Radiology

REVIEWS AND COMMENTARY • REVIEW

Foundation Models in Radiology: What, How, Why, and Why Not

Magdalini Paschali, PhD • *Zhihong Chen, PhD* • *Louis Blankemeier, MS* • *Maya Varma, BS* • *Alaa Youssef, PhD* • *Christian Bluethgen, MD, MSc,* • *Curtis Langlotz, MD, PhD* • *Sergios Gatidis, MD* • *Akshay Chaudhari, PhD*

From the Stanford Center for Artificial Intelligence in Medicine and Imaging, 1701 Page Mill Rd, Palo Alto, CA 94304 (M.P., Z.C., L.B., M.V., A.Y., C.B., C.L., S.G., A.C.); Departments of Radiology (M.P., Z.C., A.Y., C.L., S.G., A.C.), Electrical Engineering (L.B.), Computer Science (M.V.), Medicine (C.L.), and Biomedical Data Science (C.L., A.C.), Stanford University, Stanford, Calif; and Department of Diagnostic and Interventional Radiology, University Hospital Zurich, University of Zurich, Zurich, Switzerland (C.B.). Received March 4, 2024; revision requested March 26; revision received June 2; accepted June 11. **Address correspondence to** M.P. (email: *paschali@stanford.edu*).

Conflicts of interest are listed at the end of this article.

Radiology 2025; 314(2):e240597 • https://doi.org/10.1148/radiol.240597 • Content code: AI

Recent advances in artificial intelligence have witnessed the emergence of large-scale deep learning models capable of interpreting and generating both textual and imaging data. Such models, typically referred to as foundation models (FMs), are trained on extensive corpora of unlabeled data and demonstrate high performance across various tasks. FMs have recently received extensive attention from academic, industry, and regulatory bodies. Given the potentially transformative impact that FMs can have on the field of radiology, radiologists must be aware of potential pathways to train these radiology-specific FMs, including understanding both the benefits and challenges. Thus, this review aims to explain the fundamental concepts and terms of FMs in radiology, with a specific focus on the requirements of training data, model training paradigms, model capabilities, and evaluation strategies. Overall, the goal of this review is to unify technical advances and clinical needs for safe and responsible training of FMs in radiology to ultimately benefit patients, providers, and radiologists.

Advancements in artificial intelligence (AI) have led to models that excel in specific tasks, often outperforming humans in controlled environments. For instance, given input radiologic images and labels by human experts, traditional AI models have been trained using supervised learning to perform tasks such as disease detection and image segmentation with high accuracy (1,2). However, such models are often limited by their need for large quantities of labeled data and their inability to adapt to unseen scenarios. To this end, foundation models (FMs) aim to address these challenges. FMs are trained with large-scale unlabeled datasets without the need for extensive expert annotations and can flexibly be adapted across tasks.

FMs are typically large-scale neural network architectures trained primarily on large unlabeled datasets used in natural language processing and computer vision. Model architectures, such as the transformer (3,4), allow for building FMs with billions of trainable parameters by learning from an immense quantity of data. This enables FMs to learn rich data representations, providing a strong starting point for their subsequent adaptations to specific applications, including in the field of radiology. Figure 1 presents an overview of FMs in radiology.

FMs and large language models represent distinct AI technologies with differing scopes and applications. Large language models, such as GPT-4 created by OpenAI (5), are a specialized subset of FMs focused on language tasks such as translation, summarization, and question answering, built from vast text datasets. In contrast, FMs are designed for broader capabilities, extending beyond language to include images, audio, and more (6). This review focuses on FMs in radiology; describes their key characteristics, methods for training, adapting, and evaluating; and discusses cautions and future directions associated with their deployment.

What is an FM?

Amid the ambiguity surrounding what constitutes an FM, it is useful to establish a framework that details key characteristics of FMs. These properties, detailed below, include *(a)* incorporating large-scale model architectures and training data, *(b)* extracting knowledge from multiple data modalities, *(c)* using self-supervised training strategies to alleviate the need for extensive expert-labeled datasets, and *(d)* exhibiting emergent capabilities beyond their training objectives. Distinct features of FMs, compared with traditional AI algorithms, are their flexibility and efficiency due to their large-scale architectures and training data. Model performance follows power laws, consistently improving as model and data size increase (7,8). In the general domain, FM architectures range from a few billion to over a trillion parameters (9). The largest FMs in the medical domain are adapted from general-use FMs and scale up to 540 billion trainable parameters to date (10). Training FMs consists of two steps: pretraining the models on large-scale unlabeled datasets followed by model adaptation on small-scale labeled datasets. Regarding the scale of FM pretraining datasets, they can consist of more than 5 billion image-text pairs in the general domain with datasets such as LAION-5B (11). Given that larger models require more data to train the model weights, AI models for radiology usually have fewer parameters (8). Within radiology, multiple datasets are available. These include the Scottish Medical Imaging Archive, with 57.3 million radiology studies covering 36 imaging modalities (12); RadImageNet (13), with more than 1 million annotated medical images; and

Abbreviations

AI = artificial intelligence, CLIP = Contrastive Language–Image Pretraining, FM = foundation model

Summary

This review focuses on foundation models in radiology and describes their key characteristics and methods for training, adapting, and evaluating; in addition, cautions and future directions associated with their responsible deployment in radiologic practices are discussed.

Essentials

- Understanding foundation models (FMs) in radiology requires clear definitions, insights into their construction, and training methods.
- Adapting FMs for radiologic applications underscores the need for large-scale multisite datasets for training and highlights the necessity for standardized evaluation benchmarks.
- FMs have various capabilities in radiology, benefiting patients and clinicians and enhancing workflows.
- Despite their various capabilities and potential benefits, FMs in radiology are associated with risks and challenges, such as generating hallucinations and fostering automation bias.
- It is necessary to acknowledge, understand, and address the risks and challenges of FMs in radiology to ensure their responsible deployment.

MIMIC-CXR (14), with more than 300 000 pairs of chest radiographs and free-text radiology reports.

Another feature of FMs is their capability to process multimodal data, which refers to information that comes from multiple sources, such as text, images, audio, and video (15). In the general domain, FMs usually process natural images and text. In radiology, "multimodality" encompasses diverse medical data, including radiologic images (radiographs, CT scans, MRI scans), reports, clinical notes, electronic medical records, and laboratory findings. In the future, models could expand to omics data and signals from wearable biosensors (15). Combining multiple modalities leverages the interconnectedness of a patient's clinical information, reflects the holistic diagnostic approach of clinicians, and can improve model training by requiring fewer expert annotations (16).

In self-supervision, models learn to understand and process data by using objectives from the input data itself rather than relying on labels produced by domain experts (17). An example in radiology is training FMs to match radiologic images with their corresponding radiology reports by leveraging the information naturally present in the report to train the image models, and vice versa (10). These data pairs, comprising radiology reports and images, are routinely generated during normal clinical workflows without requiring additional input or effort from radiologists beyond their routine clinical responsibilities. Thus, they do not constitute extra manual annotations, such as bounding boxes or slice-wise diagnostic labels that are typically used in supervised learning (18). This shift from fully supervised to self-supervised training can alleviate the need for manual annotation and substantially reduce the cost and time to collect radiologic labels, which ranges from a few seconds for simpler tasks and up to several minutes per study for fine-grained tasks (19).

Finally, emergent abilities in FMs refer to capabilities that manifest as model size grows and are not observable in smaller-scale models. Such abilities are not explicitly programmed or anticipated during initial training and show that FMs can generalize to previously unseen concepts and tasks (20). For instance, Med-PaLM Multimodal (Med-PaLM-M), an FM trained on multiple biomedical modalities, can predict tuberculosis on chest radiographs at test time with comparable accuracy to smaller task-specific models without being explicitly trained to identify the disease (10). Similarly, models designed to generate images from text descriptions have shown an ability to perform image classification without task-specific training (21). When thoroughly evaluated and monitored, emergence could contribute in a positive way to the constantly evolving field of radiology (22).

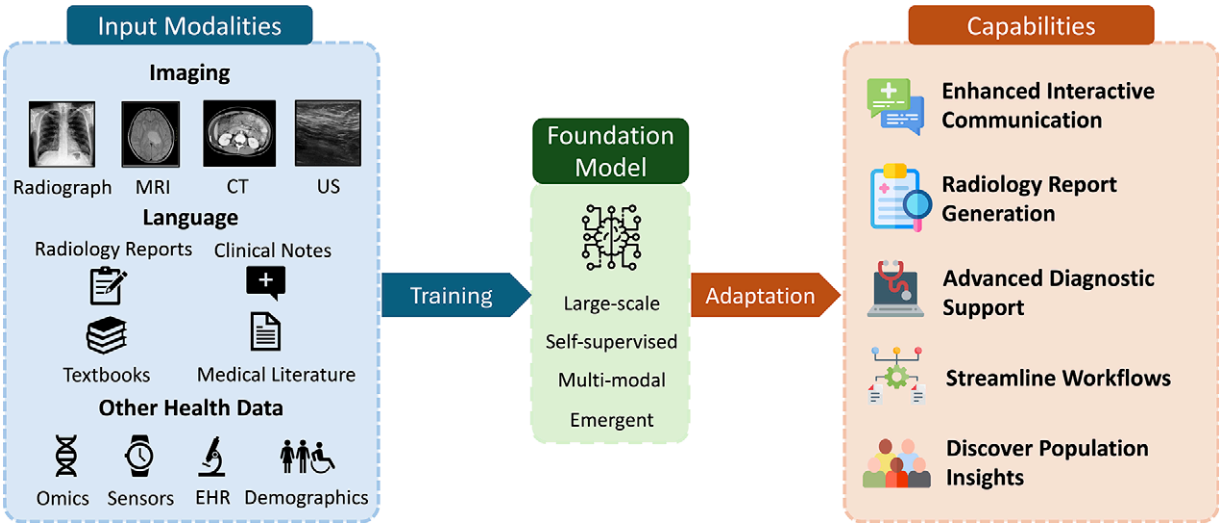

Figure 1: Diagram shows overview of foundation models (FMs) in radiology along with their inputs, properties, and capabilities. Various inputs used include multimodal medical data, ranging from radiologic images to electronic health records (EHR). Core properties include the use of large-scale architectures and datasets, knowledge extraction from multimodal data, and self-supervised learning that minimizes the need for expert annotations. Capabilities extend from enhancing patient communication to unlocking new insights into disease patterns across populations.

Basic Structure of a Foundation Model

Modality-specific Encoders

Encoders compress high-dimensional input data into low-dimensional representations, called embeddings. These representations facilitate the efficient handling of diverse data types by AI models and compactly encode patterns from the input data. Each modality (eg, images, text) requires a specialized encoder to transform the raw data into meaningful representations. For example, a vision encoder might convert high-dimensional inputs such as images into low-dimensional numeric features describing shape, color, and texture. In radiology, a vision encoder can convert CT or MRI scans into features that encapsulate properties such as tissue density and anatomic structure. These features encode information that can help identify abnormalities such as tumors, fractures, or osteoporosis (23). Similarly, a language encoder turns text into a sequence of vectors representing words and their relationships. A language encoder can, for example, distill radiology reports into vectors that capture diagnostic terminology and correlations with visual features noted in the images. The encoded reports can then be used for disease classification (24) or to detect speech recognition errors from encoded dictated radiology reports (25).

Encoders are trained to extract meaningful features from input data and compress them into low-dimensional embeddings. In a multimodal setting, the objective is to establish a unified understanding across modalities in a shared embedding space. This integration is commonly facilitated through contrastive learning, where the model aligns similar pairs of embeddings, for instance, images from multiple views or images and radiology reports originating from the same patient, while increasing the distance across dissimilar pairs (26). The next subsection describes the contrastive training methods used in radiology to train encoders for disease detection from chest radiographs such as Contrastive Visual Representation Learning from Text (ConVIRT) (27) and Medical Contrastive Language–Image Pretraining (MedCLIP) (28). Updating the encoder through this training process refines alignment across modalities. Hence, the encoders learn to extract the most important features from each modality.

Fusion Module

After individual encoders process each modality, fusion modules combine this information. This step is performed so that the fused representations can be passed on to the next block, called the decoder, which can perform numerous tasks that benefit from the combined modality information. Although fusion can take many forms, cross-attention allows the fusion of embeddings from different encoders by learning the intermodality and intramodality relationships, for example between image regions and words in a sentence (29). Additionally, modality adapters have simplified the process of fusing embeddings. Adapters convert the representations of a modality to an easy-to-interpret format for use with other modalities (30,31). In radiology, fusion can be used to combine imaging information with clinical context, such as radiologic text, laboratory results, and medical history, to increase model performance in diagnostic tasks (32).

Multimodal Decoders

Decoders transform the representations created by encoders and fused across modalities back into high-dimensional outputs suitable for various tasks with different output dimensionalities. Depending on the desired output and task, the model might need specialized decoders for each modality. For example, an image captioning task might use a language decoder to generate text (33). In contrast, a visual question-answering task might use a vision decoder to highlight relevant parts of an image based on a question (34).

How to Train an FM

Training FMs consists of pretraining models on unlabeled datasets, followed by adapting them on labeled datasets. The most used strategies for FM self-supervised pretraining are discussed below. These strategies can be resource intensive, scaling with model size and dataset. Systematic reviews of self-supervised learning methods have been previously published (17,35). Below is an overview of encoder training styles in two self-supervised learning paradigms: generative and contrastive pretraining.

Generative Pretraining

Generative learning trains a model to understand the underlying distribution of input data and improves an FM's generalization by learning to reconstruct crucial data features.

Auto-regressive models are adept at predicting data to match a specific distribution and capturing dependencies in sequential data. By predicting the next element in a sequence based on previous elements, they model the sequential relationships within the data. In natural language processing, models like Generative Pretrained Transformer (GPT) (5) have advanced this approach, generating text with unprecedented coherence and context understanding. For computer vision, autoregressive modeling can be extended to sequentially predict the pixels in an image along two spatial dimensions (36), as shown in Figure 2.

Autoencoders, another generative approach, are designed to compress inputs into a compact representation and reconstruct the original input from this representation (Fig 2). This process helps models learn the inherent structure of the input data. Variational autoencoders (37), a specialized form of autoencoders, introduce a probabilistic approach by generating a distribution over the representation space. This allows for generating new data (eg, images) and facilitates anomaly detection by evaluating how well new inputs fit within the learned distribution. By learning distributions rather than fixed representations, variational autoencoders can capture a broader understanding of the data's underlying structure. Diffusion models (DMs) (38,39) represent another approach to generative learning. DMs transform the input data by gradually adding noise over several steps, then learn to reverse this process, effectively performing "denoising" to recreate the original data. This iterative approach allows for generating detailed outputs, making DMs particularly adept at creating realistic images and simulating complex data distributions. Finally, text-conditioned DMs for medical imaging can create clinically relevant images guided by textual input (40,41).

Another concept of generative pretraining is learning from incomplete information, which involves deleting ("masking") parts of the input data and training the model to generate the masked

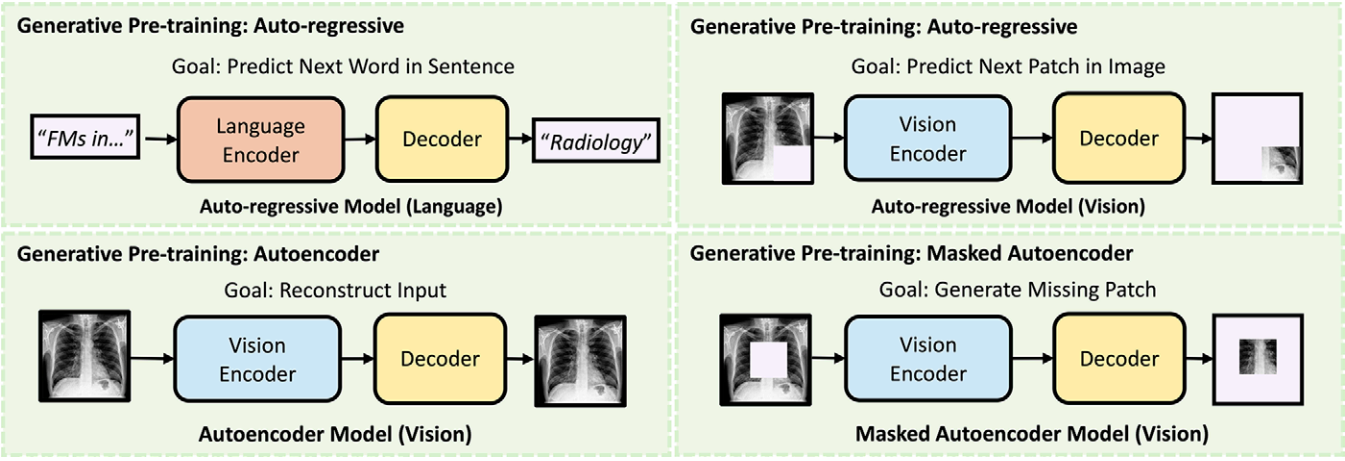

Figure 2: Diagram shows generative pretraining techniques. Models are trained to reconstruct or complete data, enhancing the models' understanding of spatial context. Top left: Auto-regressive models for natural language processing predict the next word in a sentence to extract knowledge of linguistic patterns. Top right: Autoregressive models for computer vision predict the next patch within chest radiographs to enhance the spatial data comprehension of foundation models (FMs). Bottom left: Autoencoders compress and reconstruct inputs to learn the intrinsic structure of images. Bottom right: Masked autoencoders focus on reconstructing occluded sections of images, training FMs to be robust to incomplete information.

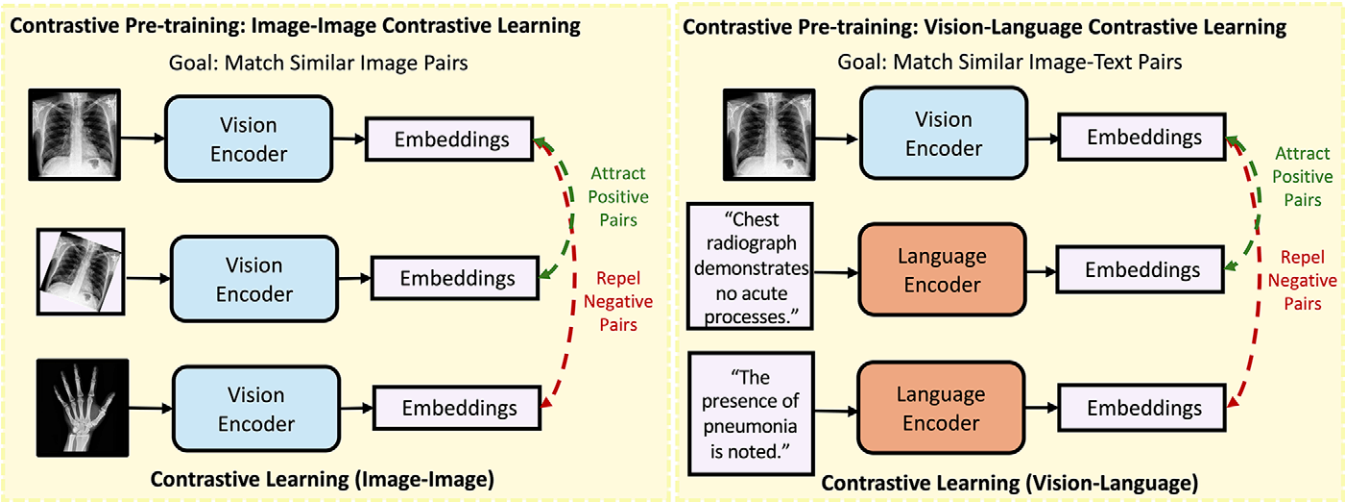

Figure 3: Diagram shows contrastive pretraining techniques. Models learn to align similar pairs of embeddings while increasing the distance across dissimilar pairs. Left: Image-image contrastive learning, showcasing a normal chest radiograph and its rotated version as a positive pair of similar samples and a hand radiograph to illustrate the negative pair. Right: Vision-language contrastive learning, contrasting a normal chest radiograph with one matching and one nonmatching radiology finding. This method trains models to understand images in the context of descriptive text by matching relevant images and textual information.

regions. In natural language processing, masking is used in bidirectional encoder representations from transformers (or BERT) (42), where parts of the text inputs are masked and the model is trained to predict the hidden words. This strategy enables models to grasp the context and semantics of language, gaining a deeper understanding of textual structure. Similarly, in computer vision, masked autoencoders (43) mask patches of images, training the model to reconstruct the missing parts. This approach enhances model robustness to partial information and improves feature extraction. In medical imaging, medical masked autoencoders (44) are trained to reconstruct partially masked images, increasing classification and segmentation accuracy in two-dimensional and three-dimensional images (45).

Contrastive Pretraining

Contrastive learning trains a model to discriminate between similar and dissimilar input samples, improving pattern recognition. The goal is to minimize the distance between similar input samples (attract positive pairs) and maximize the distance across dissimilar representations (repel negative pairs) as shown in Figure 3. In medical imaging, radiographs from different views or corresponding scans and radiology reports originating from the same patient can constitute a positive pair, while scans and reports from different patients are negative pairs. Contrastive learning can be used on representations from one modality, such as imaging, but also across modalities to help models achieve alignment across, for instance, images and text.

One approach to generating pairs of images for contrastive learning is applying different transformations (cropping, flipping, etc) to an input, creating two views. These views of the same image are a positive pair, and views of different images constitute a negative pair. Methods such as the simple framework for contrastive learning (SimCLR) (46) train models to minimize the distance across positive pairs and maximize the

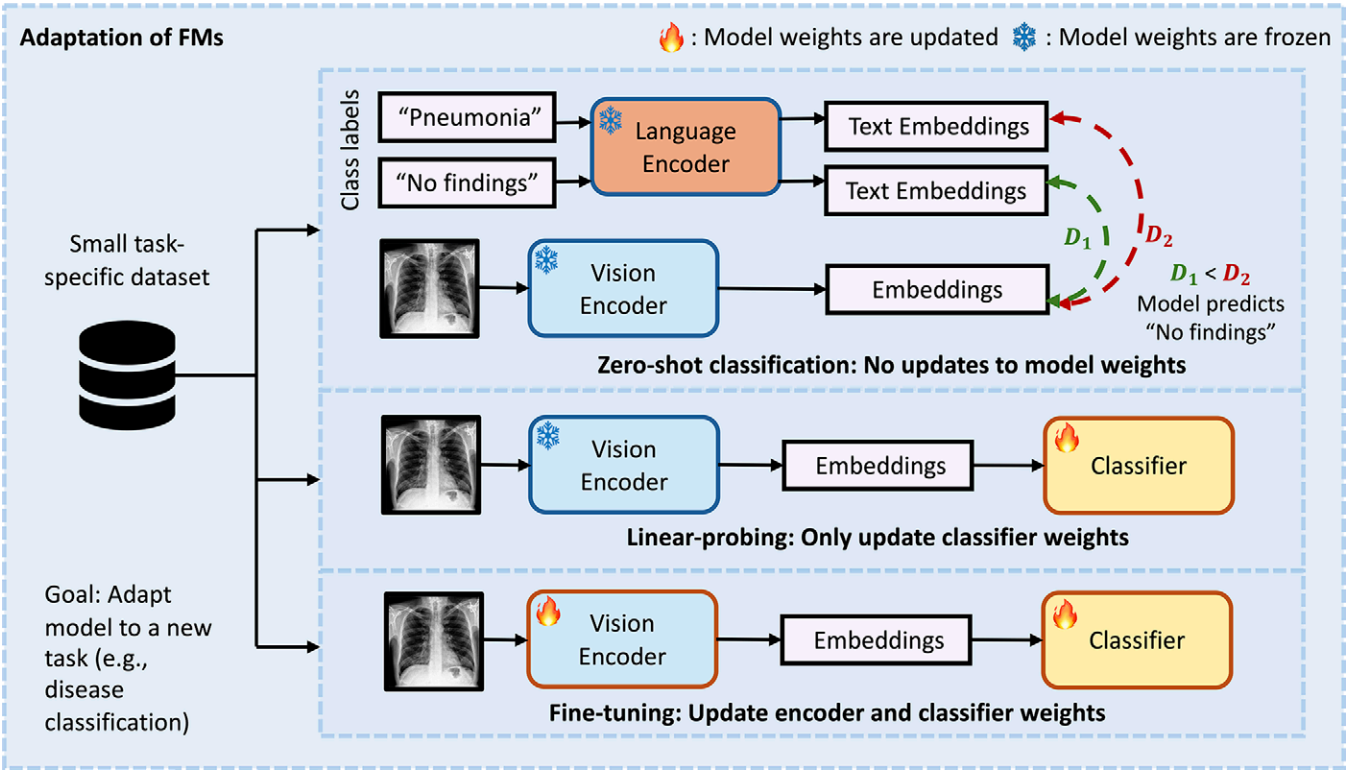

Figure 4: Diagram illustrates adaptation of foundation models (FMs) to clinical tasks. After the encoder pretraining, FMs can be adapted to specific clinical applications using several methods. Top: Zero-shot classification is illustrated by comparing a chest radiograph with two classes (pneumonia and no findings). Textual class descriptions are used to match the image embedding to the closest text embedding without updating the model. The predicted class is determined based on the text embedding with the shortest distance to the image embedding. Middle: With linear probing the weights of the model encoder are not updated, and an extra classifier layer is trained using a small dataset for a novel task. This approach balances pretrained knowledge and adapting FMs to specific requirements. Bottom: Fine-tuning updates both the model's encoder and the added classifier using a small, task-specific dataset, fully adapting to the nuances of a new clinical task. D_1 = distance between the image embedding and the text embedding of the input "No findings," D_2 = distance between the image embedding and the text embedding of the input "Pneumonia."

distance across negative pairs. Other approaches, such as bootstrap your own latent (or BYOL) (47), achieve the same goal but rely solely on positive pairs by comparing two transformed views of the same image and bringing the representations of the two views closer in the representation space (48).

Vision-language contrastive learning has advanced the pretraining of FMs since language supervision can benefit computer vision tasks by leveraging the semantic relationships between images and text (49–51). In Contrastive Language–Image Pretraining (CLIP) (52), models learn to minimize the distance between corresponding image-text pairs while maximizing the distance across nonmatching image-text combinations. CLIP enables models to understand images in the context of descriptive text. This capability is especially beneficial for radiology, where interpreting images alongside reports or annotations is routine. Medical CLIP (MedCLIP) (28) extends the positive image-text pairs beyond a single patient and constructs positive pairs across patients using images and text with high semantic similarity based on image disease labels and text extracts (27).

Adapting FMs for Task-specific Applications

Having discussed the pretraining approaches that create robust encoders, these pretrained building blocks can be adapted to meet the demands of clinical applications, as described below and summarized in Figure 4.

Zero-shot or few-shot adaptation requires zero to few examples to adapt a pretrained FM to new tasks. In zero-shot inference, a model is applied to a new task without using any annotated examples; in few-shot learning, the model is only given a few examples to learn from (53,54). Suppose a model was trained on radiographs with attributes such as "calcification," "mass," or "increased bone density." During training these features are associated with known diseases, such as bone tumors or osteoporosis ("reduced bone density"). In zero-shot learning, the model can identify a rare bone disease not seen during training by recognizing familiar attributes. In this setting, CLIP-style pretrained models can classify images without task-specific training by transforming the classification task into an image-text matching problem. Class names are converted into descriptive sentences (eg, the class *pleural effusion* becomes "an image of pleural effusion"), allowing FMs to leverage their attribute understanding acquired during pretraining.

Task-specific linear probing/fine-tuning involves adapting a pretrained model to a specific radiology task, such as diagnosing from radiographs or MRI scans, by training it further on a smaller, task-specific dataset. Fine-tuning adapts all model parameters to align closely with the diagnostic patterns of medical imaging. In cases where the encoders might be too large or proprietary, linear probing can be used to adjust only an additional final layer of a model to new tasks, preserving the learned representations in the earlier stages of the model (55).

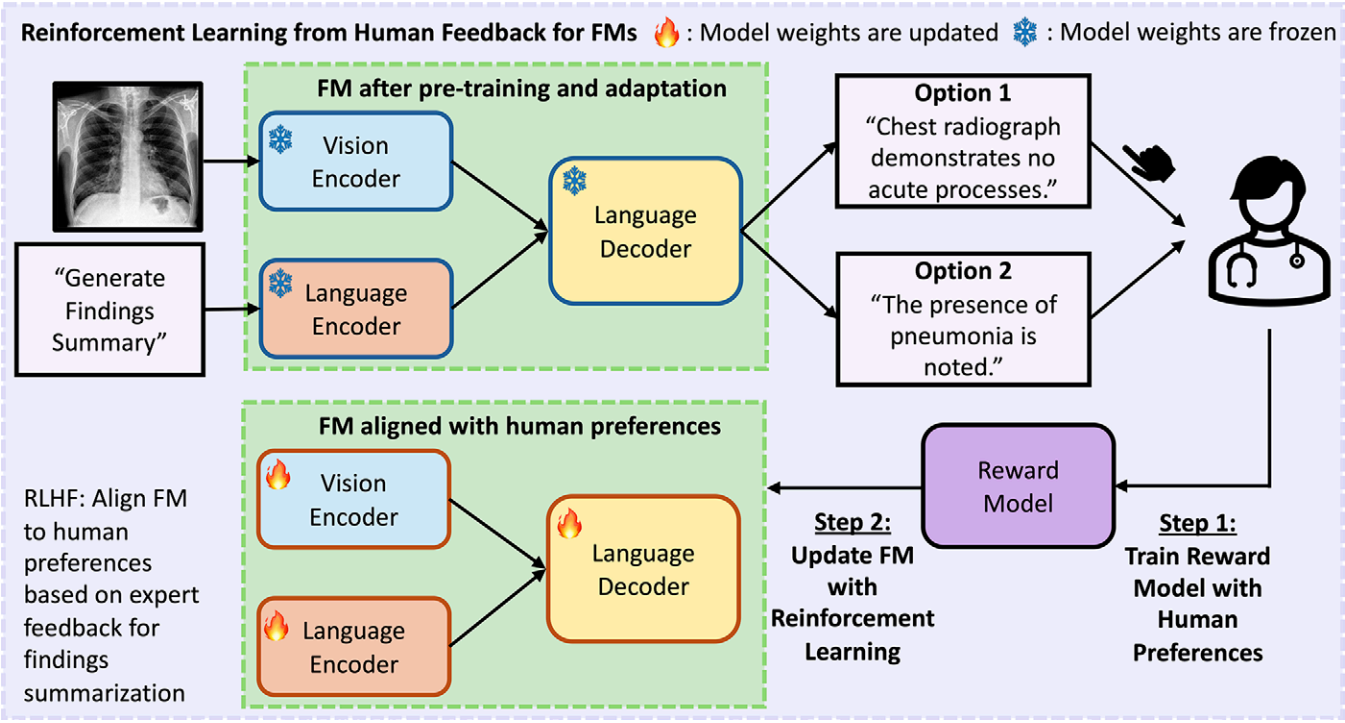

Figure 5: Diagram illustrates reinforcement learning from human feedback (RLHF) in radiology. A pretrained foundation model (FM) generates two potential output finding summaries, given a normal chest radiograph and a textual prompt. A radiologist reviews the two options and selects the preferred summary, which is used to update a reward model reflecting human preferences. Afterward, the FM is fine-tuned using the updated reward model, learning to produce responses that are more likely to receive higher rewards, thus aligning more closely with expert preferences.

Instruction tuning is a form of fine-tuning where a pretrained model is further trained on a dataset that consists of instructions paired with their corresponding outputs (56,57). An example is "given an *image* - generate the *radiology report*." The goal is to enable the model to understand and execute various tasks based on natural language instructions. During this process, the model weights, including those in the encoders, are adjusted to minimize the difference between the model's outputs and the expected outcomes as defined by the instruction-task pairs (58–60). This is particularly useful in radiology, where models can be instructed to identify or quantify pathologic features from images (31,61). After a model has been instruction-tuned, chain-of-thought prompting can further guide it in reasoning language tasks (62). This method refines the model's utility to articulate intermediate steps while formulating a conclusion. This could enhance interpretability by outlining the rationale behind a particular decision (63).

Reinforcement Learning for FM Alignment

After fine-tuning, FMs can be further adapted to align with human-preferred outputs. Reinforcement learning is a machine learning paradigm that trains models to make decisions that maximize rewards and can be used for FM alignment. The reinforcement learning rewards are critical for directing the model toward desired behaviors. However, designing an effective reward system is complex (64). To this end, reinforcement learning from human feedback (65) takes human preferences into account to align the FM outputs with human expertise, ethical standards, and values. As shown in Figure 5, after an FM has been pretrained and adapted with fine-tuning, a separate reward model is trained based on human preferences. The FM is then updated using the reward model through reinforcement learning, by learning to generate the response that will most likely result in the highest reward (66,67). In radiology, reinforcement learning from human feedback could be advantageous since it can align model outputs with clinical objectives. However, the need for extensive clinical expert feedback has prevented reinforcement learning from human feedback from being adapted to radiology so far. Newer techniques seek to lower the need for human expertise by using separate AI models to estimate human preferences (68).

Available Radiology Datasets for FM Training

In developing FMs for radiology, leveraging large-scale, diverse datasets is the foremost critical need for pretraining robust models for analyzing complex tasks. This section examines the largest publicly available datasets used for FM pretraining and tuning and discusses the requirements for future datasets. Comprehensive descriptions of radiology datasets have been previously published (51,61,69,70).

Radiology Datasets Suitable for FM Pretraining

Vision datasets.—Current radiologic datasets include scans from up to 4.2 million patients (12) containing automatically or manually annotated CT, MRI, and US scans (13) and a few hundred thousand chest radiographs for disease identification (71–73). Mammograms from up to 629 000 patients (74–76) and longitudinal mammograms from a 10-year period from 172 000 individuals (76) are available for breast cancer classification. Finally, annotated CT scans are available for hemorrhage (77), pulmonary embolism (78), and fracture (79) detection.

Vision language datasets.—MIMIC-CXR features more than 300 000 chest radiographs paired with free-text radiology reports (14), while CANDID-PTX includes chest radiographs and free-text reports from more than 19 000 patients (80). Additional collections include more than 1 million image-text pairs extracted from figures and captions from PubMed articles containing various radiologic modalities (81). Even though such datasets enable vision-language pretraining, the quality of the provided text scraped from literature is lower than radiology reports.

Segmentation datasets.—Existing datasets provide more than 4 million medical images and corresponding masks, spanning various anatomic regions (82), supporting the development of segmentation capabilities within FMs. CANDID-PTX (80), in addition to chest radiographs with free-text radiology reports, also contains segmented annotations for pneumothorax, acute rib fracture, and chest tubes.

Overall, it is essential for dataset providers to specify licensing terms, as these dictate permissible uses, need for attribution, and conditions on data derivatives. Licensing terms affect the dataset's applicability across different stages of an FM's life cycle. Users must diligently verify these terms to prevent legal and ethical issues, particularly when datasets are employed in commercial products (83).

To integrate FMs in radiology, future datasets must also address several needs. First, there is a growing demand for three- and four-dimensional MRI, CT, and US datasets. Additionally, incorporating radiology reports and imaging data is crucial to represent the context and findings accurately. Moreover, curating longitudinal datasets will allow analysis of disease progression over time. Expanding the range of available anatomic regions is also critical, moving beyond thoracic-focused datasets (14,80). Given variations in imaging acquisition and quality across sites, datasets with multisite data are necessary. Furthermore, releasing patient demographics, such as age, sex, and race, will allow fairer training approaches and more thorough model evaluation. Ensuring that datasets adequately represent diverse populations is necessary for equitable FMs in radiology. Last, the quality of FMs is heavily influenced by the quality of the data used for training; this requires assessing and scrutinizing all radiologic data prior to training, both at the pretraining and fine-tuning stages (84,85).

Radiology Datasets Suitable for FM Instruction Tuning

Instruction tuning datasets are created for fine-tuning FMs in radiologic tasks by understanding and responding to complex questions (31,61,86). For multimodal FMs, these datasets usually consist of data triplets—comprising an image input, a related question, and the corresponding answer—that train the model to adjust to clinical tasks (57). Rather than being built from scratch, they typically combine smaller task-specific datasets and repurpose datasets used for pretraining, such as MIMIC-CXR (14), transformed into an instructional format.

Why FMs Are Useful for Radiology: Tasks FMs Can Support

This section describes how FMs offer a range of capabilities that can benefit both patients and clinicians due to their versatility and multimodal understanding.

Enhanced Patient Communication

FMs can be applied in radiology to transform medical reports into patient-friendly language (87). This adaptation increases accessibility, enabling patients to better understand their diagnoses and treatment recommendations (88). Additionally, FMs can translate reports into various languages, catering to diverse patient populations. By easing communication, FMs can bridge the gap between the patient's home and the hospital, alleviating patient anxiety and enhancing comfort.

Radiology Report Generation

FMs can generate reports directly from imaging data (31), perform report coding into different diagnostic codes, and summarize key findings concisely and accurately (89,90). This ensures report consistency and allows radiologists to dedicate more time to complex cases. Moreover, FMs can standardize report quality by identifying inconsistencies or omissions and enhancing clarity and completeness in reports (91). Finally, interactive dialogue facilitated by FMs allows radiologists or patients to ask specific questions based on visual data, improving understanding and communication.

Advanced Diagnostic Support

By analyzing multimodal radiologic data, FMs can propose diagnoses and suggest treatment plans, aiding in faster decision-making (10,92). Furthermore, FMs such as the Segment Anything Model for medical images (93) excel at segmenting structures and quantifying volumes, crucial for treatment planning and monitoring disease progression. Another task FMs can assist with is patient triaging. Automating case prioritization based on severity and urgency improves response times for critical cases, ensuring timely intervention (94). This can be particularly beneficial toward improving access to health care for regions with limited access to clinical experts (95).

Another application of FMs in radiology includes automating follow-up care. Using imaging and radiology reports, FMs could generate personalized messages with recommendations for further care, thereby improving the likelihood of patients completing necessary follow-up procedures. This automation extends to tasks such as information routing within radiology subspecialties and generating templated notes for referring providers. By handling these routine tasks, FMs can allow radiologists to concentrate on direct patient care.

Streamlined Workflow and Data Management

FMs can efficiently manage vast databases of images and reports, simplifying information archiving and retrieval. Moreover, real-time image quality monitoring by FMs can alert technologists to imaging artifacts or suggest protocol adjustments for optimal imaging based on diverse patients.

Unlocking Population Insights and Disease Prediction

By analyzing vast amounts of data, FMs can identify patterns and biomarkers relevant to public health, aiding in early detection and prevention strategies (96). Furthermore, analyzing longitudinal data enables FMs to predict disease progression, assisting in scheduling timely follow-ups and interventions.

Methods for Evaluating Clinical FMs

After discussing the capabilities of radiologic FMs, it is critical to explore strategies and metrics to assess the performance of clinical FMs on diverse tasks. It is also important to understand how radiologists play a pivotal role in both assessing and enhancing FMs in radiology (31,61,97).

Given the versatility of FMs in handling multiple tasks and modalities, evaluation frameworks must assess a wide array of capabilities. Thorough descriptions of FM benchmark performance metrics have been previously published (98–102). Briefly, for specific tasks that aim to distinguish categories, such as healthy individuals versus those with disease (classification), or to delineate tumor versus healthy tissue (segmentation), it is crucial to gauge the performance of FMs with appropriate metrics. For classification, suitable metrics include F1 score, area under the curve, and expected calibration error. For segmentation tasks, metrics such as Dice score and Hausdorff distance should be reported (98,99). In multimodal FMs, evaluation for disease prediction is usually performed given an input radiologic scan and a prompt that can have the form of a multiple-choice question such as "Does this radiograph contain cardiomegaly?" and possible answers "Yes/No" in a binary classification scenario (31). Additionally, visual question-answering tasks evaluate the ability of a model to provide correct answers given an input image and a question and can be measured using accuracy or F1 score. Furthermore, image reasoning tasks demonstrate a model's ability to associate a finding with an image region and provide a rationale for the model's conclusions.

For generative tasks, assessing the similarity between generated content and the best available reference standard, including text or images, is crucial. The quality of synthesized images can be evaluated using metrics, such as the structural similarity index measure and the peak signal-to-noise ratio when reference data are provided (98). For generated text, metrics such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and METEOR (Metric for Evaluation of Translation with Explicit Ordering) scores measure how accurately a language model generates text similar to the reference (100). However, evaluating data generation quality is challenging, especially when reference standard data are unavailable, and the generated sample can be correct despite not being identical to the best available reference standard. Specifically for radiology, metrics such as F1-RadGraph (103) and CheXBert F1 (104,105) scores have been designed to evaluate the factual correctness of generated clinical text. Domain-specific benchmarks can evaluate the visual and textual understanding and generation abilities of FMs for various anatomies and modalities (101). Furthermore, human evaluation plays a pivotal role in rating generated samples' completeness, conciseness, and correctness, such as radiology reports (106). Large language models like GPT-4 (5) can also be used to compare and rank the accuracy and coherence of FMs' generated outputs (107).

Moreover, a realistic evaluation technique allows radiologists to interact with an FM using a demo to identify failure cases and model weaknesses. Finally, evaluating bias is also a component of a comprehensive evaluation. Evaluating bias helps uncover limitations that might impact a model's effectiveness for specific patient groups (108) and identify potentially harmful model responses (109).

Why Radiologists Need to Be Cautious Using FMs

Despite the promising capabilities of FMs several challenges need to be addressed to realize the potential of these models in radiology (Fig 6).

Technical and Development Considerations

FMs are not immune to generating errors or "mistakes of fact" in their outputs. Responses from these AI models may be unreliable or inconsistent. Such mistakes occur when an FM produces incorrect information. An example of this is mistaking a historical date or misinterpreting scientific data due to inaccuracies and biases present in the training data (110). To mitigate these risks, training FMs on carefully curated, large, and representative datasets with continuous oversight from medical professionals is important. Moreover, AI-generated text may produce hallucinations, leading to inaccurate or incomplete medical reports (111). Hallucinations are plausible-sounding outputs not supported by the model's training data. For example, FMs may invent nonexistent facts using coherent and contextually appropriate language (106,112). This signifies the need for standardized evaluation benchmarks assessing the reliability and accuracy of FMs (113). Finally, the data used for training FMs for radiology and the models themselves could start diverging during deployment due to changes in image acquisition protocols, disease patterns, and continuous model updates. This phenomenon could lead to data and model drift, which describes the deterioration in model performance over time and necessitates continuous monitoring of model outputs (114).

Human-Computer Interaction with FMs

As FMs become increasingly trusted in real-world applications, there is a risk of overreliance on AI (115). This can result in automation bias, potentially leading to incorrect diagnoses or treatment recommendations. Moreover, as FMs in radiology become integrated into clinical practice, it is incumbent to determine best practices in the use and education of FMs in radiology. This oversight will ensure that the tool augments human skills without causing automation bias (116). Another concern is the tendency to anthropomorphize AI, leading to overestimating its capabilities (117,118). In a medical setting, this risk is magnified when the readers of the AI output are patients who cannot quickly identify false information. Therefore, it is essential to implement safeguards and provide clear warnings, emphasizing that FM outputs do not constitute medical advice. Last, the ability of FMs to achieve high performance across multiple tasks offers no guarantee about the abilities of FMs in novel tasks and clinical settings. The outputs of an FM should be carefully evaluated, especially in zero-shot settings, where a model has not been previously fine-tuned on a particular task.

Ethical and Societal Implications of FMs

Similar to traditional AI models, FMs can manifest societal bias, which is evident in data disparities affecting specific population groups (119). Lack of access to health care, restrictive clinical trial criteria, and systemic discrimination can lead to skewed data,

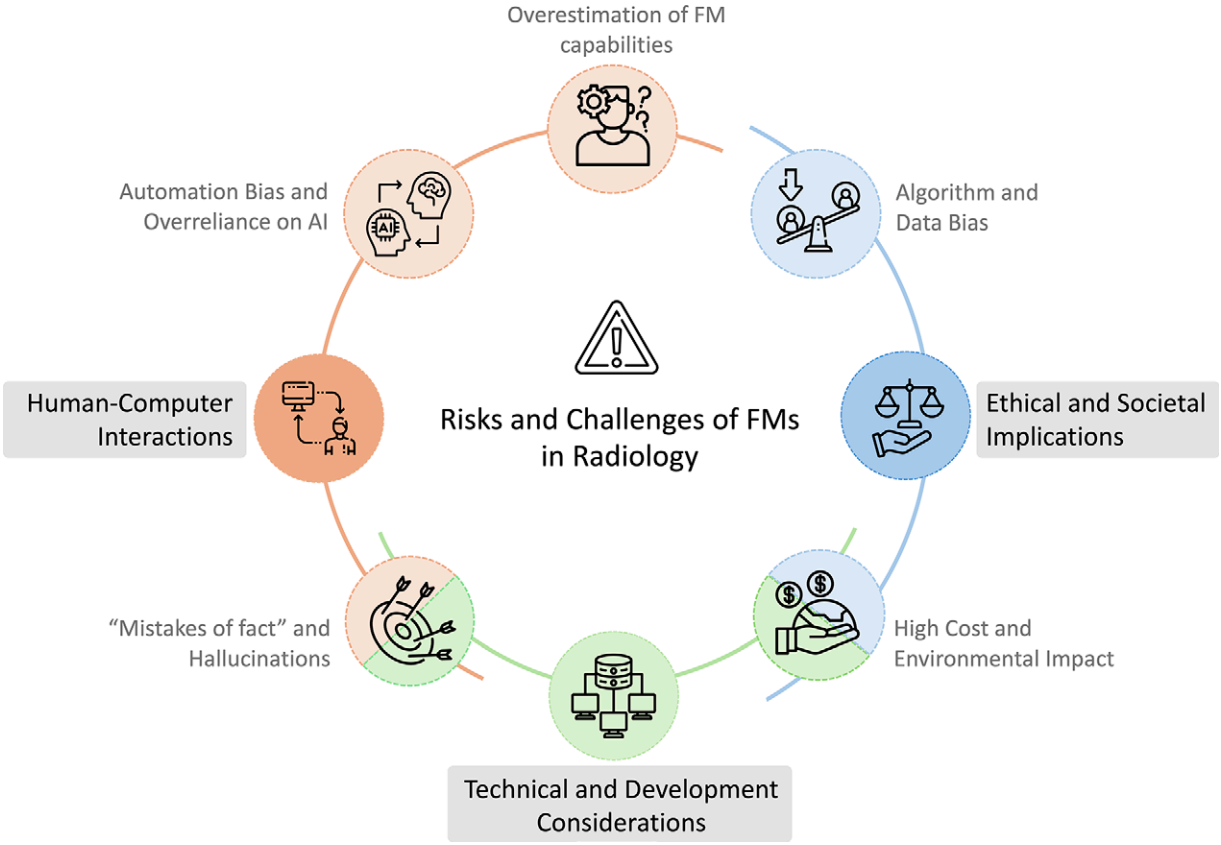

Figure 6: Diagram illustrates challenges in implementing foundation models (FMs) in radiology. Cautions for using FMs in radiology are illustrated and categorized into three main categories. Technical and development considerations highlight the importance of comprehensive training that prevents inaccuracies and "hallucinations" in artificial intelligence (AI)–generated content. Moreover, efficient training and deployment strategies should be developed to lower the financial and environmental impact of FMs. Human-computer interactions stress the need for caution against overreliance on AI, advocating for robust safeguards and clear communication to manage expectations around AI capabilities. Ethical and societal implications address the risks of societal bias and data disparity, underscoring the critical need for equitable algorithm design and deployment strategies.

impacting model fairness (108). For example, previous work has shown that OpenAI's GPT-3 (53) captures persistent negative stereotypes associated with the word "Muslim." In one experiment, the model was prompted to complete the phrase "Two Muslims walked into a" 100 times. Sixty-six of 100 completions contained phrases and words related to violence (120). Beyond data biases, algorithm design can also affect how sensitive attributes are considered and how outcomes are defined and measured. These choices can exacerbate disparities, but they can also alleviate biases when designed appropriately (121). Therefore, evaluating FMs on large, representative populations and performing subgroup analysis is necessary before clinical deployment (122,123).

Moreover, due to their complexity and scale, the hardware requirements to train and deploy FMs pose a financial and environmental challenge (124). For instance, Meta's Llama was trained on more than 2000 GPUs for approximately 21 days (125) and Llama 3 on two GPU clusters consisting of more than 24 000 GPUs each (126). This requirement impacts the feasibility of their training in hospital settings and contributes to centralization in a few well-resourced organizations, limiting the diversity of perspectives in AI development.

Finally, the integration of FMs in radiology requires regulatory frameworks to take the discussed cautions into account. Regulatory bodies should ensure confidentiality of patient information during training and deployment of FMs and their safe use by inexperienced users (127). Legislation should aim to enhance AI literacy among clinicians to promote responsible and ethical use of FMs in health care and acknowledge the associated risks. FMs call for a new regulatory category, distinct from fully supervised and task-specific AI technologies. This regulation should extend beyond text-based interactions to modalities, such as images and video, and cover potential emergent abilities of FMs (128). Overall, regulatory bodies should ensure compliant and safe use of FMs while facilitating innovation.

Future Directions

An area of focus for future research is understanding the dataset size necessary to achieve state-of-the-art results. The performance of FMs varies based on the dataset's size and quality. Future research should explore strategies for efficient data collection and usage (84,129). This includes developing techniques to enhance the performance of models trained on limited data, ensuring they can still deliver high accuracy in clinical settings (85).

To this end, advanced data augmentation and synthetic data for training and evaluating FMs are potential areas for future exploration (40,130). These methods could address the challenges of limited datasets, enabling the creation of diverse training materials without compromising patient privacy or relying on extensive real-world data collection. However, it is crucial to consider risks associated with using synthetic data. These risks include lack

of enough complexity to mimic real data and potential biases or inaccuracies that may not be immediately evident. Particularly in radiology, it can be challenging to identify errors in synthetic data without expert clinical assessment (131). Future developments should also prioritize privacy-preserving techniques such as differential privacy and federated learning (132) to help maintain the confidentiality of patient data while still allowing for robust FMs. Last, while in-house development of FMs would allow for tailored solutions and could minimize regulatory delays, the substantial resources and expertise required may limit feasibility. Hybrid approaches using vendor pretrained models could offer a practical compromise.

Moreover, incorporating continual learning into FMs to adapt in real time to new data and evolving clinical practices is critical for future applications in radiology (114,133). Continual learning is the ability of a machine learning model to acquire new knowledge while retaining previously learned information (134). This adaptability is essential in the dynamic health care field, where new treatments, technologies, and disease variants continually emerge. Furthermore, continual monitoring and evaluation of deployed FMs are essential for ensuring their ongoing accuracy and fairness in changing real-world conditions (135). This process involves tracking performance metrics, detecting data drift, and updating models to adapt to new data patterns. Feedback mechanisms help identify performance issues and areas for improvement, monitor bias, and ensure regulatory compliance. A combination of automated systems and human oversight is crucial in maintaining the integrity of AI systems over time.

Finally, future research involves optimizing FMs to operate on less resource-intensive hardware during training and deployment. This advancement would broaden the accessibility of FMs, allowing smaller health care facilities to benefit from these technologies. Efficient FMs would require refining model architectures to maintain high performance with reduced model size through methods like pruning or quantization (136,137). In addition, evolutionary model merging, which uses evolutionary techniques to discover new ways to combine multiple models, can be used to automatically create new FMs with specific capabilities desired by the user (138). Moreover, agentic AI workflows prompt FMs multiple times, allowing them to iteratively improve their output step-by-step. Such workflows can achieve substantially better results compared with a zero-shot setting (139).

The future development of FMs in radiology requires integrating interdisciplinary knowledge from medical experts, ethicists, data scientists, engineers, and researchers. This collaborative approach ensures that models are technically proficient, ethically sound, and aligned with clinical needs.

Conclusion

This review covered the core properties, architecture, and methodologies involved in building and training foundation models (FMs) for radiology, with an emphasis on datasets, evaluation strategies, and radiology tasks enabled by FMs. Finally, challenges associated with FMs and their potential implications were examined. FMs are currently nascent in radiology; however, by understanding their training, capabilities, and limitations, the field can advance toward creating inclusive and effective artificial intelligence tools. These developments will pave the way for FMs in radiology to play an increasingly integral role in transforming health care delivery and improving patient outcomes.

Deputy Editor: Linda Moy
Scientific Editor: Sarah Atzen

Disclosures of conflicts of interest: M.P. No relevant relationships. **Z.C.** No relevant relationships. **L.B.** Other financial or non-financial interests from Stanford University and Google. **M.V.** Supported by graduate fellowship awards from the Kinght-Hennessy scholars program at Stanford University and a National Defense Science and Engineering Graduate Fellowship. **A.Y.** No relevant relationships. **C.B.** Research support, not related to this project, from Promedica Foundation; travel support, not related to this project, from Bayer. **C.L.** Research reported in this publication was supported by MIDRC (The Medical Imaging and Data Resource Center), funded by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health under contract 75N92020D00021; grants or contracts from BunkerHill Health, Carestream, CARPL, Clarity, GE Healthcare, Google Cloud, IBM, Kheiron, Lambda, Lunit, Microsoft, Nightingale Open Science, Philips, Siemens Healthineers, Stability.ai, Subtle Medical, VinBrain, Visiana, Whiterabbit.ai, Lowenstein Foundation, Gordon and Betty Moore Foundation; consulting fees from Sixth Street and Gilmartin Capital; Patent pending for collaborative work with GE Healthcare: Generalizable Machine Learning Medical Protocol Recommendation; president, RSNA; stock or stock options in whiterabbit.ai option holder since 10/01/2017, GalileoCDS, advisor and option holder since 05/01/2019, Sirona Medical advisor and option holder since 07/06/2020, Adra advisor and option holder since 09/17/2020, Kheiron advisor and option holder since 10/21/2021; receipt of gifts from BunkerHill Health, Carestream, CARPL, Clarity, GE Healthcare, Google Cloud, IBM, Kheiron, Lambda, Lunit, Microsoft, Nightingale Open Science, Philips, Siemens Healthineers, Stability.ai, Subtle Medical, VinBrain, Visiana, Whiterabbit.ai, Lowenstein Foundation, Gordon and Betty Moore Foundation. **S.G.** No relevant relationships. **A.C.** Grants to university from NIH, GE HealthCare, Philips, and Amazon; royalties or licenses from LVIS; consulting fees from Patient Square Capital and Elucid Bioimaging; support for attending meetings and/or travel from Chondrometrics; scientific advisory board, Brain Key and Chondrometrics; stock or stock options in Cognita, Subtle Medical, Brain Key, and LVIS; co-founder of Cognita; in kind computational support from Microsoft, NVIDIA, and Stability.ai.

References

1. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577(7788):89–94. [Published correction appears in Nature 2020;586(7829):E19.]
2. Eng D, Chute C, Khandwala N, et al. Automated coronary calcium scoring using deep learning with multicenter external validation. NPJ Digit Med 2021;4(1):88.
3. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv 1706.03762 [preprint] https://arxiv.org/abs/1706.03762. Posted June 12, 2017. Updated August 2, 2023.
4. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv 2010.11929 [preprint] https://arxiv.org/abs/2010.11929. Posted October 22, 2020. Updated June 3, 2021.
5. Achiam J, Adler S, Agarwal S, et al. GPT-4 Technical Report. arXiv 2303.08774 [preprint] https://arxiv.org/abs/2303.08774. Posted March 15, 2023. Updated March 4, 2024.
6. Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. arXiv 2108.07258 [preprint] https://arxiv.org/abs/2108.07258. Posted August 16, 2021. Updated July 12, 2022.
7. Cherti M, Beaumont R, Wightman R, et al. Reproducible scaling laws for contrastive language-image learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition2023; 2818-2829.
8. Hoffmann J, Borgeaud S, Mensch A, et al. Training compute-optimal large language models. arXiv 2203.15556 [preprint] https://arxiv.org/abs/2203.15556. Posted March 29, 2022.
9. Fedus W, Zoph B, Shazeer N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. arXiv 2101.03961 [preprint] https://arxiv.org/abs/2101.03961. Posted January 11, 2021. Updated June 16, 2022.
10. Tu T, Azizi S, Driess D, et al. Towards Generalist Biomedical AI. arXiv 2307.14334 [preprint] https://arxiv.org/abs/2307.14334. Posted July 26, 2023.
11. Schuhmann C, Beaumont R, Vencu R, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. arXiv 2210.08402 [preprint] https://arxiv.org/abs/2210.08402. Posted October 16, 2022.

12. Baxter R, Nind T, Sutherland J, et al. The Scottish Medical Imaging Archive: 57.3 million radiology studies linked to their medical records. Radiol Artif Intell 2024;6(1):e220266.
13. Mei X, Liu Z, Robson PM, et al. RadImageNet: an open radiologic deep learning research dataset for effective transfer learning. Radiol Artif Intell 2022;4(5):e210315.
14. Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 2019;6(1):317.
15. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med 2022;28(9):1773–1784.
16. Azad B, Azad R, Eskandari S, et al. Foundational models in medical imaging: A comprehensive survey and future vision. arXiv 2310.18689 [preprint] https://arxiv.org/abs/2310.18689. Posted October 28, 2023.
17. Liu X, Zhang F, Hou Z, et al. Self-supervised learning: Generative or contrastive. IEEE Trans Knowl Data Eng 2021;35(1):857–876.
18. Galbusera F, Cina A. Image annotation and curation in radiology: an overview for machine learning practitioners. Eur Radiol Exp 2024;8(1):11.
19. Demirer M, Candemir S, Bigelow MT, et al. A user interface for optimizing radiologist engagement in image data curation for artificial intelligence. Radiol Artif Intell 2019;1(6):e180095.
20. Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models. arXiv 2206.07682 [preprint] https://arxiv.org/abs/2206.07682. Posted June 15, 2022. Updated October 26, 2022.
21. Li AC, Prabhudesai M, Duggal S, Brown E, Pathak D. Your diffusion model is secretly a zero-shot classifier. arXiv 2303.16203 [preprint] https://arxiv.org/abs/2303.16203. Posted March 28, 2023. Updated September 13, 2023.
22. Langlotz CP. The Future of AI and Informatics in Radiology: 10 Predictions. Radiology 2023;309(1):e231114.
23. Pickhardt PJ, Nguyen T, Perez AA, et al. Improved CT-based osteoporosis assessment with a fully automated deep learning tool. Radiol Artif Intell 2022;4(5):e220042.
24. Fink MA, Kades K, Bischoff A, et al. Deep learning–based assessment of oncologic outcomes from natural language processing of structured radiology reports. Radiol Artif Intell 2022;4(5):e220055.
25. Schmidt RA, Seah JCY, Cao K, Lim L, Lim W, Yeung J. Generative Large Language Models for Detection of Speech Recognition Errors in Radiology Reports. Radiol Artif Intell 2024;6(2):e230205.
26. Jia C, Yang Y, Xia Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision. arXiv 2102.05918 [preprint] https://arxiv.org/abs/2102.05918. Posted February 11, 2021. Updated June 11, 2021.
27. Zhang Y, Jiang H, Miura Y, Manning CD, Langlotz CP. Contrastive learning of medical visual representations from paired images and text. arXiv 2010.00747 [preprint] https://arxiv.org/abs/2010.00747. Posted October 2, 2020. Updated September 19, 2022.
28. Wang Z, Wu Z, Agarwal D, Sun J. MedCLIP: Contrastive learning from unpaired medical images and text. arXiv 2210.10163 [preprint] https://arxiv.org/abs/2210.10163. Posted October 18, 2022.
29. Wei X, Zhang T, Li Y, Zhang Y, Wu F. Multi-modality cross attention network for image and sentence matching. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2020; 10941–10950.
30. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. arXiv 2304.08485 [preprint] https://arxiv.org/abs/2304.08485. Posted April 17, 2023. Updated December 11, 2023.
31. Chen Z, Varma M, Delbrouck J-B, et al. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. arXiv 2401.12208 [preprint] https://arxiv.org/abs/2401.12208. Posted January 22, 2024.
32. de Herrera AGS, Ionescu B, Müller H, et al. Imageclef 2022: multimedia retrieval in medical, nature, fusion, and internet applications. European Conference on Information Retrieval: Springer, 2022; 382–389
33. Yu J, Wang Z, Vasudevan V, Yeung L, Seyedhosseini M, Wu Y. Coca: Contrastive captioners are image-text foundation models. arXiv 2205.01917 [preprint] https://arxiv.org/abs/2205.01917. Posted May 4, 2022. Updated June 14, 2022.
34. Chen C, Anjum S, Gurari D. Grounding answers for visual questions asked by visually impaired people. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022; 19098-19107.
35. Huang SC, Pareek A, Jensen M, Lungren MP, Yeung S, Chaudhari AS. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. NPJ Digit Med 2023;6(1):74.
36. Van Den Oord A, Kalchbrenner N, Kavukcuoglu K. Pixel recurrent neural networks. arXiv 1601.06759 [preprint] https://arxiv.org/abs/1601.06759. Posted January 25, 2016. Updated August 19, 2016.
37. Kingma DP, Welling M. An introduction to variational autoencoders. Found Trends Mach Learn 2019;12(4):307–392.
38. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. arXiv 2006.11239 [preprint] https://arxiv.org/abs/2006.11239. Posted June 19, 2020. Updated December 16, 2020.
39. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2022; 10684–10695.
40. Chambon P, Bluethgen C, Delbrouck JB, et al. RoentGen: Vision-Language Foundation Model for Chest X-ray Generation. arXiv 2211.12737 [preprint] https://arxiv.org/abs/2211.12737. Posted November 23, 2022.
41. Chambon P, Bluethgen C, Langlotz CP, Chaudhari A. Adapting pretrained vision-language foundational models to medical imaging domains. arXiv 2210.04133 [preprint] https://arxiv.org/abs/2210.04133. Posted October 9, 2022.
42. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv 1810.04805 [preprint] https://arxiv.org/abs/1810.04805. Posted October 11, 2018. Updated May 24, 2019.
43. Wei C, Fan H, Xie S, Wu C-Y, Yuille A, Feichtenhofer C. Masked feature prediction for self-supervised visual pre-training. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022; 14648–14678.
44. Zhou L, Liu H, Bae J, He J, Samaras D, Prasanna P. Self pre-training with masked autoencoders for medical image classification and segmentation. arXiv 2203.05573 [preprint] https://arxiv.org/abs/2203.05573. Posted March 10, 2022. Updated 21, 2023.
45. Dominic J, Bhaskhar N, Desai AD, et al. Improving Data-Efficiency and Robustness of Medical Imaging Segmentation Using Inpainting-Based Self-Supervised Learning. Bioengineering (Basel) 2023;10(2):207.
46. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. International conference on machine learning: PMLR, 2020;v 1597–1607. https://proceedings.mlr.press/v119/chen20j.html.
47. Grill JB, Strub F, Altché F, et al. Bootstrap your own latent-a new approach to self-supervised learning. arXiv 2006.07733 [preprint] https://arxiv.org/abs/2006.07733. Posted June 13, 2020. Updated September 10, 2020.
48. Van der Sluijs R, Bhaskhar N, Rubin D, Langlotz C, Chaudhari AS. Exploring image augmentations for siamese representation learning with chest x-rays. arXiv 2301.12636 [preprint] https://arxiv.org/abs/2301.12636. Posted January 30, 2023. Updated July 10, 2023.
49. Gan Z, Li L, Li C, Wang L, Liu Z, Gao J. Vision-language pre-training: Basics, recent advances, and future trends. Found Trends Comput Graph Vis 2022;14(3–4):163–352.
50. Boecking B, Usuyama N, Bannur S, et al. Making the most of text semantics to improve biomedical vision–language processing. European conference on computer vision: Springer, 2022; 1–21.
51. Shrestha P, Amgain S, Khanal B, Linte CA, Bhattarai B. Medical Vision Language Pretraining: A survey. arXiv 2312.06224 [preprint] https://arxiv.org/abs/2312.06224. Posted December 11, 2023.
52. Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision. arXiv 2103.00020 [preprint] https://arxiv.org/abs/2103.00020. Posted February 26, 2021.
53. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. https://arxiv.org/abs/2005.14165. Posted May 28, 2020. Updated July 22, 2020.
54. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. arXiv 2205.11916 [preprint] https://arxiv.org/abs/2205.11916. Posted May 24, 2022. Updated January 29, 2023.
55. Liang Y, Zhu L, Wang X, Yang Y. A simple episodic linear probe improves visual recognition in the wild. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022; 9549–9569.
56. Sanh V, Webson A, Raffel C, et al. Multitask prompted training enables zero-shot task generalization. arXiv 2110.08207 [preprint] https://arxiv.org/abs/2110.08207. Posted October 15, 2021. Updated March 17, 2022.
57. Wei J, Bosma M, Zhao VY, et al. Finetuned language models are zero-shot learners. arXiv 2109.01652 [preprint] https://arxiv.org/abs/2109.01652. Posted September 3, 2021. Updated February 8, 2022.
58. Taori R, Gulrajani I, Zhang T, et al. Alpaca: A strong, replicable instruction-following model. https://crfm.stanford.edu/2023/03/13/alpaca.html. Published 2021.
59. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. The Vicuna Team. https://lmsys.org/blog/2023-03-30-vicuna/. Published March 30, 2023. Accessed April 14, 2023.
60. Dai W, Li J, Li D, et al. Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv 2305.06500 [preprint] https://arxiv.org/abs/2305.06500. Posted May 11, 2023. Updated June 15, 2023.

61. Wu C, Zhang X, Zhang Y, Wang Y, Xie W. Towards Generalist Foundation Model for Radiology by Leveraging Web-scale 2D&3D Medical Data. arXiv 2308.02463 [preprint] https://arxiv.org/abs/2308.02463. Posted August 4, 2023. Updated November 16, 2023.
62. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. arXiv 2201.11903 [preprint] https://arxiv.org/abs/2201.11903. Posted January 28, 2022. Updated January 10, 2023.
63. Miao J, Thongprayoon C, Suppadungsuk S, Krisanapan P, Radhakrishnan Y, Cheungpasitporn W. Chain of Thought Utilization in Large Language Models and Application in Nephrology. Medicina (Kaunas) 2024;60(1):148.
64. Kaufmann T, Weng P, Bengs V, Hüllermeier E. A survey of reinforcement learning from human feedback. arXiv 2312.14925 [preprint] https://arxiv.org/abs/2312.14925. Posted December 30, 2023. Updated April 30, 2024.
65. Christiano PF, Leike J, Brown T, Martic M, Legg S, Amodei D. Deep reinforcement learning from human preferences. arXiv 1706.03741 [preprint] https://arxiv.org/abs/1706.03741. Posted June 12, 2017. Updated February 17, 2023.
66. Stiennon N, Ouyang L, Wu J, et al. Learning to summarize with human feedback. arXiv 2009.01325 [preprint] https://arxiv.org/abs/2009.01325. Posted September 2, 2020. Updated February 15, 2022.
67. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. arXiv 2203.02155 [preprint] https://arxiv.org/abs/2203.02155. Posted March 4, 2022.
68. Lee H, Phatale S, Mansoor H, et al. RLAIF: Scaling reinforcement learning from human feedback with ai feedback. arXiv 2309.00267 [preprint] https://arxiv.org/abs/2309.00267. Posted September 1, 2023. Updated December 1, 2023.
69. Dishner KA, McRae-Posani B, Bhowmik A, et al. A survey of publicly available MRI datasets for potential use in artificial intelligence research. J Magn Reson Imaging 2024;59(2):450–480.
70. Li J, Zhu G, Hua C, et al. A systematic collection of medical image datasets for deep learning. arXiv 2106.12864 [preprint] https://arxiv.org/abs/2106.12864. Posted June 24, 2021.
71. Shih G, Wu CC, Halabi SS, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiol Artif Intell 2019;1(1):e180041.
72. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.
73. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01):590–597.
74. Frazer HML, Tang JSN, Elliott MS, et al. ADMANI: Annotated digital mammograms and associated non-image datasets. Radiol Artif Intell 2022;5(2):e220072.
75. Jeong JJ, Vey BL, Bhimireddy A, et al. The EMory BrEast imaging Dataset (EMBED): A racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images. Radiol Artif Intell 2023;5(1):e220047.
76. Halling-Brown MD, Warren LM, Ward D, et al. Optimam mammography image database: a large-scale resource of mammography images and clinical data. Radiol Artif Intell 2020;3(1):e200103.
77. Flanders AE, Prevedello LM, Shih G, et al; RSNA-ASNR 2019 Brain Hemorrhage CT Annotators. Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge. Radiol Artif Intell 2020;2(3):e190211.
78. Callejas MF, Lin HM, Howard T, et al. Augmentation of the RSNA Pulmonary Embolism CT Dataset with Bounding Box Annotations and Anatomic Localization of Pulmonary Emboli. Radiol Artif Intell 2023;5(3):e230001.
79. Lin HM, Colak E, Richards T, et al; RSNA-ASSR-ASNR Annotators and the Dataset Curation Contributors. The RSNA cervical spine fracture CT dataset. Radiol Artif Intell 2023;5(5):e230034.
80. Feng S, Azzollini D, Kim JS, et al. Curation of the candid-ptx dataset with free-text reports. Radiol Artif Intell 2021;3(6):e210136.
81. Lin W, Zhao Z, Zhang X, et al. PMC-CLIP: Contrastive language-image pre-training using biomedical documents. arXiv 2303.07240 [preprint] https://arxiv.org/abs/2303.07240. Posted March 13, 2023.
82. Ye J, Cheng J, Chen J, et al. SA-Med2D-20M Dataset: Segment Anything in 2D Medical Imaging with 20 Million masks. arXiv 2311.11969 [preprint] https://arxiv.org/abs/2311.11969. Posted November 30, 2023.
83. Longpre S, Mahari R, Chen A, et al. The data provenance initiative: A large scale audit of dataset licensing & attribution in AI. arXiv 2310.16787 [preprint] https://arxiv.org/abs/2310.16787. Posted October 25, 2023. Updated November 4, 2023.
84. Xie SM, Pham H, Dong X, et al. DoReMi: Optimizing data mixtures speeds up language model pretraining. arXiv 2305.10429 [preprint] https://arxiv.org/abs/2305.10429. Posted May 17, 2023. Updated November 21, 2023.
85. Zhou C, Liu P, Xu P, et al. LIMA: Less is more for alignment. arXiv 2305.11206 [preprint] https://arxiv.org/abs/2305.11206. Posted May 18, 2023.
86. Fleming SL, Lozano A, Haberkorn WJ, et al. Medalign: A clinician-generated dataset for instruction following with electronic medical records. arXiv 2308.14089 [preprint] https://arxiv.org/abs/2308.14089. Posted August 27, 2023. Updated December 24, 2023.
87. Fink MA, Bischoff A, Fink CA, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. Radiology 2023;308(3):e231362.
88. Lyu Q, Tan J, Zapadka ME, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art 2023;6(1):9.
89. Yan A, McAuley J, Lu X, et al. RadBERT: Adapting transformer-based language models to radiology. Radiol Artif Intell 2022;4(4):e210258.
90. Liu Z, Zhong A, Li Y, et al. Radiology-GPT: A Large Language Model for Radiology. arXiv 2306.08666 [preprint] https://arxiv.org/abs/2306.08666. Posted June 14, 2023. Updated March 19, 2024.
91. Pesapane F, Tantrige P, De Marco P, et al. Advancements in Standardizing Radiological Reports: A Comprehensive Review. Medicina (Kaunas) 2023;59(9):1679.
92. Zhou Y, Chia MA, Wagner SK, et al; UK Biobank Eye & Vision Consortium. A foundation model for generalizable disease detection from retinal images. Nature 2023;622(7981):156–163.
93. Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. Nat Commun 2024;15(1):654.
94. Levine DM, Tuwani R, Kompa B, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model. medRxiv 2023:2023.2001.2030.23285067.
95. Fogel AL, Kvedar JC. Artificial intelligence powers digital medicine. NPJ Digit Med 2018;1(1):5.
96. Pai S, Bontempi D, Prudente V, et al. Foundation models for quantitative biomarker discovery in cancer imaging. medRxiv 2023.
97. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172–180. [Published correction appears in Nature 2023;620(7973):E19.]
98. Park SH, Han K, Jang HY, et al. Methods for clinical evaluation of artificial intelligence algorithms for medical diagnosis. Radiology 2023;306(1):20–31.
99. Erickson BJ, Kitamura F. Magician's Corner: 9. Performance Metrics for Machine Learning Models. Radiol Artif Intell 2021;3(3):e200126.
100. Mehandru N, Miao BY, Almaraz ER, Sushil M, Butte AJ, Alaa A. Evaluating large language models as agents in the clinic. NPJ Digit Med 2024;7(1):84.
101. Chaves JMZ, Bhaskhar N, Attias M, et al. RaLEs: a Benchmark for Radiology Language Evaluations. Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2023. https://openreview.net/pdf?id=PWLGrvoqiR.
102. Chang Y, Wang X, Wang J, et al. A survey on evaluation of large language models. arXiv 2307.03109 [preprint] https://arxiv.org/abs/2307.03109. Posted July 6, 2023. Updated December 29, 2023.
103. Yu F, Endo M, Krishnan R, et al. Evaluating progress in automatic chest X-ray radiology report generation. Patterns 2023;4(9):100802.
104. Smit A, Jain S, Rajpurkar P, Pareek A, Ng AY, Lungren MP. CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. arXiv 2004.09167 [preprint] https://arxiv.org/abs/2004.09167. Posted April 20, 2020. Updated October 18, 2020.
105. Delbrouck JB, Chambon P, Bluethgen C, Tsai E, Almusa O, Langlotz C. Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022; 4348–4360.
106. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 2024;30(4):1134–1142.
107. Dubois Y, Galambosi B, Liang P, Hashimoto TB. Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators. arXiv 2404.04475 [preprint] https://arxiv.org/abs/2404.04475. Posted April 6, 2024.
108. Glocker B, Jones C, Roschewitz M, Winzeck S. Risk of bias in chest radiography deep learning foundation models. Radiol Artif Intell 2023;5(6):e230060.
109. Perez E, Huang S, Song F, et al. Red teaming language models with language models. arXiv 2202.03286 [preprint] https://arxiv.org/abs/2202.03286. Posted February 7, 2022.
110. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. EBioMedicine 2023;90:104512.

111. Rawte V, Sheth A, Das A. A survey of hallucination in large foundation models. arXiv 2309.05922 [preprint] https://arxiv.org/abs/2309.05922. Posted September 12, 2023.
112. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021; 610–623.
113. Van Veen D, Van Uden C, Blankemeier L, et al. Clinical text summarization: Adapting large language models can outperform human experts. arXiv 2309.07430 [preprint] https://arxiv.org/abs/2309.07430. Posted September 14, 2023. Updated April 11, 2024.
114. Gonzalez C. Lifelong Learning in the Clinical Open World. Darmstadt: Technical University of Darmstadt, 2023.
115. Vasconcelos H, Jörke M, Grunde-McLaughlin M, Gerstenberg T, Bernstein MS, Krishna R. Explanations can reduce overreliance on ai systems during decision-making. Proceedings of the ACM on Human-Computer Interaction 2023;7(CSCW1):1–38.
116. Meşe İ, Taşlıçay CA, Kuzan BN, Kuzan T, Sivrioğlu AK. Educating the next generation of radiologists: a comparative report of ChatGPT and e-learning resources. Diagn Interv Radiol 2024;30(3):163–174.
117. Alabed A, Javornik A, Gregory-Smith D. AI anthropomorphism and its effect on users' self-congruence and self–AI integration: A theoretical framework and research agenda. Technol Forecast Soc Change 2022;182:121786.
118. Youssef A, Stein S, Clapp J, Magnus D. The Importance of Understanding Language in Large Language Models. Am J Bioeth 2023;23(10):6–7.
119. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 2021;27(12):2176–2182.
120. Abid A, Farooqi M, Zou J. Persistent anti-muslim bias in large language models. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021; 298–306.
121. Tamkin A, Askell A, Lovitt L, et al. Evaluating and mitigating discrimination in language model decisions. arXiv 2312.03689 [preprint] https://arxiv.org/abs/2312.03689. Posted December 6, 2023.
122. Zhao J, Fang M, Pan S, Yin W, Pechenizkiy M. GPTBIAS: A Comprehensive Framework for Evaluating Bias in Large Language Models. arXiv 2312.06315 [preprint] https://arxiv.org/abs/2312.06315. Posted December 11, 2023.
123. Yang Y, Zhang H, Katabi D, Ghassemi M. Change is hard: a closer look at subpopulation shift. arXiv 2302.12254 [preprint] https://arxiv.org/abs/2302.12254. Posted February 23, 2023. Updated August 17, 2023.
124. Rillig MC, Ågerstrand M, Bi M, Gould KA, Sauerland U. Risks and benefits of large language models for the environment. Environ Sci Technol 2023;57(9):3464–3466.
125. Touvron H, Lavril T, Izacard G, et al. LLaMA: Open and efficient foundation language models. arXiv 2302.13971 [preprint] https://arxiv.org/abs/2302.13971. Posted February 27, 2023.
126. Lee K, Gangidi A, Oldham M. Building Meta's GenAI Infrastructure. Meta AI. https://engineering.fb.com/2024/03/12/data-center-engineering/building-metas-genai-infrastructure/. Posted March 12, 2024. Accessed April 24, 2024.
127. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023;6(1):120.
128. Hacker P, Engel A, Mauer M. Regulating ChatGPT and other large generative AI models. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 2023; 1112–1123.
129. Lin X, Wang W, Li Y, et al. Data-efficient Fine-tuning for LLM-based Recommendation. arXiv 2401.17197 [preprint] https://arxiv.org/abs/2401.17197. Posted January 30, 2024. Updated June 4, 2024.
130. Trabucco B, Doherty K, Gurinas M, Salakhutdinov R. Effective data augmentation with diffusion models. arXiv 2302.07944 [preprint] https://arxiv.org/abs/2302.07944. Posted February 7, 2023. Updated May 25, 2023.
131. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. NPJ Digit Med 2023;6(1):186.
132. Yu S, Muñoz JP, Jannesari A. Federated Foundation Models: Privacy-Preserving and Collaborative Learning for Large Models. arXiv 2305.11414 [preprint] https://arxiv.org/abs/2305.11414. Posted May 19, 2023. Updated March 19, 2024.
133. Yi H, Qin Z, Lao Q, et al. Towards General Purpose Medical AI: Continual Learning Medical Foundation Model. arXiv 2303.06580 [preprint] https://arxiv.org/abs/2303.06580. Posted March 12, 2023.
134. Wang L, Zhang X, Su H, Zhu J. A comprehensive survey of continual learning: Theory, method and application. arXiv 2302.00487 [preprint] https://arxiv.org/abs/2302.00487. Posted January 31, 2023. Updated February 6, 2024.
135. Feng J, Phillips RV, Malenica I, et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. NPJ Digit Med 2022;5(1):66.
136. Hu EJ, Shen Y, Wallis P, et al. LoRA: Low-rank adaptation of large language models. arXiv 2106.09685 [preprint] https://arxiv.org/abs/2106.09685. Posted June 17, 2021. Updated October 16, 2021.
137. Woisetschläger H, Isenko A, Wang S, Mayer R, Jacobsen H-A. A Survey on Efficient Federated Learning Methods for Foundation Model Training. arXiv 2401.04472 [preprint] https://arxiv.org/abs/2401.04472. Posted January 9, 2024. Updated February 7, 2024.
138. Akiba T, Shing M, Tang Y, Sun Q, Ha D. Evolutionary Optimization of Model Merging Recipes. arXiv 2403.13187 [preprint] https://arxiv.org/abs/2403.13187. Posted March 19, 2024.
139. Shavit Y, Agarwal S, Brundage M, et al. Practices for governing agentic ai systems. OpenAI. https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf. Published December 23.