nature portfolio

Peer Review File

CLEAR: An Auditable Foundation Model for Radiology Grounded in Clinical Concepts

Corresponding Author: Professor Tianyu Han

Version 0:

Decision Letter:

Dear Professor Han,

Thank you again for submitting to *Nature Biomedical Engineering* your manuscript, "CLEAR: An Auditable Foundation Model for Radiology Grounded in Clinical Concepts". The manuscript has been seen by two experts, whose reports you will find at the end of this message.

You will see that the reviewers appreciate aspects of the work. However, they articulate concerns about the degree of support for some of the claims and about the advance that the work represents over relevant published studies, and provide useful suggestions for improvement. We hope that with substantial further effort you can address the criticisms, increase the strength of the evidence, and convince the reviewers of the merits of the study. In particular, we would expect that a revised version of the manuscript provides:

* A stronger case for the benefits over CheXzero from a conceptual perspective (ref 3)

* Improved evidence that the method will have practical benefits in a clinical setting (both refs)

When you are ready to resubmit your manuscript, please upload the revised files, a point-by-point rebuttal to the comments from all reviewers, the reporting summary, and a cover letter that explains the main improvements included in the revision and responds to any points highlighted in this decision.

Please follow the following recommendations:

*Please submit your revised manuscript in a word doc or LaTeX format.

* Clearly highlight any amendments to the text and figures to help the reviewers and editors find and understand the changes (yet keep in mind that excessive marking can hinder readability).

* If you and your co-authors disagree with a criticism, provide the arguments to the reviewer (optionally, indicate the relevant points in the cover letter).

* If a criticism or suggestion is not addressed, please indicate so in the rebuttal to the reviewer comments and explain the reason(s).

* Consider including responses to any criticisms raised by more than one reviewer at the beginning of the rebuttal, in a section addressed to all reviewers.

* The rebuttal should include the reviewer comments in point-by-point format (please note that we provide all reviewers will the reports as they appear at the end of this message).

* Provide the rebuttal to the reviewer comments and the cover letter as separate files.

We expect that you will be able to resubmit the manuscript within 25 weeks of receiving this message. If this is the case, you will be protected against potential scooping. Otherwise, we will be happy to consider a revised manuscript as long as the significance of the work is not compromised by work published elsewhere or accepted for publication at *Nature Biomedical Engineering*.

We hope that you will find the referee reports helpful when revising the work. Please do not hesitate to contact me should you have any questions.

Best wishes,
Rita

Rita Strack, Ph.D.
Chief Editor
Nature Biomedical Engineering

Reviewer #1 (Report for the authors (Required)):

This manuscript proposed CLEAR, a chest x ray (CXR) foundation model that maps images to a concept space. The method provides intepretability by the concept-based image embeddings to identify the reasoning path. It is trained contrastively on ~0.87M image–report pairs (MIMIC-CXR, CheXpert-Plus, ReXGradient). CLEAR exceeds CheXzero, BiomedCLIP, and OpenAI-CLIP on VinDr-CXR, Indiana, PadChest (macro-AUROC gains of ~2–18 pts depending on dataset/finding).

Degree of advance
Methodological/technological: Strong. CLEAR's two-stage concept→LLM embedding design is a clear departure from end-to-end VL pretraining and from classic CBMs that require hand-curated labels. The auditing pipeline (concept importance + cluster differential analysis) is well integrated.

Translational/clinical: Results are external and multicontinental. The interpretability may help real-world adoption if calibration, workflow integration, and prospective performance hold up.

Broad implications
Provides a traceable reasoning path that clinicians can interrogate, potentially improving trust, QA, and dataset curation (identifying spurious co-occurrences).

If generalized beyond CXR and coupled with calibration/thresholding, could support prospective auditing and continuous monitoring for domain shift.

Major technical comments:
1. Please clarify strict patient-level splits. Since one patient can have multiple images as well as corresponding reports. If the split is not strict, there can be potential leakage.
2. AUROC is one of the important metrics, could the authors provide calibration results for some of the analysis?
3. Could the author provide more details on the selection of the SFR-Embedding-Mistral model? It is a bit vague on how this one gives best performance. Provide systematic ablations and variance across seeds/models, and analyze robustness to small prompt changes ("no atelectasis" variants, negation handling).
4. Cluster auditing uses an ImageNet-pretrained EfficientNet embedding. Why not using the clusters from CLEAR's own embeddings?

Reviewer #3 (Report for the authors (Required)):

The submitted manuscript introduces a concept-based approach that links chest X-rays with large language model embeddings extracted from radiology reports. The authors position CLEAR as an auditable foundation model capable of mapping each prediction to specific radiological concepts. The work is methodologically considerate, clearly presented, and evaluated on multiple public datasets including PadChest, VinDr-CXR, and the IU collection. The results show moderate improvements over prior vision-language models but more importantly provide examples of how CLEAR identifies concept-level confounders, such as the influence of atelectasis on cardiomediastinum classification.

Major:
Despite its technical strength, the contribution seems incremental rather than transformative. The work builds directly on the paradigm introduced by CheXzero (Nature Biomedical Engineering, 2022), which was the first to demonstrate that chest X-ray interpretation could be learned without any human-annotated labels. The CheXzero study presented an impressive conceptual shift, showing that self-supervised contrastive learning could reach expert-level diagnostic performance while eliminating the dependency on curated labels.

In contrast, CLEAR adopts the same self-supervised framework as CheXzero and extends it by projecting image–text similarities into a semantic space defined by large language model embeddings. This design introduces a valuable interpretability layer enabling predictions to be linked to specific radiological concepts and enabling transparent auditing of model decisions. To clarify, I do not consider the modest improvement in performance a weakness, and it does not diminish the merit of the work. However, the key question is whether this enhanced interpretability justifies a separate publication in

Nature Biomedical Engineering, given that the underlying learning paradigm and experimental platform remain nearly identical to CheXzero which was already published in the same journal. From a practical perspective, the interpretability achieved through CLEAR is conceptual rather than actionable. Moreover, foundation models like CheXzero and CLEAR still rely on unstructured report text, which may encode biases or shortcuts. Even when the model exposes its concept-level reasoning, radiologists would still have to judge whether those concepts came from clinically valid findings or from dataset artifacts (e.g., position markers, devices).

Minor:
- It seems that CLEAR has replaced the validated text extraction used in CheXzero with an LLM-driven, prompt-based approach. While modern language models are flexible and capable of extracting nuanced phrases, they are also prone to hallucinations, inconsistent negation handling, or non-deterministic outputs. The manuscript currently does not describe any quality control efforts or radiologist review in response to these concerns. Since these automatically parsed observations form the foundation of the concept space, I suspect the absence of validation raises concern about the reliability of the extracted concepts and the interpretability derived from them.
- There are several overstatements including the claim that every prediction is “traceable to specific radiological observations.” Behind the scenes, CLEAR maps image embeddings to text-derived concept embeddings; that mapping reflects statistical similarity in an embedding space rather than causal or spatial traceability. In other words, the model identifies which textual concepts correlate with its latent features but does not verify whether they correspond to genuine radiological findings.

Reviewer #3 (Remarks on code availability):

I reviewed their publicly available code repository. The evaluation scripts appear technically sound and consistent with standard evaluation procedures for vision–language models. However, I did not see any code handling negation during evaluation, possibly suggesting that it may be handled implicitly by the model. If that is the case, the manuscript should clearly explain how this was achieved and verified, because implicit handling of negation is very difficult to accomplish reliably.

Version 1:

Decision Letter:

Dear Dr. Han,

Thank you for submitting your revised manuscript "CLEAR: An Auditable Foundation Model for Radiology Grounded in Clinical Concepts" (NBME-25-4035A). Having consulted with the original reviewers, I am pleased to write that we shall be happy to publish the manuscript in Nature Biomedical Engineering, pending minor revisions to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements. Please do not upload the final materials and make any revisions until you receive this additional information from us.

TRANSPARENT PEER REVIEW
Nature Biomedical Engineering offers a transparent peer review option for new original research manuscripts submitted from 17th February 2021. We encourage increased transparency in peer review by publishing the reviewer comments, author rebuttal letters and editorial decision letters if the authors agree. Such peer review material is made available as a supplementary peer review file. **Please state in the cover letter ‘I wish to participate in transparent peer review’ if you want to opt in, or ‘I do not wish to participate in transparent peer review’ if you don’t.** Failure to state your preference will result in delays in accepting your manuscript for publication.

Please note: we allow redactions to authors’ rebuttal and reviewer comments in the interest of confidentiality. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons. Reviewer names will be published in the peer review files if the reviewer signed the comments to authors, or if reviewers explicitly agree to release their name. For more information, please refer to our FAQ page.

ORCID
IMPORTANT: Non-corresponding authors do not have to link their ORCIDs but are encouraged to do so. Please note that it will not be possible to add/modify ORCIDs at proof. Thus, please let your co-authors know that if they wish to have their ORCID added to the paper they must follow the procedure described in the following link prior to acceptance: https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

Author names using non-Roman characters
Nature Portfolio journals can support presentation of author names using non-Roman characters in the HTML version of the article. If you wish to, please include author names in parentheses after the Roman-character spelling; see example online here. Currently supported scripts are: Arabic, Chinese, Cyrillic, Devanagari, Greek, Hebrew, Hangul, Japanese and Persian. You will be asked to verify the rendering is correct at proof stage.

Sincerely,
Rita

Rita Strack, Ph.D.
Chief Editor
Nature Biomedical Engineering

Reviewer #3 (Report for the authors (Required)):

First and foremost, the authors should be commended for the comprehensive revision and addressing the multiple concerns raised during the initial review. The manuscript has improved substantially, and the additional analyses and clarifications strengthen the technical presentation as well as the evidence supporting the proposed framework.

The remaining limitations relate primarily to the perceived significance of the conceptual advance and the proximity of the approach to prior work, rather than methodological weaknesses in the study in its current form. While the authors provide additional evidence supporting the interpretability and auditing capabilities of the CLEAR framework, the underlying learning paradigm remains closely related to existing vision–language approaches. However, to clarify, these remaining points reflect inherent characteristics of the current design rather than issues that could reasonably be addressed through further rounds of revision.

In sum, the authors have made a good-faith effort to respond to the reviewers' comments, and the revised manuscript presents a clearer and more rigorous account of the proposed method and its potential clinical relevance -- which should be of interest to the broader scientific community.

Reviewer #3 (Remarks on code availability):

No major issues were identified in the repository provided by the authors.

Version 2:

Decision Letter:

Dear Professor Han,

I am happy to inform you that your manuscript, "CLEAR: An Auditable Foundation Model for Radiology Grounded in Clinical Concepts", has now been accepted for publication in *Nature Biomedical Engineering*.

Over the next few weeks, the figures will be checked for production quality, the text edited to ensure that it conforms to house style, and the manuscript typeset.

Our Articles are published about 40 days after the acceptance date (we recommend that you inform your institutional press office of this timeframe), and you will be notified of the actual publication date a few days in advance. Articles can be published any working day of the week, and are pushed live shortly after 10 am London time.

Publishing agreement. You will be asked to digitally sign a publishing agreement (grant of rights). After the signed publishing agreement has been received, the proofs of the article will be sent to you for review. If you have any queries during the production process, or you cannot meet the requested deadline for returning the proofs, please contact rjsproduction@springernature.com.

Nature Biomedical Engineering is a Transformative Journal. Authors may publish their research with us through the traditional subscription access route, or make their paper immediately open access through payment of an article-processing charge. More information about publication options is available.

You may need to take specific actions to comply with funder and institutional open-access mandates. If the

work described in the accepted manuscript is supported by a funder that requires immediate open access (as outlined, for example, by Plan S) and your manuscript was originally submitted on or after January 1st 2021, then you should select the gold OA route. Authors selecting subscription publication will need to accept our standard licensing terms (including our self-archiving policies), and these will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

Acceptance of your manuscript is conditional on agreement, by all authors, with both our media embargo and confidentiality and pre-publicity policies. In particular, you may arrange your own publicity of the Article (for instance, through your institutional press office), as long as you ensure that journalists strictly adhere to the media embargo.

To assist you in disseminating the work, as soon as the Article is published you will be able to take advantage of the Springer Nature SharedIt initiative to generate a unique shareable link to the Article that will allow anyone (with or without a subscription) to read it. Recipients of the link who are subscribers will also be able to download and print the PDF.

Thank you for having submitted this work to *Nature Biomedical Engineering*.

Best wishes,
Rita

Rita Strack, Ph.D.
Chief Editor
Nature Biomedical Engineering

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

Response to Reviewers

Comments from the Chief Editor

Editorial Comment: General Comment

Dear Professor Han,
Thank you again for submitting to Nature Biomedical Engineering your manuscript, "CLEAR: An Auditable Foundation Model for Radiology Grounded in Clinical Concepts". The manuscript has been seen by two experts, whose reports you will find at the end of this message.
You will see that the reviewers appreciate aspects of the work. However, they articulate concerns about the degree of support for some of the claims and about the advance that the work represents over relevant published studies, and provide useful suggestions for improvement. We hope that with substantial further effort you can address the criticisms, increase the strength of the evidence, and convince the reviewers of the merits of the study.

We thank the Editor for the opportunity to revise our manuscript and for the constructive feedback from both reviewers. We have undertaken substantial additional work to strengthen the evidence and address all concerns raised. Key additions include a new concept intervention experiment demonstrating that CLEAR's interpretability is actionable, enabling targeted model correction that is impossible with end-to-end black-box models. We also conducted blinded radiologist evaluation studies confirming both the clinical relevance of CLEAR's concepts and the fidelity of its observation extraction pipeline, as well as comprehensive calibration analyses and systematic ablations of embedding model selection and prompt robustness. We address each comment point-by-point below.

Editorial Comment: Detailed Comment 1

In particular, we would expect that a revised version of the manuscript provides:
* A stronger case for the benefits over CheXzero from a conceptual perspective (ref 3)

We have substantially strengthened the case for CLEAR's advance over CheXzero and other end-to-end vision-language models. In our response to Reviewer 3 Major Comment 1, we present a new concept intervention experiment demonstrating that CLEAR's interpretability is *actionable*: by identifying and removing confounding atelectasis-related concepts from the enlarged cardiomediastinum (EC) classification task, we improved AUROC from 0.727 to 0.784, a targeted correction that is fundamentally impossible with any end-to-end black-box model, including CheXzero, which exhibits the same performance degradation but offers no mechanism to diagnose or correct it. This workflow, together with CLEAR's concept-level auditing and automated concept bottleneck model construction, represents a qualitative departure from the CheXzero paradigm. We have revised the manuscript (Figures 2c–2g and Discussion) to prominently present these results.

Editorial Comment: Detailed Comment 2

* Improved evidence that the method will have practical benefits in a clinical setting (both refs)

We have added several new analyses that provide concrete evidence of clinical utility. The concept intervention experiment (response to Reviewer 3 Major Comment 1) demonstrates a practical physician-in-the-loop workflow where clinicians can inspect CLEAR's concept-level reasoning, identify spurious

associations, and correct them without retraining the model, yielding measurable diagnostic improvements. In response to Reviewer 3 Major Comment 2, we conducted a blinded evaluation with three board-certified radiologists who rated 130 concept–pathology pairs. Of these, 89.8% of CLEAR's top-ranked concepts were judged diagnostically relevant and 0% were classified as dataset artifacts (Fleiss' $\kappa = 0.749$), confirming that CLEAR surfaces clinically meaningful reasoning rather than shortcuts. A separate radiologist evaluation of the observation extraction (response to Reviewer 3 Minor Comment 1) confirmed 97.3% faithfulness with 0% hallucination rate, validating the reliability of the concept space that underpins CLEAR's interpretability. We also provide comprehensive calibration analyses (response to Reviewer 1 Major Technical Comment 2) demonstrating that CLEAR maintains excellent probability calibration across all evaluation settings, a prerequisite for clinical deployment.

Comments from Reviewer 1

Reviewer 1: General Comment

This manuscript proposed CLEAR, a chest x ray (CXR) foundation model that maps images to a concept space. The method provides intepretability by the concept-based image embeddings to identify the reasoning path. It is trained contrastively on 0.87M image–report pairs (MIMIC-CXR, CheXpert-Plus, ReXGradient). CLEAR exceeds CheXzero, BiomedCLIP, and OpenAI-CLIP on VinDr-CXR, Indiana, PadChest (macro-AUROC gains of 2–18 pts depending on dataset/finding).

We thank the reviewer for this accurate summary of our work and for recognizing the performance improvements CLEAR achieves across multiple external datasets.

Reviewer 1: Detailed Comment 1

Degree of advance
Methodological/technological: Strong. CLEAR's two-stage concept→LLM embedding design is a clear departure from end-to-end VL pretraining and from classic CBMs that require hand-curated labels. The auditing pipeline (concept importance + cluster differential analysis) is well integrated.

We appreciate the reviewer's recognition of CLEAR's methodological contributions. The two-stage design leveraging LLM embeddings indeed enables concept-based interpretability without requiring manually curated concept labels, which has been a major limitation of traditional concept bottleneck models [1].

Reviewer 1: Detailed Comment 2

Translational/clinical: Results are external and multicontinental. The interpretability may help real-world adoption if calibration, workflow integration, and prospective performance hold up.

We thank the reviewer for highlighting the translational relevance of our work. We have now included comprehensive calibration analyses (see our response to Reviewer 1 Major Technical Comment 2), demonstrating that CLEAR maintains excellent calibration across all evaluation settings, with expected calibration error (ECE) values ranging from 0.007 to 0.040 on external test sets in the concept bottleneck model framework. Regarding workflow integration, the concept intervention experiment described in our response to Reviewer 3 Major Comment 1 provides a concrete example of a physician-in-the-loop audit-and-correct workflow, where identifying and removing confounding atelectasis-related concepts from EC classification improved AUROC from 0.727 to 0.784 without retraining the image encoder. These additions are reflected in the revised Results and Discussion sections of the manuscript.

Reviewer 1: Detailed Comment 3

Broad implications
Provides a traceable reasoning path that clinicians can interrogate, potentially improving trust, QA, and dataset curation (identifying spurious co-occurrences).

We agree with the reviewer's assessment. The ability to decompose predictions into specific radiological concepts enables clinicians to verify model reasoning and identify potential spurious correlations in training data, as we demonstrate through our auditing case studies.

Reviewer 1: Detailed Comment 4

If generalized beyond CXR and coupled with calibration/thresholding, could support prospective auditing and continuous monitoring for domain shift.

We thank the reviewer for this forward-looking perspective. The CLEAR framework is designed to be modality-agnostic, requiring only paired image-report data for training.

Reviewer 1: Major technical comment 1

Major technical comments:
1. Please clarify strict patient-level splits. Since one patient can have multiple images as well as corresponding reports. If the split is not strict, there can be potential leakage.

We appreciate the reviewer raising this important concern regarding train/test data splits. We confirm that our study design ensures strict patient-level separation through two mechanisms: **1. External Validation Design:** Our primary evaluation uses external test sets from entirely different institutions than the training data, making patient-level overlap impossible by design (Table R1).

We have added the following subsection to the Methods section of the revised manuscript:

Our study design ensures strict patient-level separation between training and evaluation data through two mechanisms. First, our primary evaluation uses external test sets (VinDr-CXR, PadChest, Indiana) from entirely different institutions and countries than the training data, making patient-level overlap impossible by design. Second, for the CheXpert test set, although CheXpert-Plus is used for pretraining, the official CheXpert test set contains 500 patients held out by Stanford; we programmatically verified zero patient overlap between our CheXpert-Plus training subset (64,525 patients) and the CheXpert test set (0 of 500 test patients appear in training). MIMIC-CXR uses the official PhysioNet patient-level splits. For supervised experiments (linear probing and CBM), the classifier is trained exclusively on MIMIC-CXR using frozen image encoder features, and all evaluation data come from entirely different institutions. A detailed summary of patient-level data splits is provided in Supplementary Table S1.

2. Within-Dataset Verification: For experiments using the same data source for both training and testing (e.g., linear probe trained on MIMIC-CXR evaluated on CheXpert test), we verified zero patient overlap between training and test sets.

Table R1: ***Note that this table corresponds to the Supplementary Table S1 in the revised manuscript.*** Patient-level data split summary demonstrating no overlap between training and evaluation.

Experiment	Split	Dataset	Images	Patients	Institution/Country
Contrastive Pretraining (Stage 1)					
	Train	MIMIC-CXR [2]	377,110	65,379	Beth Israel Deaconess, USA
	Train	CheXpert-Plus [3]	223,228	64,525	Stanford Hospital, USA
	Train	ReXGradient [4]	273,004	109,487	Harvard, USA
Supervised Evaluation (Linear Probing & CBM)					
	Train	MIMIC-CXR	377,110	65,379	Beth Israel Deaconess, USA
Test Data (All Experiments)					
	Test	CheXpert Test [3]	500	500	Stanford Hospital, USA
	Test	VinDr-CXR [5]	3,000	—	Vietnam
	Test	PadChest [6, 7]	7,943	7,272	San Juan Hospital, Spain
	Test	Indiana	7,466	3,851	Indiana University, USA

Specifically:

- **VinDr-CXR, PadChest, Indiana:** These test sets originate from entirely different institutions and countries than the training data, eliminating any possibility of patient overlap.
- **CheXpert Test:** Although CheXpert-Plus is used for pretraining, the official CheXpert test set contains 500 patients held out by Stanford. We programmatically verified zero patient overlap between our CheXpert-Plus training subset and the CheXpert test set (0 of 500 test patients appear in the 64,525 training patients).
- **MIMIC-CXR:** We use the official MIMIC-CXR splits which enforce strict patient-level separation as specified by PhysioNet.
- **Supervised experiments (linear probing and CBM):** For these downstream tasks, the classifier is trained exclusively on MIMIC-CXR using frozen image encoder features. Since the supervised training data (MIMIC-CXR) and evaluation data (CheXpert Test and external datasets) come from entirely different institutions, patient-level separation is maintained by design.

Reviewer 1: Major technical comment 2

2. AUROC is one of the important metrics, could the authors provide calibration results for some of the analysis?

We thank the reviewer for this important suggestion. We have analyzed calibration across all three evaluation settings: linear probing, concept bottleneck models, and zero-shot classification.

Linear Probing (CheXpert Test Set): For supervised linear probing, which produces well-distributed probability outputs across [0, 1], we report Expected Calibration Error (ECE) and Brier Score averaged over 20 random seeds (Table R2).

Table R2: ***Note that this table corresponds to the Supplementary Table S14 in the revised manuscript.*** Calibration metrics for linear probing on CheXpert test set (mean ± std over 20 seeds).

Model	ECE	Brier Score
CLEAR	**0.072 ± 0.001**	0.119 ± 0.000
CheXzero [8]	0.076 ± 0.000	**0.118 ± 0.000**
BiomedCLIP [9]	**0.072 ± 0.001**	0.135 ± 0.000
OpenAI-CLIP [10]	0.086 ± 0.000	0.142 ± 0.000

Concept Bottleneck Models: We compare CLEAR CBM against standard CBM across four external datasets (Table R3).

Table R3: ***Note that this table corresponds to the Supplementary Table S16 in the revised manuscript.*** Calibration metrics for concept bottleneck models (mean ± std over 20 seeds). ECE = Expected Calibration Error, Brier = Brier Score.

Dataset	Standard CBM		CLEAR CBM	
	ECE	Brier Score	ECE	Brier Score
CheXpert	0.043 ± 0.002	**0.099 ± 0.001**	**0.040 ± 0.001**	0.100 ± 0.000
VinDr-CXR	0.020 ± 0.000	**0.026 ± 0.000**	**0.017 ± 0.001**	0.029 ± 0.000
PadChest	0.052 ± 0.002	0.030 ± 0.000	**0.025 ± 0.001**	**0.025 ± 0.000**
Indiana	0.025 ± 0.002	0.032 ± 0.000	**0.007 ± 0.000**	0.032 ± 0.000

As shown in Tables R2 and R3, CLEAR achieves excellent calibration in supervised settings. For linear probing, CLEAR matches or outperforms all baselines (ECE: 0.072 vs. 0.076 to 0.086). For concept bottleneck models, CLEAR CBM consistently achieves better ECE than standard CBM, with the most substantial improvements on PadChest (ECE: 0.025 vs. 0.052, 52% reduction) and Indiana (ECE: 0.007 vs. 0.025, 72% reduction). Brier Scores are comparable between the two CBM variants across all datasets. These results demonstrate that CLEAR's interpretable architecture does not compromise probability calibration.

Note that Tables R2–R4 correspond to Supplementary Tables S14–S12 in the revised manuscript. We have added the following text to the linear probing Results section:

Beyond classification performance, CLEAR also demonstrates excellent probability calibration for linear probing, with an Expected Calibration Error (ECE) of 0.072 that matched or outperformed all baselines (Supplementary Table S14).

And the following text to the concept bottleneck model Results section:

CLEAR CBM also achieved ECE values of 0.007 to 0.040 across four external test sets, consistently outperforming the standard CBM (Supplementary Table S16). These results demonstrate that CLEAR's interpretable architecture does not compromise probability calibration, a prerequisite for clinical deployment.

We have also added ECE and Brier Score definitions to the Evaluation Metrics section of the Methods:

To assess probability calibration, we computed the ECE [11] and Brier Score [12] across all evaluation settings (Supplementary Tables S14–S12). ECE measures the average absolute difference between predicted probabilities and observed frequencies across probability bins, with lower values indicating better calibration. The Brier Score is the mean squared error between predicted probabilities and binary outcomes, simultaneously capturing calibration and classification accuracy.

Zero-Shot Classification: We also report zero-shot calibration for completeness (Table R4), though we note that these metrics require careful interpretation. Zero-shot CLIP-based models compute probabilities via softmax over normalized embedding similarities, which produces outputs concentrated around 0.5 (typical range: 0.45 to 0.53) rather than spanning [0, 1]. This is a well-known property of

contrastive vision-language models and results in elevated ECE values for all models in this setting. Within this framework, CLEAR achieves calibration comparable to CheXzero across all external test sets.

Table R4: ***Note that this table corresponds to the Supplementary Table S12 in the revised manuscript.*** Zero-shot calibration metrics for CLIP-based models. ECE = Expected Calibration Error, Brier = Brier Score. Zero-shot evaluation is deterministic (no training), so no std is reported. Note: predictions cluster around 0.5 due to softmax over normalized similarities.

Dataset	CLEAR		CheXzero	
	ECE	Brier Score	ECE	Brier Score
CheXpert	**0.299**	**0.248**	0.300	0.249
VinDr-CXR	**0.429**	**0.242**	0.433	0.247
PadChest	0.487	0.248	**0.486**	**0.247**
Indiana	**0.478**	**0.245**	0.481	0.248

Reviewer 1: Major technical comment 3

3. Could the author provide more details on the selection of the SFR-Embedding-Mistral model? It is a bit vague on how this one gives best performance. Provide systematic ablations and variance across seeds/models, and analyze robustness to small prompt changes ("no atelectasis" variants, negation handling).

We thank the reviewer for this important methodological point. We provide comprehensive details on embedding model selection, systematic ablations, and prompt robustness analysis below.

1. Embedding Model Selection: We evaluated four state-of-the-art embedding models for generating concept embeddings: SFR-Embedding-Mistral [13] (Salesforce, 4,096 dimensions, last-token pooling with instruction formatting), Qwen3-Embedding-8B [14] (Alibaba, 4,096 dimensions, last-token pooling), text-embedding-3-small (OpenAI, 1,536 dimensions, proprietary architecture), and Biomed-BERT [15] (Microsoft, 768 dimensions, [CLS] token pooling). To avoid selection bias from evaluating on test data, we performed model selection on the CheXpert validation set (n=200), which was held out from both training and final evaluation. Table R5 shows the results:

Table R5: ***Note that this table corresponds to the Supplementary Table S5 in the revised manuscript.*** Embedding model selection on CheXpert validation set (n=200). SFR-Embedding-Mistral achieved the highest macro AUROC and was selected for CLEAR. External test set evaluations were performed only after this selection.

Embedding Model	Macro AUROC
SFR-Embedding-Mistral	**0.789 ± 0.076**
text-embedding-3-small (OpenAI)	0.692 ± 0.172
Qwen3-Embedding-8B	0.618 ± 0.195
BiomedBERT	0.607 ± 0.195

SFR-Embedding-Mistral achieved the highest validation performance (0.789 AUROC), outperforming the second-best model (OpenAI) by 9.7 percentage points. This model was therefore selected for all subsequent experiments. External test set results (VinDr-CXR, PadChest, Indiana) were computed only after this selection, ensuring no data leakage.

2. Variance Analysis

Since LLM embedding generation is deterministic (identical inputs produce identical outputs), traditional seed-based variance does not apply. Instead, we quantify uncertainty through three complementary approaches: bootstrap resampling (n=1,000 iterations) to compute 95% confidence intervals for each pathology's AUROC, cross-model comparison showing performance spread across the four embedding architectures, and cross-dataset evaluation demonstrating consistent model ranking across three external test sets. The per-label bootstrap confidence intervals are provided in the Supplementary Information.

3. Prompt Robustness and Negation Handling

We systematically evaluated robustness to different prompt formulations, testing six variants of positive/negative prompt pairs as requested by the reviewer:

Table R6: ***Note that this table corresponds to the Supplementary Table S6 in the revised manuscript.*** Prompt variants tested for robustness analysis. Each variant represents a different way to express the presence or absence of a finding.

Variant	Positive Prompt	Negative Prompt
Standard	"{disease}"	"no {disease}"
Absence	"{disease}"	"absence of {disease}"
Present/Absent	"{disease} present"	"{disease} absent"
Not Present	"{disease}"	"{disease} not present"
Clinical	"findings consistent with {disease}"	"no evidence of {disease}"
Negative	"{disease}"	"negative for {disease}"

We evaluated all four embedding models across all six prompt variants on three external test sets. Table R7 summarizes the robustness statistics:

Table R7: ***Note that this table corresponds to the Supplementary Table S7 in the revised manuscript.*** Prompt robustness analysis across embedding models and datasets. Mean AUROC and range (min–max) computed across six prompt variants. Narrower ranges indicate greater robustness to prompt formulation changes.

Model	VinDr-CXR	PadChest	Indiana
Mean AUROC [range]			
SFR-Embedding-Mistral	**0.774** [0.760–0.788]	**0.717** [0.710–0.724]	**0.735** [0.719–0.748]
text-embedding-3-small	0.693 [0.636–0.758]	0.670 [0.645–0.703]	0.672 [0.636–0.718]
Qwen3-Embedding-8B	0.718 [0.681–0.757]	0.672 [0.621–0.717]	0.699 [0.666–0.728]
BiomedBERT	0.549 [0.534–0.571]	0.570 [0.529–0.605]	0.556 [0.525–0.582]

SFR-Embedding-Mistral demonstrates exceptional robustness, with AUROC ranges of only 0.014 to 0.029 (2 to 4% variation) across all datasets, indicating that the model reliably captures the semantic meaning of clinical concepts regardless of how presence/absence is expressed. Beyond robustness, SFR-Embedding-Mistral also achieves the highest mean AUROC across all datasets and prompt variants, confirming that our model selection was appropriate. Negation handling is effective across all tested formulations. All six negation variants, including "no atelectasis," "absence of atelectasis," "atelectasis absent," "atelectasis not present," "no evidence of atelectasis," and "negative for atelectasis," produce consistent results, demonstrating that CLEAR's LLM embedding space properly encodes semantic negation. By comparison, other models show greater sensitivity to prompt changes. OpenAI's text-embedding-3-small shows the widest AUROC range (up to 0.121 on VinDr-CXR), suggesting that some embedding models are more sensitive to prompt wording than others, further justifying our selection of SFR-Embedding-Mistral.

Note that Tables R5–R7 correspond to Supplementary Tables S5–S7 in the revised manuscript. We have updated the Concept Space Curation section to now read:

To avoid selection bias from evaluating on test data, we performed embedding model selection on the CheXpert validation set (n=200), which was held out from both pretraining and final evaluation. SFR-Embedding-Mistral achieved the highest validation macro AUROC (0.789 ± 0.076), outperforming text-embedding-3-small (0.692 ± 0.172), Qwen3-Embedding-8B (0.618 ± 0.195), and BiomedBERT (0.607 ± 0.195) (Supplementary Table S5). SFR-Embedding-Mistral was therefore selected for all subsequent experiments; external test set results were computed only after this selection. We further validated robustness to prompt formulation by testing six positive/negative prompt variants (e.g., "{disease}" / "no {disease}", "{disease} present" / "{disease} absent") across all four embedding models on three external test sets. SFR-Embedding-Mistral demonstrated the greatest robustness (AUROC variation of 0.014 to 0.029) and the highest mean performance across all conditions (Supplementary Tables S6 and S7).

Reviewer 1: Major technical comment 4

4. Cluster auditing uses an ImageNet-pretrained EfficientNet embedding. Why not using the clusters from CLEAR's own embeddings?

We thank the reviewer for this insightful methodological question. Our use of ImageNet-pretrained EfficientNetV2-S [16] embeddings for cluster auditing directly follows the MA-MONET methodology [17], which specifies using "50 principal components of the embedding of the penultimate layer of the EfficientNetV2-S model (pretrained on ImageNet)" for clustering in model auditing. We adopted this validated approach to maintain methodological consistency and enable comparability with prior work.

Beyond following established methodology, using an external embedding is principled for model auditing: an ImageNet-pretrained model provides a domain-agnostic visual feature representation independent of the model being audited, ensuring that the clustering step does not inherit the audited model's own representation biases.

We have added this justification to the Model Auditing subsection of the Methods:

This follows the MA-MONET methodology [17] to maintain comparability with prior work; using an external, domain-agnostic embedding for clustering avoids inheriting the audited model's own representation biases.

Comments from Reviewer 3

Reviewer 3: General Comment

The submitted manuscript introduces a concept-based approach that links chest X-rays with large language model embeddings extracted from radiology reports. The authors position CLEAR as an auditable foundation model capable of mapping each prediction to specific radiological concepts. The work is methodologically considerate, clearly presented, and evaluated on multiple public datasets including PadChest, VinDr-CXR, and the IU collection. The results show moderate improvements over prior vision-language models but more importantly provide examples of how CLEAR identifies concept-level confounders, such as the influence of atelectasis on cardiomediastinum classification.

We thank the reviewer for this thoughtful and balanced summary of our work. We appreciate the recognition that CLEAR provides concept-level confounder identification, which we believe represents a key advance for trustworthy medical AI.

Reviewer 3: Major Comment 1

Major:
Despite its technical strength, the contribution seems incremental rather than transformative. The work builds directly on the paradigm introduced by CheXzero (Nature Biomedical Engineering, 2022), which was the first to demonstrate that chest X-ray interpretation could be learned without any human-annotated labels. The CheXzero study presented an impressive conceptual shift, showing that self-supervised contrastive learning could reach expert-level diagnostic performance while eliminating the dependency on curated labels.
In contrast, CLEAR adopts the same self-supervised framework as CheXzero and extends it by projecting image–text similarities into a semantic space defined by large language model embeddings. This design introduces a valuable interpretability layer enabling predictions to be linked to specific radiological concepts and enabling transparent auditing of model decisions. To clarify, I do not consider the modest improvement in performance a weakness, and it does not diminish the merit of the work.
However, the key question is whether this enhanced interpretability justifies a separate publication in Nature Biomedical Engineering, given that the underlying learning paradigm and experimental platform remain nearly identical to CheXzero which was already published in the same journal. From a practical perspective, the interpretability achieved through CLEAR is conceptual rather than actionable.

We thank the reviewer for this thoughtful and important critique. We respectfully address both core concerns below: whether CLEAR represents a sufficient advance over existing end-to-end vision-language models, and whether the interpretability is actionable rather than merely conceptual.

CLEAR's interpretability is actionable. To directly address the reviewer's central concern, we conducted a new concept selection intervention experiment that demonstrates how CLEAR's interpretability can be translated into concrete, targeted model improvements, a capability that is absent in end-to-end vision-language models. We present the EC classification task as a case study, where we previously identified that atelectasis-related concepts were confounding the supervised classification (Figure R1c,e). The experiment follows a diagnose, intervene, and improve workflow. CLEAR's concept importance analysis first revealed that the top-ranked concepts driving EC classification were all atelectasis-related rather than mediastinal concepts (Figure R1e), indicating that the model had learned a spurious shortcut from label co-occurrence in MIMIC-CXR. Data auditing of the MIMIC-

CXR training set further confirmed this confounding: the most differentially expressed concepts between EC-positive and EC-negative images centered on atelectasis and effusion rather than mediastinal contours (Figure R1d). Leveraging CLEAR's concept-based architecture, we then intervened by retaining only the 49,167 clinically relevant mediastinal concepts (out of 368,294 total) for classification, requiring no retraining of the image encoder, no collection of new data, and no modification to the model architecture. Only the lightweight linear classifier was refitted on the filtered concept features. This targeted intervention improved the AUROC from 0.727 ± 0.000 to 0.784 ± 0.001 (mean $\pm$ std over 20 random seeds), a 7.8% relative improvement (Figure R1f). Crucially, the post-intervention concept importance analysis confirms that predictions are now driven entirely by clinically appropriate mediastinal concepts such as "The mediastinal contours are widened but stable due to a torturous enlarged thoracic aorta" and "Thoracic aorta enlarged in both ascending and descending portions" (Figure R1g).

This workflow is fundamentally impossible with any end-to-end black-box model. As shown in Figure R1f, CheXzero [8] exhibits the same performance degradation on EC (AUROC dropping from 0.896 zero-shot to 0.728 supervised), yet in end-to-end models there is no mechanism to identify that atelectasis concepts are the source of the problem, selectively remove the confounding influence, or verify that a correction has been applied. The only recourse would be to retrain the entire model on curated data, an approach that neither identifies the root cause nor guarantees that the same or other confounders will not persist.

Broader contributions beyond end-to-end vision-language models. Beyond the concept intervention experiment, CLEAR introduces several capabilities that are qualitatively distinct from existing end-to-end approaches. For any zero-shot prediction, CLEAR can surface the radiological observations most influential to its decision, enabling clinicians to verify whether predictions are grounded in clinically appropriate reasoning, a mechanism entirely absent in current vision-language foundation models. Through concept differential analysis, CLEAR can also systematically identify dataset-level confounders that would remain hidden in black-box models and could propagate to clinical deployment. Additionally, CLEAR enables the automated construction of concept bottleneck models that achieve radiologist-level performance using concept sets discovered through data-driven analysis rather than manual expert curation.

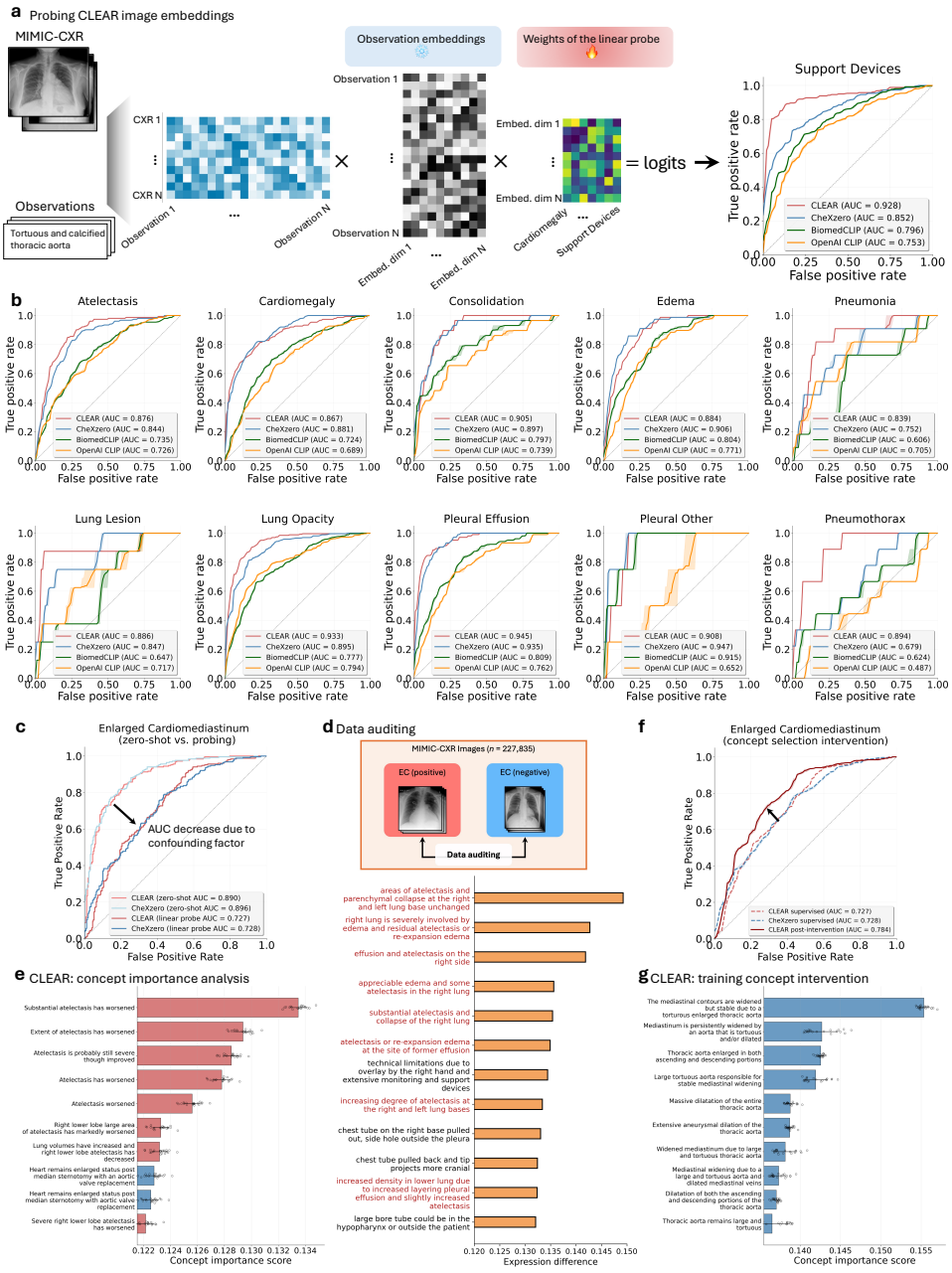

Fig. R1. ***Note that this figure corresponds to Fig. 3 in the revised manuscript.*** **Supervised linear probing experiments, data auditing, and concept intervention. (a)** Linear probing pipeline. CLEAR computes similarities between CXRs and 368,294 radiological observations, projects them via LLM embeddings, and applies learned weights for classification. For the "support devices" example, the AUROC for CLEAR (0.928) is superior to those of CheXzero (0.852), BiomedCLIP (0.796), and OpenAICLIP (0.753). **(b)** ROC curves for supervised classification on CheXpert pathologies. CLEAR (red) achieved AUROCs ranging from 0.839 (pneumonia) to 0.945 (pleural effusion) across 10 diagnostic tasks. **(c)** ROC curves showing AUROC degradation from zero-shot to supervised probing for both CLEAR and CheXzero, caused by confounding factors in the training data. **(d)** Data auditing of the MIMIC-CXR training set reveals that atelectasis-related concepts (highlighted in red) dominate the distinction between EC-positive and EC-negative cases. **(e)** Concept importance analysis (baseline): the top-ranked concepts are atelectasis-related (red bars), with only two mediastinal concepts (blue bars). **(f)** After concept intervention (retaining only mediastinal concepts), CLEAR's AUROC improves from 0.727 to 0.784, surpassing CheXzero's supervised AUROC of 0.728. **(g)** Post-intervention concept importance: predictions are now driven entirely by clinically appropriate mediastinal concepts (blue bars).

Note that Figure R1 corresponds to Fig. 3 in the revised manuscript. We have added the following paragraph to the Results section:

Critically, CLEAR's concept-based architecture enables not only the detection but also the targeted correction of such confounders through concept intervention. Leveraging the concept importance analysis above, we retained only the 49,167 clinically relevant mediastinal concepts (out of 368,294 total) for enlarged cardiomediastinum classification, requiring no retraining of the image encoder and no collection of new data; only the lightweight linear classifier was refitted on the filtered concept features. This targeted intervention improved the AUROC from 0.727 ± 0.0003 to 0.784 ± 0.0011 (mean $\pm$ std over 20 random seeds), a 7.8% relative improvement (Figure 3f). Post-intervention concept importance analysis confirms that predictions are now driven entirely by clinically appropriate mediastinal concepts such as widened mediastinal contours and enlarged thoracic aorta descriptions (Figure 3g). This diagnose-intervene-improve workflow is fundamentally impossible with any end-to-end black-box model: CheXzero exhibits the same performance degradation on enlarged cardiomediastinum (AUROC dropping from 0.896 zero-shot to 0.728 supervised), yet in end-to-end models there is no mechanism to identify the confounding source, selectively remove its influence, or verify that a correction has been applied (Figure 3f).

We have also added the following paragraph to the Discussion:

Importantly, CLEAR's interpretability is not merely conceptual but actionable: by identifying atelectasis-related concepts as confounders in EC classification, we intervened by retaining only 49,167 clinically relevant mediastinal concepts (out of 368,294 total), improving AUROC from 0.727 to 0.784 without retraining the image encoder (Figure 3f,g). This diagnose-intervene-improve workflow represents a qualitative departure from end-to-end vision-language models such as CheXzero, which exhibits the same performance degradation but offers no mechanism to diagnose or correct it. The concept intervention demonstrates a practical physician-in-the-loop paradigm in which clinicians can inspect CLEAR's concept-level reasoning, identify spurious associations, and correct them with measurable diagnostic improvements.

Reviewer 3: Major Comment 2

> Moreover, foundation models like CheXzero and CLEAR still rely on unstructured report text, which may encode biases or shortcuts. Even when the model exposes its concept-level reasoning, radiologists would still have to judge whether those concepts came from clinically valid findings or from dataset artifacts (e.g., position markers, devices).

We thank the reviewer for raising this important point about biases in report-derived models.

Clarification on CLEAR's use of report text. CLEAR's pipeline involves two stages: In the first stage (visual-language pretraining), we train image and text encoders using contrastive learning on raw impression text, following the same paradigm as CheXzero [8] and other vision-language models. In the second stage (concept annotation and inference), however, CLEAR does not operate on raw report text. Instead, the trained encoders compute similarity scores between each CXR and 368,294 individual radiological observations, which are short, discrete clinical sentences, e.g., "mild cardiomegaly", "bilateral pleural effusions", "calcified nodule in the left upper lobe", extracted and deduplicated from reports using Ministral-8B from Mistral AI.

Unlike raw impressions, these observations largely exclude document-level comparison language and non-diagnostic content. For example, a raw impression reading "In comparison with the study ___, the patient has taken a better inspiration. The diffuse areas of increased opacification in the right hemithorax have essentially cleared. Cardiac silhouette is at the upper limits of normal" is decomposed into the observations "Better inspiration", "Diffuse areas of increased opacification in the right hemithorax have cleared", and "Cardiac silhouette at the upper limits of normal", stripping the temporal comparison preamble. Across the full 368,294-observation concept space, comparison language (e.g., "in comparison with", "compared to") appears in only 0.40% of observations versus 14.2% of raw

impressions, recommendation language in 0.17% versus 2.9%, and clinical history references in 0.037% versus 0.38%. This decomposition ensures that model reasoning manifests as inspectable concept scores rather than remaining hidden in an opaque embedding, enabling bias detection and correction as we demonstrate below.

CLEAR enables detection and correction of biases from training. As we demonstrated in our response to Reviewer 3 Major Comment 1, the atelectasis confounder in EC classification was a bias that entered through the training data. In an end-to-end model, this bias would remain hidden. In CLEAR, it was detected through concept importance analysis, traced to label co-occurrence in the training set through data auditing, and corrected through concept selection (AUROC improving from 0.727 to 0.784). Similarly, our zero-shot pneumonia audit identified interstitial lung patterns and lingular atelectasis as confounders in the highest-error cluster, findings that align with established Fleischner Society guidelines [18, 19]. CLEAR provides the transparency that makes physician-in-the-loop verification possible, whereas in end-to-end models there is very limited information for radiologists to inspect.

Radiologist evaluation of concept clinical relevance. To directly quantify whether the concepts CLEAR surfaces are clinically valid or dataset artifacts, we conducted a blinded evaluation study with three board-certified radiologists (9, 10, and 15 years of experience in diagnostic radiology). For each of the 13 CheXpert diagnostic labels, we presented the top 10 highest-importance concepts identified by CLEAR's linear probing on the full concept space of 368,294 observations (Supplementary Figures S5 and S6). We chose linear probing rather than the CBM, because the linear probing concepts are drawn from the full, uncurated observation space, providing a more stringent test of concept quality. Concepts were shown in randomized order. Each radiologist independently rated every concept-pathology pair into one of three categories: Category A (diagnostically relevant) indicates that the concept directly supports the target diagnosis. Category B (clinically valid but non-specific) indicates that the concept describes a real radiological finding but is not specific to this pathology. Category C (dataset artifact or shortcut) indicates that the concept reflects imaging artifacts, position markers, devices, or institutional patterns rather than genuine diagnostic information.

Table R8: ***Note that this table corresponds to Table 3 in the revised manuscript.*** Radiologist evaluation of concept clinical relevance. Three board-certified radiologists independently rated the top 10 concepts per pathology from CLEAR's linear probing on the full 368,294-concept space. Values represent the percentage of concepts rated in each category (mean across three raters).

Pathology	A (%)	B (%)	C (%)
Atelectasis	100.0	0.0	0.0
Cardiomegaly	100.0	0.0	0.0
Consolidation	96.7	3.3	0.0
Edema	100.0	0.0	0.0
Enlarged Cardiomediastinum	20.0	80.0	0.0
Fracture	100.0	0.0	0.0
Lung Lesion	86.7	13.3	0.0
Lung Opacity	96.7	3.3	0.0
Pleural Effusion	100.0	0.0	0.0
Pleural Other	66.7	33.3	0.0
Pneumonia	100.0	0.0	0.0
Pneumothorax	100.0	0.0	0.0
Support Devices	100.0	0.0	0.0
Average	89.8	10.2	0.0
Fleiss' κ (inter-rater agreement): 0.749			

As shown in Table R8, across all 13 pathologies, not a single top-ranked concept was classified as a

dataset artifact or shortcut (Category C = 0%), indicating that none of the concepts CLEAR surfaces reflect imaging artifacts, position markers, devices, or institutional patterns. Instead, 89.8% of concepts were rated as diagnostically relevant (Category A), with the remaining 10.2% rated as clinically valid but non-specific (Category B). Ten of the 13 pathologies achieved ≥96.7% Category A ratings, and eight reached perfect 100% agreement across all three raters.

The two pathologies with elevated Category B ratings are clinically informative rather than concerning. For EC, 80% of concepts were rated clinically valid but non-specific. This result independently corroborates our concept intervention analysis (Reviewer 3 Major Comment 1): the radiologists confirmed that the top-ranked EC concepts describe real radiological findings (e.g., atelectasis, effusion) rather than artifacts, but these findings are not specific to EC. This is precisely the confounding pattern that CLEAR's concept importance analysis had already detected computationally and that we subsequently corrected through concept selection, improving AUROC from 0.727 to 0.784. For pleural other (66.7% A, 33.3% B), the moderate non-specificity reflects the heterogeneous nature of this catch-all diagnostic category, which encompasses diverse pleural abnormalities beyond effusion and pneumothorax. Inter-rater agreement was substantial (Fleiss' $\kappa = 0.749$), indicating that the three radiologists independently reached consistent conclusions about concept relevance despite evaluating 130 concept–pathology pairs in randomized, blinded order.

Note that Table R8 corresponds to Table 3 in the revised manuscript. We have added the following paragraph to the Results section:

Although CLEAR's concept space is derived from report text, which may encode biases or shortcuts, CLEAR makes these inspectable rather than hidden. A blinded evaluation with three board-certified radiologists (9, 10, and 15 years of experience) confirmed that 89.8% of CLEAR's top-ranked concepts were rated diagnostically relevant and 0% were classified as dataset artifacts across all 13 CheXpert pathologies (Fleiss' $\kappa = 0.749$ [20]; Table 3). This demonstrates that CLEAR surfaces clinically meaningful reasoning rather than shortcuts, and its transparency enables physician-in-the-loop verification that is entirely absent in end-to-end models where biases remain hidden.

Reviewer 3: Minor Comment 1

Minor:
- It seems that CLEAR has replaced the validated text extraction used in CheXzero with an LLM-driven, prompt-based approach. While modern language models are flexible and capable of extracting nuanced phrases, they are also prone to hallucinations, inconsistent negation handling, or non-deterministic outputs. The manuscript currently does not describe any quality control efforts or radiologist review in response to these concerns. Since these automatically parsed observations form the foundation of the concept space, I suspect the absence of validation raises concern about the reliability of the extracted concepts and the interpretability derived from them.

We thank the reviewer for this important concern. Since the 368,294 extracted observations form the foundation of CLEAR's concept space, their reliability is critical. We have now conducted systematic validation addressing each of the three specific concerns raised: non-deterministic outputs, inconsistent negation handling, and hallucination risk.

Determinism of the extraction process. We evaluated output reproducibility by running the full extraction pipeline (Ministral-8B-Instruct with temperature set to 0 and greedy decoding) twice on 500 randomly sampled MIMIC-CXR impression texts, using identical prompts and model weights. Of these 500 reports, 491 (98.2%) produced observation lists that were character-for-character identical across both runs (Table R9). All nine non-identical outputs involved minor wording variations or differences in observation granularity. None of the nine discrepancies fabricated a clinical observation absent from the original report or reversed the polarity of a negated statement. The residual 1.8% non-determinism despite greedy decoding is a known property of GPU-based LLM inference due to floating-point non-associativity [21, 22]. Importantly, the concept space was extracted once and used

consistently across all experiments, so this residual variability does not affect downstream results.

Table R9: ***Note that this table corresponds to the Supplementary Table S2 in the revised manuscript.*** Determinism validation of Ministral-8B observation extraction on 500 MIMIC-CXR reports (temperature = 0, greedy decoding). Each report was processed twice with identical inputs.

Metric	Count	Rate
Observation list exact match	491 / 500	98.2%
Total observations (run 1)	1,529	
Total observations (run 2)	1,525	

Negation fidelity. Among the 50 pairs selected for radiologist validation (described below), 15 (30%) contained negation language. No negation reversal was observed among these 15 pairs (0% rated incorrect; Fleiss' $\kappa = 0.934$; Table R11). Table R10 shows representative extraction outputs, demonstrating correct decomposition of conjunctive negations, preservation of implicit negation phrasing, and omission of non-diagnostic text. This is further corroborated by the prompt robustness analysis (Reviewer 1 Major Technical Comment 3), where six negation formulations produced consistent AUROC values (ranges 0.014–0.029) across three external test sets (Table R7).

Table R10: ***Note that this table corresponds to the Supplementary Table S3 in the revised manuscript.*** Representative extraction examples showing all observations produced by Ministral-8B for each impression. Negated observations are shown in *italics*.

Original Impression	All Extracted Observations
"Elevated right hemidiaphragm. Tiny left pleural effusion. No signs of pneumonia."	1. Elevated right hemidiaphragm 2. Tiny left pleural effusion 3. *No signs of pneumonia*
"Mild cardiomegaly, new since the prior study, and enlargement of the azygos compatible with volume overload without frank pulmonary edema."	1. Mild cardiomegaly 2. Enlargement of the azygos 3. *Volume overload without frank pulmonary edema*
"No previous images. The cardiac silhouette is mildly enlarged with left ventricular prominence. However, no vascular congestion, pleural effusion, or acute focal pneumonia."	1. Cardiac silhouette is mildly enlarged 2. Left ventricular prominence 3. *No vascular congestion* 4. *No pleural effusion* 5. *No acute focal pneumonia*
"Normal heart, lungs, hila, mediastinum, and pleural surfaces. No evidence of intrathoracic malignancy or infection."	1. Normal heart 2. Normal lungs 3. Normal hila 4. Normal mediastinum 5. Normal pleural surfaces 6. *No evidence of intrathoracic malignancy* 7. *No evidence of intrathoracic infection*

Extraction fidelity validation by radiologists. To directly address the concern about hallucination, we conducted a blinded validation study. We sampled 50 report-observation pairs from the extraction outputs, stratified to include 30% negated and 70% non-negated observations. Three board-certified radiologists independently rated each pair on faithfulness (faithful / partially faithful / hallucinated) and negation preservation (correct / incorrect / N/A). The results (Table R11) show 0% hallucination across all 50 pairs (97.3% faithful, 2.7% partially faithful).

Table R11: ***Note that this table corresponds to the Supplementary Table S4 in the revised manuscript.*** Radiologist evaluation of extraction fidelity. Three board-certified radiologists independently rated 50 report-observation pairs (15 negated, 35 non-negated) on faithfulness and negation preservation.

Metric	All pairs (n=50)	Negated pairs (n=15)
Faithful (%)	97.3	100.0
Partially faithful (%)	2.7	0.0
Hallucinated (%)	0.0	0.0
Negation correct (%)	–	93.3
Negation incorrect (%)	–	0.0
Negation N/A (%)	–	6.7
Fleiss' κ (negation)	0.934	

Note that Tables R9–R11 correspond to Supplementary Tables S2–S4 in the revised manuscript. We have added the following quality control paragraph to the Concept Space Curation section of the Methods:

To validate extraction reliability, we assessed three dimensions of quality control. First, determinism: running the full pipeline twice on 500 randomly sampled reports (temperature = 0, greedy decoding) yielded 98.2% character-for-character identical outputs; the 1.8% residual variability involved minor wording differences with no fabricated or negation-reversed observations (Supplementary Table S2). Second, negation fidelity: among 15 negated observation pairs evaluated by three board-certified radiologists, 0% exhibited negation reversal (Fleiss' $\kappa = 0.934$ [20]). Third, content faithfulness: in a blinded evaluation of 50 report-observation pairs, 97.3% were rated faithful and 0% were classified as hallucinated (Supplementary Tables S3 and S4).

Reviewer 3: Minor Comment 2

- There are several overstatements including the claim that every prediction is "traceable to specific radiological observations." Behind the scenes, CLEAR maps image embeddings to text-derived concept embeddings; that mapping reflects statistical similarity in an embedding space rather than causal or spatial traceability. In other words, the model identifies which textual concepts correlate with its latent features but does not verify whether they correspond to genuine radiological findings.

We thank the reviewer for this important clarification. The reviewer is correct that the mapping between image features and concept embeddings reflects statistical similarity in a learned embedding space, not a causal or spatially grounded correspondence. We acknowledge that terms such as "traceable" could be read as implying a stronger form of traceability than what is technically provided, and we have revised the manuscript accordingly.

Specifically, we have replaced language that could imply causal or spatial grounding (e.g., "traceable to specific radiological observations") with more precise formulations such as "predictions can be decomposed into weighted contributions from individual radiological observations" and "the concepts most associated with a given prediction can be surfaced for clinical inspection." These revised descriptions accurately characterize what CLEAR provides: a transparent, concept-level decomposition of each prediction into interpretable components, where each component corresponds to a specific radiological observation and carries an explicit numerical weight. Specifically, in the Abstract, we have replaced "making every prediction traceable to specific radiological observations" with:

making every prediction decomposable into weighted contributions from individual radiological observations.

In the Discussion and Methods, we have replaced "clinicians can trace predictions through concept

activations" with:

clinicians can inspect predictions through concept activations

We have also added the following paragraph to the Discussion acknowledging the distinction between statistical association and causal traceability:

We note that CLEAR's concept-level decomposition reflects statistical similarity in a learned embedding space, not causal or spatial traceability. The mapping between image features and concept embeddings is correlational: the model identifies which textual concepts are most associated with its latent features but does not verify whether they correspond to spatially localized findings in the image. Nevertheless, this correlational decomposition is actionable, as demonstrated by the concept intervention experiment (Figure 3f,g), where removing statistically identified confounders yielded measurable diagnostic improvement.

Reviewer 3: Remarks on Code Availability

I reviewed their publicly available code repository. The evaluation scripts appear technically sound and consistent with standard evaluation procedures for vision–language models. However, I did not see any code handling negation during evaluation, possibly suggesting that it may be handled implicitly by the model. If that is the case, the manuscript should clearly explain how this was achieved and verified, because implicit handling of negation is very difficult to accomplish reliably.

We thank the reviewer for the careful code review and for raising this point. Negation is in fact handled explicitly in the evaluation code, not implicitly by the model. We apologize that this was not sufficiently documented in the repository. We have now added inline comments to clarify the mechanism.

Explicit negation handling via positive/negative prompt pairs. CLEAR's zero-shot evaluation constructs two separate text prompts for each diagnostic label: a positive prompt (e.g., "atelectasis") and a negative prompt (e.g., "no atelectasis"). Both prompts are independently encoded by the text encoder, and the cosine similarity between each image embedding and both prompt embeddings is computed separately. The final prediction probability is then obtained via a softmax over the positive and negative logits:

$$P(\text{positive}) = \frac{\exp(\text{sim}(\mathbf{x}, \mathbf{t}_{\text{pos}}))}{\exp(\text{sim}(\mathbf{x}, \mathbf{t}_{\text{pos}})) + \exp(\text{sim}(\mathbf{x}, \mathbf{t}_{\text{neg}}))} \tag{1}$$

where $\mathbf{x}$ is the image embedding, $\mathbf{t}_{\text{pos}}$ and $\mathbf{t}_{\text{neg}}$ are the text embeddings for the positive and negative prompts, and $\text{sim}(\cdot, \cdot)$ denotes cosine similarity. This formulation treats the positive and negative prompts as competing hypotheses under a softmax, so the model must encode a meaningful distinction between "atelectasis" and "no atelectasis" in embedding space for accurate classification.

Code locations. The positive/negative template pair is defined as `("{}", "no {}")` in the benchmark configuration (`examples/benchmark/benchmark_base.py`). The softmax evaluation is implemented in `src/clear/zero_shot.py` (function `run_softmax_eval`), which separately computes positive and negative logits and applies the softmax normalization shown above. The same logic appears in `src/clear/pipeline.py` using PyTorch's `F.softmax` over a stacked positive/negative logit tensor. For the concept-based zero-shot evaluation (`examples/concepts/exp_zeroshot.py`), positive and negative class prompts are generated as `"{label}"` and `"no {label}"`, embedded separately by the LLM, and combined via the same softmax formulation.

Verification of negation encoding quality. As presented in our response to Reviewer 1 Major Technical Comment 3 (Table R7), we systematically tested six different negation formulations and confirmed that SFR-Embedding-Mistral produces consistent classification performance across all variants, with AUROC ranges of only 0.014 to 0.029 per dataset. This demonstrates that the embedding model reliably distinguishes positive from negated clinical statements regardless of how the negation is phrased. Additionally, our extraction fidelity analysis (Reviewer 3 Minor Comment 1) confirmed that negated observations are faithfully preserved during the concept extraction step. We have added a

description of the explicit negation handling mechanism to the revised Methods section and improved the code documentation in the repository.

References

[1] Koh, P. W. *et al.* Concept bottleneck models. In *International conference on machine learning*, 5338–5348 (PMLR, 2020).

[2] Johnson, A. E. *et al.* Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data* **6**, 317 (2019).

[3] Irvin, J. *et al.* Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 590–597 (2019).

[4] Zhang, X., Acosta, J. N., Miller, J., Huang, O. & Rajpurkar, P. Rexgradient-160k: A large-scale publicly available dataset of chest radiographs with free-text reports. *arXiv preprint arXiv:2505.00228* (2025).

[5] Nguyen, H. Q. *et al.* Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data* **9**, 429 (2022).

[6] Bustos, A., Pertusa, A., Salinas, J.-M. & De La Iglesia-Vaya, M. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis* **66**, 101797 (2020).

[7] Han, T. *et al.* Reconstruction of patient-specific confounders in ai-based radiologic image interpretation using generative pretraining. *Cell Reports Medicine* **5** (2024).

[8] Tiu, E. *et al.* Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature biomedical engineering* **6**, 1399–1406 (2022).

[9] Zhang, S. *et al.* A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI* **2**, AIoa2400640 (2025).

[10] Radford, A. *et al.* Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763 (PmLR, 2021).

[11] Naeini, M. P., Cooper, G. & Hauskrecht, M. Obtaining well calibrated probabilities using Bayesian binning into quantiles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29 (2015).

[12] Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**, 1–3 (1950).

[13] Meng, R. *et al.* SFR-Embedding-Mistral: Enhance Text Retrieval with Transfer Learning. Salesforce AI Research Blog (2024). URL https://www.salesforce.com/blog/sfr-embedding/.

[14] Zhang, Y. *et al.* Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176* (2025).

[15] Gu, Y. *et al.* Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)* **3**, 1–23 (2021).

[16] Tan, M. & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114 (PMLR, 2019).

[17] Kim, C. *et al.* Transparent medical image ai via an image–text foundation model grounded in medical literature. *Nature Medicine* **30**, 1154–1165 (2024).

[18] Lynch, D. A. *et al.* Diagnostic criteria for idiopathic pulmonary fibrosis: a fleischner society white paper. *The Lancet respiratory medicine* **6**, 138–153 (2018).

[19] Hatabu, H. *et al.* Interstitial lung abnormalities detected incidentally on ct: a position paper from the fleischner society. *The lancet Respiratory medicine* **8**, 726–737 (2020).

[20] Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychological Bulletin* **76**, 378–382 (1971).

[21] Yuan, J. *et al.* Understanding and mitigating numerical sources of nondeterminism in LLM inference. *arXiv preprint arXiv:2506.09501* (2025).

[22] Shanmugavelu, S. *et al.* Impacts of floating-point non-associativity on reproducibility for HPC and deep learning applications. *arXiv preprint arXiv:2408.05148* (2024).