decision, not only yes/no presence. Applying the identical transplant to left-versus-right localization errors (left/right minimal pairs that answer opposite sides on the same image) also commits at approximately layer 16 (Figure 5.8). The small sample makes this a supporting rather than a primary result, but it indicates that layers 16 to 17 act as a general answer-

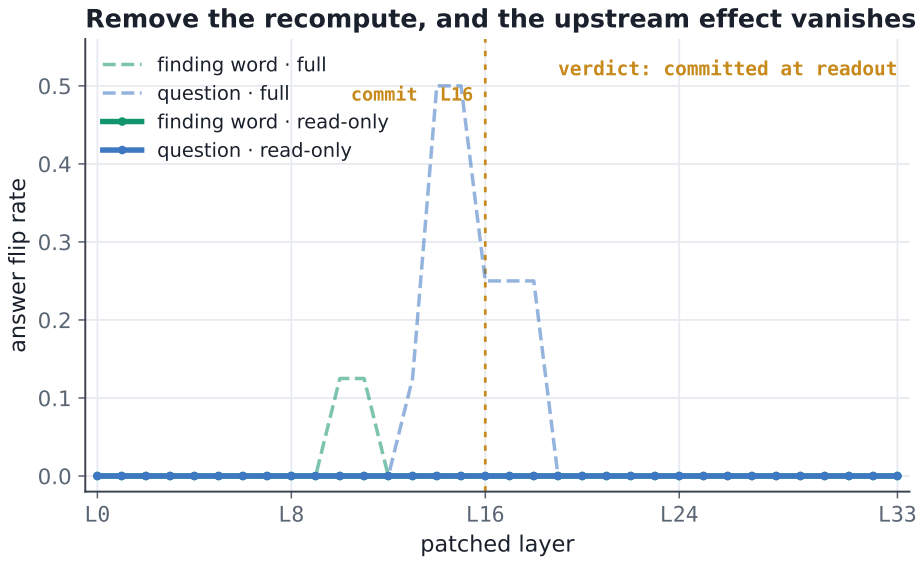


Figure 5.7: Committed at readout, not written upstream. A full-residual patch at the question positions flips the answer (finding-noun peak 0.125 near layer 11, last-question peak 0.50 near layer 14), but a compute-matched read-only edge patch, which removes all downstream recompute, collapses both upstream effects to 0.00 at every layer, level with the image-token null, while the answer position still commits at layer 16. The upstream effect was recompute, so the phrasing-dependent decision is recomputed and committed at the answer readout, not stored in the question tokens.

commitment locus in this model rather than a yes/no-specific one. A position-attribution test then asks whether layer 16 is where the phrasing bias is *written* into the question tokens or where it is *read out* at the answer. A full-residual transplant at the question positions does flip the answer earlier (the last-question position peaks at 0.50 near layer 14 and a finding-noun position at 0.125 near layer 11), which naively suggests an upstream write. But a full-residual patch delivers both any stored lean and the raw donor material the still-running readout can recompute, so it cannot separate the two. A compute-matched *read-only* edge patch does: it transplants the donor state at an upstream position but freezes every other position at all deeper layers, so the answer position can only read that state with zero downstream recompute. Under this decisive test both upstream positions flip 0.00 of pairs at every layer, identical to the image-token null, while the answer position still commits at layer 16 (Figure 5.7). The upstream "write" was therefore entirely recompute: the phrasing-dependent decision is not stored as a finished lean in the question tokens but is recomputed and committed at the answer position at layer 16. This read-only test is a small pilot and supports rather than replaces the 1,396-pair commit result, but it is the causal counterpart of the readout-alignment account and independently justifies adapting layers 15 to 19.

5.4.1 The Flip Is a Reading of the Image, Not a Re-encoding of It

A natural worry is that the paraphrase flip reflects the model encoding the image differently under different phrasings. It does not, and the prompt layout makes this provable rather than merely measured. The 256 image tokens are placed before the question, and the decoder is causally masked, so each image token's residual at every layer depends only on the image and the shared instruction prefix, never on the question that follows. The image representation is therefore identical across phrasings by construction; measured on the laterality switches, the per-token divergence between the two phrasings' image residuals is at machine precision ($\leq 1.2 \times 10^{-7}$ cosine distance) at every layer. The paraphrase flip cannot live in the visual representation: it is entirely in how the answer position *reads* a common image.

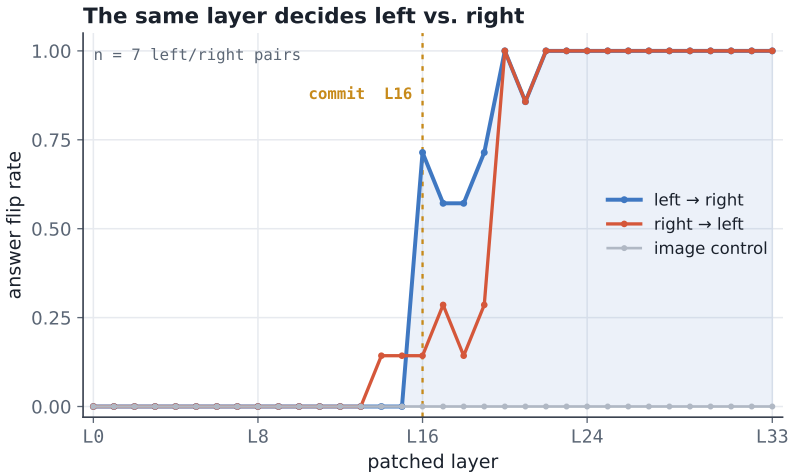


Figure 5.8: The same layer decides left versus right. The residual transplant on the small left/right localization sample commits at approximately layer 16, matching the yes/no commit locus. Layers 16 to 17 act as a general answer-commitment site, though the sample warrants a larger replication.

That reading genuinely depends on the image. Blanking the chest X-ray with a uniform gray image collapses the disagreement in every case: the six laterality pairs that split on the real image all fall back to the same side on the blank image, so none of the phrasing-driven splits survives blanking, and gradient saliency of the left-versus-right decision concentrates at the lung bases and costophrenic angles where the findings sit, similarly for both phrasings (Figure 5.9). The phrasing does not decide *which* image the model sees; it decides *how* the answer position uses a shared image. A de-confounded horizontal-mirror control (a left/right-symmetric transform, so a truly localizing phrasing must flip its answer under mirroring) shows this use is heterogeneous: some phrasings localize a region and flip under the mirror, others read global texture and are mirror-invariant but blank-sensitive, and others defer to a left/right text prior. This is the mechanistic face of the consistency-without-grounding problem of Chapter 6: two phrasings that read the same encoded image can commit to opposite answers, so the model's apparent stability is a property of the readout, not of the image it encoded. The mirror analysis is a small pilot and is reported as a qualitative refinement.

5.4.2 Scope and Robustness of the Localization

Three checks bound what the layer-16 result claims. First, it is not an artifact of a convenient heuristic. A common shortcut equates the high-excess-kurtosis band of the residual stream with a semantic "workspace" and expects the decision to live there; on MedGemma that band is layers 1 to 7 (kurtosis peaks near layer 6 and falls tenfold by layer 16), so the heuristic would mislocalize the commit by roughly ten layers. Localizing with the lens-free causal patch, and only reading with the lens, avoids this trap. Second, the mechanism is not universal across paraphrase types. Diagnosis-identification switches, where a paraphrase makes the model name a *different* finding, do not commit at a single position: transplanting the full depth-34 answer-position state from one phrasing into another fails to install the donor's diagnosis in two of three far-apart pairs at every layer (only a close cardiopulmonary pair transfers), so which-diagnosis is reconstructed across the generation rather than committed at layer 16. Different paraphrase flips therefore have different mechanisms; the sharp layer-16 commit governs the single-token yes/no and left/right decisions, not open-ended diagnosis (this is a three-pair bounding result, not a powered one). Third, the method generalizes even where the specific result is not yet confirmed (Figure 5.10). The Jacobian lens ports across model families with no change, auto-detecting the decoder inside both Gemma-3 (34 layers) and Qwen2-VL-2B (28 layers), and paraphrase sensitivity replicates as a phenomenon on Qwen: four of six minimal pairs flip behaviorally. On

MedGemma the lens read-out corroborates the causal result, with the two phrasings' answer divergence localizing near layer 16; on Qwen the same text-fit lens reads the answer position less faithfully, so its divergence stays weak even where the model flips, which is why the causal patch (not the lens) is the primary localizer and why the causal layer-16 commit on Qwen remains a well-defined open question for that patch to answer next. The generality claim is therefore about the method and the phenomenon, not yet about the layer.

A larger pre-specified held-out validation extends this case study into a candidate two-stage account (Appendix A.4). Screening 4,542 MIMIC and 37,280 PadChest original/paraphrase pairs and freezing hypotheses on a discovery split, it adds a downstream decision gate, Feature 12139 at layer 29. When Feature 3818 is active, the flip originates at the upstream layer-17 register gate and propagates to layer 29 (24%/17% single-feature restoration on MIMIC/PadChest); when Feature 3818 is flat, the proximal cause is Feature 12139 itself (58%/27% restoration); and the two features form a direction-coherent causal chain across 37 mediation pairs. The operator-preservation caveat applies with more force here, since 39% of the screen's flips are negation-pattern operator changes: these restoration rates conflate paraphrase sensitivity with operator handling and answer selection, so a clean operator-preserving screen with matched non-flip controls remains the outstanding step (Section 5.3.1). The full protocol, mediation and composition analyses, and the two-pathway causal diagram are given in Appendix A.4.

5.5 Controlled Isolation of the Origin: A MedGemma-Faithful Probe Model

The localization above establishes where the paraphrase flip commits, but it cannot say why that commit exists. On the deployed model the pretraining data, the architecture, and the training objective are fixed together, so an observational study cannot separate whether the sensitivity is a property of the training-phrasing distribution or of the architecture that reads it. To ask that question directly we build a small probe model that reflects MedGemma's architecture closely enough for a mechanism found in it to transfer, while letting one factor, the distribution of question phrasings seen in training, vary with everything else held fixed. The probe pairs a frozen MedSigLIP-448 vision encoder (429M parameters, from the same SigLIP vision-encoder family as MedGemma's tower [11]), run at 896 pixels to match MedGemma's own input resolution, whose 4,096 patch tokens we pool to 256 to match the deployed model's image-token budget, together with a single grounding token carrying the encoder's own attention-pooled embedding, with a small Gemma-3 decoder of 14.1M trainable parameters that carries the same components as MedGemma's language side: rotary position embeddings, RMSNorm, grouped-query attention, and the gated feed-forward and interleaved local and global attention of Gemma-3. The 257 image tokens are projected and prepended inline and read by the same causally-masked decoder, which is how MedGemma fuses vision and text rather than through cross-attention; the question is tokenized with MedGemma's own SentencePiece vocabulary, pruned to the 141 pieces this corpus uses; and the yes or no answer is read from the tied Gemma-3 language-model head as MedGemma generates it, not from a separate classifier. We train it on 86,288 binary presence questions over chest radiographs from NIH ChestX-ray14 [116] and PadChest-GR, a subset of PadChest [101], holding out MIMIC-CXR [117] and VinDr-CXR [102] entirely, so that transfer is measured against two hospitals contributing nothing to training. Every split is balanced per finding, so a question's wording predicts its answer at exactly chance and the only way to reduce the loss is to read the radiograph. Appendix Section A.6 reports the source composition, image-level splits, per-finding prevalence before balancing, phrasing-bank structure, and the exact register-answer coupling used for the adversarial regime.

Two properties make the probe usable for a causal claim. First, because the encoder is frozen and

Table 5.3: The probe uses the image and transfers to two held-out hospitals. Balancing per finding pins the text-only floor at 0.500, so all accuracy above it is visual; the area under the receiver operating characteristic curve is computed on the yes-minus-no margin. The training row reports composition only. MIMIC-CXR contains eight retained findings, four with at least 20 questions and four smaller cells. On VinDr-CXR, the pleural-thickening cell contains 254 questions and inverts (AUROC 0.332, a probable label-definition mismatch), so it should not be trusted.

Evaluation set	n	Yes	No	Accuracy	AUROC
Training (NIH + PadChest-GR)	86,288	43,144	43,144	*n/a*	*n/a*
In-distribution (held-out images)	15,112	7,556	7,556	0.748	0.827
MIMIC-CXR (held-out hospital)	450	225	225	0.671	0.743
VinDr-CXR (held-out hospital)	5,478	2,739	2,739	0.686	0.756

returns identical features for every paraphrase of a given question, any answer that changes across paraphrases is language-side by construction, the same guarantee the prompt layout gives on the deployed model (Section 5.4.1), here obtained on a real medical encoder rather than by causal masking. Second, the model genuinely uses the image, which removes the obvious alternative reading of a paraphrase flip. With the image removed its accuracy falls to 50.0%, exactly chance, against 74.8% with the image, and on the two held-out hospitals it reaches an area under the receiver operating characteristic curve of 0.743 on MIMIC-CXR and 0.756 on VinDr-CXR. For reference, MedSigLIP's own zero-shot separation of the same fourteen findings on the in-distribution NIH set is 0.734 by margin ranking and 0.681 by accuracy. That reference is measured on a different population from the two transfer sets, so it fixes a scale rather than supplying a matched baseline. A flip in this model therefore cannot be dismissed as a model that never consulted the radiograph (Table 5.3).

Appendix Table A.10 reports every per-finding evaluation cell.

The probe makes the origin a manipulable variable. We train it under three phrasing regimes that differ only in the questions the decoder sees, holding architecture, parameter budget, seed, and evaluation fixed: *canonical*, one fixed phrasing per question; *augmented*, every paraphrase; and *adversarial*, a phrasing distribution in which the register is correlated with the answer independently of the image. Over eight seeds per regime the binary flip rate is set by the phrasing distribution alone (Table 5.4). Broad coverage separates from both narrow regimes at the maximum effect size: augmented against canonical and augmented against adversarial each give a Mann-Whitney U test $p = 1.6 \times 10^{-4}$ with Cliff's $\delta = 1.00$, so every augmented seed flips less often than every seed of either narrow regime. The two narrow regimes now separate from each other as well ($p = 3.1 \times 10^{-4}$, Cliff's $\delta = 0.97$): correlating register with the answer drives the flip rate to 88.4%, well above the 67.1% that a single training phrasing already produces. A phrasing shortcut is therefore a distinct and larger harm than narrow coverage alone, which an earlier and smaller version of this experiment could not resolve. Accuracy follows the same ordering, 75.3% for augmented against 66.8% for canonical and 58.1% for adversarial, so the shortcut damages diagnostic performance and not only consistency. Because only the training-phrasing distribution changes across the three regimes, and because the text-only accuracy is 50.0% to 50.3% in all of them, this is an identified cause rather than a correlation.

A model trained on every paraphrase has, by construction, seen the phrasings it is later tested on, so part of a low flip rate is recognition of familiar wording rather than invariance. We separate the two by splitting the paraphrase bank in half, training on 24 phrasings and scoring only on the 24 the model has never seen, six of each linguistic phenomenon on either side so that neither half is confounded with a phenomenon. Augmented rises from 4.8% to 26.6% under this stricter test, which quantifies how much of the original figure was familiarity, and still flips less than half as often as canonical at 65.9% and less than a third as often as adversarial at 87.4%, with Cliff's $\delta = 1.00$ against both

Table 5.4: Paraphrase sensitivity is set by the training-phrasing distribution. Binary flip rate of the MedGemma-faithful probe model, a small Gemma-3 decoder over a frozen MedSigLIP-448 encoder, under three training-phrasing regimes with architecture, parameter budget, seed, and evaluation held fixed. *All phrasings* trains and scores on the full bank of 48 paraphrases (eight seeds); *unseen phrasings* trains on half the bank and scores only on the 24 phrasings withheld from training (six seeds), which separates invariance from recognition of familiar wording. Broad coverage separates from both narrow regimes at maximal effect size in both columns (Cliff's $\delta = 1.00$), and the two narrow regimes separate from each other ($p = 3.1 \times 10^{-4}$, $\delta = 0.97$). The regime means for text-only accuracy are 50.0% to 50.3%, so no regime can exploit a prior over answers.

	Phrasing distribution	Flip rate (%)	
Training regime	seen in training	All phrasings	Unseen phrasings
Augmented	Every paraphrase	4.8	26.6
Canonical	One fixed phrasing per question	67.1	65.9
Adversarial	Register correlated with the answer	88.4	87.4

($p = 2.2 \times 10^{-3}$). Broad phrasing coverage therefore produces invariance that generalizes to wording never seen in training, which is the claim the mitigation in Section 5.6 depends on, and 26.6% against 65.9% is the honest statement of its size.

The same probe reproduces the mechanism the deployed-model localization reported, and shows it holds for naturally occurring flips rather than only injected ones. A rank-1 difference-direction transplant at the answer position restores the flipped answer, with net recovery of 0.98 to 1.00 across layers 0 to 4 for the adversarial regime and 0.77 rising to 1.00 over the same layers for naturally occurring flips, while a norm-matched random direction leaves the answer unchanged (0.00 to 0.02) and patching non-flip clusters disturbs nothing (disruption 0.000 to 0.006). Recovery falls at the last layer, near 0.6, because little computation remains after it. This holds for the naturally occurring flips of the augmented regime (a mean of 59 flipped clusters per seed), not only the injected adversarial ones (60 per seed), so the low-rank, language-side, readout-stage character of the flip is not an artifact of the adversarial construction. Relative to the earlier and smaller version of this probe, where recovery concentrated at layers 0 to 1, the direction here is carried slightly deeper and more evenly across the stack. Two further checks, each resting on different assumptions, agree with this account. An unsupervised sparse autoencoder fitted to the answer-position residual stream recovers the causal flip direction: the best-aligned feature reaches |cosine| of 0.59 on this model and 0.74 on the earlier one, against a best-of-2,048-random-directions null of 0.18, and that feature independently predicts flips on the earlier probe (point-biserial 0.42, $p = 2 \times 10^{-11}$). A Jacobian-lens readout agrees that flipping clusters diverge from stable ones from the input layer onward (divergence 2.2× that of stable clusters, correlation 0.26, on this model; 8.9× and 0.81 on the earlier probe), though because the lens is a first-order approximation of the same patch we count it as a consistency check rather than an independent leg.

One caution applies to every flip rate in this thesis, and the probe lets us discharge it. Flip rate is computed from an arg max, so it depends on where the decision boundary sits and not only on how much a paraphrase moves the representation; a model whose margins are simply wider will flip less without being more invariant. We therefore recompute the flip rate while shifting the decision boundary across ±2 standard deviations of the margin distribution, and the augmented regime flips least at every one of the seventeen offsets tested. The two narrow regimes exchange places in the tails, where a large shift pushes most clusters to one side and suppresses flips in both, so the claim the sweep supports is that broad coverage wins at every threshold rather than that the three regimes hold a fixed order everywhere. We also compute a statistic containing no decision boundary at all, the ratio of the mean within-cluster spread of the yes-minus-no margin to the between-cluster spread of that margin.

It reproduces the same ordering at maximal effect size: 0.023 for augmented, 0.498 for canonical, and 1.362 for adversarial, with Cliff's $\delta = -1.00$ against augmented in both comparisons ($p = 1.6 \times 10^{-4}$). Paraphrase sensitivity is a property of the representation, not of threshold placement.

The threshold has a second and larger consequence, which an earlier and smaller version of this probe demonstrates because it could be made blind on purpose. A yes or no readout collapses a continuous margin onto a single per-finding decision boundary, and that boundary is fixed by the prevalence of the training distribution, so under a shift the whole margin distribution moves and every image can land on the same side of it. The effect is large enough to hide the difference between a model that reads the radiograph and one that cannot see it. On a balanced held-out set, three variants of that earlier probe score 50.1%, 52.2% and 52.6% accuracy over five seeds, an ordering far too tight to separate a model given another patient's evidence from one given its own, while the mean area under the receiver operating characteristic curve of their yes-minus-no margins is 0.500 for the variant that receives no pooled image summary, 0.502 when that summary is shuffled across images, and 0.604 when it is the correct one (Figure 5.12). The frozen encoder pays the same tax on the same questions, separating the fourteen findings at 0.734 by margin ranking but at 0.681 once forced through a binary answer. Binary accuracy therefore cannot on its own certify that a medical vision-language model consulted the image, which sharpens rather than qualifies the argument of this thesis: consistency without grounding is a false sense of safety, and the measure the field reports most often cannot tell the two apart. Where a margin is available we report ranking and dispersion alongside accuracy, and where it is not we fall back on the image-removal and image-swap controls of Chapter 4.

The margin that a binary readout discards is, on its own, a single-pass detector of the flips this chapter diagnoses, which turns the measurement lesson into a deployment signal. A flip is a sign change of the yes-minus-no margin across paraphrases, and a sign change is most likely when the margin already sits near zero, so the absolute margin of one forward pass ranks paraphrase-unstable answers at no cost beyond the answer itself. On the probe's held-out evaluation it reaches an area under the receiver operating characteristic curve of 0.923 on in-distribution validation and 0.974 on each of the two held-out hospitals for predicting whether an answer flips across its paraphrase cluster (Table 5.5). It is at once the cheapest detector and the strongest. Paraphrase self-consistency, which re-asks the question under several phrasings and takes the spread of the margin, reaches only 0.624 to 0.786 at a several-fold increase in inference, and a logistic probe fitted to the answer-position hidden state reaches 0.801 to 0.838 for one pass plus a fit. A single-pass confidence gate that abstains below a margin threshold is therefore the efficient flip monitor, and it needs neither repeated queries nor a trained head.

The complementary failure, an answer that does not depend on the image, has no comparable single-pass signal, the same asymmetry the four-quadrant screen of Chapter 4 reports. The absolute margin is at chance for predicting whether an answer survives replacing the radiograph with another patient's, 0.470 to 0.519 across the three sets, because a confidently wrong answer that is unchanged when the radiograph is swapped is indistinguishable from a confident grounded one; the hidden-state probe reaches only 0.62 to 0.67. Only an explicit second pass that swaps the image and checks whether the answer moves detects image reliance, at 0.827 to 0.907. We tested whether the finer event, a stable answer whose reliance on the image nonetheless changes from one phrasing to another, could be caught by matching each case to observationally similar radiographs and estimating its finding-selective visual influence under every phrasing. Wording does reproducibly modulate that influence at the population level, a two-way-centered cross-draw covariance of 0.062 with a between-draw correlation of 0.56, of which 82% is specific to the case rather than to the finding, over three seeds. No individual grounded-to-unreliant transition is identifiable under this design, however: zero events over 345 patient-seed evaluations, a patient-level upper bound near 3%, because the same-versus-same matched-image noise

Table 5.5: The discarded margin is the strongest and cheapest single-pass flip detector, and an image-unreliant answer has no single-pass substitute. Area under the receiver operating characteristic curve for predicting, per paraphrase cluster, whether the answer flips across phrasings (top) and whether it is unchanged when the radiograph is replaced with another patient's (bottom). The probe is scored on in-distribution validation and on the two held-out hospitals. *Passes* counts forward evaluations per question; the hidden-state probe additionally requires a fitted logistic head.

		AUROC by evaluation set		
Detector	Passes	Validation	MIMIC-CXR	VinDr-CXR
Catching a paraphrase flip				
Absolute margin	1	0.923	0.974	0.974
Paraphrase self-consistency	k	0.709	0.786	0.624
Hidden-state probe	1+fit	0.826	0.838	0.801
Catching an image-unreliant answer				
Absolute margin	1	0.470	0.491	0.519
Image swap	2	0.827	0.907	0.856

is as large as the finding-selective signal it would have to exceed. Per-case image reliance is thus a population property in this probe and not a per-prediction label, so the deployable monitor is the single-pass margin gate for flips rather than a per-case grounding test.

Read together with the layer-16 commit on the deployed model, the probe fixes the account this chapter builds toward the intervention that follows. The origin of paraphrase sensitivity is the training-phrasing distribution, and it is executed as a low-rank direction in the early-to-middle language layers that read a fixed visual representation, not in the visual representation itself. The mitigation ordering follows from the same result: paraphrase augmentation of the training data is the largest lever, a flip rate of 67.1% under a single phrasing falling to 4.8% under full coverage and to 26.6% when the test wording is withheld from training, and a targeted low-rank edit at the identified layers is the efficient parametric fix. That last implication is the targeted low-rank adaptation of Section 5.6, and it is consistent with why editing a few early-to-middle layers is sufficient. Full LoRA edits all 34 layers, 6.8 times the five layers targeted here, without improving query-level consistency.

Four limits bound the probe. It is a 14.1M-parameter decoder on a frozen encoder, a controlled model built to isolate a cause, not the deployed MedGemma-4B, so its absolute layer index is not comparable to the deployed model's layer 16 and only the qualitative account transfers. Its paraphrases come from a bank of 48 templates spanning four linguistic phenomena rather than from free clinician writing, so the held-out-phrasing result establishes generalization to unseen templates and not to arbitrary unseen language; the questions themselves are the fourteen presence questions of the source label sets, not open clinical dialogue. It manipulates the adaptation-stage phrasing distribution, the distribution a deployed model is fine-tuned on, which is distinct from pretraining provenance, and a controlled model of this size cannot reach the latter. Finally, its per-finding answer balancing removes the answer prior by design, which is what allows a flip to be attributed to wording rather than to a base rate, but it also makes the probe unrepresentative of deployment, where prevalence is skewed and a model can be right for the wrong reason; the deployed-model evidence for that regime is in Chapter 6 rather than here.

5.6 Targeted LoRA Fine-Tuning

5.6.1 Architecture

We insert LoRA adapters into the attention (query, key, value, output) and MLP (gate, up, down) modules of layers 15 to 19. This layer range is motivated by the mechanistic analysis: Feature 3818 manifests at layer 17, so we target a neighborhood around it. The LoRA configuration uses rank $r = 16$, scaling factor $\alpha = 32$, and dropout 0.05. This adds 4.38M trainable parameters, or 0.10% of the full model. The vision encoder remains frozen: the mechanistic analysis identified text-side circuits as the source of sensitivity, so we leave visual processing unchanged.

The choice of layers 15 to 19 is directly motivated by the mechanistic analysis: Feature 3818 manifests at layer 17, and we target a symmetric neighborhood around it. However, the layer ablation in Section 5.7 will show that this is not necessarily the optimal intervention location. We present the mechanistically-guided layers as the primary configuration and the layer ablation as a systematic exploration.

The LoRA adapters use the standard low-rank factorization: for each weight matrix $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, we add a learned perturbation $\Delta W = BA$ where $B \in \mathbb{R}^{d_{\text{out}} \times r}$ and $A \in \mathbb{R}^{r \times d_{\text{in}}}$ with $r = 16 \ll \min(d_{\text{in}}, d_{\text{out}})$. During training, only A and B are updated; the original weights W remain frozen. At inference time, the perturbation is merged: $W' = W + \frac{\alpha}{r} BA$, adding no latency overhead.

5.6.2 Combined Loss

The central design choice is the loss function. A natural first attempt is to train on consistency alone: minimize the divergence between the model's output distributions on an original question and its paraphrase. This fails catastrophically. With a pure consistency objective ($\lambda = 0$), the model converges to predicting "Yes" for every input, achieving 0% flip rate but 46.4% accuracy. This is mode collapse: the model satisfies the consistency constraint trivially by ignoring the question entirely.

We prevent mode collapse with a combined loss that adds an accuracy term:

$$\mathcal{L}_{\text{consistency}} = \frac{1}{2} \left[D_{\text{KL}}(p_{\text{orig}} \| p_{\text{para}}) + D_{\text{KL}}(p_{\text{para}} \| p_{\text{orig}}) \right] \tag{5.3}$$

$$\mathcal{L}_{\text{accuracy}} = -\log p_{\text{orig}}(y) - \log p_{\text{para}}(y) \tag{5.4}$$

$$\mathcal{L} = \mathcal{L}_{\text{consistency}} + \lambda \cdot \mathcal{L}_{\text{accuracy}} \tag{5.5}$$

where p_{orig} and p_{para} are the softmax distributions over the yes/no logits for the original and paraphrased questions respectively, y is the ground truth label, and λ controls the accuracy weight.

The consistency term uses symmetric KL divergence, which equals zero only when the two distributions match exactly. Neither the original nor the paraphrase distribution is privileged as the reference. The accuracy term applies cross-entropy against ground truth for both the original and paraphrase, ensuring the model remains discriminative. At $\lambda = 1.0$, the two objectives receive equal weight.

5.6.3 Training Procedure

Two training configurations appear in this chapter and should not be conflated. The *headline* results (Section 5.7, Table 5.6) use the operator-preserving expanded fleet trained on the patient-disjoint 656-subject training split: 4,059 same-polarity binary paraphrase pairs, two epochs, five seeds for Targeted LoRA and three for Full LoRA. The layer-ablation, prompt-template, and replication analyses that follow instead use the *original* configuration described in this section: 500 binary paraphrase pairs from the PSF-Med training split, three epochs, one randomly sampled paraphrase per question. These 500 pairs have no overlap with the test set or the FlipBank by question identity and are balanced by answer label (approximately 250 "Yes" and 250 "No") across all five transformation types; because they predate the operator-preserving cleanup, the analyses built on them are reported for their relative

patterns rather than their absolute flip rates.

Hyperparameters: learning rate 2×10^{-4} with linear warmup (100 steps), batch size 8, AdamW optimizer with weight decay 0.01, 3 epochs. Training converges rapidly; the consistency loss drops sharply during epoch 1 while accuracy stays above 80% throughout. The total training time is approximately 20 minutes on a single NVIDIA A100 80GB GPU, making this approach computationally accessible.

Each training step processes a batch of original-paraphrase pairs. For each pair, two forward passes are required, (1) the original question with image to compute p_{orig} and (2) the paraphrased question with image to compute p_{para}, followed by the corresponding two backward passes. The consistency and accuracy losses are computed from the same forward passes and combined according to Equation 5.5.

We use a 15% calibration holdout from the training split to select the best checkpoint (lowest combined loss). The remaining training data is used for gradient updates. This prevents overfitting to the training paraphrases while providing an unbiased calibration signal.

5.6.4 Data Preparation

The 500 training pairs are drawn from the PSF-Med training split, which has no overlap with the 156-question test set or the FlipBank. Each training example consists of an image, an original question with ground truth label, and one paraphrase. For each question the paraphrase is drawn at random from that question's paraphrase set; every candidate in the set has already passed the BioClinicalBERT equivalence filter used to construct PSF-Med (Chapter 3), so the training focuses on near-equivalent reformulations where the model should produce identical outputs regardless of which one is sampled.

The training data includes paraphrases from all five transformation types (lexical, syntactic, formality, scope, negation-adjacent), with the natural distribution reflecting their generation frequency: lexical (32%), syntactic (24%), formality (19%), scope (14%), and negation-adjacent (11%). We do not oversample any category. The label distribution is approximately balanced: 253 positive ("yes") and 247 negative ("no") ground truth labels.

During training, each batch of 8 pairs requires 16 forward passes (8 originals + 8 paraphrases) and the corresponding backward passes. The effective batch size for gradient accumulation is 8 pairs, not 16 examples, because the consistency loss couples the original and paraphrase within each pair. Gradient accumulation over multiple mini-batches was not necessary given the small training set and efficient LoRA parameter count.

5.6.5 Why This Design

Three design choices deserve justification. First, we target the language model only. The mechanistic analysis showed that Feature 3818 responds to text phrasing, not visual features. Freezing the vision encoder leaves the visual representation itself untouched, though Section 5.7.5 shows this is not sufficient to preserve the model's image dependence: the adapted readout leans harder on the question text even with the encoder frozen. Second, we use a combined loss rather than a pure consistency loss. The mode collapse result makes this non-negotiable: without accuracy supervision, the model degenerates. Third, we use symmetric KL rather than alternatives like margin MSE or JS divergence. Symmetric KL provides a smooth gradient signal and naturally handles the asymmetry between over-confident and under-confident predictions.

5.7 Results

5.7.1 Main Results: MIMIC-CXR

The committed split files and evaluation code establish the patient- and image-disjoint design, including the zero-overlap assertions. The per-seed evaluation JSON files named in Table 5.6, however, are not included in the repository snapshot deposited with this dissertation. The numerical results in

Table 5.6: Main results on the *patient- and image-disjoint* MIMIC-CXR held-out test. **Primary endpoint: presence-of-finding** ($n = 238$ questions, 991 same-polarity paraphrase pairs, spanning 236 held-out subjects and 238 held-out images); the all-binary set ($n = 373$, spanning 241 subjects and 243 images) is a secondary. For reference, the held-out split contains 245 subjects and 247 images in total; the difference is questions of other types or without an evaluable same-polarity paraphrase. **Zero** subject overlap and **zero** image overlap with training, asserted in code before any metric is reported. Provenance: the split is generated by `scripts/training/make_patient_disjoint_splits.py` (seed 42, subject-level partition of the 979-subject pool: 656 train / 78 val / 245 test subjects); adapters are trained on the 656-subject split (4,059 same-polarity binary pairs, two epochs; five seeds for Targeted and three for Full LoRA) by `fleet_launch_patient_disjoint.sh`; and `eval_patient_disjoint_lora.py` asserts the zero-overlap condition and writes a per-question manifest to `results/lora_fleet_patient_disjoint/eval_{targeted,full}_s*.json`. Values are mean ± std across seeds. On the primary endpoint Targeted LoRA cuts the pairwise flip rate from 8.5% to $3.5\% \pm 0.6$ (about 59%) with no observed accuracy reduction (84.5% to $84.7\% \pm 1.0$; paired difference +0.25 pp, 95% t interval $[-1.00, +1.51]$), and every seed is significant by paired McNemar (p from 2.5×10^{-7} to 1.8×10^{-3}). Full LoRA reaches a slightly lower pairwise flip rate ($2.9\% \pm 0.3$) but is statistically indistinguishable at the query level (8.0 versus 8.1; Welch $t = 0.08$ on the per-seed means, difference +0.08 percentage points, 95% CI $[-2.4, +2.6]$) and gives up about 1.2 points of accuracy ($83.3\% \pm 0.2$). Secondary results, namely the earlier question-identity-disjoint rerun and the training-size sweep, are reported in Section 5.7 and Appendix A.8.

Model	**Pairwise flip**	**Query-level flip**	**Accuracy**
Primary endpoint: presence-of-finding ($n = 238$ questions, 991 pairs)			
Base MedGemma-4B	8.5%	19.8%	84.5%
+ Targeted LoRA (5 seeds)	**3.5% ± 0.6**	8.1% ± 1.8	**84.7% ± 1.0**
+ Full LoRA (3 seeds)	2.9% ± 0.3	8.0% ± 1.1	83.3% ± 0.2
Secondary: all binary questions ($n = 373$ questions, 1,445 pairs)			
Base MedGemma-4B	8.1%	18.5%	84.5%
+ Targeted LoRA (5 seeds)	2.9% ± 0.7	6.2% ± 1.6	85.5% ± 1.1
+ Full LoRA (3 seeds)	2.2% ± 0.1	5.6% ± 0.7	84.8% ± 0.3

this subsection are preserved from the archived analysis and dissertation tables, but they cannot be regenerated from the committed repository alone. This provenance limit applies to the patient-disjoint flip rates, accuracies, intervals, McNemar tests, and clean-adapter safety re-audit.

We report a clean rerun that fixes three problems the original headline had: the training pool contained operator-changing paraphrases ("Can X be ruled out?") that teach the model the wrong same-polarity answer, the original evaluation artifact was not saved in a form that reproduced the reported counts, and the test set was disjoint from training by question identity only. The rerun repartitions the MIMIC pool at the *subject* level, so that no patient and no image is shared across the train/test boundary; trains on operator-preserving paraphrases from the disjoint 656-subject training split (4,059 same-polarity binary pairs, two epochs, five seeds); and evaluates base and adapted models on the identical paraphrases of the held-out test. The evaluation pipeline is configured to save each prediction to a per-question manifest, but those manifests are absent from the repository snapshot. Following the convention that a detection question is the clinically meaningful endpoint, presence-of-finding is the *primary* endpoint (238 questions, 991 pairs, drawn from 236 held-out subjects and 238 held-out images); the full binary set (373 questions, 1,445 pairs, which also contains view-identification and abnormality questions) is reported as a secondary.

Unit of analysis and sources of variability. Three nested units appear in the numbers that follow, and they are not interchangeable. The pairwise flip rate takes the paraphrase *pair* as its denominator (991 pairs on the primary endpoint), pairs are nested within questions (238), and questions are nested within images (238) and within subjects (236). The query-level flip rate takes the *question* as its denominator and counts a question once regardless of how many paraphrases it carries, which is why the two rates differ in magnitude and why only the query-level rate answers the clinician-facing question of whether an answer changes at all. Nesting is close to degenerate above the question on this split, since the 238 presence questions come from 238 held-out images and 236 held-out subjects, so the units where dependence actually accumulates are the pair within the question and the question within the image, not the patient. Accuracy, text-only agreement, and swap sensitivity are all question-level quantities. Separately, every $\pm$ attached to a flip rate, an accuracy, or an agreement figure in this chapter is a standard deviation across *training seeds*: it describes how much the training procedure varies from one seed to the next, and it is not an interval for the test cohort. The two quantities answer different questions and are not pooled into one figure. The intervals reported here, the $[-1.00, +1.51]$ percentage-point accuracy interval and the $[-2.4, +2.6]$ Welch interval, are likewise computed over per-seed means and are therefore also training-variability statements. Clustered intervals at the pair, question, and image levels are not computed in this chapter; the resampling helper written for that recomputation is `scripts/analysis/cluster_bootstrap.py`, and no number reported here uses it.

Across five seeds (Table 5.6), the targeted LoRA cuts the presence-only pairwise flip rate from 8.5% to $3.5\% \pm 0.6$ (about a 59% reduction, the $\pm$ being a standard deviation across the five training seeds) and the query-level rate from 19.8% to $8.1\% \pm 1.8$, with no observed accuracy reduction (84.5% to $84.7\% \pm 1.0$; paired per-question difference +0.25 percentage points, 95% t interval $[-1.00, +1.51]$ across the five seeds). We report the interval rather than claiming equivalence: the data are consistent with anything from a one-point loss to a 1.5-point gain, so they exclude a large accuracy cost but do not establish exact parity. Every seed is significant by paired McNemar (p from 2.5×10^{-7} to 1.8×10^{-3}; 37 to 41 questions flip for the base model only, against 5 to 14 for the adapted model only). On the secondary all-binary set the pairwise rate falls from 8.1% to $2.9\% \pm 0.7$ and accuracy is unchanged (84.5% to $85.5\% \pm 1.1$). The base pairwise rate (8.5%) is lower than the 14.6% reported before the operator-preserving cleanup, because the earlier figure was inflated by operator-changing pairs whose flips were correct behavior.

The generalization claim can now be read directly. The test set is disjoint from training by *patient*

and by *image*: the pool was repartitioned at the subject level, and the evaluation script asserts zero subject overlap and zero image overlap before it reports any metric (the held-out split spans 245 subjects and 247 images in total; the primary presence endpoint draws on 236 of those subjects and 238 of those images). The reduction is therefore measured on patients and studies the adapter never saw, which is what a deployment claim requires. An earlier rerun of the same pipeline was disjoint by question identity only, with 145 of its 156 questions reusing a training-seen image and subject; it reported a larger reduction (10.5% to $3.1\% \pm 0.9$ pairwise), so the seen-study setting modestly overstated the effect while the effect itself survives on unseen patients. Two limits remain: the split is disjoint by patient but not by *institution* (both sides draw from the same MIMIC source), and the out-of-distribution transfer to PadChest, a genuinely disjoint institution, is reported separately below and is weaker.

The archived analysis reports statistical significance in every seed. Because the same questions are evaluated for each model, the paired McNemar test is the appropriate one. The recorded discordant counts are 37 to 41 questions that flip only for the base model and 5 to 14 that flip only for the adapted model, giving exact two-sided p between 2.5×10^{-7} and 1.8×10^{-3} across the five seeds. The per-question artifacts from which these counts were computed are not present in the repository snapshot, as stated above.

Training-set size is no longer a concern: the result has plateaued. On the earlier question-identity-disjoint split, sweeping the training set from a 500-sample subset to 1,288 samples to that run's full expanded pool moves the targeted pairwise flip rate from 5.1% to 3.4% to 3.1%, with the last two within the seed-to-seed confidence interval, so more data no longer helps and the roughly 3% floor reflects the method rather than the data. The replication study (Appendix A.8) shows the same plateau on MIMIC while out-of-distribution transfer keeps improving with more data. The binary pool caps at 1,288 distinct questions (1,000 local images); larger-scale multi-task adapters trained on roughly 12,000 pairs are released separately.

Full LoRA (all 34 layers, three seeds) reaches a slightly lower pairwise flip rate on the primary endpoint ($2.9\% \pm 0.3$ over three training seeds versus $3.5\% \pm 0.6$ over five for the targeted adapter, each $\pm$ a standard deviation across seeds), as expected from its larger capacity, but the advantage is narrower than it first appears. The two are statistically indistinguishable at the query level (Welch $t = 0.08$ on the per-seed means, difference +0.08 percentage points, 95% CI $[-2.4, +2.6]$) ($8.0\% \pm 1.1$ versus $8.1\% \pm 1.8$), which is the level a clinician experiences, namely whether the answer changes at all for a given question; and Full LoRA gives up about 1.2 points of accuracy ($83.3\% \pm 0.2$ against a base of 84.5% and $84.7\% \pm 1.0$ for the targeted adapter). So on patient-disjoint data the targeted adapter is the better accuracy-consistency trade on its own terms, before any image-reliance argument is made. Whether Full LoRA also buys its extra pairwise consistency by leaning more heavily on the text prior is tested directly on these same adapters in Section 5.7.5.

The layer-ablation, prompt-template, cross-model, and PadChest-transfer analyses that follow were run on the original training pipeline and are reported for their relative patterns; the clean rerun above re-establishes the central headline (a roughly 59% flip-rate reduction, from 8.5% to 3.5%, with no mode collapse) on operator-preserving data.

5.7.2 Layer Ablation

We train five LoRA configurations targeting different layer ranges: early (0 to 10), a randomly chosen contiguous block (5 to 9), all (0 to 33), middle (15 to 19), and late (25 to 33). We evaluate on the 355-question validation split, which provides more statistical power than the 156-question test set. Table 5.7 reports the resulting margin differences.

The results are surprising (Figure 5.14). Early layers produce the largest margin reduction: 0.26 mean margin difference (86% reduction from the 1.87 baseline). The random subset (layers 5 to 9) achieves 0.30 (84%). All layers reach 0.34 (82%). The mechanistically-targeted middle layers achieve

Table 5.7: Layer-range ablation on the MIMIC-CXR validation split ($n = 355$). All configurations use rank 16, $\alpha = 32$, combined loss with $\lambda = 1.0$. Early layers reduce margin difference most despite not containing the mechanistically identified Feature 3818 at layer 17.

Layer Range	**Layers**	**Margin Diff**	**% Reduction**	n **Layers**
Baseline (no LoRA)	N/A	1.87	N/A	N/A
Early	0 to 10	**0.26**	**86%**	11
Random block	5 to 9	0.30	84%	5
All	0 to 33	0.34	82%	34
Middle	15 to 19	0.38	80%	5
Late	25 to 33	0.70	63%	9

0.38 (80%). Late layers perform worst at 0.70 (63%).

Early layers outperform the mechanistically-guided middle layers by 6 percentage points in relative margin reduction. This means the SAE analysis does not prescribe the optimal intervention location. Feature 3818 at layer 17 tells us where register sensitivity *manifests*, but the most effective fix acts upstream, modifying representations before the register distinction forms. This is analogous to treating a disease at its origin rather than at the site of symptoms.

The layer ablation also shows that full-model LoRA (all 34 layers) does not outperform targeted ranges. Adding more layers adds more trainable parameters but does not improve consistency. This suggests the consistency signal is learnable from a small number of layers, and the additional parameters introduce noise or overfit to training-specific paraphrase patterns.

5.7.3 Extended Block Sweep

To move beyond the five-configuration ablation, we conduct an exhaustive block sweep across all possible contiguous 5-layer windows in the 34-layer model, plus 20 randomly sampled non-contiguous 5-layer subsets. This produces 50 LoRA configurations in total (30 contiguous + 20 random), each trained with the same hyperparameters and evaluated on the same validation set.

The contiguous sweep reveals a U-shaped pattern in margin reduction (Figure 5.15): early windows (layers 0 to 4 through 4 to 8) achieve the strongest flip reduction, middle windows (10 to 14 through 18 to 22) show moderate performance, and late windows (28 to 32, 29 to 33) show the weakest performance. The global minimum in margin difference occurs at layers 1 to 5 (0.24 logits, 87% reduction from baseline), slightly outperforming the early-layer window 0 to 10 from the five-configuration ablation.

The random subsets provide a natural control. Non-contiguous sets of 5 layers drawn uniformly at random achieve margin reductions ranging from 0.28 to 0.62 logits. The best random subsets (e.g., layers {2, 5, 7, 24, 29}) approach but do not exceed the best contiguous windows, suggesting that layer adjacency is mildly beneficial but not essential.

The block sweep confirms the main finding from the five-configuration ablation: the optimal intervention location is upstream of the mechanistically-identified layer 17. It also establishes that any 5-layer window in the first half of the model (layers 0 to 16) achieves substantial flip reduction with little accuracy cost (Figure 5.16); windows in the second half are consistently weaker. This gradient (strong early intervention, weak late intervention) is consistent with the hypothesis that paraphrase sensitivity is best treated before the register-sensitive representation forms.

5.7.4 Comparison: Targeted vs. Full LoRA

A natural question is why not simply train LoRA on all 34 layers, maximizing the model's ability to learn consistency? The answer is that Full LoRA (all 34 layers, 30.5M parameters) achieves the lowest flip rate but at a cost first reported in Chapter 6: on the original-pipeline adapters its answers were

far more often invariant under image removal and far less often changed by an image swap, so more of its consistency came from the question text than from the image.

On the MIMIC-CXR flip bank (n=98), Full LoRA reaches 4.1% flip rate versus Targeted LoRA's 20.4%. But Full LoRA's text-only agreement is 80.6% versus 66.3% for Targeted LoRA. Swap sensitivity drops to 14.3% for Full LoRA versus 32.7% for Targeted LoRA. On these original-pipeline adapters, and on this n=98 flip bank alone, the additional 26.1M parameters (from 4.38M to 30.5M) buy a lower flip rate at the expense of image reliance, with the image swap rather than image removal carrying the weight of that reading.

This comparison motivates the central argument of Chapter 6: flip rate alone is an inadequate safety metric. This ordering, however, is specific to the original-pipeline adapters and does not replicate on the clean, patient-disjoint adapters, where the two are tied on image reliance and both are markedly more text-reliant than the base model (Section 5.7.5); the targeted adapter is the better intervention on accuracy and cost, not on grounding. On those original adapters, then, Full LoRA had the better flip-rate number and the worse image-reliance profile; Section 5.7.5 shows that this particular contrast does not survive a clean audit.

5.7.5 Safety Re-Audit of the Clean Adapters

The comparison just given, and the image-reliance audit of Chapter 6, both rest on adapters trained by the original pipeline. Because the mitigation result above is measured on the clean, patient-disjoint adapters, we re-ran the image-reliance audit on those exact adapters, using the same protocol as Chapter 6 and the same patient- and image-disjoint held-out set. The audit reports a slightly larger population than the flip-rate table above: text-only agreement and swap sensitivity are defined on a single question, whereas a flip rate requires at least one paraphrase to compare against. The presence endpoint therefore carries 241 questions here against 238 there, and the all-binary endpoint 382 against 373; the three and nine additional questions are those the equivalence audit left without a surviving paraphrase. The two tables are computed on the same split and the same adapters, and the small difference in n is this and nothing else.

Table 5.8: Image-reliance re-audit of the *clean, patient-disjoint* adapters, using the same protocol as Chapter 6: the text-only condition removes the image token from the prompt entirely, and the swap condition re-pairs each question with another test image. All measurements are on the patient- and image-disjoint held-out set (zero subject and zero image overlap with training). Higher text-only agreement means more text-reliant; higher swap sensitivity means more image-reliant. Two results stand out. First, the targeted and full adapters are statistically indistinguishable on both image-reliance measures, so the ordering reported in Chapter 6 for the original-pipeline adapters (targeted 66.3% text-only agreement and 32.7% swap sensitivity against full 80.6% and 14.3%) does not replicate here. Second, *both* adapters raise text-only agreement sharply relative to the base model (53.5% to about 77% on the presence-only endpoint) while slightly lowering swap sensitivity, so both buy a substantial part of their consistency from the question text rather than from more stable use of the image.

Model	**Text-only agree $\uparrow_{\text{risk}}$**	**Swap sens. $\uparrow$**	**Accuracy**
Primary endpoint: presence-of-finding ($n = 241$)			
Base MedGemma-4B	53.5%	**22.0%**	84.7%
+ Targeted LoRA (5 seeds)	76.8% ± 3.1	19.9% ± 1.2	**84.3% ± 1.0**
+ Full LoRA (3 seeds)	76.6% ± 1.7	20.9% ± 1.3	82.4% ± 0.5
Secondary: all binary questions ($n = 382$)			
Base MedGemma-4B	67.0%	18.1%	84.5%
+ Targeted LoRA (5 seeds)	69.8% ± 2.7	18.5% ± 0.4	85.2% ± 1.1
+ Full LoRA (3 seeds)	69.4% ± 1.8	19.7% ± 1.3	84.5% ± 0.4

The earlier ordering does not replicate. On the presence-only endpoint the targeted and full adapters are statistically indistinguishable on both image-reliance measures: text-only agreement is 76.8% ± 3.1

versus 76.6% ± 1.7, and swap sensitivity is 19.9% ± 1.2 versus 20.9% ± 1.3, with the full adapter marginally the more swap-sensitive of the two; each ± here is a standard deviation across training seeds, five for the targeted adapter and three for the full adapter, and not an interval for the test cohort. The contrast reported for the original-pipeline adapters (targeted 66.3% text-only agreement and 32.7% swap sensitivity against full 80.6% and 14.3%) is therefore specific to those adapters and that evaluation set. On the clean adapters, *the claim that targeted adaptation preserves image dependence relative to full adaptation is not supported.*

The more consequential observation is what both adapters do relative to the base model. Text-only agreement rises from 53.5% to about 77% for both, a 23-point increase, while swap sensitivity falls slightly (22.0% to 19.9% and 20.9%). Both interventions therefore buy a substantial part of their consistency by leaning harder on the question text rather than by making more stable use of the image. This is the dissertation's own central warning applied to its own mitigation: a low flip rate does not certify grounding, and consistency training does not become exempt from that warning by being guided by a mechanistic localization. Read together with Thrust 4, the honest summary is that the targeted adapter reduces paraphrase sensitivity without improving grounding, and partly at the expense of it.

This changes the basis of the recommendation rather than the recommendation itself. The targeted adapter remains preferable to the full adapter, but for the reason that survives a clean audit: at statistically indistinguishable query-level consistency (8.1% ± 1.8 versus 8.0% ± 1.1; (Welch $t = 0.08$ on the per-seed means, difference +0.08 percentage points, 95% CI $[-2.4, +2.6]$)) and tied image reliance, it preserves accuracy (84.3% ± 1.0 versus 82.4% ± 0.5, against a base of 84.7% on this n=241 audit set; the base moves with the evaluation set, so the primary n=238 endpoint of Table 5.6 has a base of 84.5% and a paired difference of +0.25 points) while modifying 0.1% of parameters rather than 0.72%. It is the cheaper and more accurate route to the same consistency, not a more grounded one. Any deployment of either adapter must therefore be paired with the image-reliance checks of Chapter 6, not justified by the flip-rate reduction alone.

5.7.6 Lambda Sweep

Table 5.9: Lambda sweep for the combined loss $\mathcal{L} = \mathcal{L}_{\text{consistency}} + \lambda \cdot \mathcal{L}_{\text{accuracy}}$ on the MIMIC-CXR validation split. A flat Pareto front exists for $\lambda \in [0.1, 1.0]$: flip rate ranges from 3.9% to 4.6% while accuracy ranges from 85.6% to 88.2%. Pure consistency ($\lambda = 0$) causes mode collapse; accuracy-only training yields nearly double the optimal flip rate.

λ	Accuracy (%)	Flip Rate (%)	Margin Diff
0 (consistency only)	46.4	0.0	0.11
0.1	85.6	3.9	0.18
0.3	88.2	3.9	0.25
0.5	85.9	4.6	0.31
1.0	**88.2**	**3.9**	0.37
2.0	87.3	4.6	0.45
5.0	87.3	5.9	0.51
∞ (accuracy only)	87.3	7.2	0.55

We train 8 configurations with $\lambda \in \{0, 0.1, 0.3, 0.5, 1.0, 2.0, 5.0, \text{accuracy-only}\}$, evaluated on a 153-question MIMIC-CXR validation subset (a smaller held-out set than the 355-question split used for the layer ablation in Table 5.7). The results reveal a flat Pareto front for $\lambda \in [0.1, 1.0]$ (Table 5.9): flip rate ranges from 3.9% to 4.6% while accuracy ranges from 85.6% to 88.2%. Our chosen $\lambda = 1.0$ sits at the optimal point (3.9% flips, 88.2% accuracy).

At the extremes, both objectives fail when used alone. At $\lambda = 0$ (consistency only), the model

achieves 0% flips and 46.4% accuracy: mode collapse. An accuracy-only LoRA ($\lambda \to \infty$) yields 7.2% flips at 87.3% accuracy, nearly double the optimal flip rate. At $\lambda = 5.0$, the accuracy term dominates so heavily that consistency degrades to 5.9% flips. The flat Pareto front means the method is not sensitive to the exact λ value within a reasonable range.

5.7.7 Prompt Template Stability

Table 5.10: Prompt template stability on the MIMIC-CXR test set ($n = 97$ pairs). The LoRA reduces flip rate across all five templates, confirming it modifies internal representations rather than learning surface-level associations with the training prompt. Margin-based metrics are invariant to decoding strategy (greedy, temperature 0.3, nucleus $p = 0.9$, beam $k = 4$).

Template	**Flip Rate (%)**		**Margin Diff**	
	Base	+ LoRA	Base	+ LoRA
Default	12.4	**6.2**	1.703	0.339
Instruction-first	15.5	13.4	1.049	0.436
Question-first	15.5	13.4	0.541	0.318
Bare	14.4	10.3	1.035	0.685
Role-primed	15.5	**4.1**	1.597	0.402

A concern with any fine-tuning approach is that it might learn prompt-specific patterns rather than genuine consistency. We evaluate on the 97-pair MIMIC test set across five prompt templates (Table 5.10): default (the training template), instruction-first, question-first, bare (no system prompt), and role-primed ("You are a radiologist."). We also test four decoding strategies: greedy, temperature 0.3, nucleus $p = 0.9$, and beam search $k = 4$.

Decoding strategy has zero effect on margin-based metrics. Margins are computed from a single forward pass and do not depend on sampling. The LoRA reduces flip rate across all five templates: default 12.4% $\to$ 6.2%, instruction-first 15.5% $\to$ 13.4%, question-first 15.5% $\to$ 13.4%, bare 14.4% $\to$ 10.3%, role-primed 15.5% $\to$ 4.1%. The improvement is template-general, confirming the LoRA modifies internal representations rather than learning a surface-level association with the training prompt.

5.7.8 Cross-Dataset Generalization: PadChest

Table 5.11: Cross-dataset generalization to PadChest (balanced $n = 250$: 125 yes, 125 no). The LoRA was trained exclusively on MIMIC-CXR. Margin reduction transfers strongly (75.4%); flip rate reduction is modest on this already-consistent balanced set (10.5%); accuracy improves by 6.4 percentage points.

Model	**Accuracy**	**Margin Diff**	**Flip Rate**
Base MedGemma-4B	85.2%	1.02	7.6%
+ Targeted LoRA	**91.6%**	**0.25**	**6.8%**
Change	+6.4 pp	−75.4%	−10.5%

PadChest provides an out-of-distribution stress test: different institution (Spain vs. United States), different imaging equipment, different patient demographics. On 250 balanced PadChest questions (125 yes, 125 no), the LoRA improves accuracy from 85.2% to 91.6% (+6.4 percentage points) and reduces the margin difference from 1.02 to 0.25 (75.4% reduction, Table 5.11). The base model is already fairly consistent on this balanced set (7.6% flip), so the flip rate changes little (to 6.8%); the transfer shows up in accuracy and margin sharpening rather than flip reduction.

The transfer is partial. On this balanced PadChest set the base model already flips only 7.6% of the time, so there is little flip rate to reduce; the LoRA instead transfers as a large margin sharpening

(75% reduction) and an accuracy gain. This is expected: the LoRA was trained exclusively on MIMIC-CXR data, and distribution shift reduces any fine-tuning's effectiveness. But the accuracy *increase* on PadChest is notable. The LoRA does not simply memorize MIMIC-specific patterns; it learns something about paraphrase-invariant clinical reasoning that transfers across institutions.

5.7.9 Replication Study

Table 5.12: Replication study across 5 random seeds (42, 123, 456, 789, 2024) at two training set sizes. The MIMIC-CXR flip rate is stable across seeds (low variance). Quadrupling the training data from $n = 500$ to $n = 2{,}000$ yields less than 0.5 pp additional gain, indicating the consistency signal is learnable from a few hundred examples.

Training Size	**MIMIC Flip Rate (mean ± std)**	**95% CI**
$n = 500$	4.6% ± 0.3%	[4.1%, 5.1%]
$n = 2{,}000$	4.3% ± 0.2%	[3.9%, 4.7%]

We replicate the main result across 5 random seeds (42, 123, 456, 789, 2024) at two training set sizes ($n = 500$ and $n = 2{,}000$; Table 5.12). At $n = 500$, the mean MIMIC flip rate is 4.6% ± 0.3% (standard deviation across the five training seeds), confirming low training variance. At $n = 2{,}000$, the mean drops slightly to 4.3% ± 0.2% on the same seed-to-seed basis, less than 0.5 percentage points of additional gain. The small marginal benefit of quadrupling the training data suggests the consistency signal is learnable from a few hundred examples.

5.7.10 Cross-Model Feature Validation

To assess whether Feature 3818's role is specific to Base MedGemma-4B or extends to adapted variants, we extract SAE features for three model configurations (Base, Targeted LoRA, Full LoRA) across both datasets. For each configuration, we rank features by their association with paraphrase flips using a point-biserial correlation.

On PadChest, Feature 3818 ranks #1 for both Targeted LoRA ($r_{pb} = -0.462$) and Full LoRA ($r_{pb} = -0.776$): when the feature is active, flip rate drops to 30%, versus 65% when inactive. On MIMIC, Feature 3818 ranks #2 for Base ($r_{pb} = -0.430$), confirming its role in-distribution. However, Targeted LoRA *suppresses* Feature 3818 on MIMIC (rank 33), while leaving it active on PadChest (rank 1). This dissociation reveals that the LoRA's consistency improvement on MIMIC involves learning to ignore the register gate in-distribution, but the feature reasserts itself out-of-distribution.

Layer-level analysis shows that early-layer features (layer 9) are near-identical across all models (Jaccard similarity 0.87 to 1.0 for the top 50 features), while layer 17 features diverge substantially (Jaccard 0.30 to 0.45). Layers 22 and 29 show model-specific feature usage. The LoRA's targeted intervention on layers 15 to 19 reshapes the mid-layer feature space without affecting early representations, explaining why the mechanism manifests differently across configurations.

A causal patching experiment tests whether restoring Feature 3818's activation from the original question can recover flips in the paraphrase. For Targeted LoRA on PadChest, patching recovers 16.7% (4/24) of OOD flips, versus 0% for matched control features. On MIMIC, recovery is 0% for all models, consistent with the LoRA having already suppressed the feature in-distribution. This confirms Feature 3818's causal role is OOD-specific for the adapted model.

5.7.11 SAE Validity After LoRA

A natural concern is whether the LoRA modifies activations so much that the SAE becomes invalid. This is a separate question from the base-model transfer validated in Section 5.2 ($R^2 = 0.997$ on unmodified MedGemma-4B): here we ask only whether the additional Targeted LoRA adaptation degrades reconstruction. We recompute FVU on 98 MIMIC samples for both the base model and the

LoRA-adapted model. Base FVU is $0.33\% \pm 0.02\%$. LoRA FVU is $0.50\% \pm 0.06\%$. The increase of 0.17 percentage points is negligible: both values are well below 1%. Feature 3818's mean activation barely changes (211.8 vs. 213.5). The SAE remains a valid analysis tool after LoRA adaptation, since the LoRA modifies only 5 of 34 layers at rank 16.

5.8 Discussion

The two-stage account is discussed further in Appendix A.4: the cross-dataset differences in restoration efficacy are compositional (driven by flip-direction and phenomenon mix) rather than mechanistic, and the account identifies the dominant decision-stage features rather than the complete circuit.

5.8.1 Mechanisms vs. Interventions

The most important finding is the dissociation between mechanistic diagnosis and optimal intervention. The SAE analysis identifies layer 17 as the site where register sensitivity is represented. But the layer ablation shows that early-layer interventions (0 to 10) outperform middle-layer interventions (15 to 19). The mechanism tells us *what* the model computes and *where* that computation manifests, but the best place to intervene is upstream, before the problematic representation forms.

This dissociation is not unique to our setting. In medicine, the site of a symptom is often different from the site of effective treatment. The mechanistic analysis provides a testable hypothesis about the nature of the problem (register sensitivity), even if the intervention acts at a different level.

5.8.2 Why Early Layers Outperform Middle Layers

Several hypotheses can explain the dissociation between the mechanistic diagnosis (layer 17) and the optimal intervention (layers 0 to 10):

Hypothesis 1: Prevention vs. treatment. Early-layer interventions prevent the register-sensitive representation from forming in the first place. By modifying the residual stream at layers 0 to 10, the LoRA alters the input to layers 11 to 17, changing what features are activated. Feature 3818 at layer 17 simply does not receive the input pattern that triggers register-dependent activation. This is analogous to vaccination (preventing the disease) versus treatment (addressing symptoms after they appear).

Hypothesis 2: Information bottleneck. Early transformer layers perform initial text-image integration. The register distinction (presence vs. exclusion framing) is a relatively simple syntactic property that is encoded early in processing. By layer 17, the register information has been integrated with image features and is harder to modify without disrupting other computations. Early-layer intervention modifies the register signal before it becomes entangled with visual processing.

Hypothesis 3: Gradient flow. LoRA adapters at early layers may receive stronger gradient signals for the consistency loss because they are closer to the input and further from the output softmax. This could make early-layer LoRA more efficient at learning the consistency signal, independent of mechanistic considerations.

The block sweep data is most consistent with Hypothesis 1. The U-shaped pattern (strong early, moderate middle, weak late) maps onto the typical information flow in transformers: early layers encode syntactic features, middle layers perform composition, and late layers decode to output space. Intervening at the syntactic encoding stage (where register is first represented) is more effective than intervening at the composition stage (where register has already influenced feature combinations).

This finding has broader implications for mechanistic interpretability. The common practice of intervening at the layer where a feature is detected may not be optimal. The SAE identifies *where* a computation happens, but the best intervention may be upstream, at the layer where the inputs to that computation are formed.

5.8.3 The Mode Collapse Problem

Mode collapse under pure consistency training has a straightforward explanation. The consistency loss $\mathcal{L}_{\text{consistency}}$ is minimized when $p_{\text{orig}} = p_{\text{para}}$. The easiest way to achieve this is to make both distributions identical for *all* inputs, regardless of the question or image. A model that always predicts "Yes" satisfies this constraint perfectly. The accuracy term breaks this degeneracy by requiring the model to match ground truth, which forces it to discriminate between images.

This result has practical implications. Any approach to improving paraphrase consistency, not just LoRA, needs an anchoring mechanism to prevent trivial solutions. The anchoring need not be ground-truth labels; pseudo-labels from high-confidence base model predictions, contrastive objectives in activation space, or self-distillation with a moving-average teacher could serve the same role. But some form of anchor is necessary.

5.8.4 Relation to Other PEFT Methods

The LoRA approach used in this work is one of several Parameter-Efficient Fine-Tuning (PEFT) methods. QLoRA [118] applies LoRA to quantized models, reducing memory requirements at the cost of quantization noise. Prefix tuning prepends learned continuous vectors to the input, modifying model behavior without changing any weights. Adapter layers insert small trainable bottleneck modules between frozen layers.

We chose LoRA because it offers two properties critical for our application. First, it modifies specific layer ranges, enabling the mechanistically-guided targeting that distinguishes our approach from blanket fine-tuning. Prefix tuning and adapter methods operate at different granularities that do not map cleanly onto the layer-level mechanistic analysis from the SAE. Second, LoRA adapters can be merged into the base weights at inference time ($W' = W + \frac{\alpha}{r}BA$), adding zero latency overhead, so the deployed model runs at the base model's inference speed.

A comparison with full fine-tuning is also instructive. Full fine-tuning modifies all 4B parameters and typically requires more training data and compute to avoid catastrophic forgetting. The LoRA's 0.1% parameter budget forces it to learn a low-rank perturbation that captures the consistency signal without overwriting the base model's visual processing capabilities. On the original-pipeline adapters this looked like an implicit regularizer against text-only shortcuts, but the clean patient-disjoint re-audit does not support that reading: the two adapters tie on text-only agreement at about 77%. The parameter budget buys cost and accuracy, not grounding.

5.8.5 Training Dynamics

The training dynamics reveal how the combined loss balances its two objectives. During epoch 1, the consistency loss drops sharply (from $\sim$0.8 to $\sim$0.15) as the model learns to align distributions across paraphrased inputs. The accuracy loss remains approximately flat (0.55 to 0.50), indicating that the model maintains discriminative capability throughout consistency learning. During epochs 2 and 3, both losses plateau, with the consistency loss reaching 0.05 and accuracy loss reaching 0.45.

The rapid convergence of the consistency loss (within 1 epoch) on 500 training pairs suggests that the paraphrase-invariant signal is relatively simple for the model to learn. This is consistent with the Feature 3818 analysis: the model already has the representational capacity for consistent processing, and the LoRA primarily needs to suppress the register-sensitive pathway rather than build new capabilities. The early convergence also explains why increasing the training set from 500 to 2,000 samples yields only marginal improvement (4.6% to 4.3% flip rate); the signal is learned quickly, and additional data provides diminishing returns.

5.8.6 Clinical Implications

From a deployment perspective, these results suggest two recommendations. First, evaluation of medical VLMs should include consistency metrics (flip rate, margin difference) alongside accuracy. A model with 90% accuracy and 15% flip rate gives different answers to the same clinical question one time in seven, which is unacceptable for diagnostic use. Second, targeted fine-tuning can reduce sensitivity without large compute costs. The LoRA adds 4.38M parameters, trains on 500 examples in under 20 minutes on a single A100, and generalizes partially to unseen institutions.

5.8.7 Computational Requirements

The Targeted LoRA approach is designed for accessibility. Training requires a single NVIDIA A100 80GB GPU and completes in approximately 20 minutes. The 4.38M trainable parameters require approximately 128MB of optimizer state (AdamW with first and second moment estimates). The total GPU memory footprint during training is approximately 24GB: 16GB for the frozen base model, 4GB for activations, 3GB for gradients, approximately 128MB for optimizer state, and the remainder for framework and allocator overhead.

At inference time, the LoRA adapters are merged into the base weights ($W' = W + \frac{\alpha}{r}BA$), eliminating any latency overhead. The merged model has identical architecture and inference cost to the base model. This is a key advantage of LoRA over architectural modifications (adapter layers, prefix tuning) that add parameters and latency at inference time.

The feature clamping baseline from the PSF-Med benchmark [119] adds approximately 12ms per forward pass and 128MB of memory for the SAE. While modest, this overhead is permanent (it cannot be removed at inference time) and the intervention is less effective (31% relative flip reduction vs. about 59% for LoRA on the clean patient-disjoint operator-preserving rerun). The combined approach (feature clamping + LoRA) was not explored because the LoRA alone achieves sufficient flip reduction.

Chapter 6 examines whether the LoRA's consistency improvement is *safe*: does the more consistent model still rely on the image, or has it, like the models examined in Chapter 4, achieved consistency by leaning on the question text? Section 5.7.5 has already given the answer for the clean adapters, and Chapter 6 sets it in the wider four-quadrant frame.

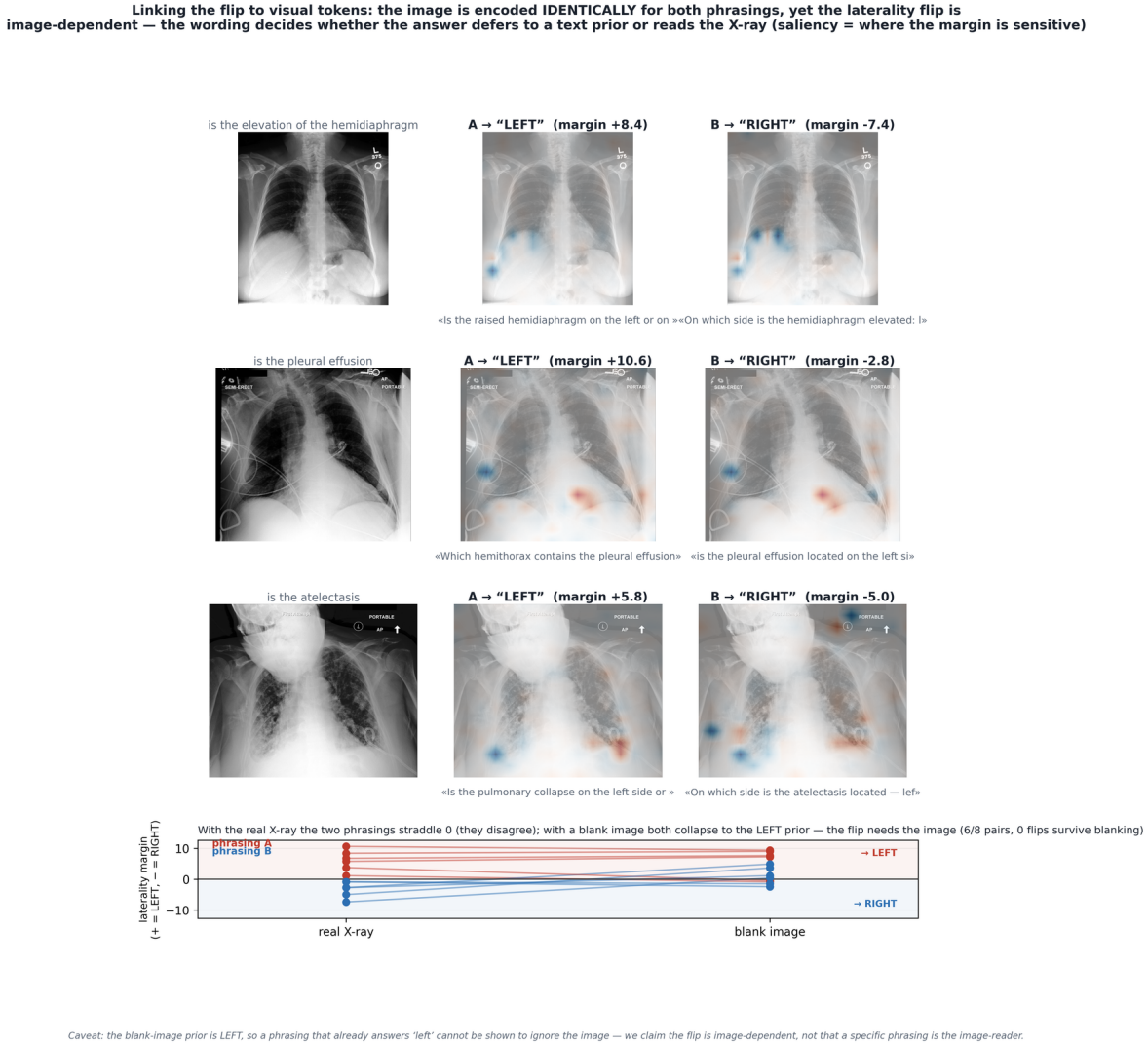


Figure 5.9: The paraphrase flip is a reading of a shared image. For three laterality switches the model answers opposite sides to two phrasings of the same question on the same chest X-ray; gradient saliency of the left-versus-right margin concentrates at the lung bases for both phrasings (top rows). Because the 256 image tokens precede the question under causal masking, their residuals are identical across phrasings to machine precision, so the flip is entirely in the readout; yet blanking the image collapses every disagreement (bottom), so the readout genuinely depends on the shared image.

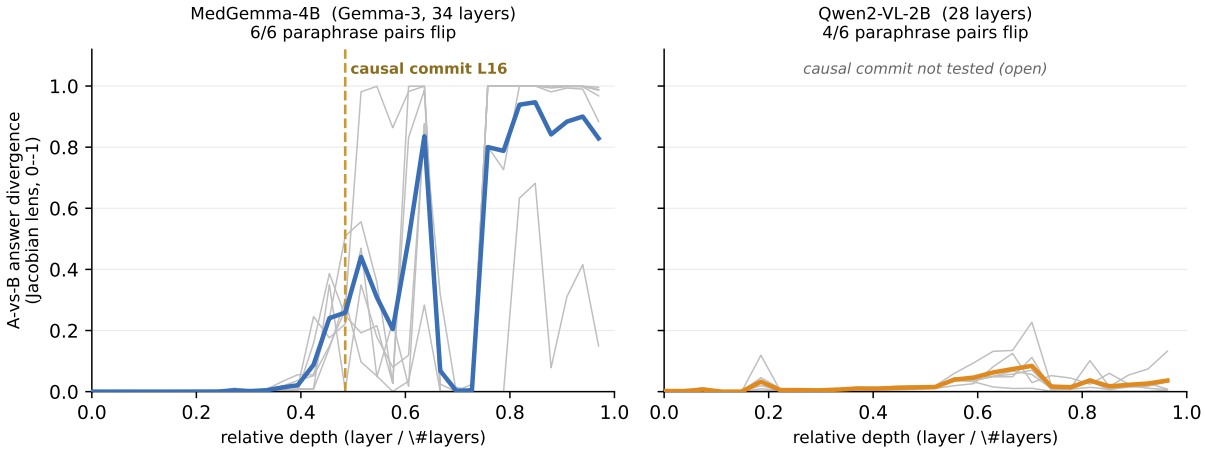


Figure 5.10: The lens and the phenomenon port across architectures; the causal localization is confirmed only on MedGemma. Per-layer Jacobian-lens divergence between two phrasings of the same question (six shared PadChest cases, thin gray; mean in color), plotted against relative depth. On MedGemma-4B (left) the divergence localizes near the causally established commit layer 16. On Qwen2-VL-2B (right) the paraphrase flip replicates behaviorally (four of six pairs), but the text-fit lens reads Qwen's answer position less faithfully, so its divergence is weak: the low signal reflects lens faithfulness, not model robustness. Establishing where Qwen commits requires the lens-free causal patch, which has not been run on Qwen. The lens is corroboration only in both models.

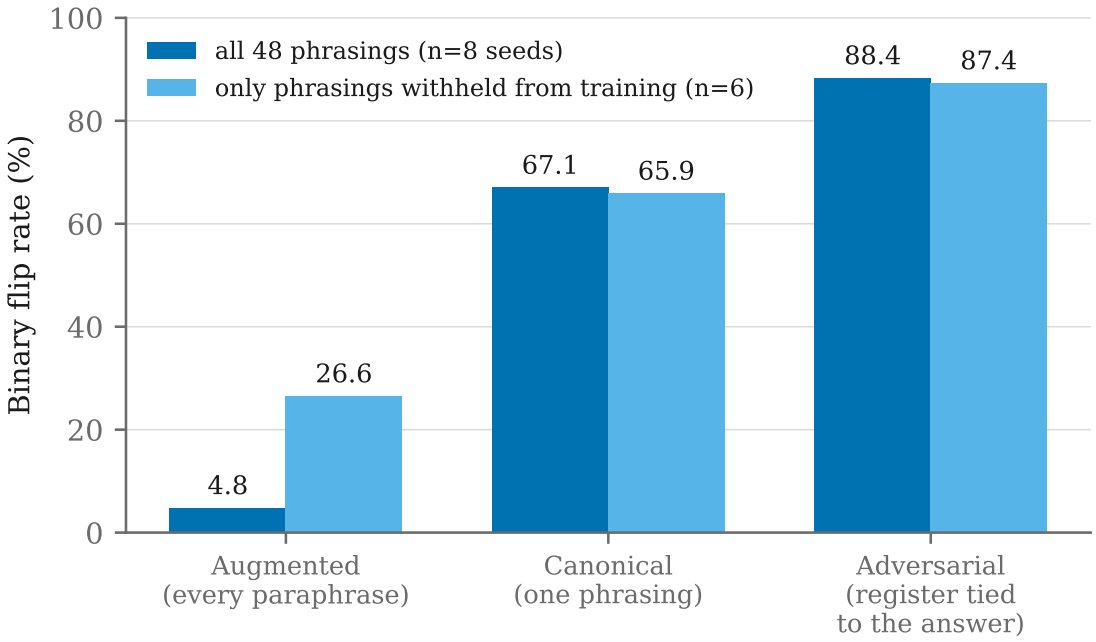


Figure 5.11: The training-phrasing distribution sets paraphrase sensitivity, and the effect survives on wording the model never saw. Binary flip rate of the probe under three training regimes with architecture, parameter budget, seed, and evaluation held fixed. Dark bars train and score on the full bank of 48 paraphrases (eight seeds); light bars train on half the bank and score only on the 24 phrasings withheld from training (six seeds), which separates invariance from recognition of familiar wording. Text-only accuracy is 50.0% to 50.3% in every regime, so none of them can exploit a prior over answers.

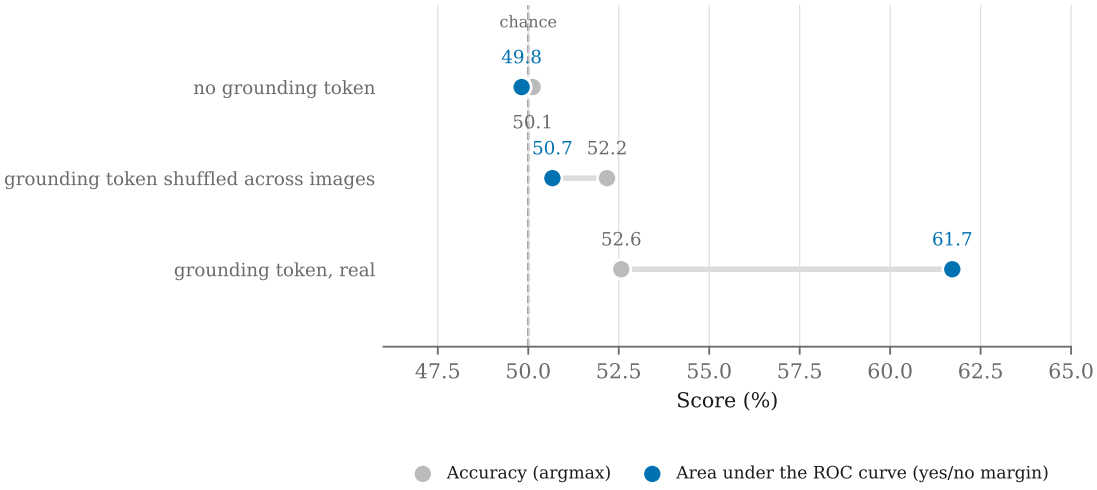


Figure 5.12: Binary accuracy cannot separate a model that reads the radiograph from one that cannot see it. Three variants of the earlier probe on a balanced held-out set: the decoder receives no pooled image summary, receives one belonging to a different radiograph, or receives its own. Accuracy (grey) is flat across all three and close to chance, while the area under the receiver operating characteristic curve of the same yes-minus-no margin (blue) separates the correct summary from the other two. The gap between the paired points is the evidence a binary readout discards.

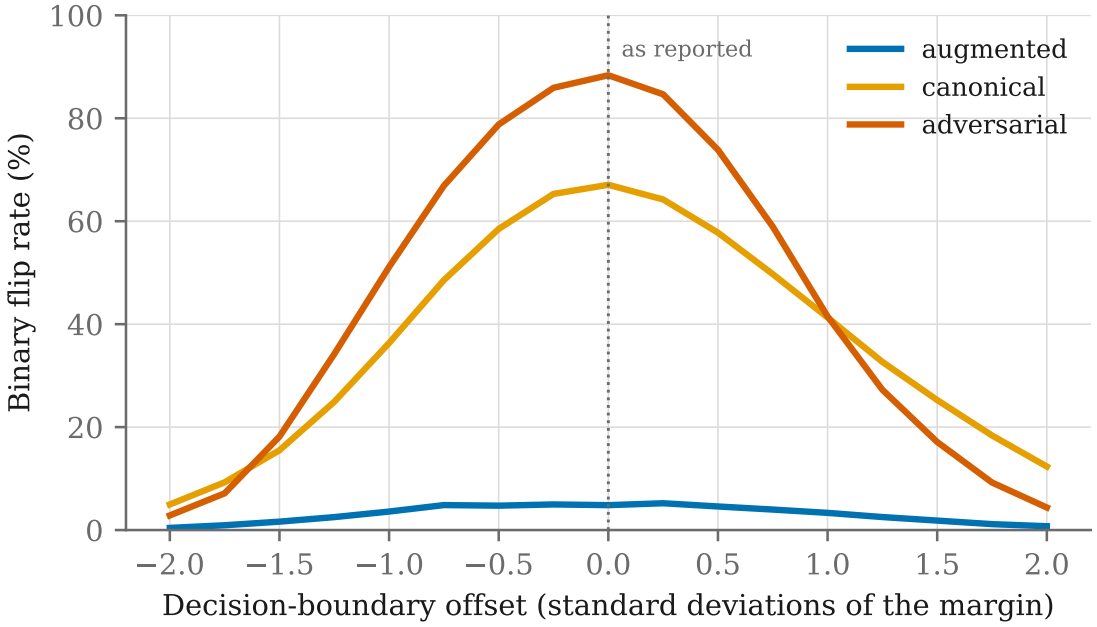


Figure 5.13: The advantage of broad phrasing coverage is not an artifact of where the decision boundary sits. Flip rate recomputed with the boundary shifted across two standard deviations of the margin distribution in each direction. The augmented regime flips least at every offset. The two narrow regimes exchange places in the tails, where a large shift pushes most clusters to one side and suppresses flips in both.

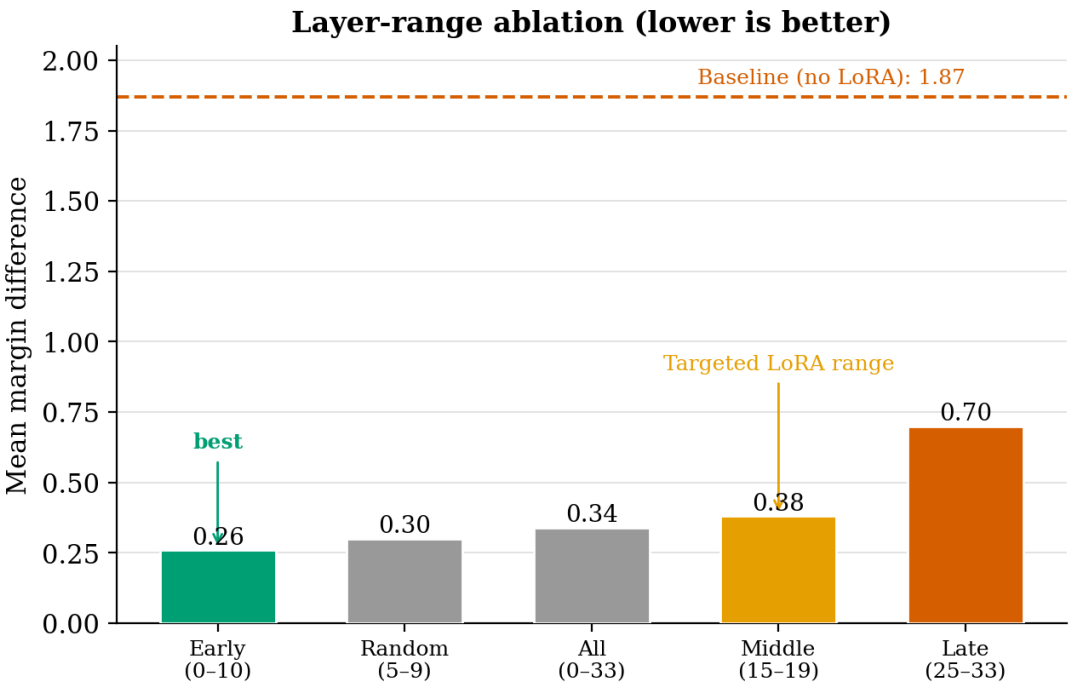


Figure 5.14: Layer ablation results. Early layers (0 to 10) achieve the largest margin reduction despite the mechanistic signal appearing at layer 17. The dissociation between where sensitivity manifests and where it is best treated parallels treating a disease at its origin rather than at the symptom site.

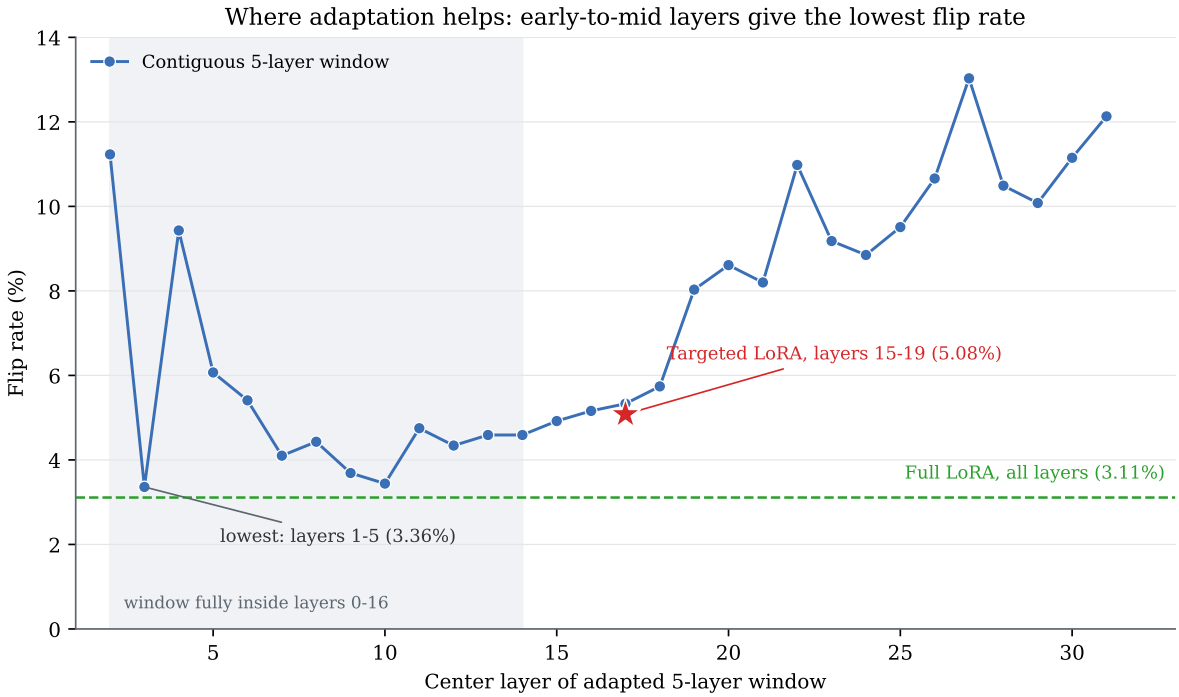


Figure 5.15: Flip rate as a function of the center layer of the adapted five-layer window (contiguous configurations, MIMIC-CXR validation set). Early-to-mid windows give the lowest flip rate, and effectiveness falls for windows centered on later layers. The mechanistically-targeted layers 15 to 19 (star) are competitive but not optimal, consistent with the dissociation between where register sensitivity manifests and where it is best treated.

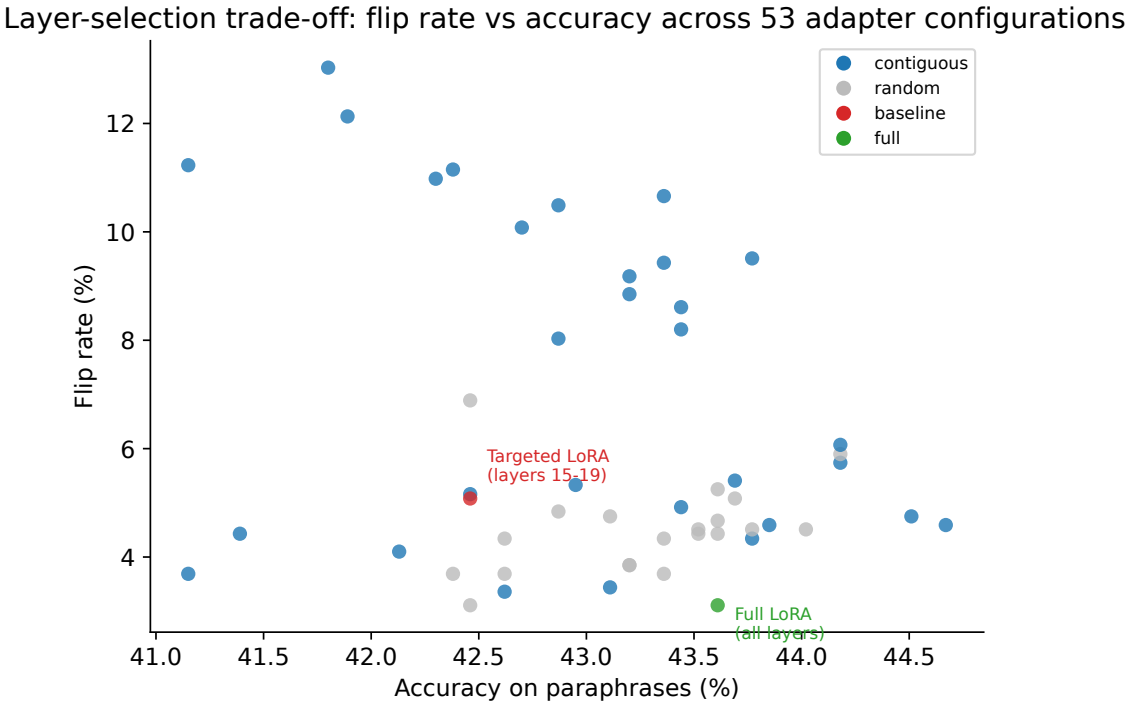


Figure 5.16: Flip rate versus paraphrase accuracy for all 50 adapter configurations in the block sweep (MIMIC-CXR validation set). The targeted layers-15 to 19 adapter and the full-model adapter are highlighted. Most configurations reach low flip rate with preserved accuracy, so the consistency gain is not unique to one layer choice.

Chapter 6

Thrust 4: Safety Evaluation

Research questions. This chapter answers **RQ4.1** (whether reducing the flip rate improves safety, or can create a false sense of safety), **RQ4.2** (whether attention heatmaps reliably indicate visual grounding), and **RQ4.3** (whether demographic disparities are larger for the worst-calibrated models).

Main contribution. A four-quadrant behavioral screen over consistency and image reliance showing that, across ten model-dataset settings, a mean of 81% of each model's consistent predictions are unchanged when the image is removed, together with convergent evidence that attention heatmaps are coarse localizers rather than faithful explanations ($\rho \approx 0$).

The preceding chapters measured paraphrase sensitivity (Chapter 3), diagnosed its origin as fragile visual grounding (Chapter 4), and developed a targeted mitigation (Chapter 5). This chapter asks the safety question: *is a more consistent model a safer model?*

The answer is not straightforward. Chapter 4 showed that consistency can come from text-only shortcuts, creating an illusion of reliability. A model whose answer does not depend on the image gives the same answer regardless of phrasing, achieving a low flip rate while its answers track the text prior rather than the visual evidence. This chapter develops a multi-phase safety evaluation protocol that tests consistency, image reliance, and attention grounding jointly, then formalizes a four-quadrant taxonomy that separates genuine consistency from text-driven consistency. Calibrated uncertainty estimates and the deployment-time machinery that builds on this taxonomy are taken up separately in Chapter 7.

Two main findings emerge. First, attention heatmaps are coarse localizers but not faithful causal explanations: attention concentrates in the annotated pathology region (true-box coverage 29.6% to 38.5%, 5.2 to 6.4× the random-box rate) and ablating that region changes the answer (causal scores 0.6 to 3.0, all intervals excluding zero), but coverage only marginally exceeds a displaced-box baseline and attention magnitude does not rank-predict causal importance ($\rho \approx 0$). Second, a four-quadrant behavioral screen that classifies predictions along consistency and image-reliance axes shows that a low flip rate does not certify grounding: averaged across ten model-dataset settings, 81% of each model's consistent predictions are image-invariant (range 45 to 100%), and the models with the lowest flip rates (Targeted LoRA, 13.2%, and LLaVA-Rad, 21.7%, on the curated PadChest flip bank) leave almost all of their consistent predictions unchanged when the image is removed. Correctness is a separate axis: on PadChest the image-invariant cell is often *more* accurate than the grounded cell, because high finding base rates make the text prior usually correct, so accuracy on this distribution is itself no evidence of grounding. One can also correlate flip rate with the image-invariant fraction and obtain $r = -0.86$, but this is largely a definitional consequence of that fraction being a subset of consistent predictions (conditioning on consistency reduces it to $r = -0.15$); the defensible statement is the level, not the correlation. The chapter also reports a fairness analysis stratified by sex and age,

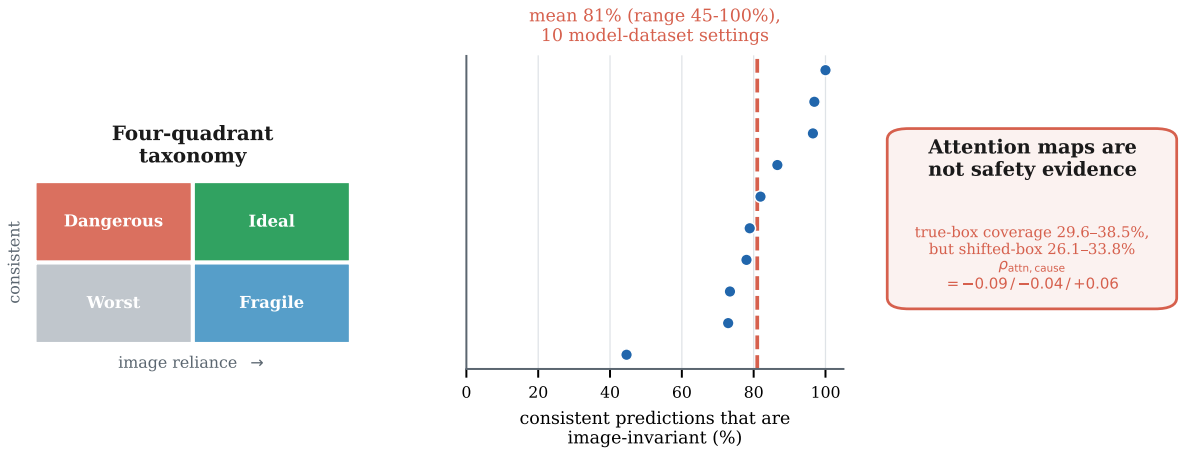


Figure 6.1: Thrust 4 at a glance. A four-quadrant behavioral screen classifies each prediction by consistency and image reliance. Across ten model-dataset settings, a mean of 81% of each model's consistent predictions are image-invariant (range 45 to 100%), so a low flip rate does not certify grounding; the often-cited $r = -0.86$ between flip rate and the image-invariant fraction is largely a definitional consequence of that overlap (it falls to $r = -0.15$ once conditioned on consistency). Attention maps are coarse localizers, not faithful explanations: true-box coverage is 30 to 38% (5.2 to 6.4× random) but only marginally beats a displaced box, and attention rank barely correlates with causal importance ($\rho \approx 0$).

and describes a proposed clinical validation and deployment-readiness assessment designed with the clinician coauthors of this work.

6.1 Safety Evaluation Protocol

6.1.1 Models and Datasets

We evaluate four model configurations on the flip-bank behavioral metrics, and a fifth on the four-quadrant screen (Table 6.4), chosen to span the consistency-grounding spectrum identified in Chapter 4:

1. **MedGemma-4B Base**: High flip rate, high image reliance (fragile grounding)
2. **Targeted Low-Rank Adaptation (LoRA)** (layers 15 to 19, original-pipeline adapter): Reduced flip rate with image reliance retained *on this evaluation set*; the clean-adapter re-audit does not reproduce that ordering (Section 5.7.5)
3. **Full LoRA** (all 34 layers): Low flip rate, high text-only agreement (potential text shortcut)
4. **LLaVA-Rad**: Lowest flip rate, near-complete text-only agreement (Chapter 4)
5. **LLaVA-Rad LoRA**: evaluated on the four-quadrant screen (Table 6.4); the flip-bank behavioral metrics above cover the four configurations listed first

The four-quadrant screen additionally includes LLaVA-Rad's own LoRA fine-tune as a fifth configuration, giving the ten model-dataset settings referenced throughout this chapter; the two LLaVA-Rad configurations are evaluated on the answer-balanced n=88 MIMIC and n=732 PadChest subsets (Table 6.4).

Evaluation uses both MIMIC-CXR and PadChest (861 binary questions, out-of-distribution). The MIMIC presence evaluation spans 196 question-disjoint binary questions (the expanded validation-plus-test split used for the calibration analysis of Chapter 7); the four-quadrant and text-only analyses in

this chapter use the 98-question subset for which text-only baselines were computed, since the image-reliance axis requires a matched no-image prediction for each item. PadChest provides 637 radiologist bounding boxes for attention grounding analysis. The 861-question PadChest flip bank is a curated high-confidence subset, distinct from the equivalence-filtered 14,935-question set used for the headline PSF-Med flip rates in Chapter 3.

All images are loaded with percentile-windowed intensity normalization to the model's input range.

6.1.2 Eight-Phase Protocol

Table 6.1: Eight-phase safety evaluation protocol. Phases 1 to 6 test behavioral properties (does the model's answer depend on the right input?); Phases 7 to 8 test explanatory properties (does the model's internal processing target the right image region?).

Phase	Name	What It Tests	Metric
1	Paraphrase Consistency (MIMIC)	In-distribution phrasing invariance	Flip rate
2	Text-Only Agreement (MIMIC)	In-distribution image reliance	Text-only agree %
3	Image Swap Sensitivity (MIMIC)	In-distribution visual dependence	Swap change rate
4	Paraphrase Consistency (PadChest)	Out-of-distribution phrasing invariance	Flip rate
5	Text-Only Agreement (PadChest)	Out-of-distribution image reliance	Text-only agree %
6	Image Swap Sensitivity (PadChest)	Out-of-distribution visual dependence	Swap change rate
7	ROI Ablation	Causal dependence on pathology region	Answer change rate
8	Attention Grounding	Attention overlap with pathology	True-box coverage

We define an eight-phase evaluation that captures failure modes invisible to accuracy-only testing:

1. **Paraphrase Consistency (MIMIC)**: Flip rate and margin difference on in-distribution questions
2. **Text-Only Agreement (MIMIC)**: Fraction of answers unchanged when the image is removed
3. **Image Swap Sensitivity (MIMIC)**: Fraction of answers that change when the image is replaced with one showing the opposite finding
4. **Paraphrase Consistency (PadChest)**: Out-of-distribution flip rate
5. **Text-Only Agreement (PadChest)**: Out-of-distribution text reliance
6. **Image Swap Sensitivity (PadChest)**: Out-of-distribution visual dependence
7. **Region of Interest (ROI) Ablation**: Whether occluding the pathology region changes the answer
8. **Attention Grounding**: Whether attention heatmaps overlap with pathology regions, tested against controlled baselines

Phases 1 to 6 test behavioral properties: does the model's answer depend on the right input? Phases 7 to 8 test explanatory properties: does the model's internal processing target the right image region?

6.2 Behavioral Safety Results

6.2.1 The Consistency-Grounding Spectrum

The Targeted and Full LoRA adapters analyzed for safety in this chapter are the original-pipeline adapters, trained before the operator-preserving cleanup of Chapter 5; the clean multi-seed fleet reported there *has* since been re-audited with the text-only and image-swap controls below (Section 5.7.5), and that audit does *not* reproduce the ordering reported here: on the clean adapters the targeted and full variants are statistically tied on both image-reliance measures, and both raise text-only agreement from 53.5% to about 77%. The region-of-interest and attention controls have not been repeated on the clean fleet. The grounding and four-quadrant conclusions below therefore describe the original-pipeline

Table 6.2: Behavioral safety across four models and two datasets. Flip rate measures paraphrase inconsistency (lower = more consistent). Text-only agreement measures the fraction of answers unchanged when the image is removed (higher = less image-reliant). Swap sensitivity measures the fraction of answers that change when the image is replaced with one showing the opposite finding (higher = more image-reliant). The four models span the consistency-grounding spectrum. Flip rates are question-level on the curated MIMIC (n=98) and PadChest (n=861) flip banks, distinct from the equivalence-filtered PSF-Med benchmark of Chapter 3. Flip rate, text-only agreement, and swap sensitivity are all computed together on these flip banks by the safety-evaluation pipeline (`results/miccai`); the four-quadrant screen (Table 6.4) is computed by an independent pipeline (`quadrants_fixed.json`) that does not carry the swap and text-only fields. For the MedGemma variants the two pipelines run the same models on the same flip banks and agree to within one question per cell (for example, Targeted LoRA on MIMIC flips 20/98 here versus 19/98 in the quadrant screen); the quadrant screen evaluates LLaVA-Rad on the answer-balanced n=88/n=732 subsets (Table 6.4). The small differences reflect borderline paraphrase predictions, not a change in any finding.

Model	Dataset	Flip Rate ↓	Text-Only Agree $\uparrow_{\text{risk}}$	Swap Sens. ↑
Base	MIMIC	42.9%	55.1%	28.6%
	PadChest	73.9%	76.9%	39.5%
Targeted LoRA (15 to 19)	MIMIC	20.4%	66.3%	32.7%
	PadChest	13.2%	89.9%	31.5%
Full LoRA (0 to 33)	MIMIC	4.1%	80.6%	14.3%
	PadChest	22.9%	76.3%	28.2%
LLaVA-Rad	MIMIC	32.7%	80.6%	21.4%
	PadChest	21.7%	89.9%	14.5%

adapters only, and must not be attributed to the clean adapters.

The four models span a clear spectrum on the curated PadChest flip bank (n = 861 question-level). Base MedGemma-4B has a 73.9% question-level flip rate but 39.5% image swap sensitivity: it is inconsistent but uses the image. Targeted LoRA reduces flips to 13.2% while keeping swap sensitivity high (31.5%). Full LoRA reaches 22.9% flips with 76.3% text-only agreement and 28.2% swap sensitivity. LLaVA-Rad reaches 21.7% flips with text-only agreement of 89.9%, tied with Targeted LoRA for the highest in the table, and the lowest swap sensitivity (14.5%). The tie is what separates the two: at the same text-only agreement, Targeted LoRA changes its answer on 31.5% of swapped images against LLaVA-Rad's 14.5%, so text-only agreement alone does not rank image reliance.

The trade-off is clearest at the extremes: Base is the most image-dependent (39.5% swap sensitivity) but the least consistent, while LLaVA-Rad is the most text-agreeing (89.9%) and least swap-sensitive. Targeted LoRA is an informative middle case: it is both the most consistent (13.2% flip) and, jointly, the most swap-sensitive of the LoRA variants, showing that consistency and some image dependence can coexist. Both LoRAs are trained on the same base model and data; the only difference is which layers are adapted.

6.2.2 Per-Finding Vulnerability Analysis

Not all clinical findings are equally susceptible to the Dangerous quadrant. We stratify Targeted LoRA's PadChest results by finding type (restricting to findings with $n \geq 15$ question-level instances) to identify which conditions are most vulnerable to text-driven consistency.

Common findings are disproportionately Dangerous. On Targeted LoRA (PadChest, findings with $n \geq 15$; Table 6.3), vertebral degenerative changes has 95.7% Dangerous predictions (22/23): the model has learned that "yes" is almost always correct for this finding and responds accordingly regardless of the image. Cardiomegaly reaches 80.4% (37/46) and aortic elongation 85.7% (36/42).

In contrast, findings that require spatial localization are rarely Dangerous for Targeted LoRA.

Table 6.3: Per-finding Dangerous fraction for Targeted LoRA on the curated PadChest flip bank (findings with $n \geq 15$ question-level instances). Dangerous = consistent across paraphrases but answer unchanged when the image is removed. Source: `results/uai/week1_analysis/per_finding_dangerous.json`. Findings ordered by Dangerous fraction; top five and bottom five shown.

Finding	n	Dangerous	Dangerous (%)
Most Dangerous			
Vertebral degenerative changes	23	22	95.7
Laminar atelectasis	15	14	93.3
Aortic elongation	42	36	85.7
Cardiomegaly	46	37	80.4
Alveolar pattern	15	12	80.0
Least Dangerous			
Aortic atheromatosis	26	4	15.4
Costophrenic angle blunting	22	3	13.6
Interstitial pattern	20	1	5.0
Vascular hilar enlargement	19	0	0.0
Infiltrates	16	0	0.0

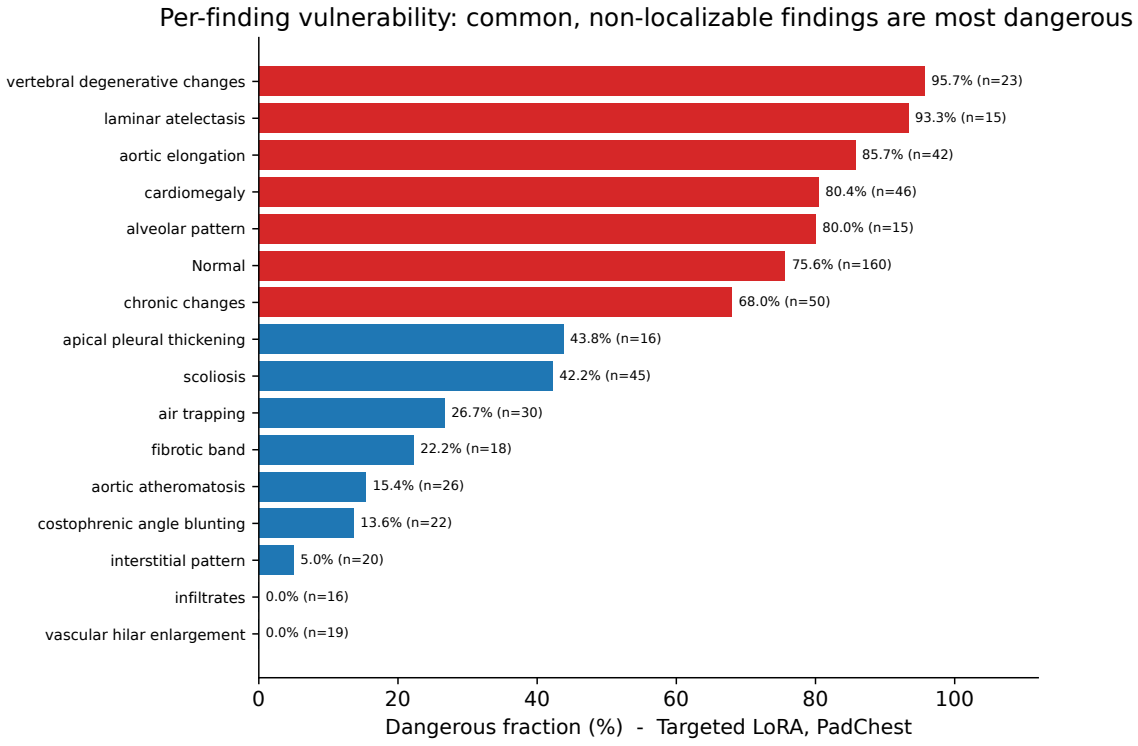


Figure 6.2: Per-finding Dangerous fraction for Targeted LoRA on PadChest (findings with $n \geq 15$). Common, diffuse findings (vertebral degenerative changes, cardiomegaly, aortic elongation) are mostly Dangerous, while findings that require localizing a specific region (vascular hilar enlargement, infiltrates) are not, because text priors alone cannot place a confident localized answer. Red bars mark findings that are Dangerous in at least half of cases.

Vascular hilar enlargement (0% Dangerous, 0/19) and infiltrates (0%, 0/16) show that when a finding requires the model to localize a specific image region, text priors alone cannot produce a confident answer. The model either uses the image (Ideal/Fragile) or produces inconsistent answers (Worst), but it does not achieve the stable text-driven pattern that defines Dangerous.

This finding-level analysis reveals that the consistency-safety paradox is not uniform: it concentrates on common, non-localizable findings where text priors are sufficient to achieve high accuracy without examining the image.

6.2.3 Four-Quadrant Behavioral Screen

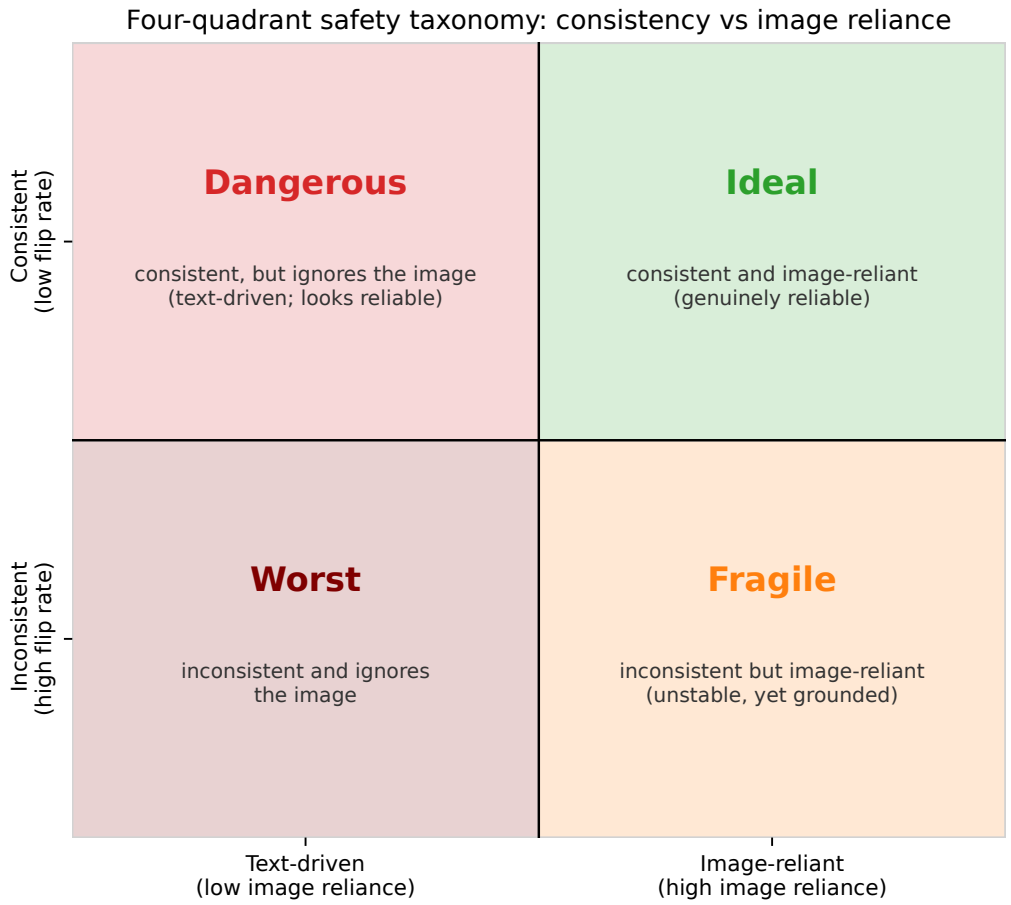


Figure 6.3: The four-quadrant behavioral screen. Predictions are classified by paraphrase consistency (vertical) and image reliance (horizontal). This is a description of behavior, not a per-prediction safety verdict: correctness is a separate axis (reported below). The Consistent-Text-driven cell (shorthand DANGEROUS) is the deployment concern, because these answers look reliable but change little when the image is removed.

We classify each prediction into one of four quadrants based on two binary criteria: (1) whether the model gives the same answer to all paraphrases of a question (consistent), and (2) whether the model's answer changes when the image is removed (image-reliant). The labels describe *behavior* (consistency $\times$ image-reliance); they are not by themselves safety or correctness verdicts, and we treat correctness as a third axis below. We retain the short mnemonics (Ideal, Fragile, Dangerous, Worst) as convenient shorthand, but the descriptive names are primary.

- **Consistent-Grounded** (IDEAL): Consistent *and* image-reliant. The model gives the same answer regardless of phrasing, and that answer changes when the image is removed.
- **Fragile-Grounded** (FRAGILE): Inconsistent but image-reliant. The model uses the image but is

sensitive to phrasing.

- **Consistent-Text-driven** (DANGEROUS): Consistent but *not* image-reliant. The model gives the same answer to every paraphrase, and that answer is unchanged when the image is removed, so the image contributes little to the decision on these items. As stated above, this is the measured behavior and not a verdict: an unchanged answer is also compatible with redundant evidence in the question text and with a margin that moved without crossing the decision boundary. The image-swap and region-of-interest controls (Sections 6.2.5 and 6.3) are what license any stronger reading. Whether these predictions are also correct is a separate question addressed below.
- **Inconsistent-Text-driven** (WORST): Neither consistent nor image-reliant. The model's answer is unchanged when the image is removed, yet it does change across paraphrases of the same question, so the answer tracks phrasing more closely than image content.

Table 6.4: Four-quadrant classification of predictions. Ideal: consistent and image-reliant. Fragile: inconsistent but image-reliant. Dangerous: consistent but not image-reliant. Worst: neither consistent nor image-reliant. Percentages computed from quadrant counts divided by total questions. MIMIC quadrants are reported on the subset with a text-only baseline. Across all ten model-dataset combinations, an unweighted mean of 81% of each model's *consistent* predictions fall in the Consistent-Text-driven cell (range 45 to 100%). We report this level rather than the $r = -0.86$ correlation between flip rate and that fraction: because the cell is by construction a subset of the consistent predictions, the correlation is largely definitional and falls to $r = -0.15$ once conditioned on consistency (Section 6.2.4). LLaVA-Rad settings use the answer-balanced LLaVA-Rad flip bank (n=88 MIMIC, n=732 PadChest; see the footnote in Chapter 3), curated to contain both answer polarities.

Model	Dataset	Ideal	Fragile	Dangerous	Worst
Base	MIMIC (n=98)	31.6%	13.3%	25.5%	29.6%
	PadChest (n=861)	3.5%	19.5%	22.5%	54.5%
Targeted LoRA	MIMIC (n=98)	21.4%	13.3%	59.2%	6.1%
	PadChest (n=861)	2.7%	7.7%	84.1%	5.6%
Full LoRA	MIMIC (n=98)	17.3%	2.0%	78.6%	2.0%
	PadChest (n=861)	16.3%	7.7%	60.6%	15.4%
LLaVA-Rad Base	MIMIC (n=88)	2.3%	18.2%	62.5%	17.1%
	PadChest (n=732)	0.0%	10.4%	79.5%	10.1%
LLaVA-Rad LoRA	MIMIC (n=88)	21.6%	10.2%	58.0%	10.2%
	PadChest (n=732)	16.0%	7.1%	56.7%	20.2%

The screen shows that a low flip rate does not certify grounding. The cleanest way to see this is to condition on consistency and ask what share of each model's *consistent* predictions are image-invariant. On PadChest that share is 86.6% for Base, 96.9% for Targeted LoRA, 78.9% for Full LoRA, and 100% for LLaVA-Rad Base: among the predictions that look reliable, the large majority are unchanged when the image is removed, and this holds regardless of the model's overall flip rate. Averaged across all ten model-dataset settings the share is 81% (range 45 to 100%).[1] As unconditioned fractions of all predictions, this reads as Targeted LoRA 84.1% Consistent-Text-driven, LLaVA-Rad Base 79.5%, Full LoRA 60.6%, and Base MedGemma-4B only 22.5% (the last is low only because Base is too inconsistent to land in a consistent cell at all, with 74% of its predictions in Fragile or Worst). Targeted LoRA is an important qualification: it lands in the Consistent-Text-driven cell most often by the text-only-agreement criterion, yet it also has the highest image-swap sensitivity of the LoRA variants (31.5%), so

[1] Four of those ten settings use the original-pipeline Targeted and Full LoRA adapters that this chapter declines to attribute to the clean multi-seed fleet of Chapter 5. Restricting the average to the six settings that do not involve them (Base and LLaVA-Rad Base and LoRA, on both datasets) gives 79.8%, so the conclusion does not depend on the disclaimed adapters.

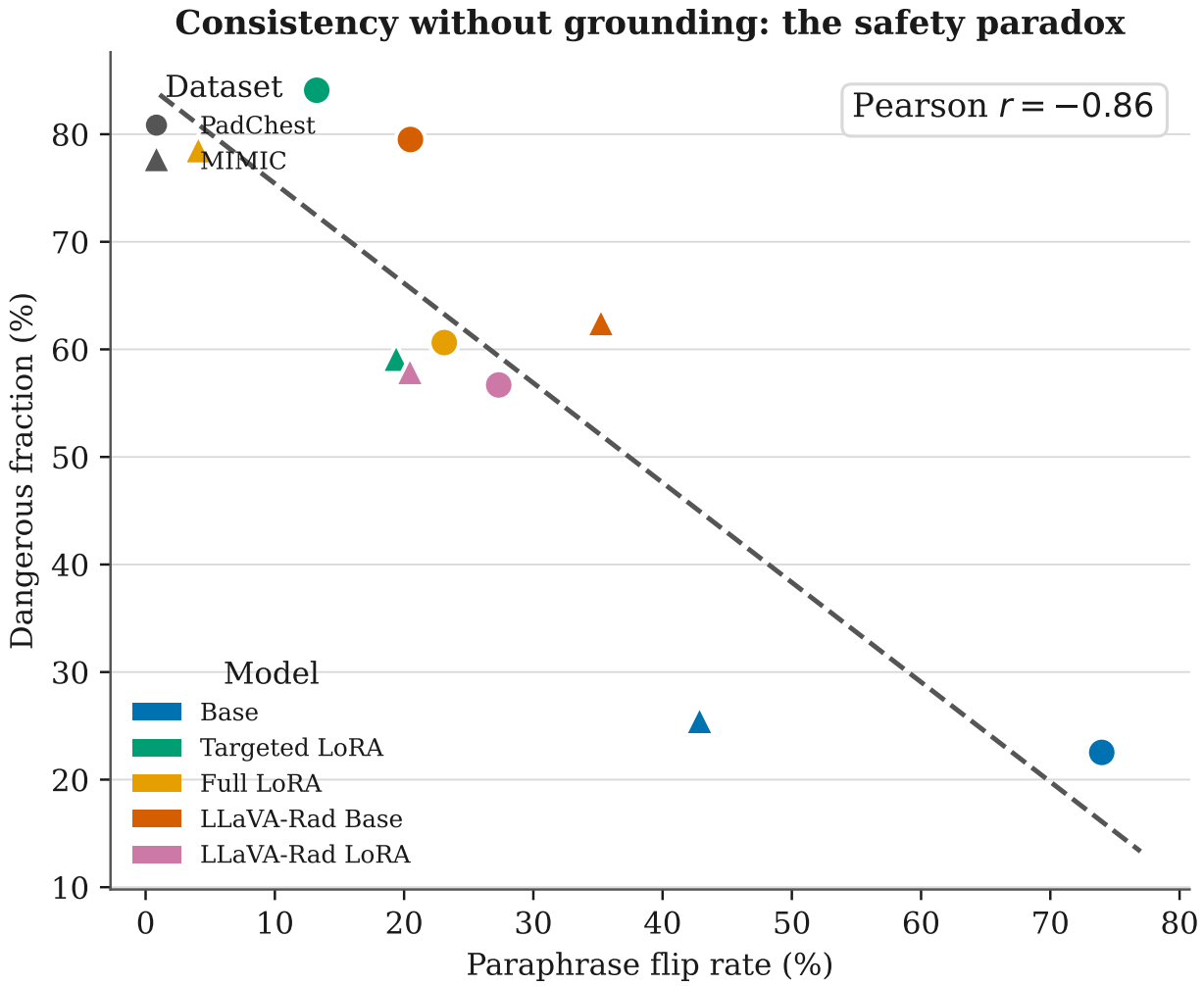


Figure 6.4: Flip rate versus the Consistent-Text-driven (DANGEROUS) fraction across ten model-dataset combinations. The two are negatively related (Pearson $r = -0.86$), but because the Consistent-Text-driven fraction is by construction a subset of consistent predictions (and the consistent fraction is $1 - \text{flip}$), this correlation is largely definitional: conditioning on consistency reduces it to $r = -0.15$ (Section 6.2.4). The defensible finding is the level, not the slope: a mean of 81% of consistent predictions are image-invariant.

a share of those predictions still change when the image is swapped. The text-only-agreement criterion can overcount on a dataset like PadChest, where high finding base rates let the text prior and the image agree; the swap and ROI evidence below refines the picture.

This does not mean the models never use the image. Invariance under image removal is a measurement, and an unchanged hard yes/no answer is also compatible with three benign readings: redundant evidence already carried by the question text, a saturated readout in which the margin moved without crossing the decision boundary, and a finding base rate high enough that the text prior is usually right anyway. The chapter's own null-image control makes the point directly, since a substantial share of Consistent-Grounded predictions are likewise unchanged under a null image (Section 6.2.5). The stronger evidence is the image-swap control, which shows the model an opposite-label image instead of removing the image, together with the region-of-interest ablation of Section 6.3; both are reported below, and every claim in this chapter that goes beyond invariance rests on them rather than on removal alone.

The unconditioned fractions can also be correlated against flip rate, which yields a strong negative relationship ($r = -0.86$). The next subsection shows that this correlation is largely a definitional consequence of how the cell is defined and should not be read causally; the defensible statement is the conditioned level above, not the correlation.

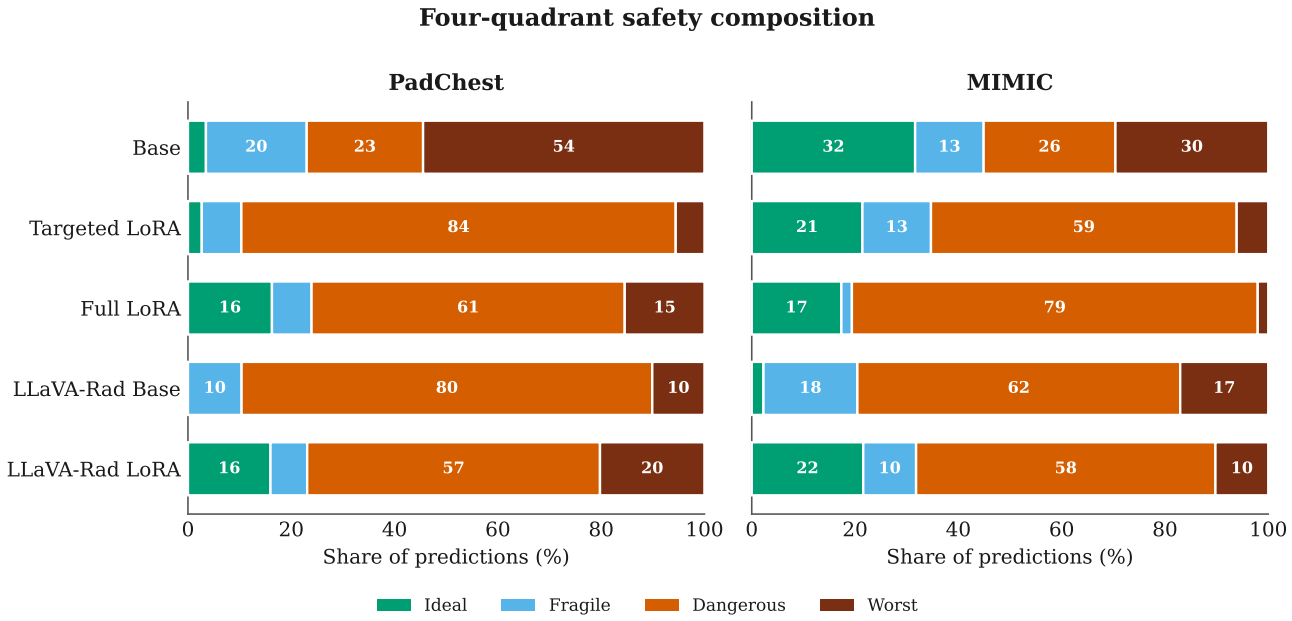


Figure 6.5: Four-quadrant behavioral screen across models and datasets. Each prediction is classified as Consistent-Grounded (IDEAL), Fragile-Grounded (FRAGILE), Consistent-Text-driven (DANGEROUS), or Inconsistent-Text-driven (WORST). Targeted LoRA and LLaVA-Rad route the largest fractions into the Consistent-Text-driven cell on PadChest (84.1% and 79.5%) despite the lowest flip rates, while MIMIC keeps a larger grounded share.

Targeted LoRA occupies an intermediate position. On MIMIC it reduces the Dangerous fraction relative to Full LoRA (78.6%) and LLaVA-Rad (62.5%) while also reducing the Fragile fraction relative to Base; on PadChest the pattern reverses, with Targeted LoRA highest of the three (84.1% against 60.6% and 79.5%). On MIMIC, it achieves 21.4% Ideal, second only to Base (31.6%), though both remain a minority of predictions.

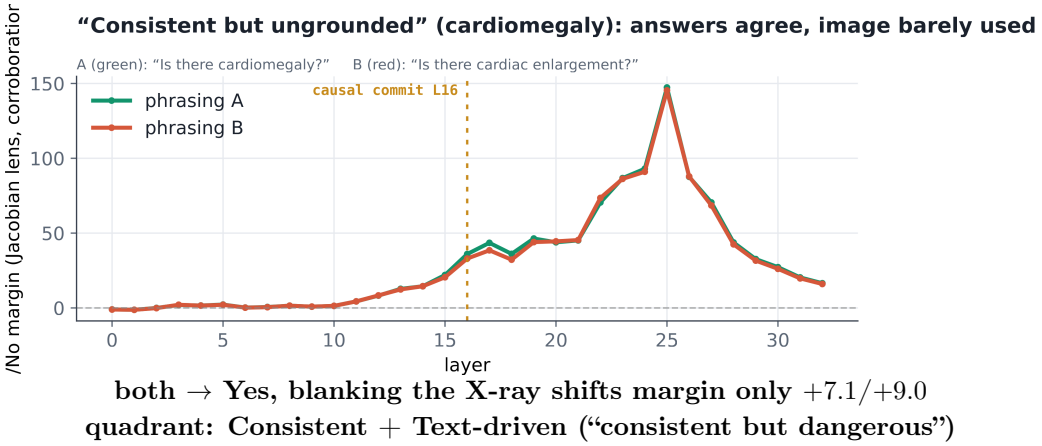


Figure 6.6: A concrete Consistent-Text-driven (DANGEROUS) case. Two clinically equivalent phrasings of a cardiomegaly question give the same answer and overlapping corroboration-lens trajectories, so the prediction is paraphrase-consistent; but blanking the image moves the margin by only +7.1 and +9.0 (against a peak near 150), so the answer is image-invariant. Consistency of this kind reflects the text prior rather than visual reasoning.

6.2.4 Interpreting the Flip-Rate Correlation

It is tempting to summarize the screen with a single correlation. Let $F(m)$ be the flip rate of model m and $D(m)$ the fraction of predictions in the Consistent-Text-driven cell. Across the ten model-dataset combinations the Pearson correlation is $r(F, D) = -0.86$ ($\rho = -0.70$). We report this number for completeness, but it must not be read causally, because it is largely an artifact of how the cell is defined.

The Consistent-Text-driven cell is by construction a subset of the consistent predictions, and the consistent fraction is exactly $1 - F(m)$. Writing $G(m)$ for the share of consistent predictions that are image-invariant gives the identity

$$D(m) = \big(1 - F(m)\big)\, G(m). \tag{6.1}$$

The factor $1 - F(m)$ alone already correlates with D at $r = +0.86$: almost all of the apparent relationship is the mechanical fact that a lower flip rate leaves more predictions in *some* consistent cell. The substantive quantity is $G(m)$, the conditional share, and G correlates with flip rate at only $r(F, G) = -0.15$ (Spearman -0.13), which is null. Conditioning on consistency thus removes the correlation. A label-permutation test does not help here, because it shuffles labels without removing the shared $1-F$ factor that couples D to F in the first place; we therefore do not rely on the correlation as evidence.

What survives conditioning is not a slope but a level: $G(m)$ is high across the board, averaging 0.81 over the ten settings with a range of 0.45 to 1.00. In words, a large majority of the predictions that any of these models makes consistently are image-invariant, whether the model flips often or rarely. That is the defensible form of the finding, and it does not depend on a ten-point correlation.

A single case makes this concrete (Figure 6.6). For a cardiomegaly question, the two phrasings “Is there cardiomegaly?” and “Is there cardiac enlargement?” produce the same confident Yes and near-identical lens trajectories at every layer, so the model is perfectly paraphrase-consistent. Yet blanking the chest X-ray shifts its yes/no margin by only +7.1 and +9.0 on a margin that peaks near 150, so the answer barely depends on the image. This is the Consistent-Text-driven cell made visible: the stability comes from the text prior, not from visual grounding.

Correctness is a genuinely separate axis, and it does not track the behavioral cell. On PadChest

the Consistent-Text-driven cell is often *more* accurate than the Consistent-Grounded cell: for Base MedGemma-4B the text-driven cell is 100% accurate versus 33% for the grounded cell, and for Targeted LoRA 100% versus 61%. High finding base rates make "yes" usually correct, so a model that answers from the text prior is frequently right on this distribution. This is precisely why the pattern is dangerous rather than merely wrong: the answers are stable, look reliable, and are often correct, yet they change little when the image is removed or replaced with an opposite-label image (a large share survive the swap unchanged, Section 6.2.5), so the accuracy should not be expected to transfer to the atypical cases where the prior is wrong. A behavioral screen therefore cannot be read as a per-prediction safety verdict; it must be paired with correctness and with the image-reliance controls below.

The practical takeaway is unchanged and does not require the correlation: any model achieving a very low flip rate should be presumed text-reliant until proven otherwise. Proof requires the text-only and swap experiments from Chapter 4, which can be run with a modest computational budget (3 to 5 additional forward passes per question).

The paradox also explains why Full LoRA, despite its excellent consistency metrics, is a poor model for clinical use on PadChest. Its 60.6% Dangerous fraction, combined with its low swap sensitivity (28.2%) and low accuracy (66.2%), means that for a majority of questions the model gives a confident, consistent answer only weakly tied to the patient's chest X-ray. A clinician relying on this model would receive stable reassurance from a system whose answers change little when the image is removed or replaced with an opposite-label image.

6.2.5 Complementary Validation of Quadrant Assignments

The four-quadrant classification relies on two binary tests: paraphrase consistency and text-only agreement. We validate these assignments using three independent methods that were not used for classification.

Kullback-Leibler (KL) divergence between image-conditioned and text-only distributions. For each prediction, we compute the KL divergence between the model's output distribution with the image (p_{img}) and without (p_{text}). Dangerous predictions should have near-zero KL divergence (the image shifts the output distribution very little), while Ideal predictions should have high divergence. KL divergence achieves AUROC = 0.76 for distinguishing Dangerous from non-Dangerous predictions, confirming that the binary text-only agreement threshold captures a real distributional difference.

Image-swap invariance. We replace the input image with one showing the opposite finding and measure whether the answer changes. Dangerous predictions are markedly more swap-invariant than Ideal predictions in every setting (the answer stays the same regardless of the image). Swapping in an opposite-label image is a stronger control than removing the image, because the model is shown contradictory visual evidence rather than no visual evidence. For the Dangerous predictions that survive the swap unchanged, this is causal evidence that the image does not drive the answer, rather than the text-only output coincidentally matching it. The remainder do change under swap and are not image-independent, so the strong reading is bounded to the swap-invariant subset.

Null-image test. As an extreme control, we replace the image with a uniform gray image (no visual content). Dangerous predictions are almost always unchanged under the null image. The control does not by itself separate the two cells, because a substantial share of Ideal predictions are also null-image invariant even though they are image-reliant by the removal criterion; the swap test above is the discriminating evidence.

These three independent controls converge: predictions in the Dangerous quadrant are invariant under image removal, under a null image, and, for a large share of them, under an opposite-label swap, with the swap the strongest of the three. The quadrant assignments are therefore not artifacts of threshold selection or definition sensitivity.

6.3 Attention Grounding Analysis

6.3.1 Attention Overlap with Pathology

Attention heatmaps, like the closely related Grad-CAM saliency maps [105], are commonly offered as evidence that a model "looks at the right place." We test this claim using PadChest's 637 radiologist bounding boxes. For each question with an associated bounding box, we extract the attention map from the decoder's self-attention, restricted to text-query-to-image-key entries and compute the fraction of attention mass that falls within the annotated pathology region (true-box coverage).

Table 6.5: Attention grounding results (PadChest, $N = 637$ bounding boxes). True-box coverage: fraction of attention mass within the radiologist-annotated pathology region. Coverage exceeds the random-far baseline by roughly 5.2 to 6.4× but only marginally exceeds the shifted-box baseline, so attention is coarsely, not precisely, localized. ROI causal score: margin change from ablating the pathology region minus a same-sized random region (95% bootstrap CI); all intervals exclude zero, so the region is causally used, most by Base and least by Targeted LoRA. Patch-rank Spearman ρ between attention weight and occlusion importance is near zero, so attention magnitude does not identify which patches matter.

Model	True Box	Shifted	Random Far	ROI Causal Score
Base	0.296	0.261	0.056	2.97 [2.67, 3.26]
Targeted LoRA	0.353	0.310	0.055	0.57 [0.51, 0.63]
Full LoRA	0.385	0.338	0.074	1.18 [1.02, 1.35]
Patch-rank ρ (attention vs. occlusion): -0.09 *Base,* -0.04 *Targeted LoRA,* $+0.06$ *Full LoRA*				

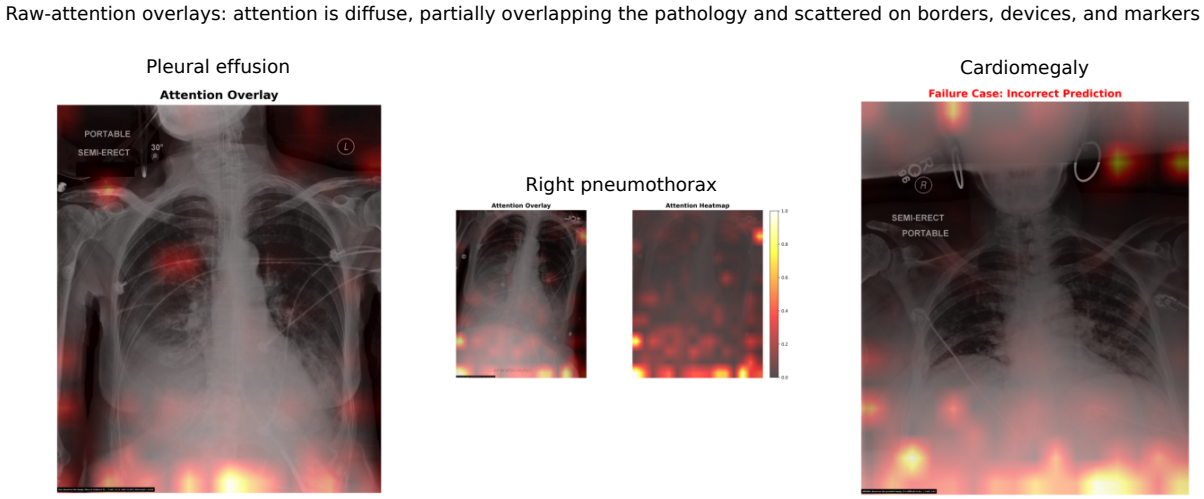


Figure 6.7: Representative raw-attention overlays for three findings. Attention is diffuse: it partially overlaps the pathology but is also drawn to image borders, support devices, and text markers. Across the dataset, true-box coverage (29.6% to 38.5%) exceeds random-box baselines by 5.2 to 6.4× but only marginally exceeds a displaced-box baseline (Section 6.3), so attention is a coarse rather than precise localizer.

Attention concentrates in the pathology region, but only coarsely. True-box coverage ranges from 29.6% (Base) to 38.5% (Full LoRA), roughly 5.2 to 6.4 times the random-far baseline (5.5% to 7.4%), so attention is preferentially directed at the annotated region rather than at random locations. Following the sanity-check methodology of Adebayo et al. [108], we compare against three baselines: shifted boxes (the true box displaced by 50 to 150 pixels in a random direction), random far boxes (same-sized boxes at random non-overlapping locations), and random-sized boxes. Coverage clears the random-far and random-sized baselines comfortably, but only marginally exceeds the shifted-box baseline (26.1%

to 33.8%): attention falls in the general anatomical neighborhood without singling out the precise pathology box.

6.3.2 Causal vs. Attentional Importance

Low attention overlap might be acceptable if attention is simply diffuse but still causally important. We test this with patch occlusion: we replace image patches with gray pixels and measure whether the model's answer changes. For each image, we rank patches by attention weight and by causal importance (answer change upon occlusion) and compute the Spearman correlation.

The correlation is near zero ($\rho = -0.09$ for Base, -0.04 for Targeted LoRA, $+0.06$ for Full LoRA). The patches that receive the most attention are not the patches whose occlusion changes the model's answer. So attention, while coarsely localized to the right region, does not identify which patches within the image are causally responsible: it is not a faithful patch-level explanation.

ROI deletion provides a complementary test at the region level. Removing the entire annotated pathology region changes the answer margin substantially: the causal grounding score (region ablation minus a same-sized random ablation) is 2.97 [2.67, 3.26] for Base, 1.18 [1.02, 1.35] for Full LoRA, and 0.57 [0.51, 0.63] for Targeted LoRA, all bootstrap intervals excluding zero. The pathology region is causally used, most by Base and least by Targeted LoRA. That ordering is telling: the consistency-trained Targeted LoRA relies on the region least, consistent with its higher text-only agreement. So the models do use the pathology region, even though their attention maps do not faithfully rank the patches within it.

6.4 Fairness Analysis

Table 6.6: Demographic fairness analysis (PadChest, softmax entropy). Targeted LoRA has the smallest between-sex calibration gap (0.012 ECE) but the widest accuracy gradient across age bins (83.3% to 96.3%, a 13-point spread with younger patients least accurate); its 2.7-point accuracy gap between sexes is slightly wider than Base's 2.0 points. Full LoRA has the largest disparities on every axis (4.3-point accuracy sex gap, 0.041 ECE sex gap, 0.042 AUGRC sex gap) and is poorly calibrated for every group. No model is uniformly the most equitable: the sex and age axes disagree.

Model	Group	n	Acc	ECE	AUGRC	Flip
Base	Male	422	77.7%	0.172	0.051	73.9%
	Female	439	79.7%	0.156	0.042	74.0%
	Age <40	48	79.2%	0.190	0.061	81.2%
	Age 40 to 60	233	76.8%	0.186	0.053	81.5%
	Age 60 to 80	416	78.8%	0.162	0.050	70.7%
	Age 80+	164	81.1%	0.141	0.030	69.5%
Targeted LoRA	Male	422	90.0%	0.106	0.018	14.7%
	Female	439	92.7%	0.118	0.011	11.8%
	Age <40	48	83.3%	0.121	0.043	18.8%
	Age 40 to 60	233	87.6%	0.092	0.023	16.3%
	Age 60 to 80	416	92.5%	0.126	0.011	12.3%
	Age 80+	164	96.3%	0.161	0.002	9.8%
Full LoRA	Male	422	64.0%	0.298	0.175	23.5%
	Female	439	68.3%	0.257	0.133	22.8%
	Age <40	48	66.7%	0.332	0.175	16.7%
	Age 40 to 60	233	63.1%	0.321	0.167	21.9%
	Age 60 to 80	416	65.9%	0.264	0.150	25.0%
	Age 80+	164	71.3%	0.249	0.129	22.0%

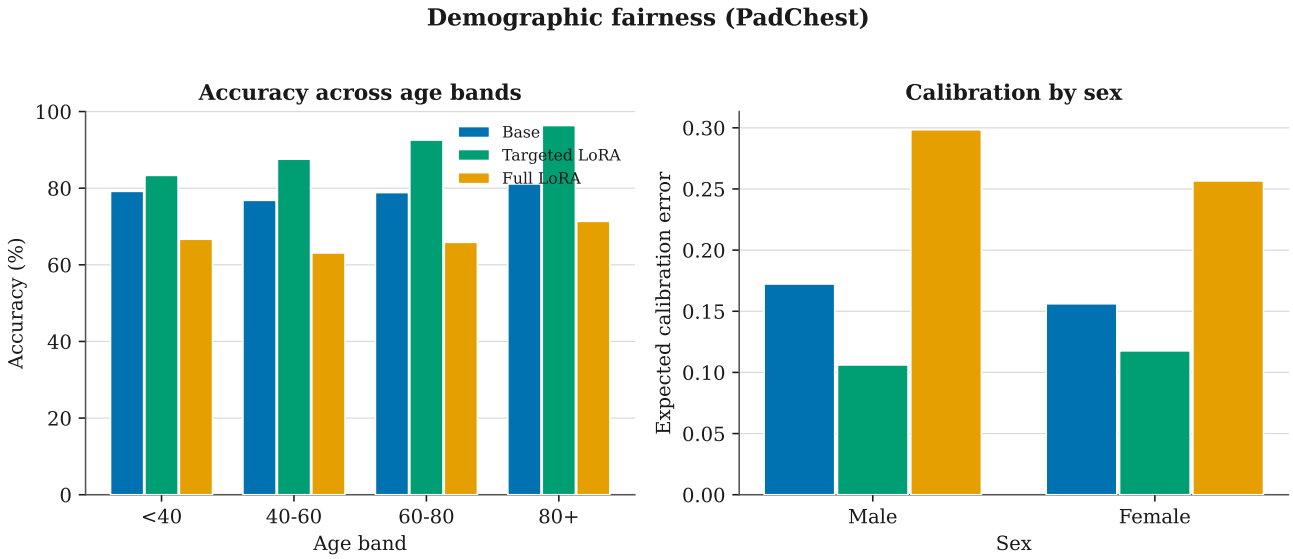


Figure 6.8: Demographic fairness on PadChest. Left: accuracy across age bands. Targeted LoRA is the most accurate overall but shows the steepest age gradient (83.3% at <40 rising to 96.3% at >80). Right: expected calibration error by sex. Full LoRA has the largest between-sex calibration gap and is poorly calibrated for every group.

We stratify results by patient demographics using PadChest's metadata (sex, age). Targeted LoRA has the smallest between-sex calibration gap (ECE differs by 0.012 between male and female patients), though its 2.7-percentage-point accuracy gap between sexes is slightly wider than Base's 2.0 points, so it is not uniformly the most equitable on this axis either. Its clearer weakness is age: accuracy climbs monotonically from 83.3% for patients under 40 to 96.3% for patients over 80, a 13 percentage point gradient, so the youngest patients are served least well. Full LoRA shows the largest disparities on every axis: a 4.3-percentage-point accuracy gap, a 0.041 ECE gap, and a 0.042 AUGRC gap between sexes, on top of poor calibration for every group (ECE at 0.249 or higher throughout). Base is the flattest across age (4.3 percentage point accuracy range) but, like Full LoRA, is poorly calibrated overall. No model is the most equitable on both axes at once: the sex and age comparisons disagree about which variant to prefer.

The fairness analysis was conducted on PadChest only, which provides patient demographics (sex and age) through its master table. The demographic analysis was not run on MIMIC-CXR: the de-identified splits used here do not carry the sex and age metadata that PadChest's master table provides, and the text-only subset would in any case be too small for stratified analysis. PadChest's 861 questions provide sufficient power for sex-stratified analysis and moderate power for age-bin analysis.

The demographic composition of the PadChest evaluation set is: 49.0% male, 51.0% female; age bins: <40 (5.6%), 40 to 60 (27.1%), 60 to 80 (48.3%), >80 (19.0%). This distribution reflects the hospital's patient population and provides adequate samples in each stratum.

A key finding is that the model with the worst overall calibration (Full LoRA) also shows the largest between-group calibration disparities. This suggests that poor calibration and demographic bias can share common causes: both reflect a model that has not learned to generalize across patient subpopulations and instead relies on shortcuts that work better for some groups than others. The pattern is not universal, though: Base is poorly calibrated but comparatively flat across age, and Targeted LoRA is well calibrated between sexes yet has the steepest age gradient.

Targeted LoRA still offers the most balanced overall profile on this evaluation: high accuracy (91.4%), the smallest between-sex gaps, and a moderate flip rate. Its image reliance is not a point in its favour: the clean-adapter re-audit in Section 5.7.5 finds it no more image-reliant than Full LoRA.

The age gradient is the caveat to that recommendation: the model's uncertainty and accuracy are least reliable for the youngest patients, who are also the least represented in PadChest ($n = 48$), so this gradient should be re-checked on an age-balanced set before deployment.

6.5 Clinician Adjudication of Paraphrase Equivalence

The preceding sections evaluate safety through automated experiments: paraphrase consistency, text reliance, attention grounding, and demographic stratification. These analyses establish what the models do, but not how clinicians interpret and act on these safety signals in practice. This section reports the one clinical input that is complete, a clinician adjudication of paraphrase equivalence on 1,200 pairs, carried out with the clinical coauthors of this work, a board-certified radiologist and a practicing anesthesiologist. It is a clinical consultation between the candidate and the clinician coauthors, not a human-subjects study. A structured deployment-readiness assessment against the failure modes identified in this dissertation has been designed but not conducted; it is described in the future work of Chapter 8.

The paraphrase equivalence underlying the flip-rate measurements was adjudicated by a three-reviewer panel that includes the radiologist and anesthesiologist coauthors. Because it validates the benchmark rather than the safety screen, that study is reported in full in Section 3.2.6: 1,200 pairs stratified over core, adversarial, and uncertain buckets, 400 of them triple-reviewed with 98.5% pairwise agreement (Cohen's $\kappa = 0.97$), against only $\kappa = 0.52$ (868/1,200) between the clinician consensus and the automated judge. The sample was drawn from the pre-regeneration audit pool and cannot restate the benchmark's flip rates. A reproducible link from the reviewed sample to the final evaluated set is not included in the committed records, so no overlap count or clinician-restricted flip-rate estimate is claimed.

6.6 Discussion

6.6.1 The Consistency-Safety Paradox

The central finding is that consistency and safety are not the same thing. A low paraphrase flip rate is not sufficient evidence of reliable visual reasoning: across the ten settings a mean of 81% of each model's consistent predictions are image-invariant, so consistency and grounding routinely come apart. A model that achieves a low flip rate by leaning on the text prior is not safer than one with a higher flip rate that actually uses visual evidence. The Consistent-Text-driven pattern is the concerning failure mode: the model appears reliable to the user (same answer every time), is often correct on the deployment distribution because the prior is usually right, yet its answers change little when the image is removed or replaced with an opposite-label image, so its accuracy should not be expected to hold on the atypical cases where the prior misleads. We deliberately state this as a level (the share of consistent predictions that are unchanged when the image is removed) rather than as the $r = -0.86$ correlation between flip rate and that share, because the correlation is largely a definitional consequence of the cell being a subset of consistent predictions (Section 6.2.4).

This paradox has implications for how medical VLMs are evaluated. Reporting flip rate alone is insufficient. Any consistency metric must be accompanied by an image-reliance test (text-only agreement, image swap sensitivity) to distinguish genuine consistency from illusory stability. Chapter 7 sharpens this evaluation requirement by showing that selective-prediction gates, when audited at the slice level, often achieve their headline accuracy by routing around the image-relevant portion of the workload.

6.6.2 Attention Maps as Safety Evidence

The attention grounding results challenge the common practice of using attention heatmaps [105] as evidence of visual reasoning, echoing broader concerns about whether attention constitutes explanation [106, 107]. Attention coverage exceeds random baselines but only marginally beats a displaced box, and the patch-rank correlation with causal importance is near zero ($\rho \approx 0$), so attention maps are coarse localizers, not faithful causal explanations: they point at roughly the right region but do not identify which patches drive the decision. Presenting attention heatmaps to clinicians as precise evidence of "what the model looked at" would therefore be misleading.

This does not mean the models never use the image. The image swap experiments from Chapters 4 and here show that some models do change their answers when the image changes. But the models' attention mechanism does not reliably target the pathology region. The image information may enter through diffuse, distributed processing that attention maps do not capture.

Chapter 7 takes up the deployment-time companion to this safety evaluation: calibrated uncertainty estimates, a multi-signal deployment gate, and an architecture-aware audit that asks what kinds of cases the gate actually admits.

Chapter 7

Thrust 5: Calibrated Uncertainty and Deployment Gating

Research questions. This chapter answers **RQ5.1** (whether single-pass predictive entropy predicts paraphrase flips as well as errors), **RQ5.2** (whether calibration and selective-prediction guarantees survive image-quality corruption and cross-site shift), **RQ5.3** (whether deployment gates admit genuinely image-grounded cases), and **RQ5.4** (whether any internal monitoring signal transfers across architecture families).

Main contribution. A single-forward-pass uncertainty signal that predicts both errors (AUROC 0.862) and paraphrase flips (AUROC 0.823), and a retrospective admission audit showing that selective-prediction gates often raise admitted-case accuracy by selecting text-answerable cases rather than image-grounded ones.

Chapter 6 established that consistency without grounding is the dominant failure mode of current medical VLMs: across ten model-dataset combinations, a mean of 81% of each model's consistent predictions are image-invariant, so a low flip rate does not certify grounding. The same chapter showed that the behavioral signals used to detect this failure (paraphrase agreement, text-only agreement, image swap sensitivity, attention overlap) all require multiple forward passes or external annotations. At deployment time, a clinical system receives one question and one image and must decide whether the model's answer is trustworthy. This chapter builds the components such a system would need and evaluates them *retrospectively*, on fixed cohorts that already carry radiologist annotations: calibrated uncertainty estimates that work from a single forward pass, a multi-signal admission rule that admits only predictions passing two independent safety checks, and an architecture-aware audit of what kind of cases that rule admits. No result in this chapter is prospective, so what follows supports a deployment-readiness audit rather than a recommendation to deploy.

Three findings define this thrust.

First, predictive entropy from a single forward pass is a unified proxy for both error detection and paraphrase vulnerability. It achieves Area Under the Receiver Operating Characteristic curve (AUROC) = 0.823 for flip prediction on Targeted Low-Rank Adaptation (LoRA) with PadChest, and AUROC = 0.862 for error detection. We call this the Uncertainty-Quantification (UQ) paraphrase bridge. The bridge replicates across architectures: LLaVA-Rad LoRA reaches AUROC = 0.830 for flip prediction on the same PadChest set.

Second, deep ensembles can degrade out-of-distribution when adapter checkpoints lack diversity. On PadChest, the MedGemma five-seed LoRA ensemble shows extreme member variance (individual-seed accuracy 52.0% to 92.1%): three of five seeds degrade, and although probability-averaging recovers aggregate accuracy to 72.6%, its selective-prediction quality collapses (Area Under the Risk-Coverage

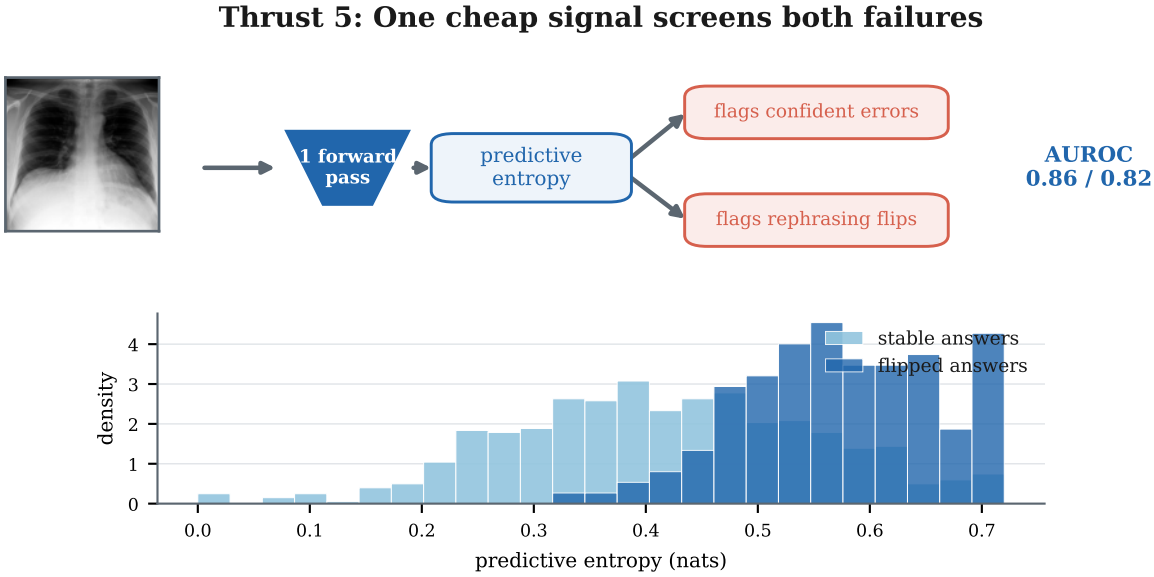


Figure 7.1: Thrust 5 at a glance. Predictive entropy from a single forward pass ranks error-prone and flip-prone answers above stable, correct ones, giving AUROC 0.86 for error detection and 0.82 for flip prediction. Confident errors sit at low entropy and are exactly what it cannot flag.

curve, AURC, 0.130 versus 0.019 for single-pass entropy, admitting only 23.8% of cases at 5% risk) while the LLaVA-Rad ensemble does not. The failure is model- and training-specific, not inherent to ensembling.

Third, gates that select on aggregate accuracy can quietly admit text-answerable rather than image-grounded cases. Across Gemma, LLaVA-Rad, and Qwen2-VL, the highest admitted-case accuracy often comes from null-image agreement rules, which are precisely the rules that screen out image-relevant questions. Any deployment claim therefore has to rest on an architecture-aware audit of the admitted subset, not only on a threshold tuned for nominal accuracy.

The chapter proceeds in four parts. Section 7.1 compares five UQ methods on clean data and under distribution shift, develops the UQ-paraphrase bridge, and decomposes total predictive entropy into aleatoric and epistemic components. Section 7.2 specifies a multi-signal admission rule, runs it as an offline readiness audit, and reports admitted-case accuracy by model and dataset. Section 7.3 sharpens the gate analysis by asking what kinds of cases the gate admits and shows that the answer is family-specific. Section 7.4 synthesizes the implications and sets out a tiered triage protocol as a candidate for prospective evaluation.

7.1 Uncertainty Quantification

The Targeted and Full LoRA adapters analysed in this chapter are the original-pipeline adapters, the same checkpoints audited in Chapter 6 and trained before the operator-preserving cleanup of Chapter 5. Where this chapter describes one adapter as more or less image-reliant than the other, that description applies to those adapters on these evaluation sets; the clean patient-disjoint re-audit of Section 5.7.5 finds the two tied on image reliance, so the ordering should not be carried over to the clean fleet.

The behavioral evaluation in Chapter 6 tells us which models and predictions are unsafe, but it requires multiple forward passes with controlled inputs (text-only, image swap, paraphrases). At inference time, a deployed model receives one question and one image. We need a signal computable from a single forward pass that flags unreliable predictions. Predictive entropy is a natural candidate.

7.1.1 Methods

We evaluate five uncertainty quantification methods:

1. **Softmax Entropy**: $H = -\sum_c p_c \log p_c$ computed from the yes/no softmax distribution. Single forward pass.
2. **Temperature Scaling**: Post-hoc calibration of softmax probabilities using a learned temperature parameter T, fit on a 15% calibration holdout.
3. **MC Dropout**: Run $K = 10$ forward passes with dropout enabled at inference time. Compute predictive entropy from the averaged distribution.
4. **Deep Ensemble**: Average predictions across 5 independently trained LoRA checkpoints (different random seeds). Available only for Targeted LoRA.
5. **Yes/No Margin**: The absolute difference $|z_{\text{yes}} - z_{\text{no}}|$ in logits before softmax. Low margin indicates uncertainty.

7.1.2 Calibration Under Clean Conditions

Table 7.1: Calibration metrics under clean (uncorrupted) conditions. Best UQ method per model shown. ECE: expected calibration error. AUGRC: area under the generalized risk-coverage curve (bounded [0, 0.5], lower = better).

Model	Dataset	Method	ECE ↓	Brier ↓	NLL ↓	AUGRC ↓
Base	MIMIC	Temp. Scaling	0.089	0.116	0.507	0.055
	PadChest	Temp. Scaling	0.136	0.161	0.508	0.047
Targeted LoRA	MIMIC	Temp. Scaling	0.096	0.120	0.385	0.042
	PadChest	Temp. Scaling	0.036	0.068	0.234	0.016
Full LoRA	MIMIC	Temp. Scaling	0.135	0.145	0.524	0.053
	PadChest	Temp. Scaling	0.229	0.271	0.893	0.156
LLaVA-Rad Base	MIMIC	N/A	N/A	N/A	N/A	N/A
	PadChest	Softmax	0.146	0.092	0.323	0.013

On clean (uncorrupted) data, calibration varies across models. On PadChest, Full LoRA has the worst ECE at 22.9%, followed by Base MedGemma-4B at 13.6% (both temperature-scaled): the model that memorizes text patterns and the model that lacks out-of-distribution calibration both fail substantially. Targeted LoRA is the best-calibrated MedGemma variant, reaching 3.6% ECE with temperature scaling. LLaVA-Rad Base has higher ECE (14.6%) but the strongest risk-coverage ranking on PadChest (AUGRC 0.013); on this set almost all of its answers are invariant under image removal (Chapter 6), so a well-ranked confidence signal is not on its own evidence that the ranking tracks visual evidence.

Temperature scaling improves calibration for most models. On MIMIC, Targeted LoRA ECE drops from 10.9% (raw softmax) to 9.6% (temperature-scaled); on PadChest it drops sharply from 11.1% to 3.6%. The calibration temperature is fit on 15% of each dataset and evaluated on the remainder.

The deep ensemble presents an instructive failure case. On MIMIC (in-distribution), the 5-seed ensemble achieves the best calibration and accuracy: ECE 9.2%, Brier score 0.091, AURC 0.030, accuracy 85.7%. On PadChest (out-of-distribution), its ECE stays modest (7.8%) but accuracy falls to 72.6% and its risk-coverage collapses (AURC 0.130 versus 0.019 for single-pass entropy): the aggregate looks calibrated while its confidence ranking is nearly useless for abstention.

Per-member performance reveals the root cause. Only seed 42 generalizes strongly to PadChest (92.1% accuracy). Seeds 123, 456, 789, and 2024 range from 52.0% to 78.7% accuracy. Probability averaging pulls the aggregate down toward the weaker members, so the ensemble (72.6%) underperforms its single best member (92.1%). This failure mode is specific to LoRA ensembles where each adapter

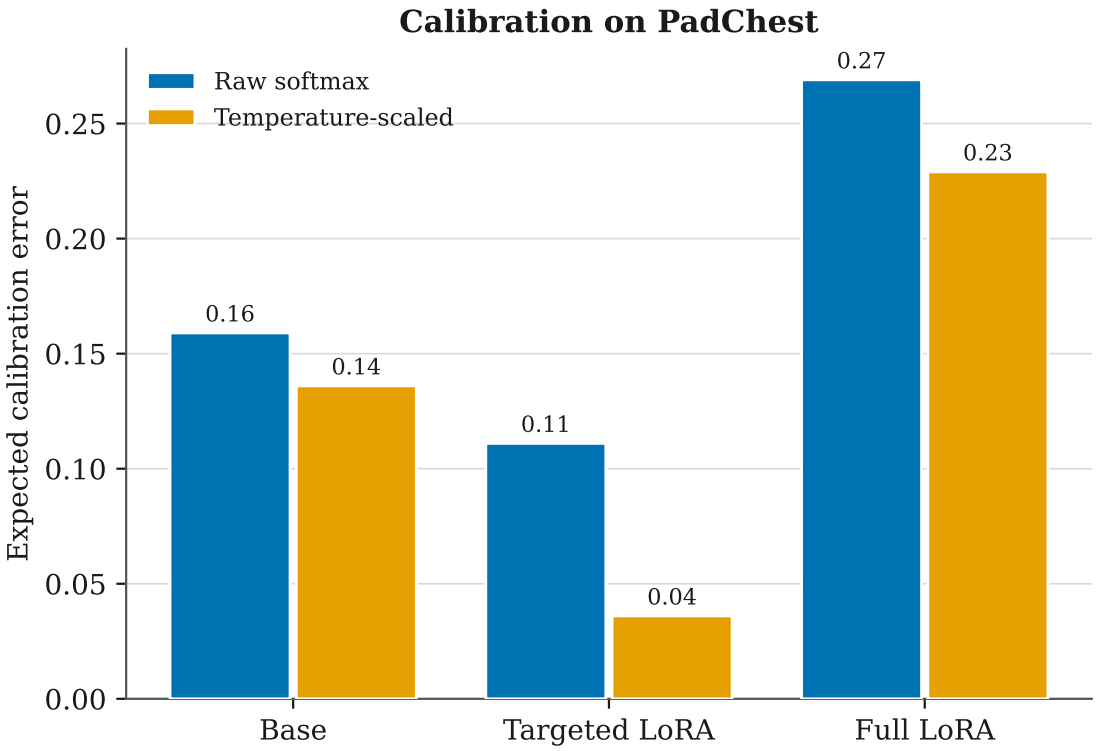


Figure 7.2: Expected Calibration Error (ECE) on PadChest under raw softmax and temperature scaling. Temperature scaling helps most for Targeted LoRA (11.1% to 3.6%); Full LoRA and Base remain the worst-calibrated MedGemma variants (22.9% and 13.6% after scaling).

was trained on the same narrow distribution (500 MIMIC-CXR pairs); the adapters have insufficient diversity to provide out-of-distribution robustness.

The MIMIC vs. PadChest comparison provides a natural in-distribution/out-of-distribution contrast. The calibration gap between the two sites is modest for most models. The figures in this paragraph are unscaled, before temperature scaling, whereas Table 7.1 reports the post-scaling values, so the two should not be read against each other: Base ECE moves from 15.9% (PadChest) to 14.5% (MIMIC) and Full LoRA from 26.9% to 12.9%, both better in-distribution, while Targeted LoRA is about even (11.1% PadChest, 10.9% MIMIC). Temperature scaling helps most on PadChest for Targeted LoRA (11.1% to 3.6%) but is nearly neutral for Base, indicating that the residual miscalibration in the base model is not a simple post-hoc scalar problem.

7.1.3 Calibration Under Distribution Shift

Table 7.2: ECE under image corruption for Targeted LoRA on PadChest. Clean ECE is 0.111 (softmax entropy). ECE remains between 0.072 and 0.119 across all conditions, showing stable calibration under distribution shift.

Corruption	Sev. 1	Sev. 3	Sev. 5
Gaussian Noise	0.111	0.104	0.092
Gaussian Blur	0.108	0.114	0.089
Contrast	0.102	0.072	0.082
Brightness	0.109	0.119	0.103
JPEG	0.114	0.112	0.108
Clean baseline		*0.111*	

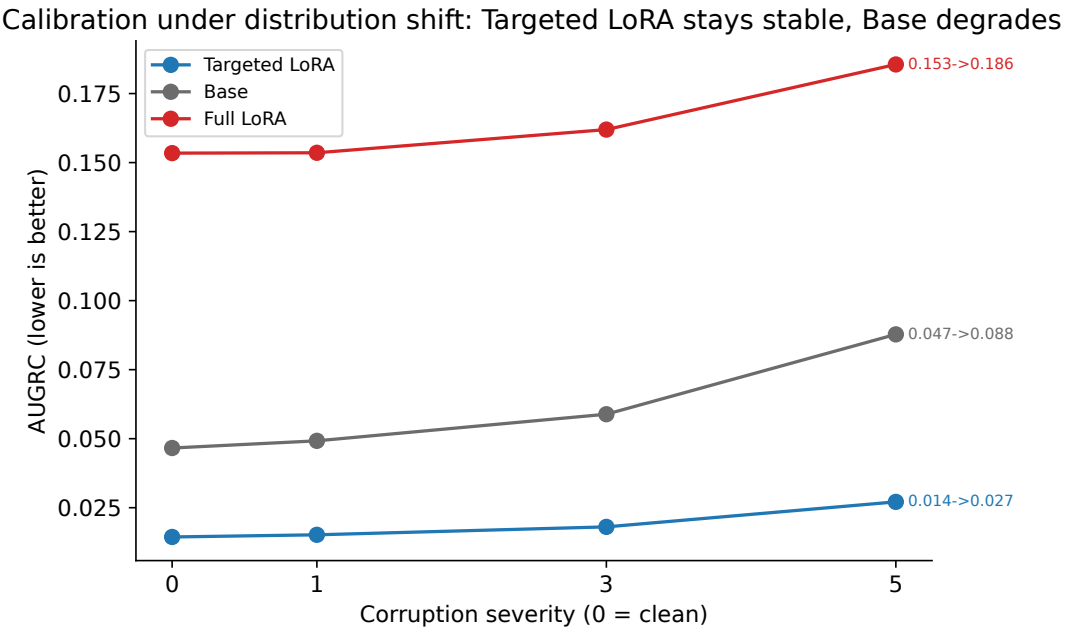


Figure 7.3: Area Under the Generalized Risk-Coverage curve (AUGRC, lower is better) on PadChest as corruption severity increases, averaged across the five corruption types. Targeted LoRA stays low and stable (0.014 clean to 0.027 at severity 5) while Base degrades (0.047 to 0.088). Full LoRA is poorly calibrated throughout (0.153 to 0.186), because its text-driven predictions barely respond to image corruption.

We test whether calibration holds under five types of image corruption at severity levels 1, 3, and 5. Brightness uses additive shifts of +20, +60, and +100 on the 0 to 255 pixel scale. Contrast factors are 0.8, 0.4, and 0.15, and Gaussian blur uses $\sigma_b = 1.0$, 3.0, and 6.0 pixels. Gaussian noise is applied at $\sigma = 10$, 35, and 70 on the same pixel scale. Joint Photographic Experts Group (JPEG) quality factors are 80, 40, and 10. Appendix Table A.15 gives the same parameterization with the corruption equations.

Each corruption is applied to all evaluation images independently, producing 15 corrupted evaluation sets per model-dataset combination. The total corruption sweep comprises 90 evaluation runs (3 models $\times$ 2 datasets $\times$ 5 corruption types $\times$ 3 severity levels), each using both softmax entropy and MC Dropout uncertainty estimation, for 180 JSONL result files.

For Targeted LoRA on PadChest, ECE remains between 7.2% and 11.9% across all conditions. Base MedGemma-4B degrades substantially: ECE rises from 15.9% (clean) to 25.3% at severity 5.

We track calibration degradation using the Area Under the Generalized Risk-Coverage curve (AUGRC) [120], a metric bounded between 0 and 0.5 where lower is better. Targeted LoRA AUGRC is stable under shift: 0.014 (clean) to 0.027 (severity-5 average). Base AUGRC degrades: 0.047 (clean) to 0.088 (severity-5 average). Full LoRA stays high but relatively flat (0.153 clean to 0.186 at severity 5), poorly calibrated throughout, and its predictions barely respond to image corruption.

7.1.4 Selective Prediction

Selective prediction allows a model to abstain on uncertain inputs, trading coverage for accuracy. We compute risk-coverage curves: at each entropy threshold, what fraction of questions does the model answer (coverage), and what is the error rate on answered questions (risk)?

On PadChest, softmax entropy and MC Dropout give nearly identical selective prediction for Targeted LoRA: at 5% risk tolerance both cover about 88% of questions (88.5% softmax, 88.4% MC Dropout), and both reach full coverage at 10% risk. Temperature scaling is close behind (86.7% at

5% risk). The extra stochastic passes of MC Dropout buy no advantage here, because adapter-only dropout barely moves the model (Section 7.1): the stochastic passes are close to repeats of the single softmax pass, so they add no ranking information. This says the probe is uninformative, not that the residual uncertainty is aleatoric. The deep ensemble covers only 23.8% at 5% risk, reflecting its OOD failure: its confidence ranking is uninformative when three of five members degrade outright and a fourth is marginal.

On MIMIC (in-distribution), the deep ensemble does provide complementary information: it achieves 78.6% coverage at 5% risk versus 56.1% for softmax entropy. The contrast between MIMIC and PadChest performance shows that UQ method selection must account for distribution shift.

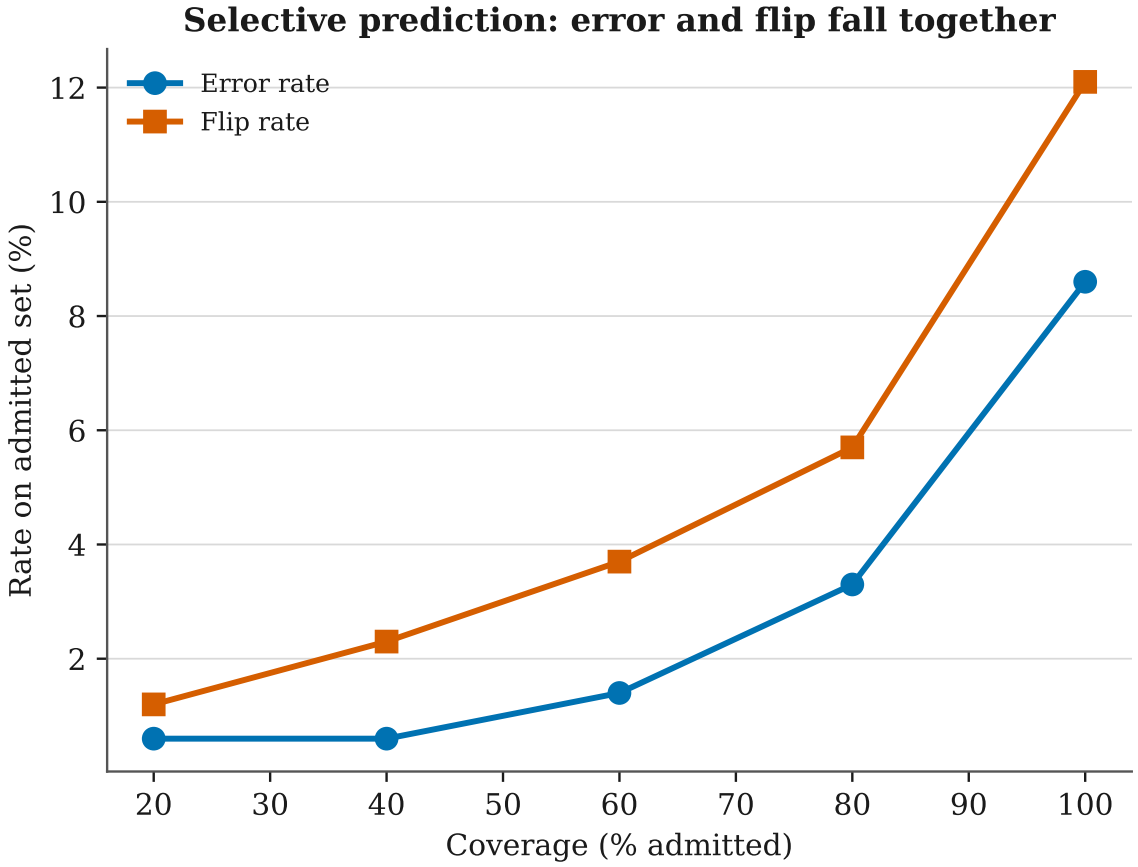


Figure 7.4: Risk-coverage trade-offs for the main selective-prediction methods. On PadChest single-pass entropy and MC Dropout give near-identical safe coverage, but the more important deployment question remains whether the admitted subset is image-grounded.

7.1.5 Entropy Decomposition: Aleatoric vs. Epistemic Uncertainty

For methods that produce multiple predictions (MC Dropout, deep ensemble), total predictive entropy H can be decomposed into aleatoric uncertainty (data noise) and epistemic uncertainty (model uncertainty), measured by mutual information (MI):

$$\mathrm{MI} = H[\bar{p}] - \frac{1}{K}\sum_{k=1}^{K} H[p_k] \tag{7.1}$$

where $\bar{p}$ is the averaged predictive distribution and p_k are individual forward-pass distributions.

For MC Dropout on Targeted LoRA (PadChest), MI is negligible: mean 0.0001 nats, with $\mathrm{MI}/H \approx$

0. The correct reading of this is narrow. MI measures the variation that *the applied perturbation* induces, and the perturbation here is dropout at $p = 0.05$ on adapters holding 4.38M of roughly 4.4B parameters; a near-zero MI therefore shows that this particular perturbation barely moves the model, so almost none of the predictive entropy is *detected* as epistemic by this probe. It does not establish that the remaining uncertainty is aleatoric, nor that it originates in ambiguous images: a probe that perturbs 0.1% of the weights cannot see epistemic uncertainty carried by the frozen 99.9%. The defensible conclusion is that adapter-only dropout is an uninformative epistemic probe for this model, which is also why it adds nothing over a single softmax pass.

The deep ensemble tells a different story. Mean MI is 0.057 nats, and MI is elevated for errors. In principle, this epistemic signal should enable better error detection. In practice, it does not: softmax entropy (single pass) achieves AUROC 0.862 for error detection on PadChest, matching MC Dropout (0.856) and beating the ensemble (0.751). The ensemble's meaningfully decomposed uncertainty is corrupted by its OOD failure, producing high MI that reflects disagreement among poorly-calibrated members rather than genuine epistemic uncertainty.

The practical implication is clear: for the LoRA-adapted models evaluated here, a single softmax forward pass provides the best uncertainty signal. MC Dropout adds cost (10× compute) for minimal benefit, and deep ensembles add cost (5× compute) with the risk of OOD catastrophe. Total entropy from one pass outperforms decomposed uncertainty from many passes.

7.1.6 The UQ-Paraphrase Bridge

The key finding of the uncertainty quantification analysis is that predictive entropy predicts paraphrase flips, not just errors. We call this the UQ-paraphrase bridge: a single uncertainty signal flags both types of unreliability.

Table 7.3: UQ-paraphrase bridge: AUROC for predicting whether a prediction will flip under paraphrasing (Targeted LoRA, PadChest). $\bar{H}_{\text{flip}}$ and $\bar{H}_{\text{stable}}$ are mean entropy (nats) of flipped and stable predictions. **These scores are not independent methods.** For a binary softmax, predictive entropy is a strictly monotone function of the absolute logit margin, and temperature scaling divides the logits by a fitted $T > 0$, which preserves that ranking. Softmax entropy, temperature-scaled entropy, and |margin| are therefore rank-equivalent and must give the *same* AUROC on the same rows, which they do: on the 861-row set softmax entropy and |margin| both give 0.8233, and on the 732-row temperature-scaling subset all three give 0.8126. †The apparent 0.823 versus 0.813 gap is entirely a difference in n, not an effect of temperature: T is fit on a 15% calibration holdout, so 129 of the 861 records are withheld, and restricting softmax entropy to the same 732 rows reproduces 0.813 exactly. ‡The |margin| row reports AUROC only; its mean columns are omitted because entropy means do not describe a margin score. Only the deep ensemble is a genuinely distinct signal, and it is the one that fails. Cross-architecture results (bottom) show the bridge generalizes to LLaVA-Rad.

Method	n	$\bar{H}_{\text{flip}}$	$\bar{H}_{\text{stable}}$	AUROC	p-value
Targeted LoRA, PadChest					
Softmax Entropy	861	0.585	0.419	0.823	4.3×10^{-29}
MC Dropout	861	0.585	0.419	0.823	4.6×10^{-29}
Deep Ensemble	861	0.502	0.473	0.552	3.7×10^{-2}
Temp. Scaling†	732	0.479	0.247	0.813	4.3×10^{-24}
|Margin|‡	861	N/A	N/A	0.823	4.3×10^{-29}
LLaVA-Rad LoRA (cross-architecture)					
Softmax Entropy (MIMIC)	167	0.602	0.299	0.905	N/A
Softmax Entropy (PadChest)	732	0.580	0.377	0.830	N/A

On PadChest, softmax entropy achieves AUROC = 0.823 for predicting whether a given prediction

will flip under paraphrasing ($p < 10^{-28}$). MC Dropout achieves AUROC = 0.823. Temperature-scaled entropy achieves AUROC = 0.813. The deep ensemble achieves only AUROC = 0.552, because three of five ensemble checkpoints degrade outright and a fourth is marginal on PadChest's out-of-distribution data. In distribution the bridge is comparably strong: on MIMIC (n = 196, validation plus test) softmax entropy reaches AUROC = 0.821 (95% CI [0.701, 0.919]).

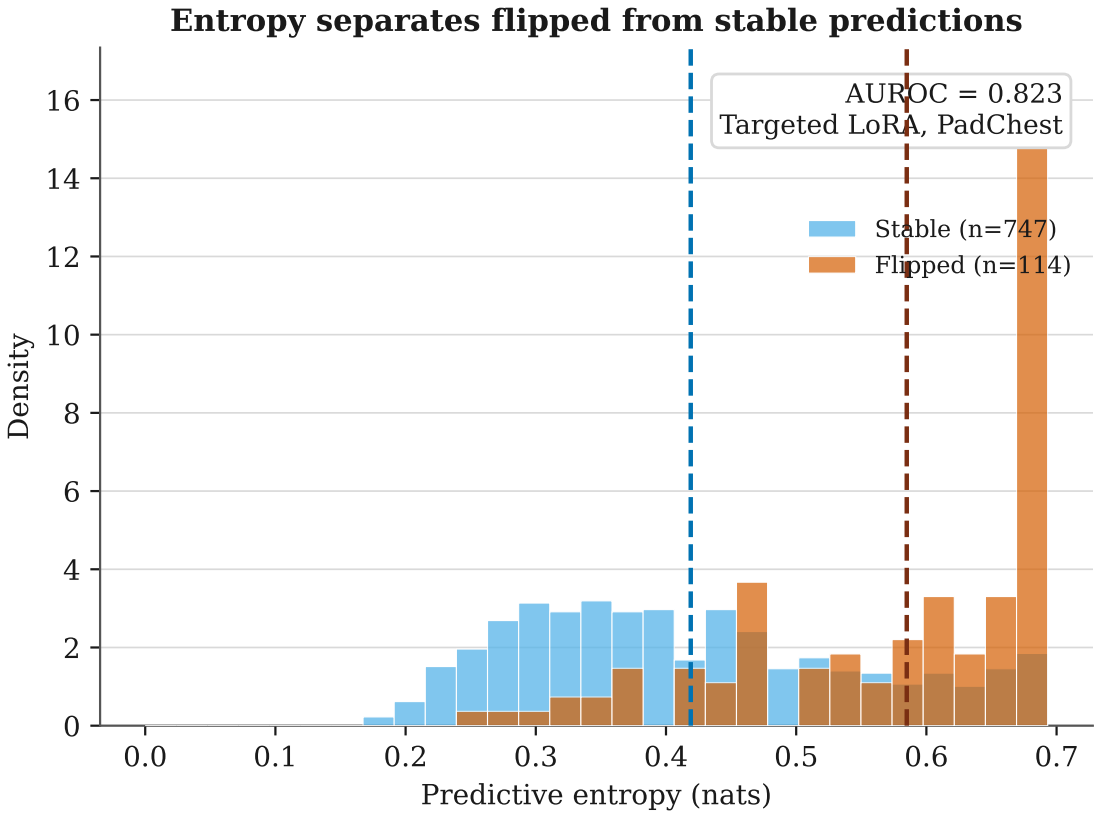


Figure 7.5: The UQ-paraphrase bridge: entropy distributions for flipped vs. stable predictions (Targeted LoRA, PadChest). Flipped predictions have higher entropy ($\bar{H}_{\text{flip}} = 0.585$ vs. $\bar{H}_{\text{stable}} = 0.419$), enabling AUROC = 0.823 for flip prediction from a single forward pass.

The bridge works because flipped predictions tend to be high-entropy. The mean entropy of flipped predictions is $\bar{H}_{\text{flip}} = 0.585$ nats, compared to $\bar{H}_{\text{stable}} = 0.419$ nats for stable predictions. When a model is uncertain about its answer, it is also uncertain about how to handle rephrased versions of the question. Entropy captures both sources of fragility through the same signal.

This has a direct practical implication. Entropy from a single forward pass can be thresholded to flag predictions for human review. Retrospectively, at a 40% coverage operating point, the error rate on admitted predictions is 0.6% and the flip rate on admitted predictions is 2.3%. Both are substantially lower than the population rates (8.6% error, 13.2% flip). The threshold is selected on the same evaluation set it is reported on, so these are in-sample operating points and not an estimate of prospective performance.

A follow-up analysis extended this to operator-preserving paraphrases (the equivalence-validated subset from Chapter 3). The all-paraphrase AUROC is 0.823 (95% CI [0.778, 0.864]); on the operator-preserving, ex-negation slice it is statistically indistinguishable at 0.826 (95% CI [0.780, 0.870]), with Error@20% / Flip@20% of 0.6% and 1.2%. Because the two confidence intervals overlap almost entirely, the bridge does not depend on adversarial reframing pairs: it holds on the equivalence-validated subset with the same strength. Across five training seeds the all-paraphrase bridge AUROC is 0.828 ± 0.021

(range 0.797-0.849), so the result is not seed-specific.

7.1.7 Geometric Intuition for the Bridge

The bridge between entropy and flip prediction has a geometric explanation. Consider the model's decision boundary in the joint space of image features and text features. For a binary question, the model produces a yes/no decision based on whether the combined representation falls on one side or the other of this boundary. Entropy is highest when the representation is close to the boundary (the model is uncertain about which side it falls on).

Now consider what happens under paraphrasing. The paraphrase changes the text features while leaving the image features unchanged. If the original representation is far from the decision boundary (low entropy), the paraphrase-induced shift in text features is unlikely to push it across the boundary. If the original representation is near the boundary (high entropy), even a small shift from paraphrasing can push it to the other side, causing a flip.

Under local linearity of the decision boundary and Lipschitz continuity of the text encoder under bounded-norm paraphrase perturbation, a smaller margin $|m|$ loosens an upper bound on the probability of crossing the boundary, and predictive entropy is a monotone decreasing function of $|m|$. We state this as an upper-bound result, not a claim that flip probability is itself monotone in entropy, which would require further assumptions about the perturbation-direction distribution we are not prepared to make.

This geometric picture predicts that the entropy-flip correlation should be strongest for models with moderate image reliance: the model must use the image enough for entropy to reflect genuine uncertainty (not just text-prior confidence), but the decision must be close enough to the boundary for text-side perturbation to matter. The predicted pattern is broadly consistent with the data: Targeted LoRA (moderate image reliance on the original-pipeline PadChest audit, AUROC = 0.823) shows a stronger bridge than Full LoRA (lower image reliance on that same audit, AUROC = 0.791) and Base (high flip rate but poorly calibrated, AUROC = 0.692).

The bridge also explains why MC Dropout does not improve flip prediction over softmax entropy (0.823 vs. 0.823 AUROC). MC Dropout captures epistemic uncertainty through variation across stochastic forward passes. But for LoRA-adapted models where only 0.1% of parameters are modified, the stochastic variation is negligible (MI $\approx$ 0.0001 nats), so the probe detects almost no epistemic uncertainty. As Section 7.1 argues, that is a statement about the probe rather than about the uncertainty: dropout on 0.1% of the weights cannot see epistemic uncertainty carried by the frozen 99.9%, so we cannot conclude that what remains is aleatoric. What follows for the bridge is narrower and sufficient: whatever MC Dropout is measuring here is already captured by a single softmax forward pass, which is why it adds nothing.

7.1.8 Cross-Architecture Validation

The bridge generalizes across model architectures. We evaluate the entropy-flip relationship on LLaVA-Rad (a different architecture, training procedure, and parameter count than MedGemma) and its LoRA-adapted variant.

On LLaVA-Rad Base (MIMIC), softmax entropy achieves AUROC = 0.884 for flip prediction (flip rate 33.5%). On PadChest, LLaVA-Rad Base still shows a strong positive bias (89.6% of predictions are "yes"), but on this curated flip bank (n=732, balanced to contain both answer polarities) it flips on 20.5% of paraphrase pairs, more than the near-zero rate it shows on the full PSF-Med benchmark (Chapter 3, where it is degenerate on the yes-dominated question mix); predictive entropy separates the flips well (AUROC = 0.952). The high AUROC reflects that the model is near-certain (entropy near zero) on the many confident-yes cases and uncertain on the rest. The two flip rates describe different evaluation sets, not a change in the model.

LLaVA-Rad LoRA provides a cleaner cross-architecture comparison. On MIMIC, it achieves AUROC = 0.905; on PadChest, AUROC = 0.830, closely matching MedGemma Targeted LoRA's 0.823. The entropy-flip relationship is not architecture-specific: it reflects a general property of how medical VLMs handle paraphrase variation. When a model is near its decision boundary (high entropy), small perturbations from rephrasing push it across.

The MedGemma ensemble illustrates that aggregate accuracy can hide member instability. On PadChest its five seeds range from 52.0% to 92.1% accuracy; only seed 42 generalizes strongly, while the other four fall below it out of distribution (78.7% and 52 to 69%). Probability-averaging recovers aggregate accuracy to 72.6%, but the ensemble's selective-prediction quality is poor (AURC 0.130 versus 0.019 for single-pass entropy, admitting only 23.8% of cases at 5% risk versus 88.5%). The instability is model- and training-specific: the MIMIC-trained adapters lack diversity on PadChest, so averaging cannot recover a well-ranked confidence signal even when it recovers the mean.

7.1.9 Conformal Prediction Under Shift

Table 7.4: Conformal prediction coverage under corruption shift (PadChest, $\alpha = 0.1$, target coverage = 90%). Calibrated on clean data, tested on corrupted images (averaged across 5 corruption types). At the standard 90% target all three models stay within about 3 points of target through severity 5; the coverage violation grows at tighter targets (see text).

Model	Clean	Sev. 3	Sev. 5
Base	91.8%	91.2%	88.7%
Targeted LoRA	93.7%	92.2%	89.9%
Full LoRA	88.7%	88.8%	87.2%
Target		*90.0%*	

Conformal prediction provides distribution-free coverage guarantees: given a desired coverage level $1 - \alpha$, the conformal prediction set contains the true label with probability at least $1 - \alpha$, under the exchangeability assumption [121]. We use the Adaptive Prediction Sets (APS) nonconformity score, which assigns higher scores to predictions with lower probability for the true class. The conformal threshold $\hat{q}$ is calibrated on the 15% holdout from each dataset using the $\lceil (n+1)(1-\alpha) \rceil / n$ quantile of calibration scores.

We test whether conformal prediction sets, calibrated on clean data, retain their empirical coverage under corruption. At a 90% target, Targeted LoRA over-covers slightly on clean data (93.7%) and lands on target under severity-5 corruption (89.9%): the empirical coverage stays near target throughout. This is a descriptive coverage result, not a formal validity guarantee, since corruption breaks the exchangeability that distribution-free conformal validity assumes. Slight over-coverage is desirable in a clinical setting, where under-coverage (missing the true diagnosis) is worse than over-coverage (including extra possibilities).

Base MedGemma-4B erodes from 91.8% (clean) to 88.7% at severity 5, a 1.3-point under-coverage at the 90% target. The violation is modest at this target but grows as the target tightens: at an 80% target, Base falls to 71.0% under severity-5 corruption (9 points under-coverage), versus 75.6% for Targeted LoRA. The under-coverage reflects the exchangeability assumption breaking down: corrupted images are not exchangeable with the clean calibration data, and the gap widens with corruption intensity (Base clean 91.8%, severity-3 91.2%, severity-5 88.7% at the 90% target).

Full LoRA holds near-constant coverage across all severities (88.7% clean to 87.2% at severity 5) because image corruptions barely move its predictions; the conformal sets retain their empirical coverage trivially when the answer distribution hardly responds to the corrupted visual input. This

highlights a subtle point: stable empirical coverage can be achieved through degenerate means (ignoring the input), and coverage alone does not imply clinical utility.

At a 95% target coverage, all three models over-cover on clean data (Targeted LoRA 95.2%, Base 95.6%, Full LoRA 94.1%) and decline only modestly under corruption: Targeted LoRA to 92.7% and Base to 92.8% at severity 5, both still above the 90% level, while Full LoRA stays nearly flat (94.1%). Calibrating conformal sets at this conservative target keeps actual coverage above 90% under shift for the two models whose predictions respond to corruption. Retrospectively, that makes the 95% target the setting worth carrying into a prospective evaluation of Targeted LoRA, because it is the only target tested here that held actual coverage above 90% at every corruption severity.

7.2 Deployment Gate: An Offline Readiness Audit

The evaluation results suggest a natural deployment protocol: admit predictions only when they pass both a consistency and an image-reliance test. What we evaluate here is that protocol run *retrospectively*, on a cohort with radiologist annotations, and it is best read as a deployment-readiness audit rather than as an online gate; the reasons are set out after the rule. To avoid ambiguity we state the rule exactly as implemented (`scripts/miccai/run_deployment_gate.py`). A prediction is admitted if and only if

$$\underbrace{\texttt{agree}(p_1, p_2)}_{\text{consistency}} \wedge \underbrace{\big(\texttt{swap\_changed} \vee \texttt{roi\_matters}\big)}_{\text{image reliance}},$$

where $\texttt{agree}(p_1, p_2)$ requires the model to give the original answer on the *first two* paraphrases, `swap_changed` is true when the answer changes under the swapped image, and `roi_matters` is true when the region-only and background-only crops yield different answers. The two image-reliance terms are a disjunction, not a conjunction, and `roi_matters` is available only on PadChest (637 of 861 records); on MIMIC the rule therefore reduces to $\texttt{agree}(p_1, p_2) \wedge \texttt{swap\_changed}$. The gate does *not* use the text-only condition.[1] Predictions failing either criterion are flagged for human review. The evaluation sets are the curated MIMIC ($n = 98$) and PadChest ($n = 861$) flip banks.

7.2.1 Why This Is an Audit and Not an Online Gate

Two properties of the rule above bound what its numbers can show. The swap partner is an arbitrary image from the same cohort, taken in rotation, not a matched opposite-finding image, so `swap_changed` measures sensitivity to image replacement in general rather than a response to finding-specific evidence. And the `roi_matters` term is available only on the 637 of 861 PadChest records with radiologist bounding boxes, where on its own it admits between 16.5% and 21.2% of cases depending on the model, so gated accuracy on PadChest is partly influenced by annotation availability, and the MIMIC evaluation rests on the swap term alone.

Two of the three inputs are available at inference time for a new patient, and one is not. Paraphrase agreement needs only the question, and `swap_changed` needs only some other image: the implementation supplies it by rotation, giving question i the image of question $i + 1$, so it never consults the true

[1] The implementation contains a third image-reliance disjunct, `question_type_uses_image`, which fires when the per-finding swap rate exceeds 0.05. It is inert on every record analysed here, and the reason is a property of the *stored data* rather than of the rule as written. The finding label exists upstream: it is populated on all 4,251 rows of the PadChest flip bank. It is dropped when `run_safety_eval.py` serialises the per-question records, so no stored record carries a `finding_label` field; the helper that builds the per-finding rates skips empty labels, and the lookup therefore always returns the 0.0 default. The two-term rule above is the rule that actually ran, which we confirmed by reproducing the admission rates in Table 7.5 exactly from the stored records. This point is load-bearing and is stated at length so that no future reader silently "repairs" it: restoring the label to the serialised records activates the third disjunct and raises Base PadChest admission from 41.6% (358 of 861) to 64.6% (556 of 861), because 71 of the 92 finding labels clear the 0.05 threshold. Every gate number in this chapter is the two-term rule's and must not be mixed with numbers recomputed after such a change, which would be measuring a three-term rule.

label and could be computed in deployment against any unrelated study. `roi_matters` is the exception. It compares a region-only crop against a background-only crop, which presupposes a radiologist bounding box for the finding in question, and that is exactly what a deployed system does not have for an undiagnosed patient. It is available for 637 of the 861 PadChest records and for none of MIMIC.

Because the image-reliance terms are a disjunction, the size of this effect scales with how often `roi_matters` is the *only* reason a prediction is admitted. Recomputing from the stored records: of Base's 358 admitted PadChest predictions, 76 (21.2%) are admitted solely on the region test; for Targeted LoRA 52 of 284 (18.3%), and for Full LoRA 38 of 230 (16.5%). Dropping the region term, which is what an online gate would have to do, lowers PadChest admission from 41.6% to 32.8% for Base, from 33.0% to 26.9% for Targeted LoRA, and from 26.7% to 22.3% for Full LoRA. The MIMIC configuration, where the rule already reduces to $\texttt{agree}(p_1, p_2) \wedge \texttt{swap\_changed}$, is deployable as written.

The honest reading is therefore split. The MIMIC numbers describe a rule a deployment could run. The PadChest numbers describe a readiness audit: they answer "on a cohort where we know where the finding is, what would this gate have admitted?", which is the right question for pre-deployment evaluation and the wrong one for runtime triage. An online variant would either drop the region term at the admission cost quantified above, or replace it with an annotation-free saliency proxy, which Chapter 6 gives reason to distrust: attention is a coarse localizer whose patch ranking does not track causal importance.

Because admission already forces the first two paraphrases to agree, the residual flip rate reported for admitted predictions is necessarily carried by paraphrases the gate never inspected (the third and later). We therefore report it as *flip rate on uninspected paraphrases*, and it answers a question worth asking: does a cheap two-paraphrase gate certify stability on rephrasings it has not seen? It does not. Requiring agreement on *all* available paraphrases instead of the first two drives that residual to 0.0% in every cell, at a lower admission rate: Base admits 13/98 (MIMIC) and 97/861 (PadChest), Targeted LoRA 23/98 and 272/861, and Full LoRA 14/98 and 206/861. The cost of certifying stability is therefore admitting far fewer predictions, which is the substantive trade-off a deployment would face.

Table 7.5: Deployment gate results. The gate admits a prediction iff the model gives the original answer on the *first two* paraphrases *and* shows image reliance (the answer changes under the swapped image, or, on PadChest only, the region-only and background-only crops disagree); it does not use the text-only condition (Section 7.2). Acc (all): unfiltered accuracy. Acc (gated): accuracy on admitted predictions. **Flip (uninspected paras)**: because admission forces the first two paraphrases to agree, this residual is carried entirely by the third and later paraphrases, which the gate never checked; it measures whether a two-paraphrase gate certifies stability on unseen rephrasings. Requiring agreement on *all* paraphrases drives it to 0.0% in every cell at a lower admission rate (Base 13/98 and 97/861; Targeted LoRA 23/98 and 272/861; Full LoRA 14/98 and 206/861). The low admission rates reflect the severity of safety failures across models.

Model	Dataset	Admitted	Acc (gated)	Acc (all)	Flip (uninsp. paras)
Base	MIMIC	22.4%	95.5%	83.7%	40.9%
	PadChest	41.6%	96.9%	78.6%	72.9%
Targeted LoRA	MIMIC	30.6%	76.7%	77.6%	23.3%
	PadChest	33.0%	96.8%	91.5%	4.2%
Full LoRA	MIMIC	14.3%	92.9%	81.6%	0.0%
	PadChest	26.7%	83.9%	66.0%	10.4%

On PadChest, the gate admits 41.6% of Base predictions at 96.9% accuracy (up from 78.6% unfiltered). For Targeted LoRA, the gate admits 33.0% at 96.8% accuracy (up from 91.5%). A majority of predictions are still rejected for failing at least one safety criterion, but the admitted subset is far more accurate than the population.

The accuracy benefit is not uniform, and it should not be stated as a blanket claim in either direction. On MIMIC the gate is a large gain for two of the three models and a wash for the third: Base admits 22.4% at 95.5% accuracy against 83.7% unfiltered, and Full LoRA admits 14.3% at 92.9% against 81.6%, while Targeted LoRA admits 30.6% at 76.7% against 77.6% unfiltered, which is no improvement at all. A rule that helps the two weaker models and does nothing for the strongest one is consistent with the reading developed in Section 7.3: what the rule removes is largely cases the weaker models answer from the text prior, and Targeted LoRA has fewer of those to remove.

A practical limitation is that the gate requires additional forward passes: the paraphrases, the swapped image, and, on PadChest, the region-only and background-only crops. It does not require a text-only pass, because the text-only term is not part of the implemented rule. The entropy-based bridge from Section 7.1.6 offers a cheaper alternative: a single-forward-pass entropy threshold that approximates the multi-pass rule. The entropy threshold admits more predictions at a lower admitted-case accuracy.

7.2.2 Gate Design Considerations

The deployment gate operates on individual predictions, not on the model as a whole. This per-prediction approach is important because no model is uniformly safe or unsafe. LLaVA-Rad, with 79.5% Dangerous predictions on PadChest and almost no Ideal predictions, is the clearest case: the gate finds almost nothing safe to admit, which is itself the correct signal for a model whose answers on this set rarely change when the image is removed. For models with a real Ideal fraction (for example Full LoRA at 16.3%), the gate admits those predictions and rejects the rest.

The gate's components can be computed in parallel. Given a question q and image I:

1. Generate K paraphrases $\{p_1, \ldots, p_K\}$ using the original PSF-Med (GPT-4) paraphrase pipeline that produced the flip-bank paraphrases (Appendix A.1).
2. Run forward passes for q and all p_k with image I (consistency check; the implemented rule inspects the first two paraphrases).
3. Run one forward pass with the swapped image I_{swap} (image-reliance check). This pass is required, not optional: on MIMIC it is the only image-reliance term available.
4. On PadChest only, run two further passes, one on the region-only crop and one on the background-only crop (region check). This step needs a radiologist bounding box and is available for 637 of the 861 records.

There is no text-only pass, because the implemented rule has no text-only term. The total computational cost is therefore $K + 2$ forward passes where only the swap test is available and $K + 4$ where the region test is also run. At the $K = 2$ used here, that is 4 forward passes on MIMIC and on the 224 PadChest records without a bounding box, and 6 on the 637 PadChest records that have one. Each pass takes approximately 100 to 150ms on an A100 GPU, so the forward passes themselves take under 1 second per question. This figure counts *only* the vision-language model's forward passes. It excludes step 1: in this work the paraphrases were generated offline by GPT-4 and reused, but an online gate would have to generate them per query, adding an external API round trip that would dominate the sub-second forward-pass budget and introduce a dependency on a third-party service. A deployment wanting the sub-second figure to hold would need to precompute paraphrases for a fixed question set, or replace the generator with a local model. Even so, the budget remains small against clinical image interpretation, which already takes minutes.

The gate can be configured in three modes of increasing strictness:

- **Consistency-only**: Admit if all paraphrases agree. Low computational cost but does not detect Dangerous predictions.
- **Consistency + text-only**: Admit if all paraphrases agree AND the answer changes when the image is removed. Detects text-reliant consistency.

- **Full gate**: Consistency + text-only + swap sensitivity. The most stringent test, requiring evidence that the answer depends on the specific image content. This mode is *not* the one evaluated below.

None of these three is the rule that was run. The implemented rule is a fourth configuration: two-paraphrase consistency combined with a disjunctive image-reliance term (swap, or the region-versus-background test on PadChest), with no text-only term at any point.

The results reported in Table 7.5 use the implemented rule of Section 7.2, namely two-paraphrase consistency combined with image reliance (swap or region), *not* the full gate: the text-only term is never evaluated. Had the full gate been run, admission would fall to roughly 2 to 11% across the six model-dataset cells, so the 41.6% and 33.0% figures should be read as the implemented gate's, not as an upper bound from the strictest configuration.

7.3 Architecture-Aware Deployment Audits

The generic deployment gate above asks whether a prediction looks safe under one fixed set of criteria. Later work sharpens the question: *what kind of cases does the gate actually admit, and does the answer depend on the model family?* Across Gemma/SigLIP, LLaVA-Rad/CLIP, and Qwen2-VL/NaViT, the answer is yes. The admitted subsets often become more accurate by selecting questions that can be answered from the text channel alone, not by isolating predictions whose answers depend on the image.

7.3.1 Selective Prediction Can Select Text-Answerable Cases

On PadChest, the strongest Gemma-family result comes from Targeted LoRA. A null-image agreement rule admits 72% of questions at 98.9% accuracy, while the strict multi-signal audit rule admits only 10% at 96.7% accuracy. That strict rule is a *different* rule from the gate of Section 7.2, not a re-parameterization of it, and the two operating points should not be read as strict and loose settings of one gate. As implemented in `scripts/analysis/mech_interp/selective_prediction_audit.py`, the strict rule admits a prediction when the yes/no logit margin exceeds 1.0, *and* all available paraphrases agree, *and* the readout cosine is at or above its 80th percentile. The Section 7.2 gate instead uses agreement on the first two paraphrases together with an image-reliance term (the swapped image, or the region-versus-background crops). The only input the two share is paraphrase agreement, and even that differs in scope (all paraphrases versus the first two): the strict rule has no swap term, no region term, and no image-reliance term of any kind, while the Section 7.2 gate uses neither the logit margin nor the readout cosine. The strict rule is not isolating a more grounded subset; its admitted predictions remain 94.4% swap-invariant and 95.6% text-agreeing. The simplest text-dominance control is both higher-coverage and at least as accurate.

The same pattern appears in Qwen2-VL with even sharper contrast. Null-image agreement reaches 99.8% accuracy at 67.6% coverage, whereas the strict behavioral gate reaches 89.2% at 17% coverage. LLaVA-Rad differs in which rule scores highest, but not in the underlying concern: the strict subset reaches 92.9% accuracy at 85% coverage, yet both strict and null-image-admitted subsets are more than 99% swap-invariant. Across families, admitted-case accuracy can rise while the admitted subset becomes almost entirely invariant to the swapped image, which is the stronger of the two image-reliance controls used here.

7.3.2 Grounded-Slice Failure Is the Safety-Relevant Test

The most safety-relevant evaluation is not full-set accuracy but behavior on slices where the image should matter. On PadChest, Targeted LoRA achieves only 2.9% accuracy on the text-disagrees slice ($n = 241$), and the strict rule raises this only to 25.0% at 2% intra-slice coverage. Qwen2-VL is more extreme: it gets 0 of 279 text-disagrees questions correct under every evaluated rule. These are precisely the cases where the image-conditioned answer departs from the text-only answer, and they reveal that the apparent safety gains come from avoiding image-relevant questions rather than solving

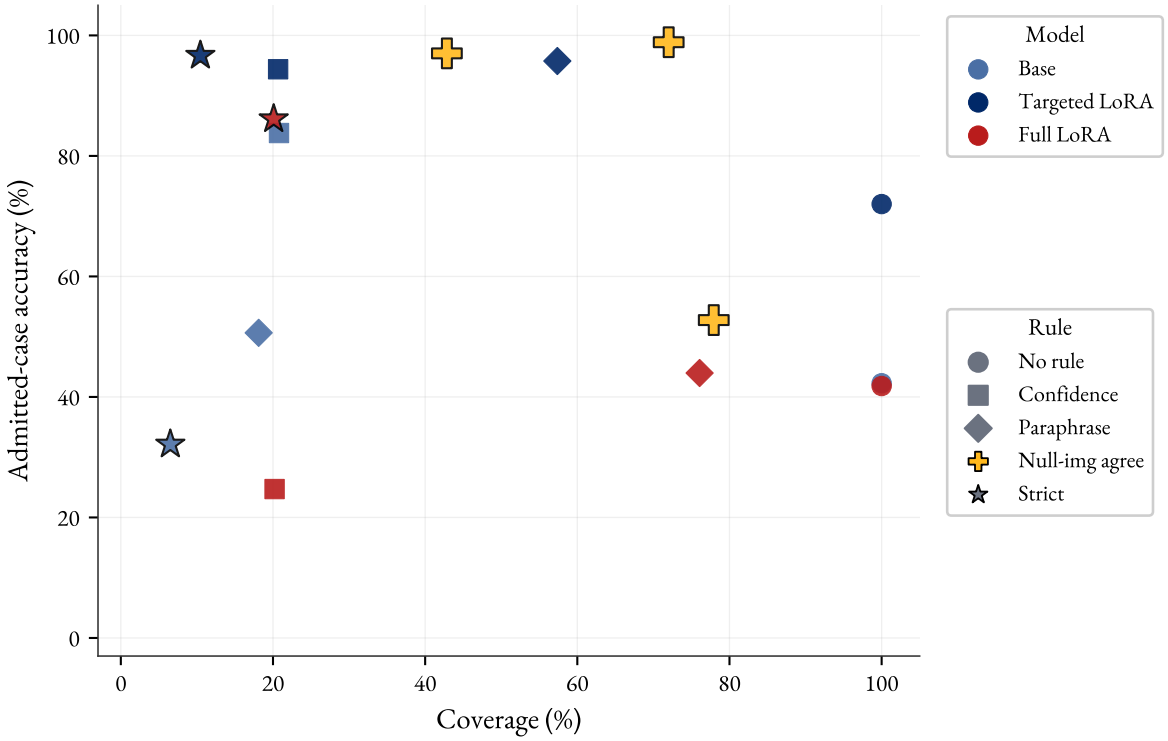


Figure 7.6: Deployment-gate trade-offs from the later architecture-aware deployment audit (Section 7.3). The highest admitted-case accuracy often comes from null-image agreement or similarly text-dominant rules, not from the stricter gate intended to preserve grounding.

them.

This severity is dataset-dependent. On MIMIC-CXR, the same text-disagrees slice is substantially less catastrophic, with image-model accuracy in the 68 to 86% range. The deployment lesson is therefore not that all medical VLMs always fail when the image matters, but that high-answerability environments amplify the shortcut. A gate that looks excellent on aggregate metrics can still be unsafe if it silently routes around the image-relevant portion of the workload.

7.3.3 No Single Monitor Transfers Across Architecture Families

The internal monitoring story is architecture-specific. Gemma-family models show a late readout-misalignment signature: transport representations remain nearly unchanged across paraphrases while the readout state diverges, and readout-cosine probes reach AUC 0.81 to 0.92. LLaVA-Rad does not follow this pattern. Its failures are broader and extend into earlier transport, while cosine probes are only weakly informative. Qwen2-VL is more opaque still: both transport and readout cosine remain near chance as flip predictors despite strong behavioral evidence of text-dominant admission.

This means there is no universal monitoring dashboard. For Gemma-family models, internal probes are useful. For LLaVA-Rad, behavioral checks and confidence are more informative than cosine probes. For Qwen2-VL, null-image and grounded-slice audits dominate because the internal probes do not explain the shortcut. A readiness audit therefore has to be repeated per family: a monitor validated on one architecture cannot be transferred to another and assumed to work.

7.3.4 A Candidate Operating Policy: Audit-Guided Escalation

The operational alternative is to separate autonomous acceptance from human-review escalation using the audit itself. For Targeted LoRA, audit-guided escalation catches 77% of candidate image-relevant cases while preserving 72% autonomous coverage. For Qwen2-VL, the null-image agreement subset reaches 99.8% accuracy at 67.6% coverage with zero errors on the image-relevant cases admitted by that policy. The key idea is not to trust the gate because it improves aggregate accuracy, but to trust it only after confirming what kinds of cases it selects and routing the remainder to review. Both figures are retrospective, computed on a fixed cohort whose annotations were already available, so they describe what such a policy would have done on that cohort and not what it would achieve online.

Read as an audit, this reframes the deployment question from a one-number thresholding problem into a workload-partitioning problem. The right question is no longer "what threshold maximizes admitted-case accuracy?" but "which cases can safely remain autonomous, and which cases must be escalated because the model's answer is likely to be text-driven?" That is the core contribution of the architecture-aware audit.

7.4 Discussion

7.4.1 Entropy as a Unified Safety Signal

The UQ-paraphrase bridge provides a practical solution to the evaluation burden of Chapter 6. Computing flip rates requires generating paraphrases and running multiple forward passes. Computing entropy requires a single forward pass. The AUROC of 0.823 on the all-paraphrase set, holding at 0.826 on the operator-preserving subset, is high enough to be worth testing prospectively as a triage signal, and it is the one result in this chapter that requires neither an annotation nor an extra forward pass. Predictions above an entropy threshold can be flagged for human review, and the threshold can be set to achieve a desired coverage-accuracy trade-off.

The bridge also suggests a deeper connection between paraphrase sensitivity and model uncertainty. A model that is uncertain about its answer (high entropy) also tends to be uncertain about how phrasing affects its answer (high flip probability). The results are *consistent with* both quantities reflecting a common instability in the model's decision boundary, though an association of this kind does not

establish that account. The geometric upper-bound argument formalizes this: under bounded-norm text-encoder perturbations and local linearity of the decision boundary, smaller margin loosens the bound on flip probability, and entropy is a monotone decreasing function of margin.

7.4.2 A Tiered Triage Protocol for Prospective Testing

The quantitative results suggest a two-tier triage protocol that balances latency and safety. It is stated here as a hypothesis to be tested prospectively rather than as a recommendation to deploy, since every operating point below was measured retrospectively on the evaluation sets of this chapter:

- **Tier 1 (Default)**: Use softmax entropy from a single forward pass. Cost: 1× base inference. Performance: AUROC 0.862 for error detection, 88.5% coverage at 5% risk on PadChest. Its latency is compatible with real-time clinical workflows, and on Targeted LoRA it is about as good as any multi-pass method evaluated here.
- **Tier 2 (Marginal)**: MC Dropout with $K = 10$ forward passes. Cost: 10× base inference. Performance: AUROC 0.856 for error detection, 88.4% coverage at 5% risk. On LoRA-adapted models the epistemic component is negligible, so MC Dropout matches single-pass entropy and does not justify its extra compute except where the decomposed uncertainty is needed for auditing.
- **Avoid**: LoRA ensemble cross-site deployment. Cost: 5× base inference. Performance: AUROC 0.751 for error detection, only 23.8% coverage at 5% risk on PadChest. The risk of out-of-distribution member collapse (Section 7.1) outweighs any in-distribution benefit.

Both tiers use Targeted LoRA as the base model. Full LoRA is excluded despite its lower flip rate on MIMIC (its PadChest flip rate, 22.9%, is higher than Targeted LoRA's 13.2%) because most of its consistent predictions are invariant under image removal (Chapter 6), so its confidence is largely being assigned to answers the image does not move. A well-calibrated uncertainty signal is only useful if the underlying predictions respond to the image.

These recommendations are architecture-aware, not universal. For Gemma-family models, internal probes such as readout cosine can be useful parts of the audit. For LLaVA-Rad and Qwen2-VL, behavioral controls dominate because the internal probes transfer poorly. Any deployment protocol should therefore begin with a family-specific audit of the admitted subset before choosing between entropy-only triage, multi-pass gating, or human-review escalation. Seen together, the entropy bridge and the architecture-aware audit answer complementary halves of the deployment question: the bridge asks whether *this* prediction is uncertain, while the audit asks which *kinds* of cases the gate admits in the first place.

7.4.3 Why the Deep Ensemble Failed

The MedGemma five-seed ensemble result is the chapter's clearest negative finding. Probability averaging across LoRA checkpoints trained on the same 500 MIMIC-CXR pairs produced an ensemble whose ranking quality was best in-distribution and worst out-of-distribution. The mechanism is mundane: only two of five seeds retain strong PadChest accuracy (92.1% and 78.7%) while the other three degrade (52 to 69%) and shift their answer prior away from the base rate, so averaging recovers a 72.6% mean but a nearly useless confidence ranking (AURC 0.130 versus 0.019 for single-pass entropy). This is not an indictment of deep ensembling in general; the LLaVA-Rad ensemble does not collapse in the same way, because its base model and training distribution provide enough diversity to survive the cross-site shift.

The lesson for deployment is that an ensemble gives nominal robustness only if its members have non-trivial diversity on the deployment distribution. For LoRA-adapted medical VLMs trained on a single hospital's data, that condition is not automatic. Per-member OOD diagnostics should be reported alongside any ensemble UQ result.

Chapter 8 synthesizes the findings across all five thrusts and discusses implications for the clinical

deployment of medical VLMs.

Chapter 8

Conclusion

8.1 Summary of Findings

This dissertation investigated paraphrase sensitivity in medical Vision-Language Models through five interconnected research thrusts. Each thrust addressed a specific question, and the findings formed a progression from measurement to diagnosis to mitigation to safety evaluation to deployment-time gating.

Thrust 1 established that paraphrase sensitivity is prevalent across medical VLMs. The PSF-Med benchmark spans three clinical populations (MIMIC-CXR in the United States, PadChest in Spain, VinDr-CXR in Vietnam) and, after a rubric-based equivalence audit and regeneration of the PadChest paraphrases, comprises an equivalence-filtered set of 8,938 MIMIC, 59,573 second-iteration PadChest, and 24,345 VinDr pairs. On the binary yes/no subset of this set, six medical VLMs flip on 6.4% to 54.7% of paraphrase pairs once two cells of degenerate yes-bias are set aside (LLaVA-Rad on PadChest, MedGemma-1.5-4B on VinDr). Restricting to binary questions changes the in-distribution ordering, so the pre-audit ranking does not carry over: scale within the MedGemma family buys no stable consistency advantage (27B is the *most* consistent of the three sizes on MIMIC but not on PadChest), and the negation-pattern category collapses after the audit rejects 93.8% of generated negation paraphrases as polarity-changing rather than equivalence-preserving. Flip rates are generally higher on the two out-of-distribution populations, but this is a cross-population difference rather than an isolated distribution-shift effect, since the sets also differ in task composition, label balance, and templates; and the two shifts are not commensurate: the US-to-Spain transition produces different per-model rankings than the US-to-Vietnam transition, and clinical deployment evaluations should not generalize from a single OOD test set.

Thrust 2 separated fragile visual grounding from text-only shortcutting as two distinct behavioral failure modes, and showed which models sit on which side, at least for the models that exhibit high flip rates. Chapter 5 then identifies the origin itself. On the curated n=107 three-backend diagnostic set (distinct from the 158-pair mechanistic FlipBank of Chapter 5), text-only and image swap experiments separated these two failure modes across three model backends spanning the consistency spectrum. MedGemma-4B shows a 42.1% flip rate with 30.8% swap sensitivity: it uses the image but is unstable. MedGemma-27B reduces flip rate to 14.0% but increases text-only agreement to 85.0%, so on this set the gain in consistency comes together with more answers that are unchanged when the image is removed and with lower sensitivity to an opposite-label image swap. LLaVA-Rad shows a 6.5% flip rate with 96.3% text-only agreement on this same 107-pair set: it returns the same answer with and without the image on almost every question, and it is the least swap-sensitive of the three backends, so its stability comes with an image contribution that is small on this set rather than with a stable reading of the image. Its flip rate is not a general consistency advantage either: on PadCh-

est in Thrust 1 its cell is separately degenerate, a near-single-class yes prior rather than a measured consistency result. Attention-bounding box analysis found that flip cases show 41% lower attention to pathology regions than consistent predictions ($p < 10^{-8}$), providing suggestive spatial evidence for the grounding-sensitivity link, though the flip labels predate the equivalence audit and may be partly confounded by paraphrase category. The diagnostic FlipBank is distinct from the headline PSF-Med run summarized for Thrust 1: the controlled set was constructed before the equivalence audit, but its qualitative ordering of models matches the four-quadrant counts in Thrust 4.

This dissociation between consistency and grounding is the central tension of the dissertation: it is a warning against reading a low flip rate as evidence of visual grounding, and it is a tendency across the models studied rather than a universal trade-off law.

Thrust 3 developed a targeted mitigation using Sparse Autoencoders and LoRA fine-tuning. Feature 3818 at layer 17 of the Gemma Scope 2 SAE behaves as an operator and register feature, separating presence framing from exclusion framing. It is the top contributor in roughly 25% of flips on a 20-pair FlipBank sample; that figure is a property of a small, operator-mixed sample rather than an estimate of the share of paraphrase flips it explains in the population, and the stricter operator-preserving reanalysis below finds it prominent but not a sufficient cause. Causal validation via activation patching confirmed that patching Feature 3818 recovers 28% of the yes-minus-no logit margin on an exemplar flip case, 3.5× more than matched control features. A larger pre-specified held-out validation then extended this into a candidate two-stage account: Feature 3818 at layer 17 acts as an upstream operator/register feature, and a downstream answer-selection feature, Feature 12139 at layer 29, converts the register shift into the flipped answer, restoring 58% of decision-stage flips on MIMIC when patched toward the original (95% CI [47.7, 67.8]; 27% on PadChest; injecting the feature's paraphrase-direction delta alone induces the flip in only 15% of cases; see the two-stage validation in the supplementary appendix); the two features form a causal chain with 100% direction coherence across 37 mediation pairs. This screen contains negation and operator-changing pairs, however, so the two-stage restoration rates conflate paraphrase sensitivity with operator handling; a clean operator-preserving validation with matched non-flip controls is still required (Section 5.3.1) before the pathway can be claimed as a paraphrase-specific circuit. The SAE transfer validation demonstrated that pretrained Gemma Scope 2 SAEs reconstruct unmodified MedGemma-4B activations without retraining ($R^2 = 0.997$ on medical inputs), and a separate check confirmed the SAE stays valid after Targeted LoRA adaptation (reconstruction FVU rises only from 0.33% to 0.50%), lowering the barrier for mechanistic interpretability of domain-specific models.

A controlled model then established where that sensitivity comes from. A compact Gemma-3 decoder over a frozen MedSigLIP-448 encoder, trained on 86,288 per-finding-balanced presence questions from NIH ChestX-ray14 and PadChest with MIMIC-CXR and VinDr-CXR held out entirely, answers every paraphrase of a question from identical image features, so any answer change is language-side by construction. Holding architecture, parameter budget, seed and evaluation fixed and varying only the training-phrasing distribution moves the flip rate from 4.8% under full paraphrase coverage to 67.1% under a single fixed phrasing and 88.4% when register is correlated with the answer, over eight seeds per regime, with Cliff's $\delta = 1.00$ against both narrow regimes and text-only accuracy pinned between 50.0% and 50.3% throughout. Accuracy follows the same ordering, 75.3%, 66.8% and 58.1%, so the shortcut costs diagnostic performance and not only consistency. Scored only on the twenty-four phrasings withheld from training, the augmented regime rises to 26.6% against 65.9% for canonical, which is the honest size of the augmentation lever once familiarity with the test wording is removed. A rank-1 difference-direction transplant at the answer position restores the flipped answer with net recovery of 0.98 to 1.00 across layers 0 to 4, against 0.00 to 0.02 for a norm-matched random direction. Because flip rate is an arg max statistic, the ordering was re-checked without any decision boundary:

the ratio of within-cluster to between-cluster margin spread is 0.023 for augmented, 0.498 for canonical and 1.362 for adversarial ($\delta = -1.00$, $p = 1.6 \times 10^{-4}$). That same continuous margin doubles as a deployment signal: its absolute value predicts a paraphrase flip in one forward pass at AUROC 0.923 in-distribution and 0.974 on each held-out hospital, outranking paraphrase self-consistency and a hidden-state probe, whereas an image-unreliant answer has no single-pass signal and per-case image reliance registers only at the population level, with no grounded-to-unreliant transition identifiable over 345 patient-seed evaluations.

Guided by this mechanistic insight, a targeted LoRA on layers 15 to 19 with a combined consistency-accuracy loss, trained on the full operator-preserving expanded pool, reduced the presence-of-finding flip rate on a patient- and image-disjoint MIMIC-CXR held-out test (pairwise 8.5% to $3.5\% \pm 0.6$ over five seeds, about 59%; paired McNemar $p < 0.002$ in every seed; zero subject and zero image overlap with training) with no observed accuracy reduction (84.5% to $84.7\% \pm 1.0$; paired difference +0.25 percentage points, 95% interval $[-1.00, +1.51]$, so a large accuracy cost is excluded but exact parity is not established) and while modifying only 0.1% of the model's parameters. On PadChest, the out-of-distribution test, the mitigation transfers as accuracy and margin rather than flip reduction: on the 250-question balanced transfer set, accuracy rises by 6.4 percentage points (85.2% to 91.6%) and the margin difference narrows by 75%, while the base model already flips little on that balanced set so the flip rate barely moves. A layer ablation revealed an important dissociation: early-layer interventions (0 to 10) outperform the mechanistically-identified middle layers (15 to 19) for flip reduction. The SAE analysis identifies *where* sensitivity manifests; the best intervention acts upstream, *before* the problematic representation forms. The combined loss is essential: training on consistency alone causes mode collapse to 46.4% accuracy. A flat Pareto front across $\lambda \in [0.1, 1.0]$ shows the method is insensitive to hyperparameter choice.

Thrust 4 showed that consistency improvements do not automatically make models safer. A four-quadrant behavioral screen classifying predictions along consistency and image-reliance axes established that a low flip rate does not certify grounding: averaged across ten model-dataset combinations, 81% of each model's consistent predictions are image-invariant (range 45 to 100%). The models with the lowest flip rates on the curated PadChest flip bank (Targeted LoRA at 13.2% and LLaVA-Rad at 21.7%) leave almost all of their consistent predictions unchanged when the image is removed. Full LoRA, which achieves 4.1% flips on MIMIC, shows 80.6% text-only agreement there. The flip rate and the image-invariant fraction do correlate at $r = -0.86$, but that correlation is largely a definitional consequence of the image-invariant cell being a subset of consistent predictions (it falls to $r = -0.15$ once conditioned on consistency), so we state the finding as a level rather than a correlation. Consistency at this level tracks agreement with the text-only answer rather than any measured increase in image reliance.

Attention heatmaps are coarse localizers but not faithful causal explanations. Attention concentrates in radiologist-marked pathology regions at 5.2 to 6.4 times the random-box rate (true-box coverage 29.6% to 38.5%), and ablating those regions changes the answer (causal scores 0.6 to 3.0, intervals excluding zero), so the models do use the pathology regions. But coverage only marginally exceeds a displaced-box baseline, and attention magnitude barely correlates with patch-level causal importance ($\rho \approx 0$), so heatmaps identify roughly the right region without pinpointing the patches that drive the decision.

Thrust 4 also reported a fairness analysis stratified by patient sex and age: Targeted LoRA has the smallest between-sex calibration gap (0.012 ECE) but the steepest accuracy gradient across age bins (13 points, youngest patients least accurate), and its 2.7-percentage-point accuracy gap between sexes is slightly wider than Base's 2.0 points; Full LoRA has the largest disparities on every axis (4.3-point accuracy sex gap, 0.041 ECE sex gap, 0.042 AUGRC sex gap) on top of poor calibration everywhere. No variant is uniformly the most equitable. The worst-calibrated model also shows the largest between-

group calibration disparities, suggesting that poor calibration and demographic bias can share common causes. A clinical-validation and deployment-readiness protocol designed with clinician collaborators (Section 6.5) is set out as future work; carrying it out would ground these technical safety metrics in clinical judgment.

Thrust 5 took up the deployment-time companion to the safety evaluation. The UQ-paraphrase bridge established that predictive entropy, computable from a single forward pass, predicts both errors and paraphrase flips: softmax entropy achieves AUROC = 0.823 for flip prediction on the PadChest flip bank, holding at 0.826 on the operator-preserving subset and replicating at AUROC = 0.830 on LLaVA-Rad LoRA on the same dataset. Under added image corruption, Targeted LoRA AUGRC remains the most stable while Base AUGRC degrades, and conformal prediction sets retain near-target empirical coverage for Targeted LoRA while the base model under-covers under shift, most visibly at tighter coverage targets (a descriptive result, since corruption breaks the exchangeability that formal conformal validity assumes). A five-seed LoRA deep ensemble shows severe member instability on PadChest (individual seeds span 52% to 92% accuracy, AURC 0.130) because three of five seeds degrade out of distribution; probability-averaging recovers aggregate accuracy to 72.6% but not a usable confidence ranking, while the LLaVA-Rad ensemble does not collapse, showing that the failure is model- and training-specific rather than inherent to ensembling.

A multi-signal gate, evaluated retrospectively on the 861-item PadChest flip bank, admits 33.0% of Targeted LoRA's predictions at 96.8% accuracy (up from 91.5% unfiltered). A later architecture-aware deployment audit then showed the operational cost of this paradox. On PadChest, Targeted LoRA reaches 98.9% admitted-case accuracy at 72% coverage under a null-image agreement rule, versus 96.7% at 10% for the strict gate. Qwen2-VL reaches 99.8% at 67.6% coverage yet fails completely on the text-disagrees slice (0/279). In both cases, the gate improves aggregate accuracy mainly by selecting text-answerable cases rather than image-grounded reliable ones. No single monitoring signal transfers across Gemma, LLaVA-Rad, and Qwen2-VL: Gemma shows late readout misalignment, LLaVA-Rad broad instability, and Qwen2-VL cosine-insensitive shortcut behavior. Audit-guided escalation catches 77% of candidate image-relevant cases at 72% autonomous coverage for Targeted LoRA.

8.2 Research Questions Revisited

Table 8.1 maps the twelve research questions that carry a headline answer to that answer and the evidence behind it; the remaining questions are answered in their own chapters rather than summarized here. Two results are deliberately reported as mixed or rejected rather than smoothed over: the best intervention layer does not match the mechanistically identified layer (RQ3.3), and the headline relationship between flip rate and the image-invariant fraction is a definitional correlation whose substance is a level rather than a slope (RQ4.1). Reporting these honestly is part of the contribution.

Table 8.1: Research questions revisited: answer and primary evidence. Mixed or rejected hypotheses (RQ3.3, and the correlational form of RQ4.1) are reported as such.

RQ	Answer	Primary evidence
RQ1.1	Paraphrase sensitivity is common (6.4 to 54.7% pairwise flips).	PSF-Med binary yes/no subset, six models, three datasets.
RQ1.3	Scale does not buy consistency; no stable ordering.	27B is most consistent on MIMIC but not on PadChest/VinDr.
RQ2.2	Across the evaluated settings, low paraphrase sensitivity can coexist with output-level image-removal invariance, so a low flip rate is insufficient evidence of grounded visual reasoning.	LLaVA-Rad 96% text-only agreement at low flips; MedGemma-4B image-dependent but flips more.

Continued on next page.

Table 8.1 continued from previous page.

RQ	Answer	Primary evidence
RQ3.1	A *candidate* two-stage mechanistic account: Feature 3818 (L17) as an operator/register feature and Feature 12139 (L29) as a downstream answer-selection feature.	Feature deltas, activation patching, and lens-free L16 localization.
RQ3.2	Partly: targeted LoRA cuts flips by about 59% (8.5%→3.5%±0.6 pairwise, 5 seeds) with no observed accuracy reduction (95% interval $[-1.00, +1.51]$ pp), but does *not* preserve visual grounding.	Clean operator-preserving rerun on a patient- and image-disjoint split (presence-of-finding primary endpoint, n=238); paired McNemar $p < 0.002$ every seed; 0.1% of parameters.
RQ3.3	**No (mixed):** the best intervention layer does not match the diagnosis layer.	Early layers (0 to 10) beat the identified L15 to 19 for flip reduction.
RQ4.1	Reducing flips can create a *false* sense of safety.	81% of consistent predictions are image-invariant (range 45 to 100%).
RQ4.2	No: attention does not reliably indicate grounding.	Coverage barely exceeds a displaced box; patch saliency vs. causal importance $\rho \approx 0$.
RQ4.3	Partially: the worst-calibrated model also shows the largest disparities.	Full LoRA has worst calibration and largest sex/age gaps.
RQ5.1	Yes: single-pass entropy predicts flips, not only errors.	AUROC 0.82 (flip) / 0.86 (error) on PadChest.
RQ5.3	No: gates admit text-answerable rather than image-grounded cases.	Gates raise admitted-case accuracy by selecting the text-answerable slice.
RQ5.4	No: no single internal monitor transfers across families.	Gemma, LLaVA-Rad, and Qwen2-VL each need different monitors.

8.3 The Central Thesis

The central finding is a dissociation: consistency and image reliance are distinct properties that must be evaluated jointly, and improvements in consistency can coexist with reduced image reliance in the evaluated models. Models can achieve consistency in two ways: by building stable representations that integrate visual evidence reliably (the goal), or by defaulting to text-based priors that the image rarely overrides (a shortcut). Current evaluation practices, which reward consistency without testing image reliance, cannot distinguish between these two paths.

This dissociation is not merely academic. A model deployed in a clinical setting that achieves 98% consistency while its answers barely move when the chest X-ray is removed or replaced appears reliable to every evaluation metric except the ones we developed here. A selective-prediction gate can make such a system look even safer by admitting only the text-answerable cases. It gives the same answer to every phrasing of the question, passes standard accuracy benchmarks (because text priors often give the right answer), and produces confident probability estimates. But on most of its consistent answers it would return the same yes or no with the image removed entirely, which is what the text-only control measures, and a subset of those answers also survives an image showing the opposite finding, which is the outcome the image-swap control measures directly (Chapter 4); the two controls are not interchangeable.

The implication is that safe deployment of medical VLMs requires evaluation along at least two axes: consistency (does the model give the same answer to equivalent questions?) and grounding (does the answer depend on the image?). Neither axis alone is sufficient. A model must be both consistent and grounded to be safe, and current models rarely achieve both simultaneously.

The five thrusts of this dissertation provide the tools to make this distinction. PSF-Med measures consistency (Thrust 1). The text-only and image-swap protocols measure image reliance, with the

swap control carrying the stronger claim of the two (Thrust 2). The targeted SAE-guided LoRA reduces flip rate at a fraction of the parameter cost with no observed accuracy reduction (95% interval $[-1.00, +1.51]$ percentage points), though a clean patient-disjoint audit shows it does not preserve image dependence (Thrust 3). The four-quadrant taxonomy classifies individual predictions along both axes (Thrust 4). Predictive entropy provides a single-forward-pass proxy for deployment-time triage, and the architecture-aware audit tests what kinds of cases a deployment gate actually admits (Thrust 5). Together, these tools form a safety evaluation and admission-audit pipeline that goes beyond accuracy and consistency to address the questions that matter for clinical use: *how much evidence is there that this model used the image to answer the question, and what kinds of cases does a gate over it actually admit?*

8.4 Contributions

This dissertation makes the following contributions to the fields of medical AI, vision-language modeling, and AI safety:

1. **PSF-Med benchmark.** A publicly available benchmark of 26,850 clinical questions and 92,856 equivalence-filtered evaluation pairs, spanning three clinical populations (MIMIC-CXR, PadChest, VinDr-CXR) and five paraphrase transformation types. To our knowledge, this is the first large-scale benchmark specifically designed to measure paraphrase sensitivity in medical VLMs (the systematic review in Chapter 2 did not identify a prior benchmark targeting this failure mode). A 122,778-pair rubric-based audit establishes a stricter core-equivalent subset in which the benchmark's central conclusion, the contrast between a low flip rate and weak image reliance, still holds. The subsequent restriction to genuine binary questions does change the in-distribution ordering, so the pre-audit ranking is not claimed to carry over. The benchmark is rubric-filtered with partial clinician validation: the clinician study evaluates the filtering procedure on a stratified pre-regeneration audit sample. The benchmark, evaluation code, and pre-computed results are released at HuggingFace (saillab/psf-med).
2. **Consistency is not evidence of grounding.** An empirical characterization of the dissociation between paraphrase consistency and visual grounding in medical VLMs, across six base models, three datasets, and four behavioral controls (image removal, opposite-label image swap, null image, and region-of-interest ablation). This finding challenges the assumption that a lower flip rate implies a safer, more image-grounded model and motivates joint evaluation of consistency and grounding; it is a tendency across the models studied, not a deterministic trade-off law.
3. **Feature 3818 mechanistic analysis.** To our knowledge, the first application of Sparse Autoencoders to diagnose paraphrase sensitivity in a medical VLM. The identification of Feature 3818 as a clinical query register gate, validated through activation patching on a single exemplar flip case with 3.5× greater margin recovery than matched control features (on the stricter operator-preserving re-analysis, single-feature ablation restores only 6 of 76 flips, so the feature is prominent rather than a sufficient cause), demonstrates that SAE-based interpretability can yield actionable insights for medical models. The SAE transfer validation ($R^2 = 0.997$ on unmodified MedGemma-4B) establishes that pretrained Gemma Scope 2 SAEs reconstruct MedGemma activations without retraining, and a separate check confirms the decomposition stays valid after Targeted LoRA (reconstruction FVU rises only from 0.33% to 0.50%).
4. **Controlled isolation of the origin.** A purpose-built model that mirrors MedGemma's architecture on a frozen MedSigLIP-448 encoder, so that every paraphrase of a question is answered from identical image features, turns the origin of paraphrase sensitivity from an observational question into a manipulable one. Varying only the training-phrasing distribution moves the flip rate from 4.8% to 67.1% to 88.4% with text-only accuracy at 50.0% to 50.3% in every regime, which identi-

fies the training-phrasing distribution as a cause. The design separates two harms that prior work conflates, narrow phrasing coverage and a phrasing shortcut ($p = 3.1 \times 10^{-4}$, Cliff's $\delta = 0.97$), states the augmentation lever on wording never seen in training (26.6% against 65.9%), and reproduces the ordering under a statistic containing no decision boundary. The same probe establishes that the discarded margin is a single-pass flip monitor (AUROC 0.923 to 0.974), while an image-unreliant answer has no single-pass substitute and per-case image reliance is identifiable only at the population level. The probe's design requirements double as training and evaluation practices for dependable medical VLMs: balance the training distribution per finding so the question text alone predicts nothing, carry an explicit grounding signal into the decoder, train on all phrasings rather than one fixed wording, and read the answer margin rather than the binary answer alone.

5. **Combined consistency-accuracy loss.** A training objective that balances symmetric KL divergence (for consistency) with cross-entropy (for accuracy), preventing the mode collapse that occurs with pure consistency optimization. The flat Pareto front across $\lambda \in [0.1, 1.0]$ shows insensitivity to the mixing coefficient.
6. **Targeted LoRA mitigation.** A parameter-efficient intervention (0.1% of model parameters, layers 15 to 19) that cuts the presence-of-finding flip rate by about 59% on a patient- and image-disjoint held-out split (a clean operator-preserving rerun over five seeds, paired McNemar $p < 0.002$ in every seed) with no observed accuracy reduction (95% interval $[-1.00, +1.51]$ percentage points; a large cost is excluded, exact parity is not established). The layer ablation and replication study provide systematic evidence for the approach's reliability, including a training-size sweep showing the in-distribution result has plateaued.
7. **Four-quadrant behavioral screen.** A classification framework for individual model predictions that distinguishes consistency whose answer changes when the image is removed (Consistent-Grounded) from consistency that persists without the image (Consistent-Text-driven), with correctness as a separate axis. The screen establishes that a low flip rate does not certify grounding (a mean of 81% of consistent predictions are image-invariant) and provides a per-prediction behavioral diagnostic, to be paired with correctness and image-reliance checks rather than read as a standalone safety verdict.
8. **Architecture-aware deployment audit.** A five-step deployment-audit framework showing that selective-prediction gates can admit text-answerable rather than image-grounded cases, that grounded-slice behavior is the safety-relevant stress test, and that family-specific monitoring is required. In the strongest Gemma-family result, audit-guided escalation catches 77% of candidate image-relevant cases at 72% autonomous coverage; in Qwen2-VL, the highest-accuracy admitted subset is also the one most dominated by the text channel.
9. **UQ-paraphrase bridge.** The finding that predictive entropy predicts paraphrase flips (AUROC = 0.823), enabling a single-forward-pass deployment signal for both error detection and consistency monitoring. This eliminates the need for paraphrase generation at inference time, but not the need to periodically audit whether the admitted subset remains image-grounded.
10. **Safety evaluation protocol.** An eight-phase evaluation pipeline that tests consistency, text reliance, image sensitivity, ROI ablation, and attention grounding across multiple models and datasets. The protocol is designed to detect the specific failure modes identified in this dissertation and is released as reproducible code.

8.5 Practical Recommendations

Five recommendations follow from this work for practitioners deploying or evaluating medical VLMs:

1. Evaluate consistency alongside accuracy. Standard medical VLM benchmarks test accuracy on fixed question sets. Adding paraphrase variants to these benchmarks is cheap (automated

generation with semantic filtering) and reveals a failure mode that accuracy testing misses. Flip rate and margin difference are simple, interpretable metrics that should be reported alongside accuracy for any medical VLM claiming clinical readiness.

2. Always test image reliance. Any model that achieves low flip rate should be tested under text-only conditions (image removed) and image swap conditions (image replaced with one showing the opposite finding). Without these controls, low flip rate is ambiguous: it is equally consistent with stable use of the image and with a text prior that the image rarely overrides. We recommend reporting text-only agreement and swap sensitivity as standard metrics, and treating the swap result as the stronger of the two: an unchanged answer after the image is removed bounds how much the image contributed, whereas an unchanged answer after the image is replaced with one carrying the opposite finding is much harder to explain by redundant evidence or by a saturated readout.

3. Use entropy for deployment-time triage, but pair it with periodic audit checks. Predictive entropy, computable from a single forward pass, predicts both errors and flips. A deployment system that flags high-entropy predictions for human review catches both types of unreliability through one mechanism. This is simpler than running paraphrase variants at inference time and provides a practical first line of defense. At a 40% coverage operating point with Targeted LoRA on PadChest, the error rate on admitted predictions drops substantially below the population rate. But entropy alone cannot tell whether the admitted subset is image-grounded, so it should be paired with periodic audits of null-image agreement, swap sensitivity, and grounded-slice accuracy.

4. Prefer targeted over full-model fine-tuning, but not because it stays grounded. On a clean patient-disjoint re-audit of the exact adapters (Section 5.7.5), the targeted and full adapters are statistically tied on both image-reliance measures, and *both* raise text-only agreement sharply relative to the base model (53.5% to about 77%). Targeted LoRA (5 layers, 0.1% of parameters) is still the better choice, because at tied consistency and tied image reliance it preserves accuracy (84.3% versus 82.4% for Full LoRA) at a fraction of the parameters. The smaller intervention is the cheaper and more accurate one, not a more grounded one: neither adapter should be deployed on the strength of its flip-rate reduction without the image-reliance checks of recommendation 2.

5. Use architecture-aware audits instead of one universal gate. Monitoring signals do not transfer cleanly across Gemma, LLaVA-Rad, and Qwen2-VL. Before deployment, every model family should be audited with the same behavioral controls on admitted subsets: null-image agreement, image swap sensitivity, and performance on grounded slices where the image and text channels disagree. A gate is only acceptable if those audits show it is admitting image-grounded cases, not merely text-answerable ones.

8.6 Limitations and Future Work

This dissertation has several limitations that bound the generality of its findings and point toward future work.

Model scope. The mitigation experiments and the sparse-autoencoder analysis focus on MedGemma-4B, a single model based on the Gemma 3 architecture; the controlled probe of Thrust 3 is a second, purpose-built model, but it is also Gemma-3 based, so the architectural scope of the mechanistic account is unchanged. The SAE transfer validation and LoRA approach should generalize to other models with shared architectures, but this has not been empirically tested. Extending to models with different backbones (LLaMA-based medical VLMs, proprietary models like GPT-4V, or encoder-only architectures) would strengthen the generality claims. The flip-rate phenomenon itself was checked on four frontier configurations in a limited comparison (Section 3.4.7); what remains untested outside Gemma-3 is the mechanistic account and the mitigation. The Feature 3818 mechanism is specific to MedGemma-4B; other models likely have different internal mechanisms for

processing clinical query register.

Task scope. The evaluation is limited to binary (yes/no) questions about chest X-rays. Open-ended questions ("Describe the findings in this image"), multi-label findings (co-occurring pathologies), and other imaging modalities (CT, MRI, pathology, dermatology) present different consistency challenges where flip rate and margin difference may not directly apply. Chest X-rays are the most common modality for medical VLM evaluation, but the generalization to other modalities is untested. The margin-based framework extends naturally to any setting with comparable output distributions, but the practical details differ.

Annotation requirements. The combined loss requires ground-truth labels for the accuracy term, limiting applicability to settings where labels are available. Exploring pseudo-label approaches (using high-confidence base model predictions as soft targets) or contrastive objectives that operate in activation space rather than output space could remove this requirement and enable consistency training on unlabeled clinical data.

Deep ensemble instability. The deep ensemble, a standard approach for uncertainty estimation, showed large seed-to-seed variance on PadChest: individual members ranged from 52% to 92% accuracy despite 87 to 89% training validation accuracy. Only one seed (92%) held its in-distribution accuracy and a second retained strong accuracy (79%), while three degraded below 70% out of distribution (52%, 67%, 69%). Probability-averaging recovered 72.6% aggregate accuracy but a poorly ranked confidence signal (AURC 0.130), leaving the ensemble unsuitable for selective prediction. This is a specific instance of out-of-distribution instability in LoRA ensembles and deserves investigation. Understanding why some seeds generalize and others collapse could improve ensemble-based uncertainty estimation for domain-shifted medical data. We document this as a negative result rather than omitting it, as it has implications for anyone attempting LoRA ensembles on medical VLMs.

What the null-image control can and cannot show. The image-reliance axis of the four-quadrant screen is measured by removing the image and checking whether the hard yes or no answer changes. An unchanged answer bounds the contribution of the image, but it does not establish that the image played no part in the decision. The same observation is produced by redundant evidence, in which the text alone already implies the answer that the image also supports; by a saturated readout, in which the decision margin moves but not far enough to cross the decision boundary; and by a base rate under which the text prior is usually right. The stronger claims therefore rest on the controls that manipulate rather than delete the evidence, namely the opposite-label image swap, the region-of-interest ablation, and the Kullback-Leibler divergence and margin comparisons between the image-conditioned and text-only distributions, all of which read the size of the shift instead of only the sign of the answer. Where this dissertation states that a model's answer does not depend on the image, that statement is scoped to the subset on which one of those controls was run.

Attention grounding. Our analysis showed that attention maps localize pathology only coarsely and do not faithfully rank causally important patches ($\rho \approx 0$ between attention saliency and patch-level causal importance). We did not develop an alternative explanation method that provides reliable grounding evidence. Causal tracing methods that identify which image patches actually influence the model's decision could provide more faithful explanations. The patch occlusion approach used in Thrust 4 is computationally expensive (requiring $O(N)$ forward passes for N patches, 256 per image for the 16×16 grid); more efficient causal explanation methods would be valuable for clinical deployment.

Ecological validity. The paraphrases in PSF-Med were generated with GPT-5-family models (GPT-4 in the superseded original construction), filtered by BioClinicalBERT embedding similarity, and retained only after a rubric-based equivalence audit. While this ensures semantic equivalence and enables large-scale evaluation, the paraphrases may not reflect the actual variation in how clinicians phrase questions. A human-authored paraphrase set, collected from practicing radiologists, would

provide stronger ecological validity. The semantic filtering may also exclude some clinically relevant but linguistically distant reformulations.

Longitudinal evaluation. All evaluations are cross-sectional: we test models at a single point in time. Clinical deployment involves ongoing monitoring, and paraphrase sensitivity might change as models are updated, retrained, or as the distribution of clinical queries shifts over time. Longitudinal consistency monitoring, analogous to concept drift detection in deployed machine learning systems, is an open problem that this work does not address.

Sample of models and statistical power. The consistency-safety finding is deliberately stated as a level, not a correlation. Because the Dangerous fraction is by construction a subset of the consistent predictions, its correlation with flip rate ($r = -0.86$) is largely a definitional identity, and it collapses to $r = -0.15$ once conditioned on consistency (Chapter 6); permutation or rank-correlation checks do not remove this coupling and are not offered as a rescue. The claim rests instead on the conditioned level: a mean of 81% of each model's consistent predictions are unchanged when the image is removed. That level is estimated from ten model-dataset combinations across five models, so a larger and more architecturally diverse set would sharpen it, and several mechanistic replications, including the small left-versus-right laterality check, are small and are reported as supporting rather than primary evidence.

Mechanistic completeness. The two-stage account identifies Feature 3818 at layer 17 as an upstream register gate and Feature 12139 at layer 29 as a downstream decision gate, and the two form a direction-coherent causal chain. These are the dominant single features, not the complete circuit: the account covers the register-active and register-flat regimes but not the ambiguous regime (36 to 49% of pairs, where flips are rare), and Feature 12139 is sufficient on its own for only about 15% of decision-stage flips, so the decision is represented across many features in superposition. A full circuit-level account of paraphrase sensitivity remains open.

Several directions follow from the findings and limitations above.

8.6.1 Cross-Architecture, Cross-Modality, and Multi-Site Generalization

The Feature 3818 analysis shows that SAEs can diagnose paraphrase sensitivity in a Gemma-based medical VLM. Testing whether query register sensitivity is a general mechanism or architecture-specific requires extending it to LLaMA-based models (LLaVA-Rad, LLaVA-Med) and encoder-only models (CheXagent); different mechanisms would motivate architecture-specific mitigations. In parallel, extending PSF-Med beyond binary chest X-ray questions to CT, MRI, pathology, and dermatology, and to open-ended and multi-label outputs, would test whether the consistency-grounding dissociation holds across modalities and output spaces. A multi-site study spanning low-resource settings, community versus academic hospitals, and portable versus fixed imaging equipment, together with cross-lingual paraphrases (PadChest's original Spanish reports are a natural testbed for queries such as "*¿Hay derrame pleural?*" versus "Is there pleural effusion?"), would map the dissociation across the full range of deployment contexts.

8.6.2 Label-Efficient and Calibration-Aware Training

The combined loss in Thrust 3 needs ground-truth labels for the accuracy term. Pseudo-labeling from high-confidence base predictions, contrastive objectives that align paraphrase representations without task labels, or self-distillation from a slowly updating teacher could enable consistency training on unlabeled clinical data without mode collapse. The poor calibration of text-reliant models under shift (Full LoRA's ECE on clean PadChest is 22.9% even after temperature scaling, 26.9% unscaled), together with the failure of temperature scaling to transfer across domains, suggests folding calibration losses into the same objective so that models are consistent, accurate, and calibrated at once rather than corrected post hoc.

8.6.3 Deployment-Time Monitoring and Efficient Uncertainty

Single-pass softmax entropy predicts both errors and flips (AUROC = 0.823) and outperforms MC Dropout and deep ensembles, which suggests uncertainty estimation for medical VLMs should prioritize total entropy over decomposed epistemic/aleatoric components; whether this holds for larger models and richer output spaces is open. Because the four-quadrant screen and deployment gate are point-in-time assessments, a production system should track flip rate, text-only agreement, and entropy over rolling windows to catch model and concept drift, and pair per-prediction flags with periodic audits of whether the admitted subset remains image-grounded.

8.6.4 Improving Visual Grounding

The most fundamental limitation is how little evidence of precise visual grounding any tested model provides: attention localizes pathology only coarsely and does not faithfully identify the causally important patches, and the consistency gains measured here did not come with any increase in image reliance. Training objectives that penalize invariance to pathology-region occlusion or supervise on radiologist bounding boxes, and architectures that separate visual feature extraction from language-driven reasoning, are the most direct routes to more transparent and modifiable grounding.

8.6.5 Regulatory and Clinical Integration

The distinction between apparent safety (low flip rate, high accuracy) and genuine safety (image-grounded consistency) is a concrete, testable criterion that current frameworks do not capture. Image-reliance testing (text-only baselines and controlled swaps) and the eight-phase protocol could be incorporated into FDA premarket assessment of machine learning-enabled devices and EU AI Act conformity assessments for high-risk medical AI, alongside the deployment gate's requirement for paraphrase agreement and image reliance.

8.6.6 Clinician Perspectives on Deployment Readiness

Section 6.5 reports a completed clinician equivalence adjudication with the clinician coauthors; the deployment-readiness consultation is designed but not yet conducted. A larger formal study with an independently recruited clinician sample would test at scale which failure modes clinicians prioritize, whether the Dangerous quadrant is treated as worse than the Fragile quadrant, what coverage thresholds are acceptable, and how entropy-based uncertainty should be surfaced in a clinical interface, questions that automated evaluation alone cannot answer.

The four scenarios. The assessment presents the clinician coauthors with four deployment scenarios drawn directly from the dissertation's findings:

1. **Paraphrase flip case**: the same image with two clinically equivalent phrasings that produce contradictory model answers (Chapter 3).
2. **Consistent but Dangerous case**: a prediction with low flip rate but high text-only agreement and low image-swap sensitivity (the Dangerous quadrant, Section 6.2).
3. **Uncertainty triage case**: a model answer accompanied by an entropy-based uncertainty flag and recommended escalation (Chapter 7).
4. **Deployment gate case**: an answer displayed only when paraphrase agreement and at least one image-reliance check pass (Chapter 7).

For each scenario, the assessment addresses which failure modes matter most in practice, which safeguards are actionable in clinical workflow, how uncertainty should be displayed to avoid misleading the reader, and in which clinical contexts the system should be permitted, restricted, or refused. The consultation question set is reproduced in Appendix C.4.

Intended outputs. This assessment is designed but not yet completed; the protocol and intended outputs are described here as declared future work. By design, the assessment would produce a ranked

list of clinically meaningful failure modes, safeguard preferences, deployment-scope boundaries, and a mapping from technical safety signals (paraphrase agreement, image-reliance checks, entropy, gating) to clinician-facing deployment requirements. Its purpose is to ground the four-quadrant taxonomy in clinical judgment (whether Dangerous predictions are treated as worse than Fragile ones), to test whether the deployment gate's admission rate (roughly 27 to 42% on PadChest) is acceptable in specific clinical contexts, and to yield design requirements for presenting entropy-based uncertainty in clinical interfaces. These outputs would inform the admission audit developed in Chapter 7.

8.7 Closing Remarks

Despite the limitations noted above, this dissertation establishes that paraphrase sensitivity is a measurable, diagnosable, and partially treatable failure mode in medical Vision-Language Models, and that treating it requires attention to the distinction between consistency and grounding. This distinction, simple in hindsight, is absent from current evaluation practices and deployment guidelines for medical AI.

The trajectory from measurement (PSF-Med, Thrust 1) through diagnosis (Feature 3818, Thrusts 2 to 3) to mitigation (Targeted LoRA, Thrust 3) to safety evaluation (four-quadrant taxonomy, Thrust 4) to deployment-time gating (UQ-paraphrase bridge and architecture-aware audit, Thrust 5) demonstrates a principled approach to addressing reliability failures in medical AI: measure the problem at scale, understand its mechanism, develop targeted interventions, and then critically evaluate whether the intervention actually improves safety or merely changes the failure mode. The consistency-safety finding (a mean of 81% of consistent predictions are image-invariant) and the gate-audit results show that the last two steps are essential. Without them, we risk deploying models that appear safer by every available metric while their answers depend only weakly on the clinical evidence they are meant to analyze.

Medical VLMs will play an increasingly important role in clinical workflows. Ensuring that they are both consistent and grounded means requiring positive evidence that the image moved the answer, from an opposite-label image swap, a region ablation, or a measured shift in the decision margin, and refusing to accept an answer that simply fails to change as that evidence. This dissertation provides the tools and evidence to make that distinction.

Appendix A

Supplementary Material

A.1 PSF-Med Benchmark Details

The subsections on generation and filtering below document the original (v1) construction pipeline of the benchmark: GPT-4 candidate generation at three to five paraphrases per question and a 0.90 embedding-similarity threshold. The benchmark as evaluated in Chapter 3 supersedes that pipeline: MIMIC-CXR and VinDr-CXR paraphrases were generated with GPT-5-mini (five per question), PadChest was regenerated with GPT-5.4-mini (six per question) under judge-informed constraints, and the MIMIC embedding prefilter is 0.95. The v1 prompt and the threshold-selection study are retained here for provenance.

A.1.1 Paraphrase Generation Prompt

The following prompt template was used with GPT-4 to generate paraphrases for each clinical question in the PSF-Med benchmark:

> You are given a clinical question about a chest X-ray. Generate 3 to 5 semantically equivalent paraphrases that preserve the clinical meaning while varying the surface form. Use the following transformation strategies:
>
> 1. Lexical substitution: Replace medical terms with synonyms or lay equivalents. 2. Syntactic restructuring: Change clause order or sentence structure. 3. Formality shift: Move between formal clinical and informal phrasing. 4. Scope variation: Adjust quantifiers ("any evidence" vs. "significant evidence"). 5. Negation-adjacent: Rephrase around negation without changing the expected clinical answer.
>
> Each paraphrase should be a complete, grammatically correct question that a clinician might naturally ask. Do not change the clinical finding being queried.
>
> Original question: [QUESTION]

Temperature was set to 0.7 to encourage diversity while maintaining coherence. Each question was processed independently; no inter-question context was provided.

A.1.2 Semantic Filtering Threshold Selection

The cosine similarity threshold of 0.90 for BioClinicalBERT filtering was selected through a manual annotation study. Two annotators (one with clinical training, one with NLP expertise) independently labeled 200 question-paraphrase pairs as "meaning-preserving" or "meaning-changing." Inter-annotator agreement was $\kappa = 0.87$ (Cohen's kappa). Using this annotated set:

The 0.90 threshold provides the best precision-recall balance for our use case, where false positives (including meaning-changing paraphrases) would inflate flip rates and false negatives (excluding valid paraphrases) reduce statistical power.

Table A.1: Precision and recall of semantic filtering at different BioClinicalBERT similarity thresholds, evaluated on 200 manually annotated pairs.

Threshold	Precision	Recall
0.80	82%	98%
0.85	89%	96%
0.90	94%	91%
0.95	98%	72%

A.1.3 Answer Parser Validation

The answer parser was validated against 500 manually audited model outputs. The validation set was stratified by model (100 outputs per model for the top 5 models) and by answer type (250 "Yes" responses, 250 "No" responses). This audit predates the current model roster: it covers MedGemma-4B, MedGemma-27B, LLaVA-Rad, RadFM, and CheXagent. CheXone has since replaced CheXagent in the evaluated set and MedGemma-1.5-4B was added later, so the 97.4% overall agreement and the per-model exclusion rates below describe those five audited models, not the full current roster (Chapter 3). Results:

Table A.2: Answer parser validation results by model and answer type. Overall agreement with human judgment is 97.4%.

Model	Agreement	Exclusion Rate	Common Error
MedGemma-4B	98.0%	5.2%	Hedge responses
MedGemma-27B	99.0%	3.2%	Equivocal "likely"
LLaVA-Rad	97.0%	4.8%	Verbose preambles
RadFM	95.0%	8.7%	Off-topic outputs
CheXagent	98.0%	6.1%	Multi-sentence hedges
Overall	**97.4%**	**5.6%**	N/A

The primary source of parser-human disagreement is hedge responses where the model expresses uncertainty without committing to yes or no (e.g., "The image is suggestive of possible pleural effusion, though further imaging may be warranted"). These cases are excluded from analysis rather than forced into binary labels.

A.1.4 Operator Change Classification

The operator change classifier uses the following lexical rules to categorize question framing:

- **Presence framing**: Contains "Is there," "Does this show," "Is [finding] present," "Can you see," "Is there evidence of"
- **Exclusion framing**: Contains "Can [finding] be ruled out," "Is [finding] excluded," "Can you exclude," "Is there no evidence of," "Can [finding] be excluded"
- **Likelihood framing**: Contains "Is [finding] likely," "How confident," "Is it possible that"
- **Neutral framing**: None of the above patterns matched

When a question-paraphrase pair changes from presence to exclusion framing (or vice versa), the expected answer polarity inverts. A model that correctly tracks this inversion should not be penalized with a flip. On the unfiltered earlier benchmark, the classifier identifies approximately 28% of apparent flips as operator changes (27 to 30% across models). On the filtered run the residual correction is small, because the equivalence audit has already rejected most operator-changing pairs as not equivalent (Chapter 3).

A.2 Model Configuration Details

A.2.1 MedGemma-4B

- **Model ID**: `google/medgemma-4b-it`
- **Architecture**: Gemma 3 backbone with MedSigLIP vision encoder
- **Parameters**: ≈4.4B total (Gemma 3 language backbone ≈4.0B, MedSigLIP vision encoder ≈0.4B; the multimodal projector is a small linear layer). Consistent with the 4.38M trainable Targeted-LoRA parameters being 0.10% of the total (Table A.3).
- **Layers**: 34 transformer layers in the language model
- **Image tokens**: 256 per image (16×16 grid from MedSigLIP)
- **Context length**: 128K-token input context (image tokens included); 8,192-token maximum output length
- **Precision**: bfloat16
- **GPU memory**: ∼8GB in bfloat16, ∼4GB with 4-bit quantization
- **Decoding**: Greedy (temperature 0)

A.2.2 MedGemma-27B

- **Model ID**: `google/medgemma-27b-it`
- **Architecture**: Gemma 3 backbone (27B) with MedSigLIP vision encoder
- **Parameters**: 27.2B total
- **GPU memory**: ∼54GB in bfloat16
- **Decoding**: Greedy (temperature 0)

A.2.3 LLaVA-Rad

- **Model ID**: `microsoft/llava-rad`
- **Architecture**: BiomedCLIP vision encoder + Vicuna-7B language model
- **Parameters**: 7.0B total
- **Note**: Requires a separate conda environment due to dependency conflicts with the MedGemma setup
- **Decoding**: Greedy (temperature 0)

A.2.4 LoRA Configurations

A.3 Sparse Autoencoder Details

A.3.1 Gemma Scope Configuration

We use the Gemma Scope SAE at layer 17 with the following configuration:

- **Width**: 16,384 features (16k)
- **Architecture**: JumpReLU SAE
- **Training data**: General web text (Gemma 3 pretraining distribution)
- **Fraction of variance unexplained (FVU) on general text**: 0.26%
- **FVU on medical text**: 0.28%
- **Mean L0 (features active per token)**: 70 ± 10 on medical inputs
- R^2 **on medical inputs**: 0.9972 ± 0.0005

A.3.2 Feature 3818 Prompt Grid

Table A.4 shows the full prompt grid used to characterize Feature 3818. All prompts were evaluated on the same pneumothorax-positive image to control for visual input.

The recorded category means are 344.5 for presence, 0.0 for exclusion, 169.5 for disambiguation, and 296.0 for the token-control prompts. Presence activations range from 302 to 386, while every

Table A.3: LoRA hyperparameters for the Targeted and Full configurations.

Parameter	**Targeted LoRA**	**Full LoRA**
Target layers	15 to 19	0 to 33
Rank (r)	16	16
Alpha (α)	32	32
Dropout	0.05	0.05
Target modules	q, k, v, o, gate, up, down	q, k, v, o, gate, up, down
Trainable params	4.38M (0.10%)	30.5M (0.72%)
Learning rate	2×10^{-4}	2×10^{-4}
Warmup steps	100	100
Batch size	8	8
Epochs	3	3
Optimizer	AdamW	AdamW
Weight decay	0.01	0.01
λ (accuracy weight)	1.0	1.0
Training time (A100)	$\sim$20 min	$\sim$45 min

exclusion prompt has zero activation. Disambiguation spans 0 to 282, and the token-control range is 272 to 342. The grid therefore shows that Feature 3818 changes across question forms even when every prompt concerns pneumothorax and uses the same image.

A.4 Candidate Two-Stage Account: Held-Out Validation

The FlipBank analysis in Chapter 5 identified Feature 3818 at layer 17 as a case study but acknowledged that it accounts for roughly 25% of observed flips. This section reports a larger-scale pre-specified held-out validation using a canonical pipeline, identifies a second causal feature, and establishes a candidate two-stage account, an operator/register feature plus a downstream answer-selection feature, that covers a substantially larger fraction of the population; it remains exploratory pending same-polarity matched controls. (Hypotheses were frozen on a discovery split before the held-out analysis but were not publicly preregistered.)

A.4.1 Canonical Evaluation Protocol

To eliminate prompt-wrapper and scoring confounds, we fixed three elements across all subsequent experiments: (i) prompt construction via the model's chat template with a standardised system instruction; (ii) Yes/No margin scoring from logits over a frozen token set; and (iii) a *margin-flip* criterion requiring sign change with non-zero margin on both sides. Under this protocol we screened 4,542 MIMIC-CXR and 37,280 PadChest original/paraphrase pairs, identifying 436 verified flips on MIMIC (9.6%) and 2,833 on PadChest (7.6%).

A.4.2 Two-Regime Classification

We partitioned verified flips by the activity of Feature 3818: pairs where the feature's absolute delta exceeds 50 units on both sides form the *3818_active* regime; pairs where Feature 3818 is exactly zero on both sides form the *3818_flat* regime; the remainder are *ambiguous*. On MIMIC the split is 27.7% active, 36.4% flat, and 36.0% ambiguous. On PadChest the split is 45.7% active, 5.2% flat, and 49.1% ambiguous. Within the flat regime, f3818 is exactly zero on 100% of pairs on both datasets, so regime assignment is unambiguous.

Feature ranking on the discovery set identified Feature 12139 at layer 29 as the top flip-associated feature in the flat regime. We froze both hypotheses, L17/#3818 for the active regime and L29/#12139

Table A.4: Recorded Feature 3818 prompt grid across presence, exclusion, disambiguation, and token-control question forms.

Prompt	F3818 Activation	Category
"Is there evidence of pneumothorax?"	302	Presence
"Is pneumothorax present?"	354	Presence
"Does this show pneumothorax?"	336	Presence
"Can you identify pneumothorax?"	386	Presence
"Can you rule out pneumothorax?"	0	Exclusion
"Is pneumothorax absent?"	0	Exclusion
"Is there no sign of pneumothorax?"	0	Exclusion
"Can pneumothorax be excluded?"	0	Exclusion
"Cannot rule out pneumothorax - do you agree?"	0	Disambiguation
"Is pneumothorax possible but uncertain?"	173	Disambiguation
"Is there questionable pneumothorax?"	223	Disambiguation
"Might there be pneumothorax?"	282	Disambiguation
"Is there radiographic indication of pneumothorax?"	272	Token control
"Are findings consistent with pneumothorax?"	284	Token control
"Does imaging suggest pneumothorax?"	286	Token control
"Is pneumothorax demonstrated?"	342	Token control

for the flat regime, before running large-scale validation.

A.4.3 Pre-Specified Held-Out Validation

Table A.5 reports the held-out validation. On the active regime, patching L17/#3818 toward the original prompt restored 24% of MIMIC flips (95% CI [16.0, 33.6], $p < 10^{-15}$, Wilcoxon signed-rank) and 17% of PadChest flips, compared to 1% and 0% for matched weak controls. On the flat regime, patching L29/#12139 restored 58% of MIMIC flips (95% CI [47.7, 67.8]) and 27% of PadChest flips, again with near-zero controls. Non-flip disruption rates were 0% in all arms.

Cross-regime specificity was asymmetric. Applying the active-regime feature to flat pairs had zero effect on both datasets. Applying the flat-regime feature to active pairs produced partial recovery: 28% on MIMIC and 54% on PadChest. This asymmetry, together with the mediation results below, is consistent with Feature 12139 being downstream of Feature 3818.

A.4.4 Mediation: 3818 → 12139

If Feature 3818 is upstream and Feature 12139 downstream, then patching 3818 at layer 17 should cause 12139 to move at layer 29. We tested this by inserting a patching hook at layer 17 (restoring the original-prompt value of 3818) and reading the resulting change in 12139 at layer 29. On 7 discovery-set MIMIC pairs and 30 held-out PadChest pairs, the direction coherence was perfect: all 37/37 pairs showed anti-correlated movement (3818↑ ⇒ 12139↓; 3818↓ ⇒ 12139↑). All pairs exceeded a minimal effect threshold ($|\Delta| > 5$). The mean downstream shift was -82 on MIMIC (-0.6%) and -350 on PadChest (-6.2%), with zero disruptions among non-flip controls (0/17 MIMIC, 0/30 PadChest; Table A.6).

Table A.5: Pre-specified held-out validation of frozen causal hypotheses on verified flips. Hypotheses frozen on 12 MIMIC discovery pairs; held-out validation sampled from the remainder of the 436-flip MIMIC screen and the 2,833-flip PadChest screen. Restore rate = fraction of flipped pairs whose original answer is recovered by single-feature patching. Signed recovery = mean fraction of the margin shift recovered (positive = toward original answer). Controls report disruption rate (fraction of stable pairs whose answer changed).

			MIMIC			**PadChest**		
Regime	**Arm**	N	**Restore**	**Rec.**	p	**Restore**	**Rec.**	p
3818_ active N/A L17/#3818 (upstream formality gate)								
	Hypothesis	100	24.0%	+0.139	$<10^{-15}$	17.0%	+0.121	$<10^{-17}$
	Weak control	100	1.0%	+0.001	0.42	0.0%	−0.003	0.96
	Non-flip disruption	50	0.0%	N/A	N/A	0.0%	N/A	N/A
	Cross-regime†	50	0.0%	0.000	1.00	0.0%	0.000	1.00
3818_ flat N/A L29/#12139 (downstream decision gate)								
	Hypothesis	100	58.0%	+0.317	$<10^{-17}$	27.0%	+0.333	$<10^{-17}$
	Weak control	100	0.0%	+0.030	$<10^{-15}$*	0.0%	+0.035	$<10^{-11}$*
	Non-flip disruption	50	0.0%	N/A	N/A	0.0%	N/A	N/A
	Cross-regime†	50	28.0%	+0.354	$<10^{-9}$	54.0%	+0.311	$<10^{-9}$

†Cross-regime: L17/#3818 applied to flat-regime flips (active rows); L29/#12139 applied to active-regime flips (flat rows). The 28%/54% cross-regime restore and the zero restore for 3818→flat are consistent with #12139 being downstream of #3818 (§A.4.4).

*Weak control shows statistically significant but negligibly small signed recovery (∼3%) with 0% restore rate, confirming specificity.

A.4.5 Sufficiency

As a complementary test, we injected the paraphrase-direction delta of L29/#12139 into the original prompt on 100 held-out flat-regime pairs. This induced the exact paraphrase answer in 15% of cases (95% CI [8.6, 23.5]). Every induced flip matched the paraphrase answer. The gap between this 15% sufficiency and the 58% restoration rate confirms that Feature 12139 is the strongest recoverable single feature at the decision stage, but not the sole contributor to paraphrase-induced answer changes.

A.4.6 Cross-Dataset Composition Analysis

The PadChest restore rates (17% active, 27% flat) are lower than MIMIC's (24%, 58%). This gap could reflect a mechanistic difference (the features matter less on PadChest) or a compositional difference (PadChest flips have different properties that make restoration harder or easier).

We fitted pair-level logistic regressions predicting restoration from dataset, flip direction, paraphrase phenomenon, margin gap, and feature delta (Table A.7). For the cross-regime arm, the raw dataset effect ($\beta = +1.10$, $p = 0.009$) attenuated by 68% and became non-significant ($\beta = -0.35$, $p = 0.59$) after conditioning on composition. Flip direction was the strongest predictor ($\beta = +2.54$, $p = 0.005$): 12139 is substantially more effective on no→yes flips, and PadChest active flips are 90% no→yes versus MIMIC's 50/50. For the flat-hypothesis arm, attenuation was 45% with the dataset effect approaching non-significance ($p = 0.056$); the strongest predictor was feature delta magnitude.

We interpret these results as showing that the 3818→12139 causal chain operates identically on both datasets, and that cross-dataset differences in effect size are primarily compositional.

A.4.7 Summary

These experiments support a two-regime mechanistic account (Figure A.1). When Feature 3818 is active, paraphrase flips arise from an upstream wording-sensitive gate at layer 17 that propagates to a downstream decision feature at layer 29. When Feature 3818 is flat, the proximal cause is Feature 12139 itself. The 3818→12139 pathway replicates across MIMIC and PadChest with 100% direc-

Table A.6: Mediation test: patching Feature 3818 at layer 17 and measuring the resulting change in Feature 12139 at layer 29. Direction coherence = fraction of pairs where 3818↑ causes 12139↓ (or vice versa). Specificity = non-flip pairs disrupted by the upstream intervention.

	MIMIC	**PadChest**
Mediation pairs tested	7	30
Direction coherent	7/7 (100%)	30/30 (100%)
All moved > threshold	7/7	30/30
Mean Δ12139	−82.3	−349.9
Mean Δ12139 (%)	−0.6%	−6.2%
Non-flip disruptions	0/17	0/30
Active-flip disruptions	0/7	1/30

Table A.7: Logistic regression predicting Feature 12139 restoration success. Adding composition variables (direction, phenomenon, margin gap, delta) attenuates the dataset coefficient, showing that the MIMIC/PadChest gap is primarily compositional. Cross-regime arm: 50 MIMIC + 50 PadChest active-regime pairs patched with L29/#12139. Flat arm: 100 MIMIC + 100 PadChest flat-regime pairs.

Arm	**Model**	β_{dataset}	p	AIC	Atten.
Cross-regime	Dataset only	+1.10	0.009	132.3	N/A
	+ direction	+0.63	0.210	131.4	43%
	+ direction, phenomenon, margin gap, delta	−0.35	0.594	114.8	68%
Flat	Dataset only	−1.32	<0.001	256.7	N/A
	+ direction	−1.27	<0.001	255.6	4%
	+ direction, phenomenon, margin gap, delta	−0.73	0.056	243.1	45%

Attenuation = $1 - |\beta_{\text{full}}|/|\beta_{\text{baseline}}|$. In the cross-regime arm, the strongest individual predictor is flip direction ($\beta = +2.54$, $p = 0.005$): Feature 12139 is more effective on no→yes flips, and PadChest active flips are 90% no→yes.

tion coherence. Cross-dataset differences in restoration efficacy are primarily compositional rather than mechanistic, driven by the distribution of flip directions and paraphrase phenomena in each dataset.

One qualification bounds this account. The canonical screen that supplies these flips is not purely paraphrastic: of its 436 verified flips, 171 (39%) are negation-pattern pairs, many of which are operator changes ("Is there X?" versus "Can X be ruled out?") whose opposite answer is correct rather than a failure. The 24% and 58% restoration rates and the 37/37 mediation result therefore conflate genuine same-polarity paraphrase sensitivity with correct operator handling and downstream answer selection, and Feature 12139's perfect direction coherence is expected of any feature that tracks the yes/no decision. The operator-preserving re-analysis of Section 5.3.1 isolates the same-polarity subset but is small; a full operator-preserving screen with matched non-flip controls is the outstanding step before the two-stage account can be presented as a clean paraphrase circuit rather than operator handling plus answer selection.

Feature 12139 is best characterised as a *downstream decision-calibration feature whose observed causal efficacy depends on the direction and phenomenon composition of the evaluated population.* It is not a dataset-specific artifact, a generic answer-bias feature, or a sufficient single-feature explanation; it is the strongest individually recoverable feature at the decision stage.

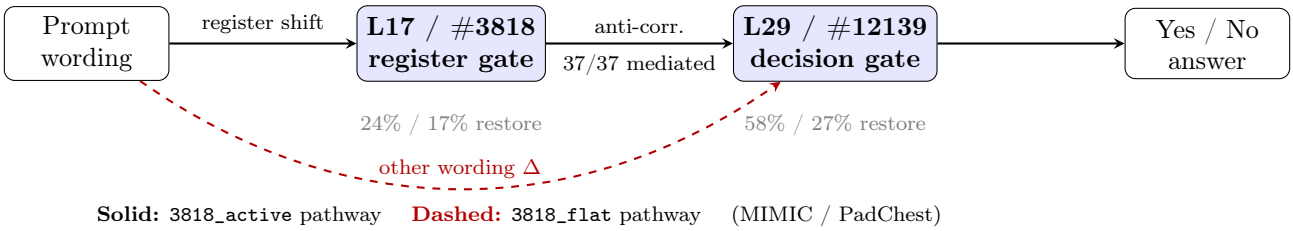


Figure A.1: Two-pathway causal model for paraphrase sensitivity. **Active pathway (solid):** register-framing changes shift Feature 3818 at layer 17, which propagates anti-correlatedly to Feature 12139 at layer 29 (24% / 17% restore on MIMIC / PadChest; 37/37 direction-coherent in mediation test). **Flat pathway (dashed):** wording changes bypass 3818 and directly perturb Feature 12139 (58% / 27% restore). Both pathways converge on the same late decision gate. Cross-dataset differences are explained by composition (direction, phenomenon mix), not mechanism (§A.4.6).

A.5 Candidate Two-Stage Account: Discussion

A.5.1 Two-Stage Mechanisms vs. Single-Feature Explanations

The canonical validation substantially revises the mechanistic picture from Section 5.3. The earlier FlipBank analysis identified Feature 3818 as a case study covering ∼25% of flips. The canonical pipeline, run at larger scale with frozen hypotheses, supports a *candidate* two-stage mechanistic account, an operator/register feature plus a downstream answer-selection feature, which remains exploratory pending same-polarity matched controls:

1. Feature 3818 at layer 17 acts as an upstream register gate, explaining flips in the `3818_active` regime (24% restore on MIMIC, 17% on PadChest).
2. Feature 12139 at layer 29 acts as a downstream decision-calibration gate, explaining flips in both the `3818_flat` regime (58% / 27% restore) and, partially, in the active regime (28% / 54% cross-regime restore).
3. The two features form a causal chain with 100% direction coherence across 37 mediation pairs spanning both datasets.

This is a stronger result than the original case study in three ways. First, the effect sizes are larger: 58% restoration for a single feature is substantial, compared to the earlier 28% exemplar-level recovery. Second, the validation is pre-specified: hypotheses were frozen on 12 discovery pairs and tested on hundreds of held-out pairs (a held-out design, though not publicly preregistered). Third, the mediation test provides causal evidence for a directed pathway, not just a correlation between feature activations and flips.

The remaining gap is informative: 12139 restores 58% of flat flips, not 100%. It indicates that paraphrase sensitivity at the decision stage involves multiple features in superposition [112], with 12139 being the largest individual contributor. The 15% sufficiency rate confirms this: Feature 12139 is necessary for many flips but sufficient for only a minority.

A.5.2 Why Cross-Dataset Differences Are Compositional

The composition analysis resolves what initially appeared to be a dataset-specific anomaly. Feature 12139 restores 58% of flat MIMIC flips but only 27% of flat PadChest flips; it restores 28% of active MIMIC flips cross-regime but 54% of active PadChest flips. These differences are not mechanistic. Pair-level logistic regressions show that conditioning on flip direction, phenomenon, margin gap, and delta magnitude attenuates the dataset coefficient by 68% (cross-regime) and 45% (flat).

The two main composition differences are:

- **Direction:** PadChest active-regime flips are 90% no→yes, while MIMIC is 50/50. Feature 12139 is roughly twice as effective on no→yes flips.

- **Phenomenon mix:** MIMIC has more `specificity_modulation` pairs, where 12139 has near-zero efficacy; PadChest is more heavily weighted toward `negation_pattern`, where 12139 is stronger.

This has a practical implication for deployment: the effectiveness of mechanistic interventions based on single features will depend on the clinical population's question distribution. A deployment-time audit should characterise the question mix before estimating expected intervention benefit.

A.5.3 Implications for LoRA Fine-Tuning

The candidate two-stage account provides retrospective insight into why the Targeted LoRA (layers 15 to 19) works. The LoRA's intervention range brackets the upstream gate (layer 17) and sits well upstream of the downstream gate (layer 29), roughly ten layers short of it. Cross-model feature validation shows that the LoRA suppresses Feature 3818 on in-distribution data (MIMIC) but not out-of-distribution (PadChest), consistent with the LoRA learning to modulate the upstream pathway rather than the downstream one. The layer ablation finding that early layers (0 to 10) outperform middle layers (15 to 19) may reflect the fact that the early intervention acts before *either* feature forms, avoiding the need to separately address both stages.

A.5.4 Limitations of the Mechanistic Account

Several caveats apply. First, the two-stage model covers the `3818_active` and `3818_flat` regimes but not the `ambiguous` regime, where both features show intermediate activity; the ambiguous regime holds 36% of MIMIC and 49% of PadChest screened pairs. The flip rate within the ambiguous regime is low ($\leq$2% of ambiguous pairs flip), so it contributes only a small minority of verified flips and the practical coverage is high, but the mechanistic coverage is incomplete. Second, the mediation test measures indirect effect through a specific pathway and does not rule out parallel pathways. Third, the PadChest mediation used 30 randomly sampled active pairs; a larger sample might reveal edge cases where direction coherence fails. Fourth, Feature 12139's sufficiency rate (15%) indicates that the decision stage is multi-feature, and the two-stage model should be understood as identifying the *dominant* features, not the complete circuit. Finally, all experiments use base MedGemma-4B-IT; the mechanism may differ in other model families or after fine-tuning.

A.6 Probe Training Data: Composition, Balancing, and Phrasing Bank

Section 5.5 describes the probe's training set in one sentence; this section gives the full distribution, because the chapter's causal claim rests on exactly what that distribution contains and what it removes. The training set is not a standalone file but the train split of a single index of 107,328 binary presence questions built once (sampling seed fixed at zero) and shared by every run in this chapter, so the regime comparisons of Table 5.4 differ only in the phrasings each run was allowed to see, never in the underlying questions.

A.6.1 Sources and Split Sizes

The index draws on two sources for training and holds two further hospitals out entirely. NIH ChestX-ray14 [116] contributes 83,328 training questions (96.6% of the training set) over 55,037 unique training images, and the PadChest-GR subset of PadChest [101] contributes 2,960 (3.4%) over 2,106 images, for 86,288 training questions over 57,143 unique images. The imbalance between sources is a property of the source material, not a design choice: the NIH release is an order of magnitude larger than the usable PadChest-GR pool (112,120 usable NIH images against 4,555 usable PadChest-GR images after label filtering), and the per-finding sampling rule described below caps each source at the same number of questions, so the smaller source simply exhausts its eligible images first. MIMIC-CXR [117] and VinDr-CXR [102] contribute nothing to training and appear only as out-of-distribution evaluations.

Splits are assigned at the image level by a deterministic hash of the image identifier, with 85% of each source's images assigned to training, so no radiograph appears in two splits and every aggregate is exactly reproducible. Table A.8 gives the four resulting splits. Two properties are worth stating plainly. First, there is no in-distribution test split: the in-distribution evaluation reported throughout this chapter is the validation split, which is also what early stopping monitored, so the honest generalization evidence is the pair of held-out hospitals rather than the validation number. Second, the training and evaluation questions are the same binary form throughout, *is there {finding}?*, with the finding drawn from the fourteen labels of Table A.9; the phrasing bank of Section A.6.4 varies only how that one question is worded.

Table A.8: The probe's question index. Every split is answer-balanced per finding, so each cell contains exactly as many yes as no answers; unique images are fewer than questions because one radiograph can carry several findings. Splits are image-level, so no radiograph appears in two splits. There is no in-distribution test split; the validation split is the in-distribution evaluation used throughout this chapter.

Split	Questions	Yes	No	Unique images
Train (NIH ChestX-ray14 + PadChest-GR)	86,288	43,144	43,144	57,143
Validation (held-out images, same sources)	15,112	7,556	7,556	10,045
MIMIC-CXR (held-out hospital)	450	225	225	355
VinDr-CXR (held-out hospital)	5,478	2,739	2,739	3,641
Total	107,328	53,664	53,664	71,184

A.6.2 Per-Finding Composition and Natural Prevalence

Table A.9 gives the training set by finding, together with the natural prevalence of each finding in the raw source data before any sampling. The fourteen findings are the presence labels shared by the two sources: atelectasis, cardiomegaly, consolidation, edema, emphysema, fibrosis, hernia, infiltrates, mass, nodule, pleural effusion, pleural thickening, pneumonia, and pneumothorax. Natural prevalence spans two orders of magnitude, from 17.7% for infiltrates down to 0.2% for hernia in the NIH data, and PadChest-GR contributes nothing at all for edema, emphysema, fibrosis, and pneumonia, whose positive pools were empty or fell below the minimum cell size. A question set drawn in proportion to the source data would therefore be dominated by negative answers and by a handful of common findings, and a model could score well on it by learning the base rates.

The construction removes exactly that confound. Within each split, source, and finding, the builder keeps an equal number of positive and negative answers, at most the smaller of the two pools. The cap is 5,000 questions per answer, per finding, per source before it is apportioned by split size. Its training shares are 4,252 per answer for NIH and 4,262 per answer for PadChest-GR, so a capped NIH finding contributes $2 \times 4{,}252 = 8{,}504$ training questions. The PadChest-GR cap does not bind, since its largest per-finding total is 876. Cells with fewer than four available questions are dropped, which is why PadChest-GR contributes zero questions for four findings. Negatives are drawn from "No Finding" radiographs and radiographs carrying other findings, so a no answer does not mean a clean scan. For rare NIH findings the positive pool binds and the count falls, to 400 for hernia at the extreme. The result is a training set that is exactly 50% yes for every finding in every split and source, but whose per-finding sizes still reflect, in compressed form, how common each finding is.

Table A.9: Training-set composition by finding, with the natural prevalence of each finding in the raw source data before sampling. Training questions are exactly answer-balanced per finding, per source, and per split, so the yes and no counts of each cell are equal; the combined column is their sum. Natural prevalence is the positive fraction over all usable images of the source (112,120 for NIH ChestX-ray14, 4,555 for PadChest-GR). A zero marks a finding that PadChest-GR does not contribute, while *n/a* marks unavailable prevalence after filtering.

	Training questions			Natural prevalence (%)	
Finding	NIH	PadChest-GR	Combined	NIH	PadChest-GR
Atelectasis	8,504	436	8,940	10.31	5.69
Cardiomegaly	4,672	876	5,548	2.48	10.93
Consolidation	7,996	76	8,072	4.16	0.90
Edema	3,900	0	3,900	2.05	*n/a*
Emphysema	4,288	0	4,288	2.24	0.07
Fibrosis	2,870	0	2,870	1.50	*n/a*
Hernia	400	160	560	0.20	1.89
Infiltrates	8,504	258	8,762	17.74	3.34
Mass	8,504	66	8,570	5.16	0.90
Nodule	8,504	346	8,850	5.65	4.35
Pleural effusion	8,504	346	8,850	11.88	4.43
Pleural thickening	5,724	380	6,104	3.02	5.12
Pneumonia	2,454	0	2,454	1.28	*n/a*
Pneumothorax	8,504	16	8,520	4.73	0.24
Total	83,328	2,960	86,288		

The evaluation splits preserve the same per-finding answer balance, but their finding mix differs because each hospital contributes only the findings retained by the construction rule. Table A.10 gives every evaluation cell. MIMIC-CXR retains eight findings: atelectasis, cardiomegaly, consolidation, edema, emphysema, pleural effusion, pneumonia, and pneumothorax. Atelectasis, cardiomegaly, edema, and pleural effusion each contain at least 20 questions and appear in the saved per-finding performance report; the other four fall below that report's threshold but remain in the 450-question evaluation total. VinDr-CXR contains 254 pleural-thickening questions, split into 127 yes and 127 no. That cell is the denominator behind the AUROC of 0.332 reported in Table 5.3.

Table A.10: Per-finding answer counts for the three probe evaluation splits. Each pair of columns reports yes and no questions separately. A zero means that the source contributes no retained cell for that finding. Every row is balanced within each split, and the column totals reproduce the reported split sizes.

	Validation		MIMIC-CXR		VinDr-CXR	
Finding	Yes	No	Yes	No	Yes	No
Atelectasis	788	788	87	87	6	6
Cardiomegaly	500	500	40	40	1,205	1,205
Consolidation	669	669	7	7	13	13
Edema	353	353	18	18	0	0
Emphysema	372	372	6	6	0	0
Fibrosis	251	251	0	0	603	603
Hernia	33	33	0	0	0	0
Infiltrates	770	770	0	0	84	84
Mass	755	755	0	0	0	0
Nodule	772	772	0	0	55	55
Pleural effusion	776	776	52	52	572	572
Pleural thickening	566	566	0	0	127	127
Pneumonia	204	204	9	9	0	0
Pneumothorax	747	747	6	6	74	74
Total	7,556	7,556	225	225	2,739	2,739

A.6.3 Answer Balancing and Its Effect

Balancing is the design decision the rest of the chapter silently depends on, so we state its effect explicitly. Because every split is exactly 50% yes per finding, a question's wording predicts its answer at exactly chance: a text-only model scores 50.0%, and the only way to reduce the training loss is to read the radiograph. This is what turns a paraphrase flip into attributable evidence. If prevalence were left in place, a flip could be produced either by sensitivity to wording or by a shift in the answer prior the wording happens to select, and the two could not be separated from the flip rate alone; with the prior removed by construction, the flip rate differences of Table 5.4 admit only the phrasing explanation. The same property is what certifies image use in Table 5.3: the 24.8-point gap between the text-only floor and the with-image accuracy is entirely visual information, and the text-only floor holds in every regime (49.4% to 50.7% across all seeds of all three phrasing regimes), so no regime can exploit a prior over answers.

The cost is symmetric and is worth stating alongside the benefit. Removing the answer prior makes the probe unrepresentative of deployment, where prevalence is skewed toward the values of Table A.9 and a model can be right for the wrong reason; Chapter 6 measures that regime on the deployed model rather than here. Balancing also fixes the per-finding decision boundary of the binary readout at the center of the margin distribution, which is the property the threshold sweep of Section 5.5 exploits when it shifts the boundary across ± 2 standard deviations to check that the regime ordering is not an artifact of threshold placement. We note one further implementation detail for completeness: the training pipeline re-balances after expanding each question to its paraphrase cluster, but because the index is already balanced per finding that second pass never changes the composition, so the balance the regimes see is exactly the balance of Table A.8.

A.6.4 The Phrasing Bank and Its Register Structure

Every training question carries a bank of 48 meaning-preserving paraphrases, twelve templates in each of four linguistic phenomena, plus four answer-inverting negation variants that are stored separately and never enter a paraphrase cluster, since a negation is an operator change and not a rewording. The four phenomena are the register axes along which clinical presence questions actually vary, and the adversarial regime of Section 5.5 is defined over them: *lexical substitution*, which swaps content words while keeping the structure (*is there evidence of {f}?*, *are there signs of {f}?*); *syntactic restructuring*,

which keeps the words and changes the grammar (*can {f} be identified?*, *is it correct that {f} is present?*); *scope quantification*, which widens or narrows the quantifier (*is there any {f}?*, *is even mild {f} present?*); and *specificity modulation*, which moves the clinical register up and down (*is {f} visible on this chest radiograph?*, *anything like {f} on here?*). The bank replaces an earlier eight-template design whose vocabulary had collapsed to 22 content words, on which a low flip rate would have reported template memorization rather than robustness.

The bank also carries the split that separates invariance from recognition. The first six templates of each phenomenon are eligible for training and the last six are withheld, so the unseen-phrasing evaluation trains on 24 phrasings and scores only on the 24 the model never saw, with six templates of each phenomenon on either side and therefore no phenomenon confounded with the split. The augmented regime trains on the original question plus all 48 paraphrases, expanding its 86,288 training questions to 4,228,112 examples; canonical trains on the original phrasing alone.

A.6.5 The Adversarial Register and Answer Coupling

The adversarial regime is the one place where the phrasing distribution is made to carry answer information, and its construction is simple enough to state exactly. For a yes answer, 90% of training paraphrases are drawn from lexical substitution and specificity modulation; for a no answer, 90% from scope quantification and syntactic restructuring; the remaining 10% are drawn from the other two phenomena, so the correlation is strong but not deterministic and every phenomenon still appears with every answer. Nothing else changes: the questions, the images, the balance, and the evaluation are those of the other regimes. The coupling is arbitrary in the sense that no linguistic property makes lexical substitution inherently affirmative; it is precisely the kind of accidental regularity a corpus can carry when, for example, one report template is used predominantly for confirmed findings and another for rule-outs.

Its measured effect is what identifies register as a distinct harm. The flip rate rises to 88.4% against 67.1% for canonical ($p = 3.1 \times 10^{-4}$, Cliff's $\delta = 0.97$), accuracy falls to 58.1%, and the grounding gap, the accuracy lost when the image is removed, collapses in the same order as the coupling strengthens: 0.252 for augmented, 0.164 for canonical, 0.081 for adversarial. A model trained on answer-correlated registers therefore both flips more and reads the image less, which is the probe-scale version of the consistent-but-blind failure this thesis warns against, produced here by a data distribution rather than observed after the fact.

A.6.6 Earlier Probe Dataset

The earlier 1,841-question probe used a different source pair: 980 MIMIC-CXR presence questions and 861 PadChest questions over 1,775 radiographs. It retained the source answer distribution rather than imposing per-finding balance, and the resulting finding-name shortcut produced 68.8% text-only accuracy. This is the dataset behind Figure 5.12, the sparse-autoencoder cosine of 0.74, the point-biserial correlation of 0.42, the Jacobian-lens analysis, and the register disagreement results. The regime runs for that probe record scalar aggregates per seed rather than per-question predictions, so a per-register flip breakdown is not recoverable from the stored results alone. Its saved disagreement analysis shows the adversarial coupling highest for scope quantification at 0.171 and negation pattern at 0.131, and lowest for the original phrasing at 0.038, against 0.012 to 0.062 in the augmented regime. We treat this pattern as supporting evidence because the earlier probe is smaller, retains a text-only shortcut, and reports cluster disagreement rather than the flip rate used for the 86,288-question probe.

A.7 Uncertainty Quantification Details

A.7.1 MC Dropout Configuration

MC Dropout uses $K = 10$ stochastic forward passes with dropout probability $p = 0.05$ applied to the LoRA adapters (matching the adapter training dropout). Predictive entropy is computed from the averaged predictive distribution:

$$\bar{p}(y|x) = \frac{1}{K} \sum_{k=1}^{K} p_k(y|x) \tag{A.1}$$

$$H[\bar{p}] = -\sum_{c \in \{Y,N\}} \bar{p}(c) \log \bar{p}(c) \tag{A.2}$$

The choice of $K = 10$ balances computational cost with uncertainty estimate quality. At $K = 10$, the coefficient of variation of the entropy estimate is approximately 0.05, meaning the estimate is stable across different random dropout masks.

A.7.2 Temperature Scaling

Temperature scaling learns a single scalar $T > 0$ that divides the logits before softmax:

$$p_T(y|x) = \text{softmax}(z/T) \tag{A.3}$$

The temperature is fit on a 15% calibration holdout by minimizing the negative log-likelihood. For the models evaluated:

Table A.11: In-domain learned temperature parameters for each model-dataset combination (each temperature is fit and evaluated on the same distribution; the cross-domain transfer setting is analyzed separately in Section A.11).

Model	**MIMIC T**	**PadChest T**
Base	1.42	2.18
Targeted LoRA	1.15	1.38
Full LoRA	1.89	3.41
LLaVA-Rad Base	0.98	1.12

$T > 1$ indicates the model is overconfident (logits too extreme); $T < 1$ indicates underconfidence. Base MedGemma-4B and Full LoRA are substantially overconfident, especially on PadChest ($T > 2$). LLaVA-Rad is approximately calibrated ($T \approx 1$), consistent with its text-driven behavior producing well-calibrated text priors. Targeted LoRA is moderately overconfident, with lower T values reflecting better baseline calibration.

A.7.3 Corruption Types for Distribution Shift

Five corruption types (brightness, contrast, Gaussian blur, Gaussian noise, and JPEG compression) at three severity levels (1, 3, 5) were used to test calibration under distribution shift. The exact parameter for each type and severity is specified in Table A.15, and the application formulas are given in Section A.11.

These corruptions simulate realistic image degradation that medical VLMs may encounter in deployment: poor lighting conditions, scanner calibration issues, image compression artifacts in PACS systems, and electronic noise in low-dose acquisitions.

A.7.4 AUGRC Computation

The Area Under the Generalized Risk-Coverage curve (AUGRC) [120] is computed as follows. For a set of N predictions with associated uncertainty scores $u_1, \ldots, u_N$ and correctness indicators $c_1, \ldots, c_N$:

1. Sort predictions by uncertainty (ascending): $u_{\pi(1)} \leq u_{\pi(2)} \leq \cdots \leq u_{\pi(N)}$.
2. For each coverage level k/N, compute the generalized risk, the joint probability of accepting and erring: $GR(k) = \frac{1}{N} \sum_{i=1}^{k} (1 - c_{\pi(i)})$. This equals the coverage k/N times the selective risk $\frac{1}{k} \sum_{i=1}^{k} (1 - c_{\pi(i)})$.
3. AUGRC $= \frac{1}{N} \sum_{k=1}^{N} GR(k) \approx \int_0^1 GR(c)\, dc$.

AUGRC is bounded between 0 (perfect selective prediction) and 0.5, with random ordering yielding an AUGRC of approximately half the error rate (since $GR(c) \approx c \cdot \text{err}$ under random ordering, so $\int_0^1 c\,\text{err}\, dc = \text{err}/2$). Lower values indicate that the uncertainty scores better discriminate between correct and incorrect predictions. Unlike AUROC, AUGRC is a proper scoring rule for selective prediction because it accounts for the ordering of predictions within each coverage level.

A.8 Replication Study Details

A.8.1 Seed-Level Results

Table A.12 reports the full results of the replication study across 5 random seeds and 2 training set sizes for the Targeted LoRA configuration.

Table A.12: Full replication results. MIMIC flip rate (FR), accuracy (Acc), and margin difference (MD) across 5 seeds and 2 training set sizes. Superseded question-disjoint run from the original pipeline, retained for seed-variance provenance; the patient-disjoint headline result is 8.5% → 3.5% (Section 5.7).

	$n = 500$			$n = 2000$		
Seed	FR	Acc	MD	FR	Acc	MD
42	4.4%	82.3%	0.33	4.2%	83.1%	0.30
123	4.8%	81.7%	0.36	4.5%	82.8%	0.32
456	4.6%	82.0%	0.35	4.3%	83.0%	0.31
789	4.3%	82.5%	0.32	4.1%	83.2%	0.29
2024	4.9%	81.5%	0.37	4.4%	82.7%	0.33
Mean	4.6%	82.0%	0.35	4.3%	83.0%	0.31
Std	0.3%	0.4%	0.02	0.2%	0.2%	0.02

The low standard deviation across seeds (0.3% for flip rate at $n = 500$) indicates that the training procedure is stable and the results are not sensitive to random initialization. Training-set size matters up to a point and then plateaus. In the earlier rerun that was disjoint by question identity only (156 questions; superseded by the patient-disjoint result of Section 5.7), the held-out pairwise flip rate falls from 5.1% at 500 samples to 3.4% at 1,288 and to 3.1% ± 0.9 on that run's expanded pool of 4,853 paraphrase pairs (five seeds), with no accuracy cost. The last two points sit within the seed-to-seed confidence interval, so the in-distribution result is data-saturated by roughly one to two thousand samples. Out-of-distribution transfer benefits more from additional data (the PadChest mean flip rate nearly halves from $n = 500$ to $n = 2000$ in the original-pipeline sweep). The binary pool caps at 1,288 distinct questions drawn from 1,000 local images; larger multi-task adapters (≈12,000 pairs) are released separately for larger-scale claims.

A.8.2 PadChest Transfer by Seed

On PadChest, the replication results show more variance:

Table A.13: PadChest transfer results across seeds ($n = 500$ training set). Higher variance than MIMIC reflects the out-of-distribution challenge. Superseded question-disjoint (original-pipeline) run; the 8.5% mean here is unrelated to the 8.5% patient-disjoint MIMIC baseline of Section 5.7.

Seed	Flip Rate	Accuracy	Margin Diff
42	7.8%	69.4%	0.35
123	8.5%	67.2%	0.42
456	9.1%	66.8%	0.48
789	8.2%	68.6%	0.39
2024	8.8%	67.5%	0.44
Mean	8.5%	67.9%	0.42
Std	0.5%	1.1%	0.05

The increased variance on PadChest (0.5% vs. 0.3% flip rate standard deviation) reflects the challenge of out-of-distribution generalization. Different random seeds produce LoRA adapters that capture slightly different aspects of the consistency signal, and these differences are amplified when the model encounters PadChest's distribution shift. Seed 42's PadChest accuracy in this replication split (69.4%) and the seed 42 value reported in the deep-ensemble analysis (92.1%, Table A.14) are measured on different PadChest evaluation sets and are therefore not directly comparable.

A.9 Deep Ensemble Failure Analysis

The deep ensemble for uncertainty quantification was constructed from 5 independently trained Targeted LoRA checkpoints (seeds 42, 123, 456, 789, 2024). On MIMIC-CXR, the ensemble performs well: 85.7% accuracy. On PadChest the seeds diverge sharply: two retain strong accuracy while three degrade out of distribution.

Table A.14: Deep ensemble seed-level performance on PadChest. Seeds 42 and 123 remain functional; seeds 456, 789, and 2024 degrade out of distribution despite 87 to 89% training validation accuracy. "Yes" rate is the fraction of positive predictions against an 81.4% positive prevalence.

Seed	PadChest Acc	"Yes" Rate	Train Val Acc	Status
42	92.1%	75.1%	86.9%	Functional
123	78.7%	61.3%	86.9%	Functional
456	52.0%	34.1%	87.3%	Degraded
789	67.1%	49.0%	88.6%	Degraded
2024	68.5%	50.4%	88.9%	Degraded

The degraded seeds do not lose accuracy uniformly; they shift their answer prior away from the PadChest base rate. Against the 81.4% "yes" prevalence, seed 456 predicts "yes" only 34.1% of the time and seeds 789 and 2024 sit near 50%, so they miss many true-positive findings. This under-prediction, rather than the yes-collapse seen with pure consistency training ($\lambda = 0$), occurs at the standard $\lambda = 1.0$, suggesting that certain random initializations lead to training trajectories that overfit to the MIMIC-CXR training distribution and transfer poorly to PadChest.

The failure pattern has several characteristics:

- All three degraded seeds achieve normal training and validation accuracy (87 to 89%), providing no early warning of the out-of-distribution degradation.
- The degradation is specific to PadChest; all seeds perform normally on MIMIC-CXR.
- Seed 42 is the strongest PadChest member (92.1%), but there is no a priori way to identify it

without evaluating on PadChest.

- Probability-averaging over all 5 seeds recovers 72.6% aggregate accuracy but a poorly ranked confidence signal (AURC 0.130 versus 0.019 for single-pass entropy), so the ensemble is unsuitable for selective prediction even though its mean accuracy is acceptable.

We document this as a negative result with practical implications. Researchers attempting LoRA ensembles for medical VLMs should expect and test for out-of-distribution seed degradation. A simple mitigation is to validate each seed on an out-of-distribution holdout before including it in the ensemble, but this requires access to such data at training time.

A.10 Dataset Access and Ethics

A.10.1 Data Sources

- **NIH ChestX-ray14**: Obtained from the openly released NIH Clinical Center chest radiograph set. Its access terms differ from the three agreement-controlled sources below.
- **MIMIC-CXR**: Access requires completing the CITI Data or Specimens Only Research course and signing a data use agreement through PhysioNet. Images are used in compliance with the PhysioNet Credentialed Health Data License 1.5.0, which permits use for scientific research and does not permit redistribution of the data.
- **PadChest and PadChest-GR**: Obtained from the Medical Imaging Databank of the Valencia Region (BIMCV) under the PadChest Dataset Research Use Agreement, which requires registration, grants use for research purposes only, and prohibits reproducing or publishing any portion of the dataset without prior written permission from BIMCV. The images originate from Hospital Universitario de San Juan de Alicante, and the dataset was approved by that hospital's institutional review board. We use the images together with the structured labels and bounding box annotations; no patient-identifying information is accessed. The probe training set uses PadChest-GR.
- **VinDr-CXR**: Access requires the same CITI training and a signed data use agreement through PhysioNet. Images are used in compliance with the PhysioNet Credentialed Health Data License 1.5.0, on the same terms as MIMIC-CXR.

Each source is de-identified. NIH ChestX-ray14 is obtained from the open NIH Clinical Center release, while MIMIC-CXR, PadChest and PadChest-GR, and VinDr-CXR are obtained under their respective access agreements. The PSF-Med release (Section C.5) publishes questions, paraphrases, judge labels, and audit verdicts, but no images. Reproducing the image-dependent results requires obtaining NIH ChestX-ray14 from its open release and the remaining corpora directly from PhysioNet and BIMCV under their respective agreements.

A.10.2 Image Preprocessing

PadChest images are distributed as 16-bit grayscale DICOM-derived files. Before each model's standard image processor, we convert them to 8-bit RGB by windowing to the 0.5th to 99.5th intensity percentile of each image and linearly rescaling that window to $[0, 255]$, which preserves diagnostic contrast across the wide 16-bit dynamic range. MIMIC-CXR and VinDr-CXR images are 8-bit and are passed through unchanged. All flip-rate, accuracy, calibration, and grounding results reported in this dissertation use this preprocessing.

A.11 Extended Uncertainty Quantification Details

A.11.1 Corruption Parameterization

Table A.15 specifies the exact parameters used for each corruption type and severity level in the calibration under shift experiments (Chapter 7).

All corruptions are applied to the input image (8-bit RGB, values in $[0, 255]$) before the model's

Table A.15: Corruption parameters by type and severity level.

Corruption	Parameter	Sev. 1	Sev. 3	Sev. 5
Gaussian noise	σ (0 to 255)	10	35	70
Gaussian blur	σ_b (px)	1.0	3.0	6.0
Contrast	factor α	0.8	0.4	0.15
Brightness	δ (0 to 255)	+20	+60	+100
JPEG compression	Quality Q	80	40	10

standard preprocessing. Gaussian noise is additive $I' = I + \mathcal{N}(0, \sigma^2)$ with σ in pixel units. Contrast scales around the image mean: $I' = \alpha \cdot I + (1 - \alpha) \cdot \bar{I}$. Brightness is additive: $I' = I + \delta$. JPEG compression saves and reloads the image at quality Q. Values are clipped to $[0, 255]$ after corruption.

A.11.2 MC Dropout Configuration

The committed PadChest evaluation uses $K = 10$ stochastic passes and $p_{\text{drop}} = 0.05$ on the LoRA adapters. Under that configuration, mean mutual information is 0.0001 nats, error-detection AUROC is 0.856, and coverage at 5% risk is 88.4%. The corresponding single-pass softmax values are 0.862 AUROC and 88.5% coverage. A broader $K \times p_{\text{drop}}$ sensitivity script is included in the repository, but its output file is not committed, so no sensitivity-grid results are reported here.

A.11.3 Temperature Scaling Transfer Analysis

We test whether a temperature parameter learned on one distribution transfers to another. All figures in this subsection are for Targeted LoRA. The unscaled baselines used here, 15.0% expected calibration error on MIMIC-CXR and 6.1% on PadChest, are computed on this transfer protocol's own fitting and evaluation splits, which are not the splits behind the in-domain figures of Section 7.1 (10.9% and 11.1%); the two sets of numbers are therefore not directly comparable, for the same reason the temperatures differ from the in-domain fits discussed below. Four protocols are evaluated:

- **ID→ID** (MIMIC→MIMIC): The standard setting. $T^* = 0.98$, ECE improves from 15.0% to 12.1%.
- **ID→OOD** (MIMIC→PadChest): $T^* \approx 1.0$, ECE 6.3% (barely different from unscaled 6.1%). The temperature learned on MIMIC is uninformative for PadChest.
- **OOD→OOD** (PadChest→PadChest): $T^* = 1.02$, ECE 6.3%. Calibrating directly on PadChest does not help because the model's miscalibration is not a simple temperature issue.
- **OOD→ID** (PadChest→MIMIC): $T^* = 1.01$, ECE 14.8%. Slightly worse than ID→ID.

Across all four transfer protocols the re-fit temperature stays near 1.0 (0.98 to 1.02), so a temperature estimated under this transfer setup applies virtually no correction. This is a different fitting protocol from the per-model, per-dataset in-domain fits in Table A.11, where the temperatures are larger (up to 3.41 for Full LoRA on PadChest): those in-domain temperatures reduce in-domain calibration error but do not transfer across the MIMIC/PadChest shift, and re-fitting directly under shift collapses the scalar back toward 1.0 without closing the calibration gap. Either way, the residual calibration error is structural (wrong representations, not wrong confidence scaling) and cannot be fixed by post-hoc temperature adjustment. The implication for deployment is that temperature scaling should not be relied upon as a calibration strategy for medical VLMs under distribution shift.

A.11.4 Deep Ensemble Per-Member Diagnostics

Table A.16 reports per-member accuracy for the 5-seed deep ensemble on PadChest. Each member is a Targeted LoRA adapter trained with the same hyperparameters but a different random seed. Only seed 42 retains high accuracy (92.1%); the other four range from 52.0% to 78.7%. The combined ensemble predicts "yes" on only 54.7% of cases against an 81.4% yes prevalence, so it under-predicts the positive class; probability-averaging drags aggregate accuracy down to 72.6%, below seed 42 alone.

Table A.16: Per-member accuracy and positive-prediction rate of the 5-seed deep ensemble on PadChest (N = 861; yes prevalence 81.4%).

Metric	Seed 42	Seed 123	Seed 456	Seed 789	Seed 2024
Accuracy (%)	92.1	78.7	52.0	67.1	68.5
"Yes" rate (%)	75.1	61.3	34.1	49.0	50.4

Seeds 456, 789, and 2024 under-predict the positive class on PadChest (34 to 50% "yes" versus the 81.4% prevalence), and seed 123 is only mildly positive (61.3%); all four therefore miss many true-positive findings and lose accuracy. Only seed 42 tracks the prevalence (75.1% yes) and achieves high accuracy (92.1%). The mean inter-member disagreement is 20.1%, confirming that the seeds have learned qualitatively different solutions.

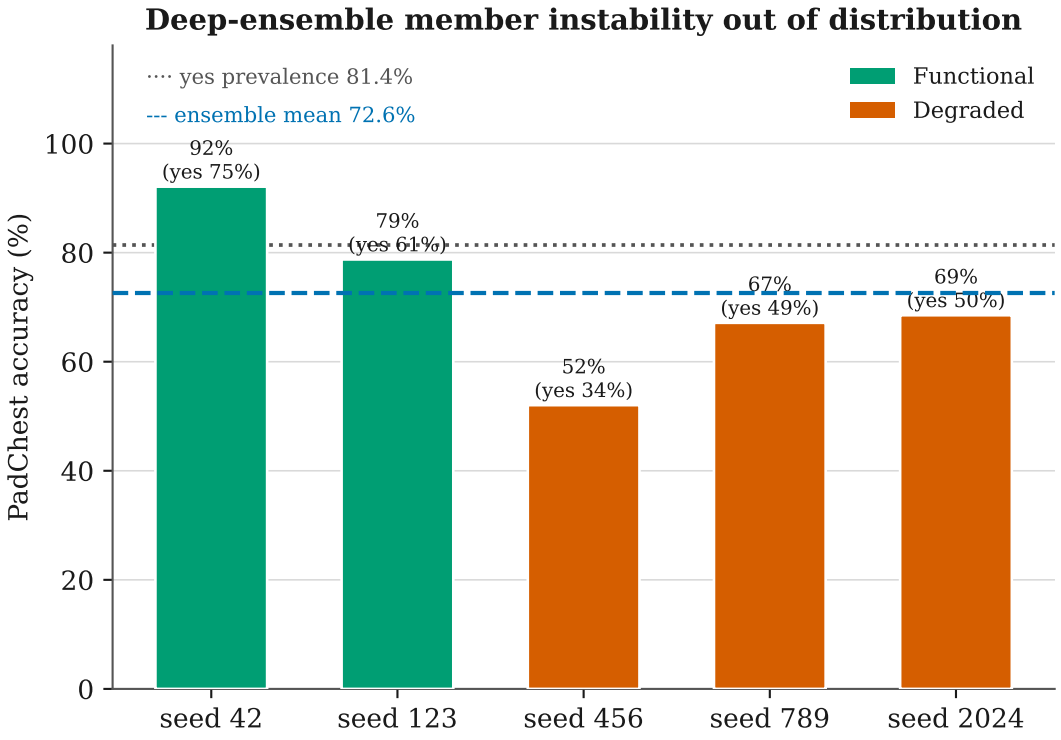


Figure A.2: Deep-ensemble member instability on PadChest. Two seeds remain functional (green); three degrade out of distribution (orange) by shifting their positive-prediction rate away from the 81.4% base rate. Probability averaging recovers a 72.6% mean but below the best single member (92.1%).

The ensemble's probability-averaging aggregation dilutes the one strong member: the four weaker members pull the consensus down to 72.6% accuracy, below seed 42's 92.1%. Alternative aggregation strategies give similar results (logit averaging 72.6%, majority vote 71.3%) because the fundamental problem is member quality, not aggregation method. The more consequential failure is not aggregate accuracy but ranking: the ensemble's risk-coverage collapses (AURC 0.130 versus 0.019 for single-pass entropy), so it cannot be used for selective prediction even though its mean accuracy looks acceptable.

This failure mode is specific to the narrow training distribution (500 MIMIC-CXR binary pairs). On MIMIC (in-distribution), all five seeds achieve 85 to 89% validation accuracy and the ensemble succeeds (85.7% accuracy, best AURC). The implication is that LoRA ensembles require diverse training distributions to provide OOD robustness.

A.11.5 Bootstrap Statistical Methodology

All confidence intervals in Chapter 6 are computed using $B = 2{,}000$ non-parametric bootstrap resamples with the following protocol:

1. For each resample $b \in \{1, \ldots, B\}$, draw N samples with replacement from the evaluation set.
2. Compute the metric of interest (AUROC, ECE, flip rate, etc.) on the resampled data.
3. Report the 95% CI as the 2.5th and 97.5th percentiles of the bootstrap distribution.

For paired method comparisons (e.g., softmax vs. MC Dropout AUROC), we use the paired bootstrap: both methods are evaluated on the same resampled data, and the CI is computed on the difference in metrics. This accounts for correlations between methods evaluated on the same predictions.

For entropy distribution comparisons (flipped vs. stable predictions), we use the Mann-Whitney U test, which is distribution-free and appropriate for the non-normal entropy distributions observed. All reported p-values for entropy comparisons use this test.

Multiple comparison correction uses Bonferroni adjustment across the 5 UQ methods. All reported p-values remain below 10^{-5} even after correction, so no results change significance.

Unit of analysis and sources of variability. The resampling unit in step 1 above is the individual observation, so every interval reported in this dissertation assumes independence at the finest level of the data. The data are in fact nested: a paraphrase pair sits inside a question, a question inside an image, and an image inside a patient. Which unit a statistic belongs to therefore has to be read off the statistic itself. A pairwise flip rate has the paraphrase pair as its denominator; a query-level flip rate, an accuracy, a text-only agreement, and a swap-sensitivity figure all have the question as their denominator; a region-of-interest coverage figure has the annotated box as its denominator. Above the image, the nesting is weak on the sets used here: the 637 region-of-interest boxes come from 579 images and 574 patients, and the 861-item PadChest flip bank comes from 795 images and 790 patients, so image-level and question-level clustering are the units that carry dependence and patient-level clustering adds almost nothing. Separately, a $\pm$ reported alongside a LoRA result in Chapter 5 is a standard deviation across training seeds and describes how much the training procedure varies from seed to seed; it is not an interval for the test cohort, and the two are never pooled into a single figure. Clustered intervals have not been recomputed for the numbers in this dissertation. The resampling helper written for that recomputation is `scripts/analysis/cluster_bootstrap.py`, which resamples whole clusters and reports the design effect against the independent bootstrap; no number in this dissertation is currently derived from it.

A.11.6 Conformal Prediction Under Distribution Shift

We implement split-conformal prediction with the Adaptive Prediction Sets (APS) nonconformity score [121]. The calibration set is 15% of PadChest (clean images); the test set is the remaining 85% under corruption.

At 90% target coverage on clean PadChest, Targeted LoRA achieves 93.7% empirical coverage with mean prediction set size 1.03 (out of 2 possible classes): slightly conservative. Under severity-5 corruptions, coverage settles at 89.9%, on target, with set size 1.05. Base MedGemma-4B erodes from 91.8% (clean) to 88.7% (severity 5), a modest 1.3-point under-coverage at this target. Full LoRA holds near-constant coverage (88.7% clean to 87.2% at severity 5) because image corruptions do not affect its text-driven predictions; its conformal validity is degenerate, holding trivially because the corruptions do not change the inputs the model actually uses.

At 95% target coverage, Targeted LoRA achieves 95.2% empirical coverage with mean set size 1.09, declining to 92.7% under severity-5 corruption. The distribution-free guarantee holds conditional on exchangeability between calibration and test data; corruption violates exchangeability, but empirical coverage stays above 90% for the image-attentive models. The coverage violation is clearest at tighter

targets: at an 80% target, Base falls to 71.0% under severity-5 corruption versus 75.6% for Targeted LoRA.

A.11.7 Computational Resources

All experiments were conducted on a shared computing cluster with NVIDIA A100 80GB GPUs, and every run was logged and tracked with Weights & Biases. Total GPU time across all experiments, including development, failed, and exploratory runs not reflected in the final saved artifacts:

- PSF-Med benchmark evaluation and equivalence audit (6 models, 3 datasets, roughly 93k paraphrase pairs; rubric-based judge audit of 122,778 pairs; PadChest v2 regeneration): ∼1,300 GPU-hours
- Uncertainty quantification pipeline (5 models, 2 datasets; softmax entropy, Monte Carlo dropout, temperature scaling, deep ensembles across 5 seeds, and text-only baselines): ∼950 GPU-hours
- Corruption robustness sweep (5 corruption types, 3 severities, across models, datasets, and uncertainty methods): ∼600 GPU-hours
- Safety evaluation and architecture-aware deployment audit (multiple model families; image-swap, text-only, and null-image controls): ∼525 GPU-hours
- Attention grounding and causal analyses (attention extraction, patch occlusion, region-of-interest ablation, and laterality across models): ∼525 GPU-hours
- LoRA training, layer and block sweeps, and the seed replication study: ∼450 GPU-hours
- Mechanistic interpretability (SAE screening and transfer, activation and residual-transplant patching, canonical validation, and lens fitting): ∼450 GPU-hours
- Total: ∼4,800 GPU-hours (approximately 200 GPU-days)

A.11.8 Reproducibility

All code, data splits, and pre-computed results are available in the project repository. Key reproducibility artifacts:

- PSF-Med benchmark data: HuggingFace dataset `saillab/psf-med`
- Evaluation scripts: `scripts/evaluation/evaluate_psf.py`
- LoRA training: `scripts/training/train_medgemma_lora_targeted.py`
- Safety pipeline: `scripts/miccai/run_all.sh`
- Environment specification: `pyproject.toml` (Python 3.12+)
- Experiment tracking: Weights & Biases (upload via `scripts/uai/upload_wandb.py`)
- Random seeds for all experiments are reported in the text

A.12 Deployment Gate Algorithm

Algorithm A.17 states the admission rule exactly as evaluated retrospectively in Chapter 7; every gate number in that chapter comes from this rule. Algorithm A.18 then presents the proposed online gate procedure, including all three gate modes; it is a design proposal for prospective deployment and has not been evaluated.

The algorithm processes predictions in order of increasing cost. Step 2 (entropy filter) requires no additional forward passes. Step 4 (consistency) requires K forward passes. Step 6 (text-only) requires 1 additional pass. Step 8 (swap) requires 1 additional pass but also requires access to a swapped image, which may not always be available.

In practice, we recommend $K = 3$ for deployment, giving a total of 5 to 6 forward passes per prediction at approximately 600 to 900ms on an A100 GPU. The entropy threshold τ_H can be set based on the desired operating point: higher thresholds admit more predictions at lower accuracy, lower thresholds refer more predictions at higher accuracy on admitted ones.

Table A.17: Evaluated retrospective admission rule (the rule behind every gate number in Chapter 7)

Algorithm: Retrospective admission rule (evaluated)

Input: Question q, Image I, Model M, Paraphrases p_1, p_2 (first two of K)
Output: ADMIT or REFER
1. $a_0 \leftarrow M(q, I)$; $a_1 \leftarrow M(p_1, I)$; $a_2 \leftarrow M(p_2, I)$
2. **if** $a_1 \neq a_0$ **or** $a_2 \neq a_0$ **then return** REFER [consistency]
3. $a_{\text{swap}} \leftarrow M(q, I_{\text{swap}})$, where I_{swap} is an arbitrary cohort image taken in rotation
4. **if** $a_{\text{swap}} \neq a_0$ **then return** ADMIT [random-image sensitivity]
5. **if** I has a radiologist box (PadChest only, 637 of 861 records) **and** the region-only and background-only crops give different answers **then return** ADMIT [ROI disjunction]
6. **return** REFER

Table A.18: Deployment Gate Pseudocode

Algorithm: Per-Prediction Deployment Gate

Input: Question q, Image I, Model M, Gate mode $\in \{\text{consistency}, \text{consistency+text}, \text{full}\}$, Number of paraphrases K, Entropy threshold τ_H
Output: Decision $\in \{\text{ADMIT}, \text{REFER}\}$, Answer a, Confidence c

1. Compute base prediction: $a_0, p_0, H_0 \leftarrow M(q, I)$
2. **Quick filter:** If $H_0 > \tau_H$, return (REFER, a_0, H_0)
3. Generate K paraphrases: $\{p_1, \ldots, p_K\} \leftarrow$ paraphrase generator(q)
4. **Consistency check:**
   For each k: compute $a_k \leftarrow M(p_k, I)$
   If $\exists\, k : a_k \neq a_0$, return (REFER, a_0, H_0)
5. If gate mode is "consistency": return (ADMIT, a_0, H_0)
6. **Text-only check:**
   Compute $a_{\text{text}} \leftarrow M(q, I_{\text{blank}})$
   If $a_{\text{text}} = a_0$, return (REFER, a_0, H_0) [text-reliant]
7. If gate mode is "consistency+text": return (ADMIT, a_0, H_0)
8. **Image swap check:**
   Compute $a_{\text{swap}} \leftarrow M(q, I_{\text{swap}})$
   If $a_{\text{swap}} = a_0$, return (REFER, a_0, H_0) [image-invariant]
9. return (ADMIT, a_0, H_0)

A.13 Cross-Modal Feature Analysis Details

The cross-modal analysis (extending the Feature 3818 investigation of Chapter 5) jointly extracts SAE features from the text pathway and visual attention metrics from the vision pathway to test whether Feature 3818's influence crosses modality boundaries.

A.13.1 Winsor-CAM Methodology

Standard Grad-CAM [105] computes importance weights as the global average of gradients with respect to feature maps. This approach is sensitive to outlier gradients, which are common in deep transformers. We use Winsor-CAM, a variant that applies Winsorization (clipping values at the 5th and 95th percentiles) to the gradient maps before computing importance weights. This produces smoother, more stable attention heatmaps.

For MedSigLIP (the vision encoder in MedGemma-4B), Winsor-CAM operates on 27 transformer layers, each producing a 64×64 grid of attention values. The per-layer importance score is computed as:

$$\alpha_l = \text{Winsorize}_{5,95}\left(\frac{1}{HW}\sum_{h,w}\frac{\partial y}{\partial A^l_{h,w}}\right) \tag{A.4}$$

where $A^l_{h,w}$ is the activation at spatial position (h, w) in layer l, and y is the target logit.

The final Winsor-CAM heatmap is a weighted combination across layers:

$$\text{CAM}(h, w) = \text{ReLU}\left(\sum_l \alpha_l \cdot A^l_{h,w}\right) \tag{A.5}$$

A.13.2 Joint SAE-Visual Extraction

For each prediction, we jointly extract:

1. **SAE features** at layers 9, 17, 22, and 29 of the language model (residual stream at the final token position). These layers sample early, middle, and late processing stages.
2. **Winsor-CAM metrics**: visual entropy ($H_{\text{vis}} = -\sum_{h,w} p_{h,w} \log p_{h,w}$ where p is the normalized heatmap), top-k coverage (fraction of attention in the top 5% of spatial locations), and cosine similarity between original and paraphrase heatmaps.

The joint extraction runs on GPU and takes approximately 2 seconds per prediction (1.5s for SAE feature extraction across 4 layers, 0.5s for Winsor-CAM computation across 27 vision layers).

A.13.3 Feature Ranking by Flip Association

For each SAE feature at each layer, we compute three association metrics:

- **Point-biserial correlation** r_{pb} between feature activation and flip outcome (binary). Negative correlation means the feature's activation is associated with non-flip (consistent) predictions.
- **Mean activation difference** between flip and non-flip groups, normalized by pooled standard deviation (Cohen's d).
- **Bootstrap stability**: the fraction of 500 bootstrap resamples in which the feature ranks in the top 20 by $|r_{pb}|$. Features with stability > 0.8 are considered stable.

Features are ranked by the product $|r_{pb}| \times$ stability, which balances effect size and stability. Feature 3818 ranks #1 on PadChest for both Base and Targeted LoRA configurations, and #2 on MIMIC for Base, confirming its cross-dataset relevance.

A.13.4 Causal Patching Protocol

The causal patching experiment tests whether restoring Feature 3818's activation from the original question can undo paraphrase-induced flips. The protocol is:

1. For each flip pair (q, p), extract the SAE feature vector $\mathbf{f}(q)$ from the original question and $\mathbf{f}(p)$ from the paraphrase at layer 17.
2. Compute the difference in Feature 3818: $\Delta f = f_{3818}(q) - f_{3818}(p)$.
3. Decode this difference through the SAE decoder to obtain a direction in residual stream space: $\Delta \mathbf{h} = \mathbf{W}_{\text{dec}}[\Delta f \cdot \mathbf{e}_{3818}]$.
4. Run the paraphrase forward pass, but at layer 17, add $\Delta \mathbf{h}$ to the residual stream before continuing.
5. Compare the patched output with the original output. If the patched paraphrase now matches the original answer, the flip is "recovered."

Control features are selected by matching Feature 3818's mean activation magnitude on the same dataset. For each control feature, the same patching protocol is applied. The recovery rate for controls provides a baseline against which Feature 3818's causal role can be assessed.

A.14 Per-Corruption AUGRC Values

Table A.19 reports AUGRC values for Targeted LoRA on PadChest across all corruption types and severity levels. These values are plotted as degradation curves in the main text.

Table A.19: AUGRC values for Targeted LoRA (PadChest) by corruption type and severity. Lower is better; clean baseline AUGRC = 0.014.

Corruption	**Sev. 1**	**Sev. 3**	**Sev. 5**
Gaussian noise	0.016	0.021	0.029
Gaussian blur	0.015	0.014	0.021
Contrast	0.016	0.026	0.040
Brightness	0.014	0.016	0.026
JPEG compression	0.014	0.013	0.019
Mean	0.015	0.018	0.027

Targeted LoRA AUGRC rises only modestly under corruption (0.014 clean $\rightarrow$ 0.027 at severity 5), staying an order of magnitude below the base and Full LoRA models. Its ranking of uncertain predictions is largely preserved even as the image degrades, so selective prediction remains effective under shift.

For comparison, Base MedGemma-4B AUGRC on PadChest rises steadily with corruption: 0.047 (clean) $\rightarrow$ 0.049 (severity 1) $\rightarrow$ 0.059 (severity 3) $\rightarrow$ 0.088 (severity 5). Full LoRA AUGRC stays high throughout, from 0.153 (clean) to 0.186 (severity 5), because image corruptions barely affect its text-driven predictions.

Appendix B

Extended Model Details and Hyperparameters

B.1 MedGemma-4B Architecture

MedGemma-4B-IT is built on the Gemma 3 architecture with the following key specifications:

- **Language backbone**: Gemma 3 with 34 transformer layers, hidden dimension 2,560, 8 query attention heads and 4 key-value heads (grouped query attention), head dimension 256
- **Vision encoder**: MedSigLIP with 27 vision transformer layers, patch size 14×14, producing 256 image tokens per input
- **Total parameters**: 4.0B (language) + 0.4B (vision) = 4.4B
- **Context length**: 128K-token input context (image tokens included); 8,192-token maximum output length
- **Training data**: Medical image-text pairs from curated clinical datasets (PubMed, radiology reports, pathology descriptions)
- **Instruction tuning**: The "IT" suffix indicates instruction tuning on conversational medical QA data

The vision-language integration uses a linear projection layer that maps MedSigLIP output embeddings to the Gemma 3 input embedding space. Image tokens are interleaved with text tokens in the input sequence, with special delimiter tokens marking the image region. The model processes image and text tokens jointly through all 34 transformer layers, with no separate vision-language fusion module.

B.2 LLaVA-Rad Architecture

LLaVA-Rad is built on the LLaVA framework with radiology-specific adaptations:

- **Language backbone**: Vicuna-7B (LLaMA-based, 32 transformer layers)
- **Vision encoder**: BiomedCLIP (PubMedBERT + ViT-B/16, trained on PMC-15M)
- **Total parameters**: ∼7B
- **Projection**: Two-layer MLP projecting vision features to language embedding space
- **Training**: Two-stage: (1) feature alignment on radiology image-text pairs, (2) instruction fine-tuning on radiology QA

The architectural differences from MedGemma are significant for interpreting the robustness-grounding trade-off. LLaVA-Rad's BiomedCLIP vision encoder was pretrained on biomedical image-text pairs (PMC-15M), while MedGemma's MedSigLIP was trained on a mixture of medical image-text pairs and natural-image data, so both encoders carry medical-domain pretraining and differ in curation and scale rather than in medical exposure alone. However, the diagnostic results in Chapter 4 show

that LLaVA-Rad achieves its consistency through text priors rather than visual grounding, suggesting that the vision encoder quality is less important than how the model integrates visual and textual information.

B.3 LoRA Hyperparameter Selection

The Targeted LoRA configuration was selected through a systematic hyperparameter search:

Table B.1: LoRA hyperparameter search space and selected values. The search covered rank, alpha, dropout, and learning rate. The selected configuration balances flip rate reduction against accuracy preservation.

Parameter	**Search Range**	**Selected**	**Rationale**
Rank (r)	{4, 8, 16, 32}	16	Best flip-accuracy trade-off
Alpha (α)	{16, 32, 64}	32	$\alpha/r = 2$ standard ratio
Dropout	{0.0, 0.05, 0.1}	0.05	Minimal regularization needed
Learning rate	{1e-4, 2e-4, 5e-4}	2e-4	Fastest convergence
Weight decay	{0.0, 0.01, 0.1}	0.01	Standard AdamW setting
Warmup steps	{0, 50, 100}	100	Smooth training start
Batch size	{4, 8, 16}	8	Memory-efficient on A100
Epochs	{1, 3, 5, 10}	3	Convergence by epoch 2

Rank 16 was chosen because rank 4 and rank 8 produced higher flip rates on the original-pipeline question-disjoint MIMIC validation split (6.1% and 5.2% vs. 4.4% at rank 16; this split is superseded for headline claims by the patient-disjoint evaluation of Section 5.7), while rank 32 did not improve over rank 16 (4.5%) despite doubling the parameter count. In LoRA, α scales the low-rank update by α/r [113]; we set $\alpha/r = 2$, a common default in parameter-efficient fine-tuning practice rather than a convention established by the original paper. Higher α values ($\alpha = 64$, $\alpha/r = 4$) produced minor accuracy degradation without further flip rate improvement.

B.4 Evaluation Metrics: Formal Definitions

For completeness, we provide formal definitions of all evaluation metrics used throughout the dissertation.

Question-level flip rate. For a model M, question q, and paraphrase set $P(q) = \{p_1, \ldots, p_K\}$:

$$\text{FlipRate}(M) = \frac{1}{|Q|} \sum_{q \in Q} \mathbf{1}\left[\exists\, k : M(q, I) \neq M(p_k, I)\right] \tag{B.1}$$

A question is counted as flipped if *any* paraphrase produces a different answer. This is a conservative measure: it counts a question as unstable even if only one of several paraphrases causes a flip.

Pair-level flip rate. The fraction of individual question-paraphrase pairs that disagree:

$$\text{PairFlipRate}(M) = \frac{1}{\sum_q |P(q)|} \sum_{q \in Q} \sum_{p \in P(q)} \mathbf{1}\left[M(q, I) \neq M(p, I)\right] \tag{B.2}$$

This is less sensitive to the number of paraphrases per question and provides a complementary view.

Margin difference. For a binary classifier with yes/no logits z_{yes} and z_{no}, the margin is $m = z_{\text{yes}} - z_{\text{no}}$. The margin difference between original and paraphrase is:

$$|\Delta m| = |m(q, I) - m(p, I)| \tag{B.3}$$

This captures the magnitude of instability even when the binary answer does not change (e.g., the margin shifts from +5 to +0.1, staying positive but becoming highly uncertain).

Expected Calibration Error (ECE). Given B equal-width bins by predicted confidence:

$$\text{ECE} = \sum_{b=1}^{B} \frac{|B_b|}{N} \left|\text{acc}(B_b) - \text{conf}(B_b)\right| \tag{B.4}$$

where $\text{acc}(B_b)$ is the empirical accuracy of predictions in bin b and $\text{conf}(B_b)$ is the mean predicted confidence. We use $B = 10$ bins throughout, following standard practice [122].

Brier score. For binary predictions with predicted probability $\hat{p}$ and true label $y \in \{0, 1\}$:

$$\text{Brier} = \frac{1}{N} \sum_{i=1}^{N} (\hat{p}_i - y_i)^2 \tag{B.5}$$

Bounded $[0, 1]$, lower is better. Decomposes into calibration and refinement terms.

AUGRC. The Area Under the Generalized Risk-Coverage curve [120] integrates the generalized risk (the joint probability of accepting and erring) rather than the selective risk, bounded $[0, 0.5]$, lower is better. For a selective risk $r(c)$ at coverage c, the generalized risk is $c \cdot r(c)$ and:

$$\text{AUGRC} = \int_0^1 c\, r(c)\, dc. \tag{B.6}$$

Under random ordering $r(c) \approx \text{err}$, so $\text{AUGRC} \approx \text{err}/2$; a well-ranked uncertainty signal pushes it below this baseline. Weighting the risk by coverage makes AUGRC less dominated by the low-coverage tail than AURC.

Appendix C

Reproducibility Information

C.1 Software Environment

The repository requires Python 3.12 or later. Exact package versions are stored with individual result artifacts because the experiments were run across more than one environment. The committed uncertainty-quantification artifact records the following stack:

- **Python**: 3.13.11
- **PyTorch**: 2.6.0 with CUDA 12.4
- **Transformers**: 4.57.3 (Hugging Face)
- **PEFT**: 0.18.1 (Hugging Face, for LoRA)
- **SAE library**: Custom implementation based on Gemma Scope weights
- **Evaluation**: scikit-learn 1.8.0, scipy 1.17.0

C.2 Hardware

- **Training**: Single NVIDIA A100 80GB GPU, 64GB system RAM
- **Inference**: NVIDIA A100 80GB or NVIDIA A6000 48GB
- **Multi-GPU**: 8× A100 for the corruption sweep (distributed across GPUs)
- **SAE extraction**: Single A100, approximately 4GB GPU memory for SAE inference

C.3 Random Seeds and Reproducibility

All experiments use fixed random seeds for reproducibility. The primary seed is 42 for the deployed-model main results. The replication study (Chapter 5) uses seeds 42, 123, 456, 789, and 2024 to characterize variance. The probe question index uses sampling seed 0. Its all-phrasing regime comparison uses eight seeds per regime, while the unseen-phrasing comparison uses six. PyTorch, NumPy, and Python random number generators are all seeded, and CUDA deterministic mode is enabled (though this may produce small numerical differences across GPU architectures).

Model weights are loaded from Hugging Face model hub checkpoints with specific commit hashes recorded in the experiment configurations. The trained LoRA adapters are saved as PEFT checkpoints and can be loaded with `PeftModel.from_pretrained()` for exact reproduction.

C.4 Clinical Consultation Materials

This section reproduces the question set prepared for the clinical deployment-readiness consultation with the clinician coauthors described in Section 6.5. That consultation is designed but has not been conducted, so everything in this section is a protocol rather than a record: no responses were collected, and nothing here reports a result. It is distinct from the clinician adjudication of paraphrase equivalence (Section 3.2.6), which was carried out and is reported there.

C.4.1 Consultation Format

Objective. To establish what evidence, safeguards, and interface behavior clinicians require before trusting a medical VLM for chest X-ray question answering in a clinical workflow.

Participants. The clinical coauthors of this work: a board-certified radiologist with chest-imaging experience and a practicing anesthesiologist.

Format. A structured consultation would be organized around the question set below and the four deployment scenarios derived from the dissertation's results: (1) paraphrase flip case, (2) consistent but Dangerous case, (3) uncertainty triage case, (4) deployment gate case.

Intended output. By design, the consultation would produce a ranked list of clinically meaningful failure modes, safeguard preferences, deployment-scope boundaries, and a mapping from technical safety signals to clinician-facing requirements. These are the protocol's intended outputs, not findings.

C.4.2 Consultation Question Set

Framing. The consultation centers on how clinicians think about safe deployment of AI systems that answer questions about chest X-rays, covering workflow, trust, and deployment requirements. It does not involve decisions about real patients.

Section 1: Background and Workflow

1. What is your current role and specialty?
2. How often do chest X-rays affect your clinical decisions?
3. Walk me through how chest imaging information usually enters your workflow.
4. Have you used any AI-assisted tools in clinical work before? If yes, what kind?

Probes: Do you review images directly, the report, or both? When do you seek radiology confirmation? Where do time pressure and uncertainty show up?

Section 2: Baseline Trust Requirements

5. If an AI system could answer questions about chest X-rays, what would it need to demonstrate before you would consider using it?
6. What kinds of failure would make you reject such a system immediately?
7. What kinds of mistakes are tolerable versus intolerable?

Probes: inconsistency across equivalent questions, confident wrong answers, answers that ignore the image, misleading explanations, hidden uncertainty.

Section 3: Stimulus A: Paraphrase Flip

Present a mock case where the system gives different answers to two equivalent clinician questions about the same image.

8. What is your reaction to this behavior?
9. How serious is this failure relative to ordinary model inaccuracy?
10. Would this change how much you trust the tool overall?

Probes: Is this a deployment blocker? Would it matter less if the final recommendation were correct most of the time?

Section 4: Stimulus B: Consistent but Dangerous

Present a mock case where the model is consistent but behaves the same when the image is removed or swapped.

11. Which seems worse to you: inconsistency across phrasings, or consistency that appears detached from the image?
12. Would a low flip rate reassure you here, or not?

13. If the model is often correct but may be using text shortcuts, how would that affect acceptable use?

Probes: screening versus decision support, low-risk versus high-risk settings, whether correctness without grounding is acceptable.

Section 5: Stimulus C: Uncertainty / Entropy Triage

Present a mock interface where the model provides an answer plus a high-uncertainty flag that recommends human review.

14. Would this uncertainty signal be useful in practice?
15. What would you want the system to do when uncertainty is high?
16. Would you prefer the system to abstain, warn, or still answer with a disclaimer?

Probes: acceptable false alarms, acceptable misses, how much extra review burden is acceptable.

Section 6: Stimulus D: Deployment Gate

Present a mock gate that only displays an answer when paraphrase agreement and at least one image-reliance test pass.

17. Does this kind of gate make the system more trustworthy, less trustworthy, or just more opaque?
18. Would a fail-closed policy be acceptable if it means the tool only answers a minority of cases?
19. What coverage would feel too low to be useful?

Probes: usefulness in triage, usefulness in overnight or resource-limited settings, whether the abstained cases are exactly the ones clinicians most want help with.

Section 7: Explanation and Interface Design

20. What sort of explanation, if any, would be useful with this type of tool?
21. Are attention heatmaps helpful, misleading, or mostly irrelevant for your decision to trust the tool?
22. What would be the minimal interface you would actually want to see?

Probes: plain-language rationale, uncertainty band, escalation recommendation, provenance or audit trail, why a case was blocked by the gate.

Section 8: Governance and Deployment Boundaries

23. In what clinical settings, if any, would you allow this system to be used?
24. Who should be accountable for setting thresholds for uncertainty, gating, or abstention?
25. What monitoring would you want after deployment?

Probes: local calibration, specialty-specific thresholds, mandatory chart review, audit logs, incident reporting.

Section 9: Closing

26. If you could change one thing about this system before deployment, what would it be?
27. Which of these safeguards matters most: consistency testing, image-grounding checks, uncertainty flags, or fail-closed gating?
28. Is there anything important we did not ask that would affect whether clinicians should trust a system like this?

Quick ranking prompt (if time permits): Please rank from most to least important for safe deployment: (a) consistent answers across equivalent questions, (b) evidence that the answer depends on the image, (c) uncertainty flagging, (d) explanation display, (e) automatic abstention or gating.

C.5 Data Availability

- **NIH ChestX-ray14**: Available through the openly released NIH Clinical Center chest radiograph set

- **MIMIC-CXR**: Available through PhysioNet with credentialed access (CITI training and a signed data use agreement); redistribution is not permitted
- **PadChest and PadChest-GR**: Available from BIMCV under the PadChest Dataset Research Use Agreement (registration required, research use only); redistribution is not permitted without prior written permission
- **VinDr-CXR**: Available through PhysioNet with credentialed access, on the same terms as MIMIC-CXR
- **PSF-Med benchmark**: Released at HuggingFace (saillab/psf-med)
- **Paraphrases**: Generated paraphrases and filtering metadata included in the PSF-Med release
- **LoRA checkpoints**: Available upon request for reproducibility
- **Evaluation code**: Released as part of the PSF-Med benchmark repository

Bibliography

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. "Intriguing Properties of Neural Networks". In: *International Conference on Learning Representations (ICLR)*. 2014. DOI: 10.48550/arXiv.1312.6199.

[2] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. "Exponential Expressivity in Deep Neural Networks Through Transient Chaos". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2016. DOI: 10.48550/arXiv.1606.05340.

[3] Yingshu Li, Yunyi Liu, Zhanyu Wang, et al. "A Comprehensive Study of GPT-4V's Multimodal Capabilities in Medical Imaging". In: *medRxiv* (2023). DOI: 10.1101/2023.11.03.23298067.

[4] Jonas Roos, Ron Martin, and Robert Kaczmarczyk. "Evaluating Bard Gemini Pro and GPT-4 Vision Against Student Performance in Medical Visual Question Answering". In: *JMIR Formative Research* 8 (2024), e57592. DOI: 10.2196/57592.

[5] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. "Llava-med: Training a large language-and-vision assistant for biomedicine in one day". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 28541–28564.

[6] Michael Moor, Qian Huang, Shirley Wu, et al. *Med-Flamingo: A Multimodal Medical Few-Shot Learner*. 2023. DOI: 10.48550/arxiv.2307.15189.

[7] Sedigheh Eslami, Christoph Meinel, and Gerard de Melo. "PubMedCLIP: How Much Does CLIP Benefit Visual Question Answering in the Medical Domain?" In: *Findings of the Association for Computational Linguistics: EACL 2023*. 2023, pp. 1181–1193. DOI: 10.18653/v1/2023.findings-eacl.88.

[8] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, et al. "A Multimodal Biomedical Foundation Model Trained from Fifteen Million Image–Text Pairs". In: *NEJM AI* 2.1 (2025). DOI: 10.1056/AIoa2400640.

[9] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, et al. "A Visual–Language Foundation Model for Pathology Image Analysis Using Medical Twitter". In: *Nature Medicine* 29.9 (2023), pp. 2307–2316. DOI: 10.1038/s41591-023-02504-3.

[10] Kai Zhang, Rong Zhou, Eashan Adhikarla, et al. "BiomedGPT: A Generalist Vision-Language Foundation Model for Diverse Biomedical Tasks". In: *Nature Medicine* (2024). DOI: 10.1038/s41591-024-03185-2.

[11] Andrew Sellergren et al. "MedGemma Technical Report". In: *arXiv preprint* (2025). DOI: 10.48550/arXiv.2507.05201.

[12] Eric J. Topol. "As Artificial Intelligence Goes Multimodal, Medical Applications Multiply". In: *Science* 381.6663 (2023), adk6139. DOI: 10.1126/science.adk6139.

[13] Nur Yildirim, Hannah Richardson, Maria Teodora Wetscherek, et al. "Multimodal Healthcare AI: Identifying and Designing Clinically Relevant Vision-Language Applications for Radiology". In: *Proceedings of CHI 2024*. 2024, pp. 1–22. DOI: 10.1145/3613904.3642013.

[14] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J. Topol. "AI in Health and Medicine". In: *Nature Medicine* 28 (2022), pp. 31–38. DOI: 10.1038/s41591-021-01614-0.

[15] Iryna Hartsock and Ghulam Rasool. "Vision-Language Models for Medical Report Generation and Visual Question Answering: A Review". In: *Frontiers in Artificial Intelligence* 7 (2024), p. 1430984. DOI: 10.3389/frai.2024.1430984.

[16] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. "Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6904–6913.

[17] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. "Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4971–4980.

[18] Feiyang Yu, Mark Endo, Rayan Krishnan, et al. "Evaluating Progress in Automatic Chest X-Ray Radiology Report Generation". In: *Patterns* 4 (2023). DOI: 10.1016/j.patter.2023.100802.

[19] Zishan Gu, Fenglin Liu, Changchang Yin, and Ping Zhang. "MedVH: Toward Systematic Evaluation of Hallucination for Large Vision Language Models in the Medical Context". In: *Advanced Intelligent Systems* 8.1 (2025), p. 2500255. DOI: 10.1002/aisy.202500255.

[20] Yongpei Ma, Pengyu Wang, Zhuoran Duan, Adam G Dunn, Usman Naseem, and Jinman Kim. "Bridging the Semantic Gaps: Improving MVQA Consistency with LLM-Augmented Question Sets". In: *Research Square* (2025). DOI: 10.21203/rs.3.rs-6867575/v1.

[21] Daniel Schwabe, Katinka Becker, Martin Seyferth, Andreas Klaß, and Tobias Schaeffter. "The METRIC-Framework for Assessing Data Quality for Trustworthy AI in Medicine: A Systematic Review". In: *npj Digital Medicine* 7.1 (2024), p. 203. DOI: 10.1038/s41746-024-01196-4.

[22] Sara Ketabi, Matthias W. Wagner, Birgit Betina Ertl-Wagner, Greg A. Jamieson, and Farzad Khalvati. "Exploring Radiologists' Expectations of Explainable Machine Learning Models in Medical Image Analysis". In: *arXiv preprint* (2026). DOI: 10.48550/arXiv.2604.11700.

[23] Dennis Pierantozzi, Luca Carlini, Mauro Orazio Drago, Chiara Lena, Cesare Hassan, Elena De Momi, Danail Stoyanov, Sophia Bano, and Mobarak I. Hoque. "When to Trust the Answer: Question-Aligned Semantic Nearest Neighbor Entropy for Safer Surgical VQA". In: *arXiv preprint* (2025). DOI: 10.48550/arxiv.2511.01458.

[24] Mohammad Asadi, Jack W. O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Rajabalifardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. "MIRAGE: The Illusion of Visual Understanding". In: *arXiv preprint* (2026). DOI: 10.48550/arXiv.2603.21687.

[25] Ankit Pal, Jung-Oh Lee, Xiaoman Zhang, Malaikannan Sankarasubbu, Seunghyeon Roh, Won Jung Kim, Meesun Lee, and Pranav Rajpurkar. "ReXVQA: A Large-Scale Visual Question Answering Benchmark for Generalist Chest X-Ray Understanding". In: *Biocomputing 2026*. 2026, pp. 251–264.

[26] Jesse Thomason, Daniel Gordon, and Yonatan Bisk. "Shifting the Baseline: Single Modality Performance on Visual Navigation & QA". In: *arXiv preprint* (2018). DOI: 10.48550/arXiv.1811.00613.

[27] Kim-Celine Kahl, Selen Erkan, Jeremias Traub, et al. "SURE-VQA: Systematic Understanding of Robustness Evaluation in Medical VQA Tasks". In: *arXiv preprint* (2024). DOI: 10.48550/arXiv.2411.19688.

[28] Lukas Buess, Matthias Keicher, Nassir Navab, Andreas Maier, and Soroosh Tayebi Arasteh. "From Large Language Models to Multimodal AI: A Scoping Review on the Potential of Generative AI in Medicine". In: *Biomedical Engineering Letters* 15.5 (2025), pp. 845–863. DOI: 10.1007/s13534-025-00497-1.

[29] Ji Seung Ryu, Hyunyoung Kang, Yuseong Chu, and Sejung Yang. "Vision-Language Foundation Models for Medical Imaging: A Review of Current Practices and Innovations". In: *Biomedical Engineering Letters* 15.5 (2025), pp. 809–830. DOI: 10.1007/s13534-025-00484-6.

[30] Daan Schouten, Giulia Nicoletti, Bas Dille, Catherine Chia, Pierpaolo Vendittelli, Megan Schuurmans, Geert Litjens, and Nadieh Khalili. "Navigating the Landscape of Multimodal AI in Medicine: A Scoping Review on Technical Challenges and Clinical Applications". In: *Medical Image Analysis* 105 (2025), p. 103621. DOI: 10.1016/j.media.2025.103621.

[31] Lu Tang, Jinxu Li, and Sophia Fantus. "Medical Artificial Intelligence Ethics: A Systematic Review of Empirical Studies". In: *Digital Health* 9 (2023), p. 20552076231186064. DOI: 10.1177/20552076231186064.

[32] Manar Aljohani, Jun Hou, Sindhura Kommu, and Xuan Wang. "A Comprehensive Survey on the Trustworthiness of Large Language Models in Healthcare". In: *Findings of the Association for Computational Linguistics: EMNLP 2025*. 2025, pp. 6720–6748. DOI: 10.18653/v1/2025.findings-emnlp.356.

[33] Aofei Chang, Le Huang, Parminder Bhatia, Taha Kass-Hout, Fenglong Ma, and Cao Xiao. "MedHEval: Benchmarking Hallucinations and Mitigation Strategies in Medical Large Vision-Language Models". In: *arXiv preprint* (2025). DOI: 10.48550/arxiv.2503.02157.

[34] Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. "Cares: A comprehensive benchmark of trustworthiness in medical vision language models". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 140334–140365.

[35] Andrea C. Tricco, Erin Lillie, Wasifa Zarin, et al. "PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation". In: *Annals of Internal Medicine* 169.7 (2018), pp. 467–473. DOI: 10.7326/M18-0850.

[36] Christian Bluthgen, Dave Van Veen, Cyril Zakka, et al. "Best Practices for Large Language Models in Radiology". In: *Radiology* 315.1 (2025), e240528. DOI: 10.1148/radiol.240528.

[37] Mourad Ouzzani, Hossam Hammady, Zbys Fedorowicz, and Ahmed Elmagarmid. "Rayyan—a Web and Mobile App for Systematic Reviews". In: *Systematic Reviews* 5 (2016), p. 210. DOI: 10.1186/s13643-016-0384-4.

[38] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports". In: *Scientific data* 6.1 (2019), p. 317.

[39] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (2019), pp. 590–597.

[40] Zijie Cheng, Ariel Yuhan Ong, Siegfried K. Wagner, David A. Merle, et al. "Understanding the Robustness of Vision-Language Models to Medical Image Artefacts". In: *npj Digital Medicine* 8 (2025), p. 727. DOI: 10.1038/s41746-025-02108-w.

[41] Marc Sebastian Huppertz et al. "Revolution or Risk? Assessing the Potential and Challenges of GPT-4V in Radiologic Image Interpretation". In: *European Radiology* 35.3 (2024), pp. 1111–1121. DOI: 10.1007/s00330-024-11115-6.

[42] Tao Tu et al. "Towards Generalist Biomedical AI". In: *NEJM AI* 1.3 (2024), e2300138. DOI: 10.1056/aioa2300138.

[43] Xu Han, Linghao Jin, Xuezhe Ma, and Xiaofeng Liu. "Light-Weight Fine-Tuning Method for Defending Adversarial Noise in Pre-Trained Medical Vision-Language Models". In: *Findings of EMNLP 2024*. 2024, pp. 10784–10799. DOI: 10.18653/v1/2024.findings-emnlp.633.

[44] Yuzhe Yang, Yujia Liu, Xin Liu, et al. "Demographic Bias of Expert-Level Vision-Language Foundation Models in Medical Imaging". In: *Science Advances* 11.13 (2025), eadq0305. DOI: 10.1126/sciadv.adq0305.

[45] Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, et al. "Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards". In: *Findings of EMNLP 2022*. 2022, pp. 4348–4360. DOI: 10.18653/v1/2022.findings-emnlp.319.

[46] Ekaterina Mozhegova, Asad Masood Khattak, Adil Khan, et al. "Assessing the Adversarial Robustness of Multimodal Medical AI Systems: Insights into Vulnerabilities and Modality Interactions". In: *Frontiers in Medicine* 12 (2025). DOI: 10.3389/fmed.2025.1606238.

[47] Zehui Liao, Shishuai Hu, Ke Zou, Huazhu Fu, Liangli Zhen, and Yong Xia. "Vision-Amplified Semantic Entropy for Hallucination Detection in Medical Visual Question Answering". In: *MICCAI 2025*. 2025, pp. 669–679. DOI: 10.1007/978-3-032-04971-1_63.

[48] Zahra Mahdavi, Zahra Khodakaramimaghsoud, Hooman Khaloo, et al. "Med-VCD: Mitigating Hallucination for Medical Large Vision Language Models Through Visual Contrastive Decoding". In: *Computers in Biology and Medicine* 200 (2026), p. 111347. DOI: 10.1016/j.compbiomed.2025.111347.

[49] Runhe Lai, Xinhua Lu, Kanghao Chen, Qichao Chen, Wei-Shi Zheng, and Ruixuan Wang. "Hierarchical Vision-Language Learning for Medical Out-of-Distribution Detection". In: *MICCAI 2025*. 2025, pp. 230–239. DOI: 10.1007/978-3-032-04971-1_22.

[50] Lie Ju, Sijin Zhou, Yukun Zhou, Huimin Lu, Zhuoting Zhu, Pearse A. Keane, and Zongyuan Ge. "Delving into Out-of-Distribution Detection with Medical Vision-Language Models". In: *MICCAI 2025*. 2025, pp. 133–143. DOI: 10.1007/978-3-032-04971-1_13.

[51] Dung Nguyen, Minh Khoi Ho, Huy Ta, et al. "Localizing Before Answering: A Benchmark for Grounded Medical Visual Question Answering". In: *Proceedings of IJCAI 2025*. 2025, pp. 7670–7678. DOI: 10.24963/ijcai.2025/853.

[52] Jinlong He, Pengfei Li, Gang Liu, and Shenjun Zhong. "Parameter-Efficient Fine-Tuning Medical Multimodal Large Language Models for Medical Visual Grounding". In: *2025 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2025, pp. 1–5. DOI: 10.1109/isbi60581.2025.10981029.

[53] Yanyuan Chen, Dexuan Xu, Yu Huang, Songkun Zhan, Hanpin Wang, Dongxue Chen, Xueping Wang, Meikang Qiu, and Hang Li. "MIMO: A Medical Vision Language Model with Visual Referring Multimodal Input and Pixel Grounding Multimodal Output". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025, pp. 25131–25141. DOI: 10.1109/cvpr52734.2025.02303.

[54] Usman Naseem, Matloob Khushi, and Jinman Kim. "Vision-Language Transformer for Interpretable Pathology Visual Question Answering". In: *IEEE Journal of Biomedical and Health Informatics* 27.4 (2023), pp. 1681–1690. DOI: 10.1109/jbhi.2022.3163751.

[55] Md Imran Hossain, Ghada Zamzmi, Peter Mouton, Yu Sun, and Dmitry Goldgof. "Enhancing Concept-Based Explanation with Vision-Language Models". In: *IEEE CBMS 2024*. 2024, pp. 219–224. DOI: 10.1109/cbms61543.2024.00044.

[56] Pengxiang Wang and Junbo Wu. "Large Vision-Language Models-based Human-Understandable Explanations for Medical Image Analysis". In: *2025 4th International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*. IEEE, 2025, pp. 709–714. DOI: 10.1109/icicml67980.2025.11333439.

[57] Sergio Tascon-Morales et al. "Consistency-Preserving Visual Question Answering in Medical Imaging". In: *MICCAI 2022*. 2022, pp. 386–395. DOI: 10.1007/978-3-031-16452-1_37.

[58] Shuning He, Da Ren, Haiwei Pan, Kejia Zhang, and Qing Li. "Advancing Reliable Medical VQA: Evaluation and Enhancement of Model Robustness". In: *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2025, pp. 5637–5642. DOI: 10.1109/bibm66473.2025.11356296.

[59] R. Patrick Xian, Noah R. Baker, Tom David, Qiming Cui, A. Jay Holmgren, Stefan Bauer, Madhumita Sushil, and Reza Abbasi-Asl. "Robustness Tests for Biomedical Foundation Models Should Tailor to Specifications". In: *npj Digital Medicine* 8 (2025), p. 557. DOI: 10.1038/s41746-025-01926-2.

[60] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. "Hallucination of Multimodal Large Language Models: A Survey". In: *arXiv preprint* (2024). DOI: 10.48550/arXiv.2404.18930.

[61] Muhammad Haseeb Shah and Heriberto Cuayáhuitl. "Systematic Analysis of Vision-Language Models for Medical Visual Question Answering". In: *Multimodal Technologies and Interaction* 10.2 (2026), p. 16. DOI: 10.3390/mti10020016.

[62] Dimple Khatri and Sanjan TP Gupta. "Towards Comprehensive Benchmarking of Medical Vision Language Models". In: *Briefings in Bioinformatics* 26 (2025), pp. i24–i25. DOI: 10.1093/bib/bbaf631.035.

[63] Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. "Vision Language Models Are Blind". In: *Computer Vision – ACCV 2024*. 2025, pp. 293–309. DOI: 10.1007/978-981-96-0917-8_17.

[64] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. "On Calibration of Modern Neural Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. 2017, pp. 1321–1330.

[65] Melanie Bernhardt, Fabio De Sousa Ribeiro, and Ben Glocker. "Failure Detection in Medical Image Classification: A Reality Check and Benchmarking Testbed". In: *Transactions on Machine Learning Research* (2022). DOI: 10.48550/arxiv.2205.14094.

[66] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, et al. "Detecting Hallucinations in Large Language Models Using Semantic Entropy". In: *Nature* 630.8017 (2024), pp. 625–630. DOI: 10.1038/s41586-024-07421-0.

[67] Peiran Wang, Linjie Tong, Jian Wu, Jiaxiang Liu, and Zuozhu Liu. "Fair-MoE: Medical Fairness-Oriented Mixture of Experts in Vision-Language Models". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*. Springer Nature Switzerland, 2025, pp. 186–196. DOI: 10.1007/978-3-032-04971-1_18.

[68] Jinming Xue, Bin Wang, Pingping Wang, Dongmei Niu, and Benzheng Wei. "Mitigating Bias Catastrophic Inheritance in Medical Large Vision-Language Models with Logit Fairness Adjustment". In: *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2025, pp. 1913–1918. DOI: 10.1109/bibm66473.2025.11356416.

[69] Judy Wawira Gichoya, Imon Banerjee, Ananth R. Bhimireddy, et al. "AI Recognition of Patient Race in Medical Imaging: A Modelling Study". In: *The Lancet Digital Health* 4.6 (2022), e406–e414. DOI: 10.1016/S2589-7500(22)00063-2.

[70] Md. Sarwar Kamal, Sonia Farhana Nimmy, and Wafa Alharbi. "Explainable Vision-Language Model for Personalized Medicine". In: *Companion Proceedings of the ACM Web Conference 2025*. ACM, 2025, pp. 1908–1915. DOI: 10.1145/3701716.3717738.

[71] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. "OmniMedVQA: A New Large-Scale Comprehensive Evaluation Benchmark for Medical LVLM". In: *CVPR 2024*. 2024, pp. 22170–22183. DOI: 10.1109/cvpr52733.2024.02093.

[72] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. "Are We on the Right Way for Evaluating Large Vision-Language Models?" In: *Advances in Neural Information Processing Systems*. Vol. 37. 2024, pp. 27056–27087.

[73] Ruotao Zhang, Constantine Gatsonis, and Jon Arni Steingrimsson. "Role of Calibration in Uncertainty-Based Referral for Deep Learning". In: *Statistical Methods in Medical Research* 32.5 (2023), pp. 927–943. DOI: 10.1177/09622802231158811.

[74] Yuzhe Yang, Haoran Zhang, Judy Wawira Gichoya, et al. "The Limits of Fair Medical Imaging AI in Real-World Generalization". In: *Nature Medicine* 30.10 (2024), pp. 2838–2848. DOI: 10.1038/s41591-024-03113-4.

[75] Ben Glocker, Charles Jones, Mélanie Bernhardt, and Stefan Winzeck. "Algorithmic Encoding of Protected Characteristics in Chest X-Ray Disease Detection Models". In: *eBioMedicine* 89 (2023), p. 104467. DOI: 10.1016/j.ebiom.2023.104467.

[76] An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. "Vision Language Models Are Biased". In: *arXiv preprint* (2025). DOI: 10.48550/arXiv.2505.23941.

[77] Mingquan Lin, Tianhao Li, Yifan Yang, et al. "Improving Model Fairness in Image-Based Computer-Aided Diagnosis". In: *Nature Communications* 14 (2023), p. 6261. DOI: 10.1038/s41467-023-41974-4.

[78] Baptiste Vasey, David A. Clifton, Gary S. Collins, et al. "DECIDE-AI: New Reporting Guidelines to Bridge the Development-to-Implementation Gap in Clinical Artificial Intelligence". In: *Nature Medicine* 27.2 (2021), pp. 186–187. DOI: 10.1038/s41591-021-01229-5.

[79] Viknesh Sounderajah, Ahmad Guni, Xiaoxuan Liu, et al. "The STARD-AI Reporting Guideline for Diagnostic Accuracy Studies Using Artificial Intelligence". In: *Nature Medicine* 31.10 (2025), pp. 3283–3289. DOI: 10.1038/s41591-025-03953-8.

[80] Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, et al. "Reporting Guidelines for Clinical Trial Reports for Interventions Involving Artificial Intelligence: The CONSORT-AI Extension". In: *Nature Medicine* 26.9 (2020), pp. 1364–1374. DOI: 10.1038/s41591-020-1034-x.

[81] Samantha Cruz Rivera, Xiaoxuan Liu, An-Wen Chan, et al. "Guidelines for Clinical Trial Protocols for Interventions Involving Artificial Intelligence: The SPIRIT-AI Extension". In: *Nature Medicine* 26.9 (2020), pp. 1351–1363. DOI: 10.1038/s41591-020-1037-7.

[82] Gary S. Collins, Karel G. M. Moons, Paula Dhiman, et al. "TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods". In: *BMJ* 385 (2024), e078378. DOI: 10.1136/bmj-2023-078378.

[83] Miao Xiong, Zhiyuan Hu, Xinyang Lu, et al. "Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs". In: *arXiv preprint* (2023). DOI: 10.48550/arxiv.2306.13063.

[84] Anastasios N. Angelopoulos, Stuart R. Pomerantz, Synho Do, et al. "Conformal Triage for Medical Imaging AI Deployment". In: *medRxiv* (2024). DOI: 10.1101/2024.02.09.24302543.

[85] Ayush Noori, Adam Rodman, Alan Karthikesalingam, et al. "A Global Log for Medical AI". In: *arXiv preprint* (2025). DOI: 10.48550/arxiv.2510.04033.

[86] Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, et al. "Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation". In: *arXiv preprint arXiv:2403.08002* (2024).

[87] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. "Towards Generalist Foundation Model for Radiology by Leveraging Web-scale 2D&3D Medical Data". In: *arXiv preprint arXiv:2308.02463* (2023).

[88] Yabin Zhang, Chong Wang, Yunhe Gao, Jiaming Liu, Maya Varma, Justin Xu, Sophie Ostmeier, Jin Long, Sergios Gatidis, Seena Dehkharghani, Arne Michalson, Eun Kyoung Hong, Christian Bluethgen, Haiwei Henry Guo, Alexander Victor Ortiz, Stephan Altmayer, Sandhya Bodapati, Joseph David Janizek, et al. "A Reasoning-Enabled Vision-Language Foundation Model for Chest X-ray Interpretation". In: *arXiv preprint arXiv:2604.00493* (2026). URL: https://arxiv.org/abs/2604.00493.

[89] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. "Contrastive Learning of Medical Visual Representations from Paired Images and Text". In: *Machine Learning for Healthcare Conference* (2022), pp. 2–25.

[90] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. "GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-efficient Medical Image Recognition". In: *IEEE International Conference on Computer Vision (ICCV)*. 2021, pp. 3922–3931.

[91] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, et al. "Making the Most of Text Semantics to Improve Biomedical Vision-Language Processing". In: *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 1–21.

[92] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. "A dataset of clinically generated visual questions and answers about radiology images". In: *Scientific Data* 5 (2018), p. 180251.

[93] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. "SLAKE: A Semantically-Labeled Knowledge-Enhanced Dataset for Medical Visual Question Answering". In: *arXiv preprint arXiv:2102.09542* (2021).

[94] Tony Lee, Haoqin Tu, Chi Heem Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin Somerville Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, and Percy Liang. "VHELM: A Holistic Evaluation of Vision Language Models". In: *arXiv preprint arXiv:2410.07112* (2024).

[95] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. "Med-HALT: Medical Domain Hallucination Test for Large Language Models". In: *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*. 2023, pp. 314–334.

[96] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. "Large Language Models Encode Clinical Knowledge". In: *Nature* 620.7972 (2023), pp. 172–180.

[97] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. "Cycle-Consistency for Robust Visual Question Answering". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 6642–6651.

[98] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. "Beyond Accuracy: Behavioral Testing of NLP Models with CheckList". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 4902–4912.

[99] Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. "Adversarial Example Generation with Syntactically Controlled Paraphrase Networks". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*. 2018, pp. 1875–1885.

[100] Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. "ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs". In: *Findings of the Association for Computational Linguistics: EMNLP 2024* (2024). arXiv:2410.12405.

[101] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vaya. "Padchest: A large chest x-ray image dataset with multi-label annotated reports". In: *Medical image analysis* 66 (2020), p. 101797.

[102] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. "VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations". In: *Scientific Data* 9.1 (2022), p. 429.

[103] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. "Publicly Available Clinical BERT Embeddings". In: *arXiv preprint arXiv:1904.03323* (2019).

[104] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. "Shortcut Learning in Deep Neural Networks". In: *Nature Machine Intelligence* 2.11 (2020), pp. 665–673.

[105] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 618–626.

[106] Sarthak Jain and Byron C Wallace. "Attention is not Explanation". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. 2019, pp. 3543–3556.

[107] Sarah Wiegreffe and Yuval Pinter. "Attention is not not Explanation". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 2019, pp. 11–20.

[108] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. "Sanity Checks for Saliency Maps". In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018.

[109] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. "Textual Explanations for Self-Driving Vehicles". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 563–578. DOI: 10.1007/978-3-030-01216-8_35.

[110] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. "Multimodal Explanations: Justifying Decisions and Pointing to the Evidence". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 8779–8788.

[111] Callum McDougall, Arthur Conmy, János Kramár, Tom Lieberum, Senthooran Rajamanoharan, and Neel Nanda. *Gemma Scope 2 — Technical Paper*. Tech. rep. JumpReLU sparse autoencoders and skip-transcoders for all layers of Gemma 3 (270M, 1B, 4B, 12B, 27B). Google DeepMind, 2025. URL: https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/gemma-scope-2-helping-the-ai-safety-community-deepen-understanding-of-complex-language-model-behavior/Gemma_Scope_2_Technical_Paper.pdf.

[112] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. "Sparse Autoencoders Find Highly Interpretable Features in Language Models". In: *arXiv preprint arXiv:2309.08600* (2023).

[113] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. "LoRA: Low-Rank Adaptation of Large Language Models". In: *International Conference on Learning Representations*. Originally arXiv:2106.09685, June 2021. 2022.

[114] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. "Towards Monosemanticity: Decomposing Language Models With Dictionary Learning". In: *Transformer Circuits Thread* (2023). Anthropic.

[115] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. "Locating and Editing Factual Associations in GPT". In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022, pp. 17359–17372.

[116] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 2097–2106.

[117] Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. *MIMIC-CXR Database (version 2.1.0)*. PhysioNet. 2024. DOI: 10.13026/4jqj-jw95.

[118] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. "QLoRA: Efficient Finetuning of Quantized LLMs". In: *Advances in Neural Information Processing Systems* 36 (2023).

[119] Binesh Sadanandan, Vahid Behzadan, Lekshmy Jayan, and Arun Gopinatha Kurup. "PSF-Med: A Clinician-Audited Benchmark for Paraphrase Sensitivity in Medical Vision-Language Models". In: *MMFM-BIOMED Workshop, CVPR 2026 (also under review, ICMLA 2026)* (2026).

[120] Jeremias Traub, Till J. Bungert, Carsten T. Lüth, Michael Baumgartner, Klaus H. Maier-Hein, Lena Maier-Hein, and Paul F. Jaeger. "Overcoming Common Flaws in the Evaluation of Selective Classification Systems". In: *Advances in Neural Information Processing Systems* 37 (2024).

[121] Anastasios N Angelopoulos and Stephen Bates. "Conformal Prediction: A Gentle Introduction". In: *Foundations and Trends in Machine Learning* 16.4 (2023), pp. 494–591.

[122] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. "Obtaining Well Calibrated Probabilities Using Bayesian Binning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. 2015.