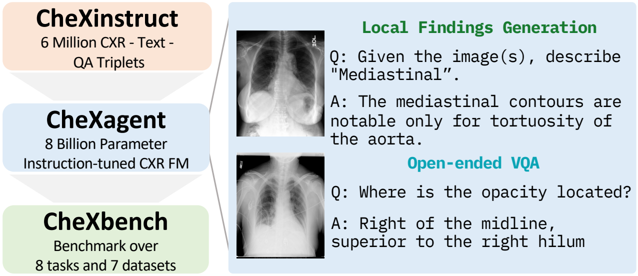
# CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation

**Zhihong Chen[1*], Maya Varma[1*], Jean-Benoit Delbrouck[1*], Magdalini Paschali[1]**
**Louis Blankemeier[1], Dave Van Veen[1], Jeya Maria Jose Valanarasu[1], Alaa Youssef[1]**
**Joseph Paul Cohen[1], Eduardo Pontes Reis[1], Emily B. Tsai[1], Andrew Johnston[1]**
**Cameron Olsen[1], Tanishq Mathew Abraham[2], Sergios Gatidis[1]**
**Akshay S Chaudhari[1], Curtis Langlotz[1]**
[1]Stanford University [2]Stability AI
{zhihongc,mvarma2,jbdel,paschali,akshaysc,langlotz}@stanford.edu

## Abstract

Chest X-rays (CXRs) are the most frequently performed imaging test in clinical practice. Recent advances in the development of vision-language foundation models (FMs) give rise to the possibility of performing automated CXR interpretation. In this work, we present (i) *CheXinstruct* - a large-scale instruction-tuning dataset curated from 28 publicly-available datasets; (ii) *CheXagent* - an instruction-tuned FM capable of analyzing and summarizing CXRs; and (iii) *CheXbench* - a novel benchmark designed to systematically evaluate FMs across 8 clinically-relevant CXR interpretation tasks. Extensive quantitative evaluations and qualitative reviews with five expert radiologists demonstrate that CheXagent outperforms previously-developed general- and medical-domain FMs on CheXbench tasks by up to 97.5%.[1]


CheXinstruct
6 Million CXR - Text - QA Triplets
CheXagent
8 Billion Parameter Instruction-tuned CXR FM
CheXbench
Benchmark over 8 tasks and 7 datasets
Local Findings Generation
Open-ended VQA

Q: Given the image(s), describe "Mediastinal".

A: The mediastinal contours are notable only for tortuosity of the aorta.

Q: Where is the opacity located?

A: Right of the midline, superior to the right hilum

Figure 1: Overview of the proposed pipeline: CheXinstruct is a curation of datasets for instruction-tuning across various CXR tasks, CheXagent is our clinical FM for CXR interpretation, and CheXbench is our comprehensive FM evaluation benchmark. Two example CXR interpretation tasks include local findings generation and open-ended visual question answering (VQA).

## Introduction

Foundation models (FMs) have recently emerged as a powerful class of models capable of performing a diverse range of reasoning and comprehension tasks (Bommasani et al. 2021). In this work, we present the following three components, also summarized in Fig. 1, to help create capable and robust FMs for chest X-ray (CXR) interpretation:

1. *CheXinstruct* is an instruction-tuning dataset with 6M instruction-image-answer triplets designed to improve the ability of FMs to interpret CXRs. We collect instructions from 34 tasks and 65 unique datasets, spanning categories including coarse- and fine-grained image understanding, question answering, and text generation.
2. *CheXagent* is an instruction-tuned foundation model with 8B parameters capable of analyzing images, understanding text, and generating responses. Our methodology for developing CheXagent includes training (1) a clinical LLM capable of understanding radiology reports, (2) a vision encoder capable of reading CXRs, and (3) a network to bridge the vision and language modalities. We then perform instruction-tuning using data from CheXinstruct.
3. *CheXbench* is a novel benchmark designed to rigorously evaluate FMs across two evaluation axes: image perception and textual understanding. We introduce 8 tasks across 7 CXR datasets, and we evaluate performance using close-ended multiple-choice predictions as well as open-ended text generation.

We use CheXbench to compare CheXagent with six previous FMs from both general and medical domains. We further provide an evaluation of potential model bias and highlight performance disparities across demographic factors of sex, race and age to improve model transparency.

## Data: CheXinstruct

CheXinstruct seeks to cover a broad range of tasks to support training CXR FMs. These tasks can either (i) improve the abilities of FMs to understand CXRs or (ii) improve clinical decision making. This dataset is comprised of five categories of tasks, each categorized by their specific capabilities:

- Coarse-grained Image Understanding, which defines the overall understanding of CXRs, e.g., view classification (Johnson et al. 2019), and disease classification (Wang et al. 2017; Irvin et al. 2019; Reis et al. 2022; Pavlova et al. 2022; Holste et al. 2023; Bannur et al. 2023; Jaeger et al. 2014; Bustos et al. 2020; Shih et al. 2019).
- Fine-grained Image Understanding, which defines the localized understanding of CXRs, e.g., abnormality detection (Nguyen et al. 2022; Pham, Tran, and Nguyen 2022),

[*] Equal contributions.

[1] Our project is at https://stanford-aimi.github.io/chexagent.html.

Table 1: Comparison between CheXagent and general domain and medical domain FMs on CheXbench. For image perception tasks, we report accuracy; For Findings Generation and Findings Summarization, we report RadGraph Score and Rouge-L, respectively.

| **Task** | **Dataset** | **General-domain FMs** | | **Medical-domain FMs** | | | | **CheXagent** |
|---|---|---|---|---|---|---|---|---|
| | | **BLIP-2** | **InstructBLIP** | **XrayGPT** | **MedFlamingo** | **RadFM** | **LLaVA-Med** | **(Ours)** |
| View Classification | MIMIC-CXR | 28.8 | 25.3 | 24.0 | 25.0 | 28.5 | 23.8 | 97.5 |
| | CheXpert | 38.0 | 34.0 | 33.0 | 39.0 | 37.0 | 30.0 | 96.7 |
| Binary Disease Classification | SIIM | 53.0 | 54.0 | 50.0 | 50.0 | 50.0 | 49.0 | 64.0 |
| | RSNA | 50.0 | 60.0 | 50.0 | 50.0 | 50.0 | 44.0 | 81.0 |
| | CheXpert | 51.5 | 53.2 | 51.5 | 48.5 | 55.8 | 47.6 | 76.0 |
| Single Disease Identification | OpenI | 40.2 | 40.2 | 45.4 | 39.0 | 42.2 | 43.8 | 47.0 |
| | MIMIC-CXR | 25.6 | 22.6 | 24.1 | 25.6 | 27.2 | 26.7 | 30.3 |
| | CheXpert | 21.3 | 19.5 | 23.7 | 26.0 | 26.6 | 26.0 | 29.6 |
| Multi Disease Identification | OpenI | 48.5 | 54.4 | 57.7 | 46.1 | 52.8 | 53.9 | 55.6 |
| | MIMIC-CXR | 30.0 | 25.3 | 39.0 | 14.7 | 22.3 | 28.7 | 55.3 |
| | CheXpert | 4.3 | 6.1 | 3.9 | 7.1 | 23.6 | 2.1 | 52.1 |
| Visual Question Answering | Rad-Restruct | 41.2 | 42.4 | 38.6 | 45.5 | 48.5 | 34.9 | 57.1 |
| | SLAKE | 74.3 | 86.4 | 52.4 | 64.8 | 85.0 | 55.5 | 78.1 |
| Image-Text Reasoning | OpenI | 47.9 | 52.6 | 52.4 | 54.7 | 54.0 | 45.8 | 59.0 |
| Findings Section Generation | CheXpert | - | - | 9.0 | 1.7 | 5.1 | 4.2 | 14.6 |
| Findings Summarization | MIMIC-CXR | - | - | - | - | - | - | 40.3 |

abnormality grounding (Boecking et al. 2022), and foreign object detection (Xue et al. 2015).

- Question Answering, which defines the ability to respond to a question, e.g., close-ended and open-ended visual question answering (VQA) (Zhang et al. 2023; Pellegrini et al. 2023; Ben Abacha et al. 2019; Bae et al. 2023), difference VQA (Hu et al. 2023), and text QA.
- Text Generation, which defines the ability to generate radiology report sections, including a description of the findings (Demner-Fushman et al. 2012; Vayá et al. 2020; Pelka et al. 2018), impression generation (Feng et al. 2021), findings summarization (Chen et al. 2023), and local findings generation (Johnson et al. 2019).
- Miscellaneous: This category defines the miscellaneous abilities that are critical for a CXR FM, e.g., report evaluation (Yu et al. 2023; Miura et al. 2021), and natural language explanation (Kayser et al. 2022).

## Model: CheXagent

The aim of CheXagent is a model that can "see" images $x_I$ and/or "read" text $x_T$ and generate "responses" $y$. Our training process for CheXagent involves the following stages: **Stage 0: Train a clinical LLM**: Our starting point is Mistral-7B-v0.1 (Jiang et al. 2023) due to its proven robust reasoning abilities in diverse benchmarks. To infuse the model with comprehensive medical and clinical knowledge, we utilize five distinct text sources for training: (i) PMC articles, (ii) MIMIC-IV, and (iii) Wikipedia. **Stage 1: Train a vision encoder for CXR**: Our model architecture reflects that of (Li et al. 2023c). For training purposes, we utilize datasets comprising image-text pairs, specifically from MIMIC-CXR, PadChest, and BIMCV-COVID-19. **Stage 2: Train a vision-language bridger**: Following the training of the clinical LLM and the CXR vision encoder, we focus on developing a bridger model, $\mathcal{M}_b$. This model is designed to map visual data to the corresponding language (semantic) space. For training $\mathcal{M}_b$, we employ the same datasets as in Stage 1. **Stage 3: Instruction tuning**: Upon completing Stage 2, The model is trained on CheXinstruct, taking into account two primary factors: (i) reserving certain task-dataset pairs exclusively for evaluation purposes, and (ii) determining optimal dataset ratios to ensure balanced training across different capabilities.

## Evaluation: CheXbench

CheXbench is structured with two evaluation axes, crafted to assess crucial aspects of CXR interpretation: image perception and textual understanding. For the former, we introduce 6 tasks across 7 datasets: View Classification, Binary Disease Classification, Single Disease Identification, Multi-Disease Identification, Visual-Question-Answering, and Image-Text Reasoning; For the latter, we introduce 2 tasks: Findings Section Generation and Findings Summarization.

In our study, we employ CheXbench to compare CheXagent against two general-domain instruction-tuned FMs, InstructBLIP and BLIP2, which achieve state-of-the-art performance in previous research (Li et al. 2023a). Additionally, we compare CheXagent with four medical FMs: XrayGPT, MedFlamingo, RadFM, and LLaVA-Med (Thawkar et al. 2023; Moor et al. 2023; Li et al. 2023b; Wu et al. 2023). This comparison aims to provide a comprehensive understanding of CheXagent's performance in relation to both general and medical-specific models.

Table 1 provides results on CheXbench. For image perception tasks, CheXagent demonstrates superior performance across image perception tasks, achieving an average improvement of 97.5% over general-domain FMs and an average improvement of 55.7% over medical FMs; For text understanding tasks, CheXagent outperforms all medical FMs on CheXpert on findings section generation and also achieve promising performance on findings summarization.

## Conclusion

In this work, we design a complete scheme for training CXR FMs by introducing CheXisnstruct, CheXagent, and CheXbench. Experimental results demonstrate the effectiveness of this scheme.

# References

Bae, S.; Kyung, D.; Ryu, J.; Cho, E.; Lee, G.; Kweon, S.; Oh, J.; Ji, L.; Chang, E. I.; Kim, T.; et al. 2023. EHRXQA: A Multi-Modal Question Answering Dataset for Electronic Health Records with Chest X-ray Images. *arXiv preprint arXiv:2310.18652*.

Bannur, S.; Hyland, S.; Liu, Q.; Pérez-García, F.; Ilse, M.; de Castro, D. C.; Boecking, B.; Sharma, H.; Bouzid, K.; Schwaighofer, A.; et al. 2023. MS-CXR-T: Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing.

Ben Abacha, A.; Hasan, S. A.; Datla, V. V.; Demner-Fushman, D.; and Müller, H. 2019. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019.

Boecking, B.; Usuyama, N.; Bannur, S.; Castro, D. C.; Schwaighofer, A.; Hyland, S.; Wetscherek, M.; Naumann, T.; Nori, A.; Alvarez-Valle, J.; et al. 2022. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, 1–21. Springer.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

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

Chen, Z.; Varma, M.; Wan, X.; Langlotz, C.; and Delbrouck, J.-B. 2023. Toward Expanding the Scope of Radiology Report Summarization to Multiple Anatomies and Modalities. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 469–484. Toronto, Canada: Association for Computational Linguistics.

Demner-Fushman, D.; Antani, S.; Simpson, M.; and Thoma, G. R. 2012. Design and development of a multimodal biomedical information retrieval system. *Journal of Computing Science and Engineering*, 6(2): 168–177.

Feng, S.; Azzollini, D.; Kim, J. S.; Jin, C.-K.; Gordon, S. P.; Yeoh, J.; Kim, E.; Han, M.; Lee, A.; Patel, A.; et al. 2021. Curation of the candid-ptx dataset with free-text reports. *Radiology: Artificial Intelligence*, 3(6): e210136.

Holste, G.; Wang, S.; Jaiswal, A.; Yang, Y.; Lin, M.; Peng, Y.; and Wang, A. 2023. CXR-LT: Multi-Label Long-Tailed Classification on Chest X-Rays. *PhysioNet*.

Hu, X.; Gu, L.; An, Q.; Zhang, M.; Liu, L.; Kobayashi, K.; Harada, T.; Summers, R. M.; and Zhu, Y. 2023. Expert Knowledge-Aware Image Difference Graph Representation Learning for Difference-Aware Medical Visual Question Answering. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4156–4165.

Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 590–597.

Jaeger, S.; Candemir, S.; Antani, S.; Wáng, Y.-X. J.; Lu, P.-X.; and Thoma, G. 2014. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6): 475.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Johnson, A. E.; Pollard, T. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Peng, Y.; Lu, Z.; Mark, R. G.; Berkowitz, S. J.; and Horng, S. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Kayser, M.; Emde, C.; Camburu, O.-M.; Parsons, G.; Papiez, B.; and Lukasiewicz, T. 2022. Explaining chest x-ray pathologies in natural language. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 701–713. Springer.

Li, B.; Wang, R.; Wang, G.; Ge, Y.; Ge, Y.; and Shan, Y. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023b. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023c. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Miura, Y.; Zhang, Y.; Tsai, E.; Langlotz, C.; and Jurafsky, D. 2021. Improving Factual Completeness and Consistency of Image-to-Text Radiology Report Generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5288–5304.

Moor, M.; Huang, Q.; Wu, S.; Yasunaga, M.; Zakka, C.; Dalmia, Y.; Reis, E. P.; Rajpurkar, P.; and Leskovec, J. 2023. Med-flamingo: a multimodal medical few-shot learner. *arXiv preprint arXiv:2307.15189*.

Nguyen, H. Q.; Lam, K.; Le, L. T.; Pham, H. H.; Tran, D. Q.; Nguyen, D. B.; Le, D. D.; Pham, C. M.; Tong, H. T.; Dinh, D. H.; et al. 2022. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *Scientific Data*, 9(1): 429.

Pavlova, M.; Tuinstra, T.; Aboutalebi, H.; Zhao, A.; Gunraj, H.; and Wong, A. 2022. COVIDx CXR-3: a Large-Scale, open-source Benchmark dataset of chest X-ray images for computer-aided COVID-19 Diagnostics. *arXiv preprint arXiv:2206.03671*.

Pelka, O.; Koitka, S.; Rückert, J.; Nensa, F.; and Friedrich, C. M. 2018. Radiology Objects in COntext (ROCO): a multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, 180–189. Springer.

Pellegrini, C.; Keicher, M.; Özsoy, E.; and Navab, N. 2023. Rad-ReStruct: A Novel VQA Benchmark and Method for Structured Radiology Reporting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 409–419. Springer.

Pham, H. H.; Tran, T. T.; and Nguyen, H. Q. 2022. VinDr-PCXR: An open, large-scale pediatric chest X-ray dataset for interpretation of common thoracic diseases. *PhysioNet*.

Reis, E. P.; de Paiva, J. P.; da Silva, M. C.; Ribeiro, G. A.; Paiva, V. F.; Bulgarelli, L.; Lee, H. M.; Santos, P. V.; Brito, V. M.; Amaral, L. T.; et al. 2022. BRAX, Brazilian labeled chest x-ray dataset. *Scientific Data*, 9(1): 487.

Shih, G.; Wu, C. C.; Halabi, S. S.; Kohli, M. D.; Prevedello, L. M.; Cook, T. S.; Sharma, A.; Amorosa, J. K.; Arteaga, V.; Galperin-Aizenberg, M.; et al. 2019. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1): e180041.

Thawkar, O.; Shaker, A.; Mullappilly, S. S.; Cholakkal, H.; Anwer, R. M.; Khan, S.; Laaksonen, J.; and Khan, F. S. 2023. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*.

Vayá, M. D. L. I.; Saborit, J. M.; Montell, J. A.; Pertusa, A.; Bustos, A.; Cazorla, M.; Galant, J.; Barber, X.; Orozco-Beltrán, D.; García-García, F.; et al. 2020. BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients. *arXiv preprint arXiv:2006.01174*.

Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; and Summers, R. M. 2017. 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*, 2097–2106.

Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*.

Xue, Z.; Candemir, S.; Antani, S.; Long, L. R.; Jaeger, S.; Demner-Fushman, D.; and Thoma, G. R. 2015. Foreign object detection in chest X-rays. In *2015 IEEE international conference on bioinformatics and biomedicine (BIBM)*, 956–961. IEEE.

Yu, F.; Endo, M.; Krishnan, R.; Pan, I.; Tsai, A.; Reis, E. P.; Fonseca, E. K. U. N.; Lee, H. M. H.; Abad, Z. S. H.; Ng, A. Y.; et al. 2023. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9).

Zhang, X.; Wu, C.; Zhao, Z.; Lin, W.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.